Towards the Automatic Processing of Language Registers: Semi-supervisedly Built Corpus and Classifier for French

Gwénolé Lecorvé, Hugo Ayats, Benoît Fournier, Jade Mekki, Jonathan Chevelu, Delphine Battistelli, Nicolas Béchet

▶ **To cite this version:**

Gwénolé Lecorvé, Hugo Ayats, Benoît Fournier, Jade Mekki, Jonathan Chevelu, et al.. Towards the Automatic Processing of Language Registers: Semi-supervisedly Built Corpus and Classifier for French. International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Apr 2019, La Rochelle, France. pp.480-492, 10.1007/978-3-031-24337-0_34 . hal-02064694

HAL Id: hal-02064694
https://hal.science/hal-02064694

Submitted on 9 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards the Automatic Processing of Language Registers: Semi-supervisedly Built Corpus and Classifier for French

Gwénolé Lecorvé[1] Hugo Ayats[1] Benoît Fournier[1] Jade Mekki[1,3]
Jonathan Chevelu[1] Delphine Battistelli[3] Nicolas Béchet[2]

[1] Univ Rennes, CNRS, IRISA / 6, rue de Kerampont, Lannion, France
[2] Université de Bretagne Sud, CNRS, IRISA / Campus Tohannic, Vannes, France
[3] Université Paris Nanterre, CNRS, MoDyCo / av. République, Nanterre, France
firstname.lastname@irisa.fr, delphine.battistelli@parisnanterre.fr

Abstract. Language registers are a strongly perceptible characteristic 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 relies on a small initial seed of expert data. After massively retrieving web pages, it iteratively alternates the training of an intermediate classifier and the annotation of new texts to augment the labeled corpus. The approach is applied to the casual, neutral, and formal registers, leading to a 750M word corpus and a final neural classifier with an acceptable performance.

1 Introduction

The language registers provide a lot of information about a communicator and the relationship with the recipients of her/his messages. Their automatic processing could show whether two persons are friends or are in a hierarchical relation, or give hints about someone's educational level. Modeling language registers would also benefit in natural language generation by enabling to modulate the style of artificial discourses. However, language registers are still poorly studied in natural language processing (NLP), particularly because of the lack of large training data. To overcome this problem, this paper presents a semi-supervised, self-training, approach to build a text corpus labeled in language registers.

The proposed approach relies on a small set of manually labeled data and a massive collection of automatically collected unlabeled web pages. Text segments extracted from these web pages are iteratively labeled using a classifier—a neural network—trained on the labeled data. For a given iteration, text segments that are classified with a high confidence are added to the training data, and a new classifier is then trained for a next iteration. Through this process, we expect to label as many segments as possible, provided that the classifier accuracy remains good enough when augmenting the training data. In practice, this process is applied on a set of $400,000$ web pages, and results in a corpus of about 750 million words labeled in casual, neutral and formal register. Alongside, when testing on 2

different test sets, the final classifier performs accuracies of 87 % and 61 %. The set of descriptors used includes 46 characteristics of various natures (lexical, morphological, syntactical ...) questions of a preliminary expert analysis.

This paper is a first step in the process of modelling language registers in NLP. As such, the will of the authors is to report about first experiments, popularize the issue of language registers, and provide a baseline for future improvements and tasks that are more elaborate. Especially, the presented semi-supervised process is not new and could be improved. Likewise, the associated classifier could also probably benefit from various sophistications.

In this paper, Section 2 presents a state of the art about language registers, while Section 3 details the semi-supervised approach. Then, Sections 4, 5, and 6 describe the collected data, the classifier training, and the resulting corpus.

2 State of the Art and Positioning

The notion of register refers to the way in which linguistic productions are evaluated and categorized within the same linguistic community [1, 2]. A register is characterized by multiple specific features (more or less complex terms, word order, verb tenses, length of sentences, etc.) and can be compared to others, sometimes with in a given ordering (e.g., formal, literary, neutral, casual, slang...). Such a partitioning depends on the angle from which linguistic communities are observed, for instance, the influence of the communication media [3] or the degree of specialization [4, 5]. Its granularity may also vary [6–8]. In this work, we consider 3 registers : casual, neutral and formal. This choice is primarily motivated by pragmatism, as this division is relatively consensual and unambiguous for manual labeling, while not prohibiting possible refinements in the future. The neutral register involves a minimal set of assumptions about any specific knowledge of the message recipient, and is therefore based on the grammar and vocabulary of the language, with no rare constructions and terms. On the contrary, the formal register assumes the recipient to have a high proficiency, whereas the casual one allows voluntary or faulty deviations of the linguistic norms. In this paper, we do not focus on the sociolinguistics, but seek to enable NLP on language registers by building a large labeled corpus.

Registers got very little attention in NLP. [9, 10] proposed to classify documents as formal or informal. In [11], the authors train a regression model to predict a level of formality of sentences. In these papers, features are derived from a linguistic analysis. Although these features are a good basis for our work, they are designed for English and do not apply for French. Moreover, they work on few data (about 1K documents) whereas we expect to build a large corpus ($>$ 100K documents). More generally, the study of language register shares similarities with authorship attribution [12, 13] and the analysis of new media like blogs, SMSs, tweets, etc. [14–18], where research is backed by the release of reference corpora. Such a corpus does not exist for the language registers.

Automatic style processing methods are all based on a set of relevant features derived from the texts to be processed. Due to its historical importance,

author attribution work can list a wide range of features. As indicated by [12], an author's preferences or writing choices are reflected at several levels of language. The most obvious—and most studied—is the lexical level, e.g.through the length of words and sentences in a text, the richness of its vocabulary or frequencies of words and word n-grams [19, 20]. In this respect, it is generally accepted in the community that grammatical words (prepositions, articles, auxiliaries, modal verbs, etc.) are of significant interest while the others (nouns, adjectives, etc..) should be avoided for style processing [21, 22], according to a principle of orthogonality between style and meaning. This principle emphasizes the need to abstract some elements of meaning, otherwise the analysis risking to be be biased by the text's topic. Nevertheless, semantics can prove to be useful, for example through the frequencies of synonyms and hypernyms, or the functional relations between propositions (clarification of a proposition by another, opposition, etc.) [23, 22]. Moreover, whatever their meaning, some specific words explicitly testifies to the fact that the text is of a specific style [24], especially in the case of language registers. Syntactically, the use of descriptors derived from morphosyntactic and syntactic analyzes is very widely used to characterize the style [21, 25, 26]. Finally, other work has been interested in graphical information by considering n-gram of characters, types of graphemes (letter, number, punctuation, capital letters, etc.) or information compression measures [21, 27, 28]. In our work, a preliminary linguistic study was conducted in this sense, leading to a set of descriptors for the 3 registers considered, as detailed in the description of the trained classifier in Section 4. Before that, next section introduces the overall semi-supervised joint construction of the corpus and classifier.

3 Proposed Approach

The semi-supervised process used to build the labeled corpus is schematized in Figure 1. This process follows a self-training approach where seed data is augmented with automatically labeled texts. This approach has been experimented in various other NLP tasks [29, 30]. In our approach, the corpus used for data augmentation is collected from the web. Queries are submitted to a search engine. These queries are derived from two specialized lexicons, one for the casual register, and the other for the formal one. Then, the collected texts are filtered using a neural network classifier to extract the most relevant ones for each of the 3 considered registers. Since the classifier requires labeled training data and data labeling requires a classifier, the approach is iterative. That is, a first classifier is initially trained on a small initial manually annotated seed. This first classifier makes it possible to select texts whose predicted register is considered as reliable. These texts are added to those already labeled, and a new iteration starts. In the end, this process results in a set of categorized texts and a classifier.

Note that the use of the Internet is not an originality of our work since many similar examples exist in literature, for example [31] (although our collection process is not iterative here) or [32] for the collection of thematic pages. Then, self-training approaches is known to potentially degrade along iterations, due to

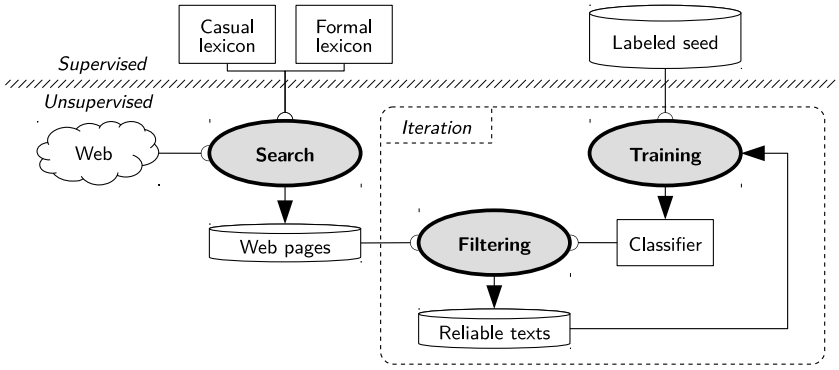

Fig. 1. Overview of the semi-supervised process.

classification mistakes, i.e., the augmented data progressively moves away from the original target task. An important objective in our work was to make sure that classification accuracy would not drop along the iterations.

The considered classes are *casual*, *neutral*, and *formal*. The assumption is also made that some texts do not belong to any of the three registers, either because they are badly formed (foreign language, SMS style, non-natural text, etc.) or because the register is not homogeneous (e.g. user comments). Our condition on classification reliability makes it possible to model this.

4 Data

In practice, the lexicons on which the collection of web pages is based are words and expressions automatically retrieved from a backup of the French version of Wiktionary. For a given register, only the unambiguous words belonging to a register are considered, that is to say the terms having all their meanings annotated as belonging to the same register. Precisely, the terms annotated as slang, casual, popular and vulgar were grouped within the casual lexicon and those categorized as literary and formal within the formal lexicon, each thus totaling respectively 6,000 and 500 entries. Equal numbers of queries are built for each register by randomly combining selected elements of the associated lexicon. Queries are empirically bound from 2 to 6 words in order to ensure a non-zero number of results and a minimal relevance for the returned pages. Web requests are made using the Bing API. In total, 12K requests are submitted, each limited to a maximum of 50 hits. Online dictionaries were excluded at the time of the request in order to only retrieve pages where the searched terms are in context, and not isolated in a definition or an example. 76 % of the queries returned at least one hit and 49 % reached the maximum limit, be it for casual or formal queries. This is results in a collection of 400K web pages.

The textual content of the web pages is extracted automatically thanks to a dedicated tool that looks for the central textual part of the page. It excludes

titles, menus, legal notices, announcements, etc. but includes comments if they have enough linguistic content and conform to the standard editorial style (punctuation, not abbreviation of words...). The cleaned texts were segmented on the paragraph boundaries into pieces of about $5,000$ characters to avoid a lack of homogeneity within long web pages (eg forums) and not to introduce training biases related to text length disparities. Furthermore, non-French textual segments were excluded, resulting in 825K segments, representing 750M words. While all web pages are supposed to contained register-specific terms from their query, segmentation also enables introducing segments where none is present.

The seed collection of hand-tagged texts gathers 435 (about 440K words) segments from novels, journals[1] and web pages[2]. Segments were jointly labeled by 2 qualified annotators and are balanced over the 3 registers. The seed texts have been selected such that there is no ambiguity about their register. As such, they can be regarded as stereotypical. Examples are given in the appendix. This seed is divided into training (40%, i.e., 174 segments), development (20%) and test (40%) sets. In addition, a second test set of 139 segments is randomly sampled from the collected corpus of web segments. These segments were labeled as follows: 27 as casual (19 %), 69 as neutral (50 %), 38 as formal (27 %), and 5 as none (4 %) because ambivalent[3]. This second set represents a more realistic situation since our seed has been designed to be unambiguous. The distribution over the registers and the presence of the "none" class illustrate this difference.

Text segments are described by 46 global features derived from related work in French linguistics [33–36]. The exhaustive list is given in Table 1. These features are relative frequencies of various linguistic phenomena covering lexical, phonetic, morphosyntactic and syntactic aspects. We address a few remark regarding these features. First, it can be noticed that no lexicon exists for the neutral register. Then, some words may be ambiguous regarding their membership in a registry. Thus, two feature variants are considered for register-specific words frequencies. The first weights the frequency of a word by the number of acceptations identified as belonging to the given register divided by the total number of its acceptations. The other variant is stricter. It only counts a word if all its acceptations are identified as belonging to the register. The case of the phrases or expressions does not require this duality because they are generally less ambiguous. Finally, most of the features denote well-known phenomena highlighting deviations to the norm of the language, for instance through mappings of non-written usages (especially speech) into the written language or syntactic mistakes. The lexical richness is also part of this deviation since there is an infinity to diverge from the norm. Hence, the casual register is rich.

In practice, feature were extracted using dedicated dictionaries, orthographic and grammatical analyzes (LanguageTool), and *ad hoc* scripts.

---

[1] Among which: *Kiffe kiffe demain* (Faïza Guène), *Albertine disparue* (Marcel Proust), *Les Mohicans de Paris* (Alexandre Dumas), The Bridge-Builders (Rudyard Kipling), *Les misérables* (Victor Hugo), and archives from the newspaper *L'Humanité*.

[2] These web pages do not come from the automatically collected set.

[3] Often because of mixed narrative and active parts.

Table 1. Features used by the classifier.

Lexicon
– Casual words weighted by their number of acceptations as casual: 7 828 items
– Formal words weighted by their number of acceptations as formal : 565 items
– Purely casual words (all acceptations are casual) : 3 075 items
– Purely formal words (all acceptations are formal) : 166 items
– Casual phrases : 3 453 items
– Formal phrases : 143 items
– Animal names : 78 items
– Onomatopoeia (e.g., "*ah*", "*pff*"...) : 125 items
– SMS terms (e.g., "*slt*", "*lol*", "*tkt*"...) : 540 items
– Lexical and syntactic anglicisms
– Unknown words
– Word "*ça*" ("it"/"this")
– Word "*ce*" ("it"/"this")
– Word "*cela*" ("it"/"this")
– Word "*des fois*" ("sometimes")
– Word "*là*" ("there"/"here")
– Word "*parfois*" ("sometimes")
Phonetics
– Vowel elision ("*m'dame*", "*p'tit*"...)
– Elision of 'r' ("*vot'*", "*céleb'*"...)
– Written liaisons "*z*" ("*les zanimaux*")
Morphology
– Repeating syllables ("*baba*", "*dodo*"...)
– Repeating vowels ("*saluuuut*")
– Word endings in "*-asse*"
– Word endings in "*-iotte*"
– Word endings in "*-o*"
– Word endings in "*-ou*"
– Word endings in "*-ouze*"
Morphosyntax
– All verb tenses and modes
– All types of subject pronouns
– Verb groups (French peculiarity)
Syntax
– Double possessive form (e.g., "*son manteau à lui*", "his coat belonging to him")
– Structure "*c'est ... qui*" ("it's ... who/which")
– Use of "*est-ce que*" (specific French interrogative form)
– Conjunction "*et*" ("and")
– Shortened negative forms without "*ne*" (e.g., "il vient pas")
– Other, uncategorized, syntactic irregularities

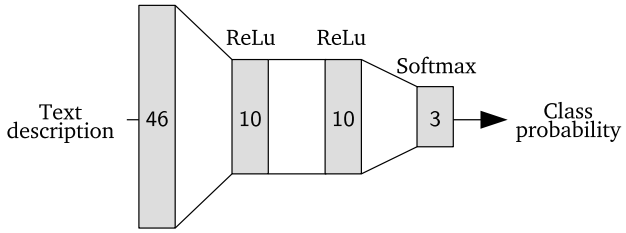

Fig. 2. Architecture of the neural network classifier.

5 Classifier training

As illustrated in Figure 2, the classifier is a multilayer neural network, fed by the 46 global features of a segment, predicting probabilities for each register (softmax layer of size 3). 2 dense hidden layers of size 10 are considered, respectively with leaky ReLU[4] and *tanh* activation function. No extensive tuning on the development set has been performed on this architecture but the use of a simple architecture and global features is voluntary since the seed data is small[5]. The reliability of a prediction is directly given by the majority class probability, and a threshold is applied to decide whether to validate the segment's label or not.

The experiments were conducted in Python using Keras and TensorFlow. Apart from learning the first model on the seed, successive classifiers are trained by batch of 100 instances over 20 epochs using the optimization algorithm *rmsprop* and the mean absolute error as loss function. At each iteration, the newly selected segments among the web data are injected into the training set for 80% and the development set for the rest. The test set is never modified in order to measure the progress of the classifier throughout the process.

The figure 3 shows the accuracy evolution through iterations on the test set and the manually labeled sample of the collected segments. Different selection thresholds are reported, ranging from 0.8 to 1 (i.e., the classifier is sure of its predictions). These high values are justified by the high accuracy of 87 % on the seed. Overall, the classifier is rather stable despite the insertion of new data, regardless of the dataset, showing that the inserted data is relevant. Still, the results on the labeled sample are lower than those on the test set, although still better than a random or naive classification. This is not an overfitting on the seed since this data is completely diluted once the training corpus is augmented with the selected web segments. Instead, we think that this discrepancy is because the automatically collected data is less clean and less stereotypical than the seed. Hence, their register is more difficult to predict. Regarding data selection, the strictest threshold value, 1, leads to deteriorate the results, while thresholds 0.9 and 0.99 produce the best results.

[4] Parameter α set to 0.1.

[5] Especially, word-based models, e.g., RNNs, could not reasonably be applied here due to this initially limited amount of data. However, such models will be studied in the future using the final corpus.

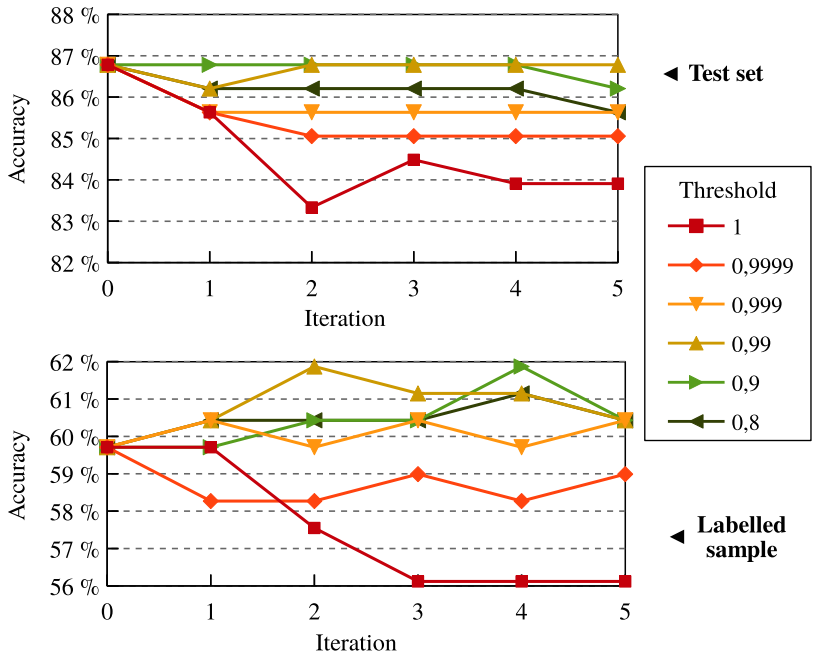

Fig. 3. Evolution of accuracy on the test set (top) and the labeled sample (bottom).

Table 2. Recall, precision, and F-measure for each register after 5 iterations with a threshold set to 0.99) on the test set and the labeled sample.

	Test set			**Labeled sample**		
	Casual	Neutral	Formal	Casual	Neutral	Formal
Recall	.90	.78	.93	.53	.72	.45
Precision	.84	.90	.87	.52	.64	.61
F-measure	.87	.83	.90	.52	.68	.52

Table 2 shows the recall, precision, and F-measure at the end of the process for the threshold 0.99, on the test set and on the labeled sample. On the test set, it appears that the results are relatively homogeneous between registers, the lowest F-measure being for the neutral register because of a lower recall. Conversely, the results on the labeled sample—which are worse as previously highly—show that the neutral register is the one best recognized as the casual and formal registers present weak F-measures. For the first one, difficulties seem to be global with precision and recall just greater than 0.5. For the second, the weak results come from a high proportion of false negatives (low recall). Therefore, while the results are encouraging, they also call for further improvements. Especially, attention should concentrate on maximizing precision, i.e., avoiding false positives, because they tend to distort the convergence of the semi-supervised process.

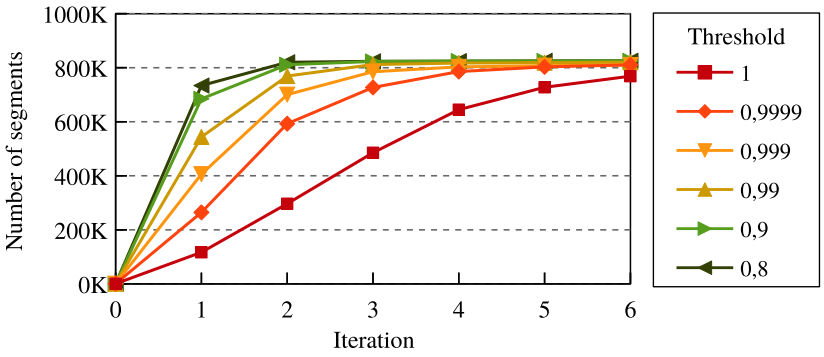

Fig. 4. Size of the labeled corpus for each iteration.

6 Automatically Labeled Corpus

Figure 4 illustrates the size evolution of the labeled corpus for different selection thresholds. First, it appears that all or almost all the collected segments end up the process with a label. Given the already mentioned complexity of the data (noise, ambiguity), this again urges on further developments on false positives and a stopping criterion. Then, it appears that the process is quick, converging in a few iterations. For example, 89% labels are validated at the end of the first pass for a selection threshold of 0.8.

At the end of the process for the threshold of 0.99, texts labeled as casual come for 68 % from casual queries and, therefore, for 32 % from formal queries. These ratios are 47/53 % and 37/63 % for the neutral and formal registers. Hence, the lexicons are appropriate to initiate the process since they do not fully confine the collected pages in the register of their original query. However, a manual analysis shows that a considerable number of texts classified as casual but coming from formal queries (and vice versa) should not be. Hence, some phenomena are still poorly understood and the method should be refined. For instance, it is likely that the model excessively trusts some features, especially register-specific terms. One solution may be to introduce a dropout mechanism when training the neural network.

Finally, as detailed by Figure 5, a deeper analysis of the corpus evolution shows that the class distribution automatically evolves. Starting from the initially balanced setting of the seed, the proportions of the casual, neutral, and formal registers in the final step corpus are 28 %, 48 %, and 24 %—consistently what is observed on the labeled sample of the web segments. 2 examples of automatically labeled segments are given in Table 3.

7 Conclusion

In this paper, we have presented a semi-supervised process that jointly builds a text corpus labeled in language registers and an associated classifier. Based on

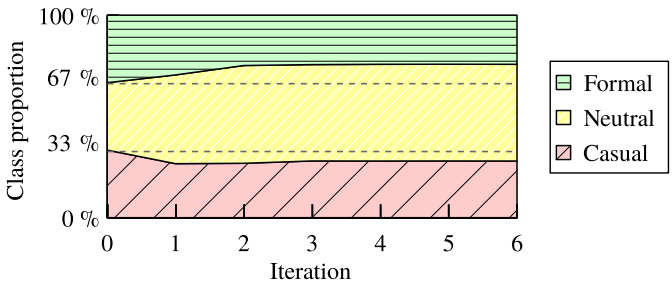

Fig. 5. Evolution of class distribution in the labeled corpus (percentage on the number of texts, threshold = 0.99).

Table 3. Excerpts from 2 texts automatically labeled as neutral and formal.

Neutral	**Formal**
`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.`
Yes, Mr. Adrien Richard, if you like better, the director of the factory, but we only call him Mr. Adrien because we went to school together and he also calls us our first name.	*Moreover, we find the same disdainful distinction with regard to professionals and their "vile salary" which does not prevent them from dying "in hospitals", at the doctor Muret's.*

a large set of text segments and a few initial expert resources, the result of this approach is a corpus of 825K textual segments representing a total of about 750M words. The classifier achieves a good accuracy of 87 % on the test set, and more modest results on a manually labeled subset of the collected segments. These results seem to demonstrate the validity of the approach, while also highlighting the need for refinements and for a less stereotypical seed.

Among the lines of future work, questions about classification uncertainty and about the model's over-confidence will be dealt in priority. The use of scaled memberships (instead of binary ones), of an "undetermined" label, and of dropout or cross validation during training should help in these perspectives. An increased attention to false positives (e.g., within the objective function of the neural network) should also be paid. Moreover, the selection criterion for new segments could be improved, e.g., by combining the classifier's output probability with the probability to make an error. Finally, in the long term, advanced studies of language registers will be conducted based on the labeled corpus (discriminative features, local features instead of global ones, sequence models, etc.).

Aknowledgements

This work has benefited from the financial support of the French National Research Agency (ANR) through the TREMoLo project (ANR-16-CE23-0019).

A Supplementary material

The following supplementary material can be downloaded at `ftp://ftp.cicling.org/in/CICLing-2019/CICLing_58.zip`:

- Exhaustive list of features.
- CSV data (seed and web segments).
- Examples of raw seed and automatically labeled web texts.

References

1. Ure, J.: Introduction: approaches to the study of register range. International Journal of the Sociology of Language **1982** (1982)
2. Biber, D., Conrad, S.: Register, genre, and style. Cambridge University Press (2009)
3. Charaudeau, P.: Le discours d'information médiatique: la construction du miroir social. Nathan (1997)
4. Borzeix, A., Fraenkel, B.: Langage et travail (communication, cognition, action). CNRS éd. (2005)
5. Moirand, S.: Les discours de la presse quotidienne. Observer, analyser, comprendre. Puf (2007)
6. Sanders, C.: Sociosituational variation. Cambridge: Cambridge University Press (1993)
7. Biber, D., Finegan, E.: Sociolinguistic perspectives on register. Oxford University Press on Demand (1994)
8. Gadet, F.: Niveaux de langue et variation intrinsèque. Palimpsestes **10** (1996)
9. Peterson, K., Hohensee, M., Xia, F.: Email formality in the workplace: A case study on the enron corpus. In: Proceedings of the Workshop on Languages in Social Media. (2011)
10. Pavlick, E., Tetreault, J.: An empirical analysis of formality in online communication. Transactions of the Association of Computational Linguistics **4** (2016)
11. Sheikha, F.A., Inkpen, D.: Automatic classification of documents by formality. In: IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE). (2010)
12. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the Association for Information Science and Technology **60** (2009)
13. Iqbal, F., Binsalleeh, H., Fung, B.C., Debbabi, M.: A unified data mining solution for authorship analysis in anonymous textual communications. Information Sciences **231** (2013)
14. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: Proceedings of the AAAI spring symposium: Computational approaches to analyzing weblogs. Volume 6. (2006)
15. Kobus, C., Yvon, F., Damnati, G.: Normalizing sms: are two metaphors better than one? In: Proceedings of COLING. (2008)

16. Gianfortoni, P., Adamson, D., Rosé, C.P.: 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. (2011)
17. Eisenstein, J.: What to do about bad language on the internet. In: Proceedings of HLT-NAACL. (2013)
18. Cougnon, L.A., Fairon, C.: SMS Communication: A linguistic approach. Volume 61. John Benjamins Publishing Company (2014)
19. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. ACM Sigmod Record **30** (2001)
20. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In: Proceedings of EMNLP. (2006)
21. Koppel, M., Schler, J.: Exploiting stylistic idiosyncrasies for authorship attribution. In: Proceedings of IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis. Volume 69. (2003)
22. Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., Levitan, S.: Stylistic text classification using functional lexical features. Journal of the Association for Information Science and Technology **58** (2007)
23. McCarthy, P.M., Lewis, G.A., Dufty, D.F., McNamara, D.S.: Analyzing writing styles with coh-metrix. In: Proceedings of the FLAIRS Conference. (2006)
24. Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., Tambouratzis, D.: 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** (2004)
25. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing **22** (2007)
26. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications **41** (2014)
27. Marton, Y., Wu, N., Hellerstein, L.: On compression-based text classification. In: Proceedings of the European Conference on Information Retrieval (ECIR). Volume 3408. (2005)
28. Escalante, H.J., Solorio, T., Montes-y Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proceedings of HTL-ACL. (2011)
29. McClosky, D., Charniak, E., Johnson, M.: Effective self-training for parsing. In: Proceedings of HLT-NAACL. (2006)
30. He, Y., Zhou, D.: Self-training from labeled features for sentiment analysis. Information Processing & Management **47** (2011)
31. Baroni, M., Bernardini, S.: Bootcat: Bootstrapping corpora and terms from the web. In: Proceedings of LREC. (2004)
32. Lecorvé, G., Gravier, G., Sébillot, P.: On the use of web resources and natural language processing techniques to improve automatic speech recognition systems. In: Proceedings of LREC. (2008)
33. Gadet, F.: La variation, plus qu'une écume. Langue française (1997)
34. Gadet, F.: Is there a french theory of variation? International Journal of the Sociology of Language **165** (2003)
35. Bilger, M., Cappeau, P.: L'oral ou la multiplication des styles. Langage et société (2004)
36. Ilmola, M.: Les registres familier, populaire et vulgaire dans le canard enchaîné et charlie hebdo: étude comparative. Master's thesis, University of Jyväskylä, Finland (2012)