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

Jade Mekki[1,2] **Gwénolé Lecorvé**[1,3] **Delphine Battistelli**[2] **Nicolas Béchet**[4]
[1]Univ Rennes, CNRS, IRISA / Lannion-Vannes, France
[2]Universtité Paris Nanterre, CNRS, MoDyCo / Nanterre, France
[3]Orange Labs / Lannion, France
[4]Université de Bretagne Sud, CNRS, IRISA / Vannes, France
firstName.lastName @ {irisa.fr, orange.com, parisnanterre.fr}

Abstract

The casual, neutral, and formal language registers are highly perceptible in discourse productions. However, they are still poorly studied in Natural Language Processing (NLP), especially outside English, and for new textual types like tweets. To stimulate research, this paper introduces a large corpus of 228,505 French tweets (6M words) annotated in language registers. Labels are provided by a multi-label CamemBERT classifier trained and checked on a manually annotated subset of the corpus, while the tweets are selected to avoid undesired biases. Based on the corpus, an initial analysis of linguistic traits from either human annotators or automatic extractions is provided to describe the corpus and pave the way for various NLP tasks. The corpus, annotation guide and classifier are available on http://tremolo.irisa.fr.

1 Introduction

Language registers are of particular interest in (socio-)linguistics because (1) they are a highly perceptible characteristics of discourse productions ; (2) they represent a significant source of information about the writer/speaker, the relationship between interlocutors, or other elements of the communication context ; (3) they are a concept known to all (advantage when running perceptual tests). Among the possible perceptions of this phenomenon, the partitioning into casual, neutral, and formal registers is probably the most used as it is found in many situations of everyday life. While corpora like GYAFC—where these variations are referred to as "formality level"—have recently popularized the domain (Rao and Tetreault, 2018), it is still poorly studied overall in Natural Language Processing (NLP), especially outside English. Moreover, current work largely focuses on textual types for which registers are already known from the linguistic literature[1] whereas many new types, with their peculiarities, arise from the *Computer-Mediated Communications* (*CMCs*) (e.g., SMS, tweets...). Therefore, the analysis of CMC corpora in terms of language registers is a challenge both in terms of descriptive linguistics and applications in NLP.

As part of the TREMoLo project focusing on language registers[2], this paper tries to go beyond these limits and presents the corpus TREMoLo-Tweets, gathering 228,505 tweets (6M words), in French, with multi-label annotations among the *casual*, *neutral* and *formal* registers. The annotations come from a CamemBERT (Martin et al., 2020) model fine-tuned on a manually annotated subset (a.k.a. *seed*) of the whole corpus.

After a state of the art in Section 2, the corpus creation is presented in Section 3. Then, Section 4 provides first linguistic conclusions derived from statistics on manually and automatically-derived linguistic traits. Finally, possible tasks opened by the proposed corpus are listed in the conclusion.

2 Background and Motivation

Notion of registers. In sociolinguistics, the notion of language registers refers to the linguistic varieties associated with particular communication situations (Todorov, 2013). A key idea is that a language register can be characterized by specific patterns (Ferguson, 1982; Ledegen and Léglise, 2013). While the use of *"level"*, *"style"* or *"genre"* co-exist (Gadet, 1996; Bourquin, 1965; Joos, 1967), the term *"register"* tends to prevail (Biber, 1991; Sanders, 1993; Ure, 1982). Based on these points, we use the term *"register"* defined as a variation of linguistic forms, at different levels of analysis

[1] For instance, classically, insults are associated with the casual register while long sentences with subordinates are associated with the formal one.

[2] https://tremolo.irisa.fr

Proceedings of Recent Advances in Natural Language Processing, pages 950–958
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_108

of the language, with respect to a given standard. This standard corresponds to the intersection of an "objective norm" (the grammatical rules) and a "subjective norm" (the rules of actual usage) (Gadet, 2007). Following this definition, a text is considered as formal when it completely conforms to the objective and subjective norms, neutral when it partially conforms to both, and casual when the objective norm is not followed.

Related work. In (Biber and Conrad, 2019; Biber, 1991), the use in the corpus of *a priori* defined linguistic features is quantitatively studied according to different axes: oral *vs.* written, formal *vs.* informal, etc. The purpose is to identify feature co-occurrences according to these axes.

For English, (Peterson et al., 2011; Pavlick and Tetreault, 2016) propose techniques to classify texts into formal *vs.* informal from a corpus of emails while (Sheikha and Inkpen, 2010) uses regression to predict a level of formality from a corpus of formal/informal texts.

For French, in (Lecorvé et al., 2019), the authors jointly study a classification task and the construction of a corpus of web pages[3] annotated using an iterative semi-supervised approach.

The quality of the previously mentioned annotated corpora can be questioned from the perspective of language registers because (1) the composition of these corpora shows different biases by mixing text types or restricting the topics to a particular domain, (2) the manual annotations do not follow an annotation guide. In this paper, we propose to address these issues by (1) only focusing on tweets with a large range of domains, and (2) providing an annotation guide that is grounded on a linguistic analysis of language registers and CMCs.

Why choosing tweets? The constitution of a corpus of written texts representative of the real use of language registers presents two major difficulties. First, the strong link between some registers and some types of texts (e.g., the formal register associated with novels of classical literature, the casual register with discussion forums, and the neutral one with journalistic dispatches). Second, the oral and written modalities are intuitively associated with the casual and formals registers, respectively (Gadet, 2000; Rebourcet, 2008). To address these issues, CMCs—which are defined as "*any human communication that occurs through the use of two or more electronic devices*" (McQuail, 2010)—were chosen as their instantaneous nature can cause a "spoken-writing" style (or so-called "*parlécrit*" in French ; Jacques (1999)). More precisely, Tweets are selected since they are CMCs and have a 280-characters limit, imposed by Twitter, which homogenizes the framework. The rather short length of tweets also prevents from texts where several registers could be present but not mixed (i.e, two distinct portions of a long text).

3 Corpus Creation

The corpus TREMoLo-Tweets is drawn from tweets collected in such a way as to cover the targeted spectrum of language registers while minimizing some unwanted biases. After various filterings and cleanings, a subset of these tweets was manually labeled. From a portion of these labeled tweets (training set), a CamemBERT classifier was fine-tuned to generalize the labels to the whole corpus. The result was validated on another part of the manually annotated tweets (test set). This section details the collection of the tweets, the labeling process, and the experimental validation.

3.1 Collection of the Tweets

Tweets have been collected by submitting queries to the Twitter API. Hence, the design of these queries is a key aspect. Here, the chosen strategy relies on the trending topics—which are the most used hashtags at a given time. Since they refer to striking events that are commented on by many users, we believe that they cover many different language functions and registers. Moreover, the varied nature of these events leads to equally diverse domains, which should enable to separate the notions of register and topic. In complement, the tweets were restricted to an unique geographical area (Paris) to minimize the impact of potential dialects. Tweets were collected on 10 different dates over a period of one month (August, 2020). For each date, $2,000$ tweets were retrieved on average for each of the 50 top trending topics on that day.

Non-French tweets were removed using the `langdetect` Python library. Tweets with a probability < 0.9 for French were discarded. This arbitrary value is fixed in order to keep texts with the presence of some interesting non-French terms (e.g.,"*lol*", "*stan*"). Truncated tweets were removed by spotting the "horizontal ellipsis" characters. The

[3] 400,000 web pages collected from queries composed of casual, neutral and formal lexicons.

corpus counts 228,505 tweets (6,201,339 words). It has been standardized by the CamemBERT tokenizer, and morphosyntactically annotated by Talismane[4] (Urieli and Tanguy, 2013).

3.2 Labeling of a Seed

Out of the entire corpus, 4,000 tweets have been randomly selected to be manually annotated in language register (named the *seed*). In the remainder, these tweets are referred to as the *seed*. Possible labels are the targeted registers (casual, neutral, formal) and an extra one to identify tweets that are badly encoded or incomprehensible. Multiples labels can be given to one tweet to reflect the copresence of several registers. The objective is that each tweet of the *seed* is annotated with a degree of belonging for each of the 4 considered classes (i.e., summing to 100,%).

Annotation guide. An annotation guide is built to frame the annotators' work and, hence, the consistency of the final corpus. To do so, it defines the considered registers, following the principles outlined in Section 2, and gathers a set of linguistic descriptors interesting for the analysis of this corpus. These descriptors (detailed in Section 4.2) reflect peculiarities from the literature about language registers as well as CMCs. It is important to highlight that the annotation guide does not link the descriptors with specific registers. This is just a way to suggest potentially interesting aspects to be looked at. The annotator must then justify her/his labeling by selecting at least one of these linguistic descriptors. This annotation guide is given in the supplementary material (in French).

Manual annotation. The labeling of the seed has involved 4 experts[5] such that each tweet has been annotated by 2 of them. For a given tweet, each annotator must indicate which register(s) (at least one) is (are) present and rank them according to their predominance[6]. These choices had to be justified by at least one descriptor from the annotation guide. These annotations are released with the corpus, and their analysis is provided in Section 4.

In a post-processing phase, rankings are converted into degrees of belonging. For a given tweet, let R denote the set of registers r reported as present, $rank(r)$ the rank of each of them, and its backward counterpart as $rank^{-1}(r) = 1 + Card(R) - rank(r)$. Then, the degree of r is defined as the backward rank normalized by the sum of all ranks, i.e.:

$$degree(r) = \frac{\sum_{r \in R} rank^{-1}(r)}{\sum_{1 \leq i \leq Card(R)} i} \qquad (1)$$

To illustrate this conversion, let one consider that a tweet labeled with the neutral register as rank 1, and casual as rank 2. Then, the resulting degrees would be $\frac{2}{3} = 67\%$ and $\frac{1}{3} = 33\%$ for the neutral and casual registers, respectively. The degree would be 0 for the two others (formal and bin).

Agreement/disagreement between annotators. Given that all tweets are annotated by 2 experts, only those which are proposed by both of them are considered, and their degree is the average of the degrees from each annotator. In 976 tweets, the 2 annotators totally disagree (i.e. no shared label). Then, a third annotation is done by a new external annotator, and a given label is kept as soon as 2 annotators out of the 3 propose it. If no agreement can still be found for some tweets, they are discarded. Finally, 3,269 tweets remained.

Overall, the agreement between annotators is more significant for the casual and neutral registers (73% and 76%, respectively) than for the formal register (36%). Regarding the bin register, the agreement is perfect (100%). In detail, it appears that (1) for the casual register most of the divergences are with the neutral register, (2) for the neutral register with the casual register, and (3) for the formal register with the neutral register.

Overview of the annotated seed. The results of the manual annotation are dominated by the neutral register (51% of the seed, i.e., 1,698 tweets), followed by the casual (39%, i.e., 1,345 tweets), the formal (10%, i.e., 340 tweets), and finally the bin (almost 0%, i.e., 18 tweets). On average, when a label is present, its degree is high: 87% for the neutral register, 92% for the casual register, 87% for the formal register, and 98% for the bin register. Only 131 tweets have at least 2 registers present, against 3,138 with a single register. Even if the agreement policy is playing a role, this result shows that the tweets are not very nuanced in terms of registers. The short length can explain this phenomenon.

[4] The accuracy is 88.5% on the French TreeBank.

[5] Ph.D. students or researchers from CMCs or NLP.

[6] Equal ranks are permitted.

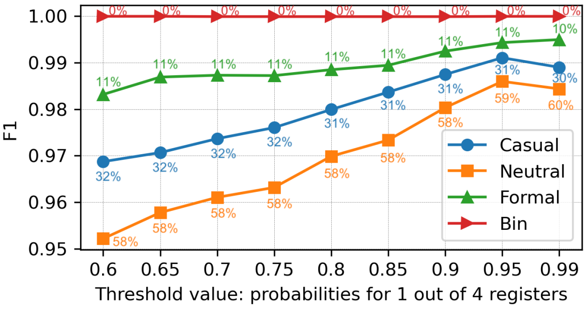

Figure 1: $F1$ and proportion for each register on the test set of the seed after data augmentation with various values of T_1.

3.3 Automatic labeling of the whole corpus.

To label the full corpus in registers, a CamemBERT model[7] is chosen to perform multi-label classification. In a first time, this model is fine-tuned on the sole seed. fine-tuned on 90% of the manually labeled seed. The idea is to use this model to label the whole corpus, and select some of these automatically annotated tweets in order to augment the model's training data. Then, a new fine-tuning is performed and the whole corpus is definitely labeled. This section focuses on data preparation, the filtering strategy, and the experimental results.

Multi-label classification from the seed. Multi-label classification is preferred to multi-dimensional regression to help the model distinguish strong differences between the registers. To do so, the degrees of belonging to each register are converted into binary labels. Tweets are labeled with a given register if and only if the associated degree is greater than or equal to 50%. The model is fine-tuned of 90% of the seed while the remaining 10% are for the evaluation. The model is the CamemBERT *base* version. Fine-tuning is performed with learning rate of 10^{-4} and during 8 epochs. As a result, F1 values obtained for the casual, neutral, formal, and bin classes are 0.85, 0.84, 0.95, and 0.99, respectively.

Training data augmentation. To improve the performance, the training set from the seed is augmented by selecting automatically labeled tweets from the non-seed part of the corpus (i.e., the other remaining tweets). This is implemented by filtering the tweets for which one of the predicted labels is a probability greater than a threshold T_1. The label probabilities of the selected tweets are then binarized in the same way as the seed, and a new fine-tuning is performed based on the augmented training data. Figure 1 shows the F1 values after this second training. These results demonstrate that data augmentation is worth it, and rather robust across the values for T_1 (all values range in $[0.95, 1]$). Best values for T_1 seem to range in $[0.9, 0.99]$. Percentages below each point refer to the proportion of the tweets labeled for each register in the whole corpus. It appears that data augmentation did not really change these proportions compared to the distribution in the seed.

To deepen this study of robustness, another series of experiments was conducted to test the number of new samples required to efficiently perform data augmentation. This is noted as another threshold T_2 on the number of tweets added. To do so, label probabilities provided by the initially trained classifier are sorted in descending order. The conclusions from these experiments are that 6% (about $14,000$ tweets) of the whole corpus is enough to obtain good labeling results.

These various results and their stability, which follow the trend in the manual labels, tend to indicate that the final labels of the whole corpus are of good quality.

4 Linguistic analysis

This section presents a first linguistic analysis of the corpus TREMoLo-Tweets in terms of language registers. After presenting the underlying linguistic descriptors, this section reports which of these descriptors have mostly been selected by the annotators, and how they appear on the whole corpus using systematic automatic extractions.

4.1 Linguistic descriptors

A set of 47 linguistic descriptors is made by updating a list, from a study that has already identified features in the scientific literature on language registers (Mekki et al., 2018), with those specific features to CMCs (examples Table 1). Among the 47 descriptors: 15 are syntactic, 9 morphological, 9 lexical, 6 discursive, 5 lexico-syntactical, and 3 phonological. They were chosen to help the annotators make their decision, and assert that their labeling is motivated. Among them, some elements are specific to tweets : (Paveau, 2013) calls them "*technomorphems*"[8]. One of our contributions is to integrate them instead of discarding them as in (Go et al., 2009; Pak and Paroubek, 2010; Agarwal et al., 2011). The main technomorphems are:

[7] Simpler models were tried but obtained lower accuracy.

[8] Forms that arise from digital discourses.

ID	Linguistic descriptors	Casual	Neutral	Formal
1	Absence of classic final punctuation	**63%**	40%	3%
2	Idiomatic expression	**41%**	12%	9%
3	Absence of an expected item	**35%**	16%	1%
4	Modalizing expression	**33%**	23%	11%
5	Electronic spelling	**27%**	2%	0%
6	Contiguous repetition of items	**25%**	5%	1%
7	Shortening of words	**23%**	9%	2%
8	Foreign language borrowing	**19%**	8%	3%
9	Removal of certain letters due to elision or apocope	**17%**	2%	0%
10	"*ça*" preferred to "*cela*"	**16%**	10%	2%
11	Interjection	**15%**	2%	0%
12	Insult	**14%**	1%	1%
13	Onomatopoeia	**11%**	1%	0%
14	Character repetition	**11%**	1%	0%
15	"*il*" replaced by "*y*"	**9%**	0%	0%
16	Capital letters used outside their conventional usage	**9%**	3%	1%
17	Verb "*aller*" for the construction of the future tense	**8%**	7%	1%
18	Discriminative termination	**7%**	1%	1%
19	*Verlan* (i.e. reversing the terms syllabe by syllabe)	**6%**	0%	0%
20	"*est-ce que*" for interrogatives sentences	6%	**19%**	0%
21	"*tu*" preferred to "*vous*" / "*on*" preferred to "*nous*"	12%	**14%**	4%
22	Present as the only tense used	3%	**13%**	3%
23	Absence of classical punctuation	12%	**13%**	1%
24	No subject/verb inversion in an interrogative sentence	7%	**10%**	2%
25	Hashtag syntactically independent	3%	**9%**	4%
26	Hashtag with no syntactic relation	1%	**7%**	5%
27	Doubled element	5%	**6%**	4%
28	Diversity of verbal tenses	2%	25%	**62%**
29	Several sentences with classical punctuation	2%	25%	**56%**
30	Hashtag syntactically integrated	5%	19%	**49%**
31	Presence of the double negation	1%	18%	**41%**
32	"*vous*" preferred to "*tu*" / "*nous*" preferred to "*on*"	4%	9%	**39%**
33	Speech citation	2%	8%	**38%**
34	Text structured by punctuation	1%	18%	**28%**
35	Mention of the user's identifier integrated in a phrase	3%	12%	**26%**
36	Presence of subject/verb inversion	0%	3%	**20%**
37	Diversity of logical connectors	0%	4%	**20%**
38	Pictogram that highlights information	1%	8%	**17%**
39	Pictogram in the replacement function	1%	3%	**6%**

Table 1: **Ratio of usage of each linguistic descriptor in the justifications of annotators when manually labeling the seed. Descriptors with all ratios lower than or equal to 5% are not reported.**

the *hashtags*, and the *pictograms*.

Hashtags are defined as one or more contiguous words preceded by a # sign (e.g., "#MerryChristmas"). Some typologies of hashtags emphasize their indexing function (Jackiewicz and Vidak, 2014) (e.g., "#Tokyo2020"). In addition to this, we assume that their syntactic integration, that is their use as a standard lexeme, also brings variety to the language registers.

Pictograms refer to both "emoticon"[9], and "emoji"[10]. The (Magué et al., 2020)'s typology of 3 functions has been used and adapted to our analysis: (1) the replacement function (when a pictogram replaces a syntagm); (2) the illustration function (when it has a referential function); (3) the modalization function, (when it indicates the emotion or the cognitive position of the author wrt to his/her statement). Then a 4^{th} function has been added: (4) the framing/structuring function (when a pictogram surrounds or points at information).

4.2 Human justifications on the seed

For a given tweet, a descriptor is manually selected by an annotator when it mainly motivates the attribution of a register.

The casual register seems to be marked by the absence of classical punctuation (#1), idiomatic expressions (#2), and modalization expressions (#4). Here, the expressive role of the absent punctuation seems to have been taken over by other linguistic objects (e.g., pictograms).

The neutral register is marked by the absence

[9] Graphic signs looking 'like' an emotion (Beccucci, 2018)

[10] Symbols listed in a database (ibid.)

ID	Linguistic descriptors	Casual	Neutral	Formal
1	Absence of classical final punctuation	**65%**	40%	12%
35	Mention of the user's identifier integrated in a phrase	**40%**	35%	23%
11	Interjection	**29%**	15%	13%
21	"*tu*" preferred to "*vous*" / "*on*" preferred to "*nous*"	**24%**	15%	8%
31	Presence of the double negation	**23%**	18%	22%
14	Character repetition	**16%**	10%	8%
18	Discriminative termination	**9%**	7%	8%
15	"*il*" replaced by "*y*"	**6%**	2%	2%
5	Electronic spelling	**7%**	1%	1%
4	Modalizing expression	20%	**22%**	**22%**
28	Diversity of verbal tenses	96%	89%	**98%**
37	Diversity of logical connectors	39%	38%	**58%**
29	Several sentences with classical punctuation	17%	28%	**57%**
33	Speech citation	30%	25%	**48%**
8	Foreign language borrowing	37%	37%	**41%**
16	Capital letters used outside their conventional usage	22%	31%	**36%**
36	Presence of subject/verb inversion	25%	20%	**30%**
32	"*vous*" preferred to "*tu*" / "*nous*" preferred to "*on*"	12%	11%	**23%**
25	Hashtag syntactically independent	6%	12%	**17%**
34	Text structured by punctuation	4%	9%	**15%**
30	Hashtag syntactically integrated	3%	6%	**11%**
26	Hashtag with no syntactic relation	3%	6%	**11%**

Table 2: **Presence of each linguistic descriptor in the whole corpus using automatic and systematic extraction. Descriptors with all ratios lower than or equal to 5% are not reported.**

of classical punctuation (#1), the diversity of the verbal tenses (#28), and the presence of several sentences with classical punctuation (#29). The neutral register seems less clear-cut (notably with different uses of the punctuation marks).

The formal is also characterized by the presence of several sentences with classical punctuation marks (#29), the syntactic integration of hashtags (#30), and pictograms that highlight information (#38). Therefore, technomorphems in the formal registers show that CMC-specific items have been integrated into the French standard.

4.3 Automatic extraction on the whole corpus

In order to analyze the whole corpus, symbolic rules were implemented to automatically spot the presence of the linguistic descriptors. Let one note that 5 descriptors could not be implemented since they refer to complex notions. This automatic extraction is not selective (all descriptors present are taken from the tweet) unlike the manual extraction from the seed which is selective (only descriptors that contribute the most to the register are taken from the tweet). The overview of these exhaustive extractions is provided in Table 2.

To characterize a register r according to the other registers (noted o), the importance of each descriptor noted d observed in r is measured by computing a growth rate (GR) as the ratio between relative frequencies of d in r as opposed to o:

$$GR(d,r,o) = \begin{cases} \infty, & iff_o(d) = 0 \\ \frac{f_r(d)}{f_o(d)}, & otherwise\ , \end{cases} \tag{2}$$

where f_x denotes the relative frequency in a register as reported in Table 3. The relative frequencies for the register "other" (f_o) is computed by merging all tweets that are not of register r. If $GR(d,r,o) > 1$, d is considered as emergent.

Table 3 reports for each register the descriptors with the highest growth rates. Interestingly, some rare descriptors appear whereas they were previously skipped in Table 2 ($\leq 5\%$). Then, it appears that the growth rates are lower for the neutral register than for the casual and the formal ones. This shows the fuzzy limits with the other registers. On the contrary, the casual register has high values, which means that it is characterized by unambiguous specific traits. The presence of *technomorphems* in the emergent descriptors, for the casual and formal registers, confirms the integration of the Twitter-specific elements to the French standard. However, they are used differently by register.

For the neutral register, a commercial application uses hashtag indexing functions:

Le jeu #MonstrumGame de @X sort [...] le 23 octobre (*@X's #MonstrumGame comes out [...] on October 23rd*).

The pictogram seems to replace classic punctuation marks at the end of the sentence:

ID	Casual	*GR*	C vs. Others	Example (*translation*)
5	Electronic spelling	7.00	7.00% / 1.00%	Ha allez ooooh !!! (*Eh lollll*)
15	"*il*" replaced by "*y*"	3.00	6.00% / 2.00%	Y'en a le 25 (*Thr's some the 25*)
40	Pattern "*juste*"	2.50	0.50% / 0.20%	Juste comme ça (*Just like that*)
	Neutral	*GR*	**N vs. Others**	**Example (*translation*)**
3	Absence of an expected item	1.50	0.10% / 0.07%	ils ø vont quand même pas (*Still not*)
16	Capital letters used outside their conventional usage	1.04	31.00% / 29.00%	17 juillet pour OM DÉVELOPPEMENT (*July 17 for OM DEVELOPMENT*)
25	Hashtags syntactically independent	1.04	12.00% / 11.50%	[...] . #MondayMotivation
	Formal	*GR*	**F vs. Others**	**Example (*translation*)**
38	Pictogram that highlights information	7.33	2.20% / 0.30%	🔴 #X banni de #Facebook (🔴 *#X banned from #Facebook*)
34	Text structured by punctuation	2.14	15.00% / 7.00%	VIDEO. Crise des transports : (*VIDEO. Traffic crisis :*)
30	Hashtag syntactically integrated	2.20	11.00% / 5.00%	les #ViolencesPolicieres ne sont pas (*#PoliceViolences aren't*)

Table 3: **Top 3 automatic descriptors w/ highest growth rate for each register against the others in the whole corpus.**

> @X @X Je pense que ça fait référence à des dates de sortie 🤔 En septembre ça tombe sur des Vendredi (*I think it refers to release dates* 🤔 *In September it falls on Fridays*)

For the formal register, hashtags are syntactically integrated:

> Les violences vécues en #France ne sont pas des #incivilites (*Violence experienced in #France aren't #incivilities*)

The pictograms are used with their framing/structuring function which brings a kind of verticality to the tweet:

> 🔥 L' #Amazonie brûle ! 🔥 Partout en France, les citoyen.nes aux côtés de @X demandent des actes au gouvernement Macron [...] (🔥 *The Amazon is burning !* 🔥 *Everywhere in France, citizens alongside @X demand actions from the Macron government [...]*)

The casual register seems more used for dialogue between users and pictograms are used for their modalization function to provide extra-linguistic information: they compensate for the lack of paraverbal information such as prosody.

> "@X La France part en couille et 1 autre con jardine au Liban 😡 😡" (*@X France is going to the dogs and the other idiot is gardening in Lebanon* 😡 😡)

Moreover, marks of orality are found:

> @X Bah là j'ai pas encore test le son mais en tout cas niveau confort y'a pas photo [...] (*Bah there I did not test the sound yet but in any case level of comfort th's not photo [...]*)

Thus, these first analyses highlight the corpus quality, and the relevance of the set of linguistic descriptors for CMC data. Likewise, the analysis of the registers identifies different linguistic functions on Twitter (argumentative, commercial, conversational speech).

5 Conclusion

In this paper, we presented the corpus TREMoLo-Tweets which gathers 228,505 tweets labeled in casual, neutral and formal registers. For this purpose, a seed was manually annotated with multiple labels, following an annotation guide derived from a linguistic analysis of the corpus. Using a CamemBERT model and data augmentation, the whole corpus is entirely labeled with an experimentally demonstrated high reliability. Furthermore, statistics on linguistic descriptors are reported to demonstrate the richness of the corpus.

The labels, linguistic descriptors, and the large size of the corpus pave the way to future tasks:

- **Standard NLP tasks on a seldomly studied style factor.** Classification (predicting the registers of a given text) and natural language generation (style transfer).

- **Data mining.** How to rank the descriptors to discriminate the registers against each other? How to reconstruct the features of interest from the raw words? Justifications given by the annotators can be used as a reference.

- **Interdisciplinary work.** Discovery of new fundamental knowledge about language registers by crossing NLP and sociolinguistics, like in (Abitbol et al., 2018) where study the linguistic variations are studied according to the writers' geographical areas and economic social status.

Acknowledgements

This work was supported by the TREMoLo project (ANR-16-CE23-0019) of the National Research Agency (ANR).

References

Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference*, pages 1125–1134.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38.

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

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

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

Charles A Ferguson. 1982. Simplified registers and linguistic theory. *Exceptional language and linguistics*, pages 49–66.

Françoise Gadet. 1996. Niveaux de langue et variation intrinsèque. *Palimpsestes*, 10.

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

Françoise Gadet. 2007. *La variation sociale en français*. Editions Ophrys.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Agata Jackiewicz and Marko Vidak. 2014. Étude sur les mots-dièse. In *shs Web of Conferences*, volume 8, pages 2033–2050. EDP Sciences.

Anis Jacques. 1999. Internet, communication et langue française.

Martin Joos. 1967. *The five clocks*, volume 58. New York: Harcourt, Brace & World.

Gwénolé Lecorvé, Hugo Ayats, Benoît Fournier, Jade Mekki, Jonathan Chevelu, Delphine Battistelli, and Nicolas Béchet. 2019. Towards the automatic processing of language registers: Semi-supervisedly built corpus and classifier for french.

Gudrun Ledegen and Isabelle Léglise. 2013. Variations et changements linguistiques.

Jean-Philippe Magué, Nathalie Rossi-Gensane, and Pierre Halté. 2020. De la segmentation dans les tweets: signes de ponctuation, connecteurs, émoticônes et émojis. *Corpus*, (20).

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Denis McQuail. 2010. *McQuail's mass communication theory*. Sage publications.

Jade Mekki, Delphine Battistelli, Gwénolé Lecorvé, and Nicolas Béchet. 2018. Identification de descripteurs pour la caractérisation de registres. In *Proceedings of Rencontres Jeunes Chercheurs (RJC) of the CORIA-TALN conference*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Marie-Anne Paveau. 2013. Genre de discours et technologie discursive. tweet, twittécriture et twittérature. *Pratiques. Linguistique, littérature, didactique*, (157-158):7–30.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics*, 4(1).

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Languages in Social Media*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Séverine Rebourcet. 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.

Carol Sanders. 1993. *French today: language in its social context*. Cambridge University Press.

Fadi Abu Sheikha and Diana Inkpen. 2010. Automatic classification of documents by formality. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*.

Tzvetan Todorov. 2013. *Mikhaïl Bakhtine. Le principe dialogique. Suivi de: Ecrits du Cercle de Bakhtine*. Le Seuil.

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

Assaf Urieli and Ludovic Tanguy. 2013. L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions: études de cas avec l'analyseur talismane. In *20e conférence du Traitement Automatique du Langage Naturel (TALN)*, pages publication–en.