
Automatic characterization of French language registers: illustration on tweets

Jade Mekki* — Nicolas Béchet** — Delphine Battistelli*** — Gwénolé Lecorvé*, ****

* Univ Rennes, CNRS, IRISA, Lannion-Vannes, France

** Univ Bretagne Sud, CNRS, IRISA, Vannes, France

*** Univ Paris Nanterre, CNRS, MoDyCo, Nanterre, France

**** Orange Research, Lannion, France

ABSTRACT. *This article presents a methodological approach to automatically characterize language registers in French. The method of emerging sequential patterns is described and the obtained results demonstrated first on a corpus of tweets, then more broadly on registers in French. As both a premise and a result of the present approach, the definition presented in this paper for the notion of a language register highlights the notion of a linguistic norm.*

KEYWORDS: *Language register, French, linguistic norm, sequential emergent patterns, tweets.*

TITRE. *Caractérisation automatique de registres en français : illustration dans un corpus de tweets*

RÉSUMÉ. *Cet article présente notre méthodologie pour caractériser automatiquement les registres de langue en français. Nous décrivons la méthode des motifs séquentiels émergents utilisée à cette fin et montrons les résultats obtenus sur un corpus de tweets ainsi que, de manière plus générale, sur les registres en français. À la fois prémisse et résultat de notre approche, notre définition de la notion de registre de langue met l'accent sur celle de norme linguistique.*

MOTS-CLÉS : *registre de langue, français, norme linguistique, motifs séquentiels émergents. tweets,*

Introduction

A language speaker knows that there are usually multiple ways to express and convey information. This aspect intuitively recognized by speakers is part of a phenomenon known as *language registers* which are typically described with terms such as casual, formal, colloquial, etc.

This phenomenon is noticeable at different linguistic levels, notably at lexical and syntactic levels (e.g. *vinasse* (*plonk*) vs. *vin de mauvaise qualité* (*poor-quality wine*), *c'est dû à ...* (*due to*) vs. *cela est dû à ...* (*this is due to*)).

- (1) #MonPireDate J'ai pleuré parce que pensais à mon ex 😞 Désolé si vous lisez ça 😞 (#MyWorstDate I cried because I was thinking about my ex 😞 Sorry if you see this 😞)
- (2) @X Adama Traoré n'a jamais été condamné pour viol. Vos propos sont diffamatoires. (@X Adama Traoré was never convicted of rape. Your comments are defamatory.)

Textual materials such as tweets constitute a challenge for the analysis of language registers because they combine traditional and new register linguistic features. Thus, some interesting linguistic questions arise, for example: Is a tweet viewed as casual just because it includes (a lot of) pictograms, as in Example (1) ¹?

Several sociolinguistic studies examined language registers to understand which human frameworks (sociological, cultural, etc.) they correspond to. Other linguistic studies have examined how language registers can be identified within textual units. Whether in sociolinguistics or general linguistics, these approaches reveal four main limitations: (i) the data sets are typically too small for natural language processing tasks and do not allow for broad generalizations; (ii) different linguistic levels are rarely analyzed together despite the importance of the relationship between these levels; (iii) language registers are usually studied in isolation despite the fact that they are often identified by contrasting them with one another; and (iv) most research focuses on traditional media.

In this paper, we address these four limitations by using a pattern mining technique. We propose to detail one of our main methodological contributions, which is using a pattern mining technique. It consists in extracting emergent sequential patterns that capture different levels of linguistic analysis and that permit distinguishing registers two by two.

This paper is structured as follows: in Section 1 we present several works done previously on the characterization of language registers and lay out our proposition for the definition for this complex phenomenon. In section 2, we describe the way it is applied to a corpus of French tweets. Section 3 details the method for discovering emerging sequential linguistic patterns that we used to characterize three kinds of registers. Section 4 provides and discusses the obtained results.

1. In this paper, all the examples are anonymized with @X for the users and url_path for URLs.

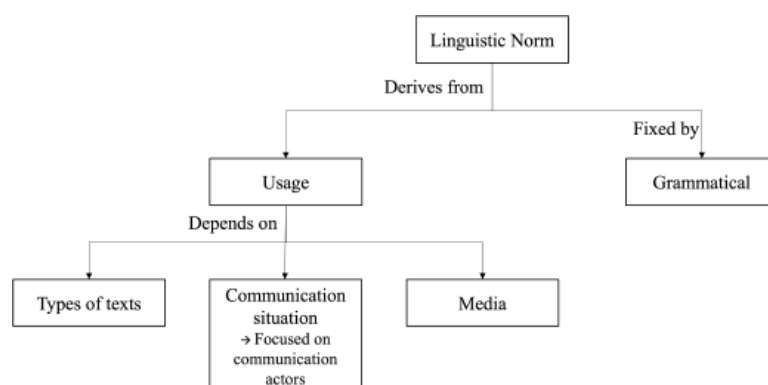


Figure 1. *Linguistic norms.*

1. The notion of language register and the role of linguistic norms

As Labov (1988) highlighted with the concept of *language variations* and, also Halliday (1985) before him with a focus on the question of variety of situations, language registers refer to the recognition of different ways to express the same idea. While language registers are intuitively identifiable, there is no unified definition in scientific literature. This lack of consensus partly stems from the influence of linguistic norms on how registers are defined.

A linguistic norm corresponds to a set of rules to follow. The content of these rules differs depending on the entity in charge of establishing them and on the idea of what a *good enough* language should be. Frei (1971) proposes a *grammar of mistakes* for French, meaning a set of linguistic features considered as errors. Examples include forms like *ils croyent* instead of *ils croient* or *il finissa* instead of *il finit*, which illustrate how speakers may create analogies or seek clarity in language use.

As a result, several types of norms exist in the literature. As compiled from three major works on French language (Gadet, 2007; Paveau and Rosier, 2008; Heller *et al.*, 2013), two types of norm can be distinguished: the *prescriptive norm* which edicts grammatical rules to be followed and the *objective norm* which derives rules from language usage.

Several works have focused on the concept of linguistic norm. In Gadet (1997), it is demonstrated that linguistic variations are about linguistic norms at various levels, such as: the phonological level (e.g. elision of *u*: *t'es mort* instead of *tu es mort*), the morphological level (non-standard word endings: *politicard*), the lexical level (borrowings from foreign languages: *je suis dead*) and the syntactic level (non-inversion of subject/verb in an interrogative sentence: *Tu vas bien ?*). In Mekki *et al.* (2018), a broad study establishes a state-of-the-art list of 72 various levels of linguistic features as they have already been identified in the literature about French registers.

The main limitations of these studies lie in the absence of a formalized significance criterion for validating whether a feature is characteristic of a register, as well as in the reliance on manual identification of such features. To address this, our work introduces the concept of *emergence* to formalize the significance criterion. Emergence involves comparing the frequencies of a given feature in texts from a target register with those from a source register. A feature is considered emergent if it appears more frequently in the target register than in the source register.

It should be noted that in the Anglophone corpus linguistics literature, the term *register* came into use mainly through the work of Douglas Biber from Biber (1991) to Egbert *et al.* (2022), in addition to Halliday's work (Halliday and Hasan, 1989). In his work, Biber defined a register as "a linguistic variety associated with a particular situation of use (including particular purposes of communication)" (Biber and Conrad, 2019). The emphasis on context found in the definitions by Ferguson (1982) and Ure (1982) is also present in Biber's approach. To study registers, Biber quantitatively observed the variation of certain manually selected linguistic features in a large corpus along different axes: oral/written, formal/informal, etc. Its goal is to identify co-occurrences of linguistic features along these axes. For example, in Biber and Conrad (2019), one of the analyses focuses on the behavior of linguistic features in newspapers and academic papers. A higher presence of the nominalization phenomenon is observed in academic papers, and a higher presence of attribute adjectives in newspapers.

From a methodological perspective, our approach diverges from Biber's by employing sequential pattern extraction with no prior assumptions about the linguistic features to be analyzed. In contrast, as noted by Branca-Rosoff (1999) and Poudat and Landragin (2017), Biber manually selected specific features for comparison, often without explicit justification. By not presetting which features are relevant and by considering all levels of linguistic analysis simultaneously, our methodology avoids the biases associated with manual feature selection and allows for the discovery of emergent patterns within the data.

Our definition of language registers. In an automatic approach, we consider categories of registers rather than a continuum as in Gadet (1996). To define language registers, we pursue a complementary approach that integrates both *prescriptive* and *objective* norms, as previously introduced. Our definition of language registers is grounded in this way (see Figure 1). We propose to consider a text as belonging to one of three distinct registers defined as follows: **casual register**, when either the grammatical or the usage norm or both are not followed; **standard register**, when a textual unit partially conforms to the grammatical and usage norms; **formal register**, when the textual unit completely conforms to the grammatical and usage norms. Tweets (3) to (5) illustrate these three types of language registers.

- (3) Bosh il lui a fait un sal boulot à kaaris il lui a montré que maintenant **y’a** plus de petit et grand (*Bosh did a great job for kaaris, he showed him that now there’s no such thing as small and big.*)²
- (4) [12:36 PM] Il a osé me frapper. **Il se rend pas** compte. ([12:36 PM] *He dared to hit me. He doesn’t realize.*)
- (5) [#LeSaviezVous] Le ministère a confirmé que les résultats de la C.L.A.S. électorale qui s’est tenue le 4 mars 2020 sont valides. La responsable #UNSA Police Yvelines est la nouvelle vice-présidente de la CLAS78 ! **url_path** ([#DidYouKnow The Ministry has confirmed that the results of the elective C.L.A.S. held on March 4, 2020 are valid. The #UNSA Police Yvelines leader is the new vice-president of CLAS78! url_path])

Tweet (3) is perceived as casual. It does not follow the usage norm since no tweet-specific elements (*i.e.* hashtags) are used. It does not follow grammatical norms, with misspellings such as *y’a* for *il y a*. Tweet (4) is perceived as standard because it respects the usage norm with the use of a hashtag, without fully respecting the grammatical norm with the incomplete form of the negation *Il se rend pas*. Tweet (5) is perceived as formal, as it respects the two norms: the norm of usage, with the use of hashtags to index its content and the insertion of URLs to link to additional informative content; and the grammatical norm, in a perfect manner without deviation.

2. Our approach: corpus and methodology

This section presents the corpus we explored and the main steps of the methodology we followed to characterize the three kinds of language registers defined above.

2.1. Corpus

In order to investigate register-specific linguistic patterns in French, we have to look for a corpus adapted to this purpose. Different works involving French (Lecorvé *et al.*, 2018; Mekki *et al.*, 2021b) have proposed corpora illustrating registers. However, the association in these works between text types and contained registers distorts this illustration: casual register and forum posts, standard register and press, formal register and literature. To avoid this bias, we chose a single type of text: tweets. Their short format allows us to have the same unity when labeling them into registers and when characterizing each register. For this reason, we built up a large corpus of tweets whose automatic labeling is learned and then generalized from a manually annotated sub-corpus called *the seed*. This work has been published in Mekki *et al.* (2021b), so we’ll be more succinct on this step. The corpus and its annotation guide are available to the scientific community³. In total, the corpus comprises 228,270 tweets, for 6,201,339 words.

2. All the tweets are translated from French to English, the translations are intended to transcribe meaning, not exact expressions.

3. <https://hal.science/hal-03218217/>.

The manual annotation protocol has two distinctive points: it is based on a ranking system that hierarchizes the presence of registers in the same tweet and it integrates linguistic elements specific to tweets (such as user identifiers, hashtags, URLs, or pictograms) instead of discarding them (Agarwal *et al.*, 2011; Pak and Paroubek, 2010; Go *et al.*, 2009). Each time the annotator assigns a rank⁴, it must be justified by the presence of at least one feature from the list presented in our annotation guide⁵. Each rank r is then transformed into a register proportion. For a text annotated with r_1 =casual, r_2 =formal and r_3 =standard, we obtained casual 50% (3/6), formal 33% (2/6) and standard 17% (1/6).

To manually label a sub-corpus for use as an initial dataset (*a seed*), 4,000 tweets have been randomly selected from the corpus of tweets. Each tweet has been annotated by two expert annotators. Only labels that were present in the intersection of the two annotations were retained. The final annotation is the average of the A1 and A2 annotations (i.e. the mean of the register proportions). In the end, 3,269 manually annotated tweets are retained which represent 82% of the tweets initially selected to form the seed. The results of manual annotation are dominated first by the standard register (51% of the seed), then the casual (39%), and finally the formal (10%). To compensate for the small proportion of the seed with respect to the corpus of tweets to be labeled (1.4%), we take an iterative approach with several training cycles, where the seed (i.e. the training dataset) is increased with each training cycle by including the automatically labeled data once it is judged reliable. This approach is based on a semi-supervised learning technique. The classifier is learned by fine-tuning a pre-trained CamemBERT language model ("base" version) (Martin *et al.*, 2019) on our data. The quality of labeling is guaranteed by two indicators: the first is a quantitative measure, the F-scores (0.99 for the casual and formal registers, and 0.98 for the standard register); the second is the distribution of registers relatively similar to that of the manually annotated seed (59% standard, 31% casual, and 10% formal).

2.2. Global view of the processing chain

Figure 2 illustrates the entire processing chain. It starts with delimiting our research object (the French language registers) (A), then building a corpus of tweets to illustrate them (B), from which emerging sequential patterns are mined (C), before analyzing them and extracting new linguistic features characteristic of the registers (D). As (A) and (B) have already been published (Mekki *et al.*, 2018; Mekki *et al.*, 2021b), the next sections focus on step (C) in Section 3 and on step (D) in Section 4.

We aim to determine whether a feature distinguishes a target register A from a source register B by observing three constraints: discovering features that can be composed of various levels of language analysis; not making any *a priori* linguistic

4. Note that not assigning a rank means that the register is not present in the tweet.

5. The annotation guide, which details the protocol and the complete list of linguistic features, is available online: <https://hal.science/hal-03218217/>.

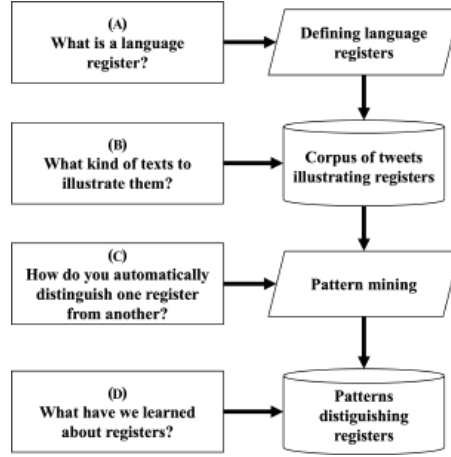


Figure 2. Main steps for characterizing French language registers.

assumption on the features to be extracted; mining a large dataset to obtain varied features. These sets of features can be used to better understand the emergence of new norms in tweets or as learning features for register prediction tasks on other types of text in natural language processing. To discover these features from the TREMoLo-Tweets corpus, we chose to use closed and emergent sequential pattern mining (Dong and Li, 1999), whose objective is the discovery of knowledge from contrasting datasets. Emergent sequential patterns discover regularities in sequential symbolic data.

Moreover, their formalization enables data to be represented by describing them via various linguistic features. While this emergent sequential pattern mining technique has advantages for our task, it also has four major drawbacks:

1) the reliability of results: without a truth base against which to compare the patterns extracted, how can we know whether they are interesting, *i.e.* truly relevant for characterizing an *A* register from a *B* register?

2) the exponential algorithmic complexity: for knowledge discovery, how can we reduce the search space (*i.e.* all possible patterns) without running the risk of missing interesting patterns?

3) the redundancy and abundance of discovered patterns: how to reduce the set of patterns while minimizing the exclusion of relevant patterns? In other words, how can we be sure to get only the most interesting patterns?

4) the manual selection of interesting patterns: how to automatically select interesting patterns without *a priori* assumptions on expected patterns? *i.e.* how to determine the most interesting motifs without deciding in advance what results to expect.

These disadvantages are not specific to the task of characterizing language registers, they are common to all pattern-mining approaches (Fournier-Viger *et al.*, 2017). One of our contributions is proposing a processing chain that overcomes these drawbacks one by one: for (i), by using an artificial language in which we knew which patterns were interesting; for (ii), we limited algorithmic complexity without reducing the search space by imposing expected pattern types, but by playing on the representation of tweets with a limited number of linguistic features; for (iii), to reduce the set of patterns discovered and automatically select interesting patterns; finally for (iv) we grouped similar patterns before automatically selecting a representative pattern per group. While our work on the reliability of results (i) is not extensively described in this paper⁶, we detail the work lifting locks (ii), (iii), and (iv) in the following sections.

3. Discovery of emerging sequential patterns characteristic of registers

In this section, we present our emerging sequential pattern mining protocol, which ensures acceptable algorithmic complexity. To keep complexity manageable, users constrain pattern mining in scientific studies. Constraining pattern mining means imposing criteria that patterns must meet to be extracted, such as containing a particular value or having a minimum or maximum length. In our case, we have chosen to illustrate tweets synthetically when converting them into sequences, to avoid reducing our search space (i.e. all patterns possible from a given set of texts) with constraints during pattern mining (i.e. the conditions that patterns must meet to be extracted) in order not to miss any interesting patterns. We begin by giving an overview of our approach, then detail the transformation of tweets into sequences, before successively introducing closed and emergent sequential pattern mining.

3.1. Global vision

To characterize a target from a source register, we convert tweets into a *target sequential database* D_t and a *source sequential database* D_s . To retain only patterns that occur frequently in D_t and D_s , only patterns with a frequency (called *support*) exceeding a user-defined threshold are retained: these are called *frequent sequential patterns*. From the frequent sequential patterns in D_t and D_s , sequential pattern mining discovers the *emergent sequential patterns* of D_t for D_s .

An emergent pattern is one whose ratio of supports in D_t and D_s , respectively, is greater than a given threshold. This ratio, called *growth rate*, aims to discover emergent patterns in a target register relative to a source register because they are more frequent in D_t than in D_s .

6. The paper (Mekki *et al.*, 2020) details the work carried out.

3.1.1. Transforming tweets into sequences

An S sequence is an ordered sequence of sets called itemsets composed of items. For example, the sequence $S = \langle (a, b, c)(a, d)(a, b) \rangle$ is a sequence of three itemsets, each composed of three, two, and two items respectively. These sequences are then stored in sequential databases.

Three kinds of objects must be instantiated when we transform tweets into sequences: the sequence, the itemset, and the item. We have chosen the entire tweet as a textual segment represented by a sequence, the word⁷ as a textual unit described by an itemset and five fixed linguistic features for each word, i.e. five items for each itemset (the lexical level with the word lemma, the morphological level with sub-word units and its morphological characteristics, the morphosyntactic level is described via the grammatical category, the syntactic level describes the syntactic function). A sentence like *Girls are asleep.* has been transposed into a sequence of 3 itemsets:

- itemset 1 : (lemma: girl, pos: noun, morpho: plural, syntax: subject, subword: _Girl, subword: s_)
- itemset 2 : (lemma: be, pos: verb, morpho: plural, syntax: root, subword: _are_)
- itemset 3 : (lemma: asleep, pos: adjective, morpho: plural, syntax: modifier, subword: _a, subword: sleep_)

In this example, the symbol *lemma* precedes the word’s lemma, *pos* its grammatical category, *morpho* its morphological characteristics, *syntax* its syntactic function and *subwords* its subwords. Each tweet is tagged with Talismane (URIELI, 2012) (no particular mislabeling effect was observed), then transformed into a sequence in this way before being stored in a database. As there are 228,270 tweets, the database contains 228,270 sequences. Finally, this database is divided into three sub-databases, each representing, respectively, casual, standard, and formal language registers.

3.2. Emergent sequential pattern mining

We aim to use two sequential databases, representing two language registers, to discover patterns that distinguish them. To achieve this, we filtered the patterns in two stages: the first selects the interesting patterns in each database; the second selects the interesting patterns comparatively, retaining patterns significantly more present in one database than in another.

Selection of frequent and closed sequential patterns

Selecting *frequent and closed sequential patterns* enabled us to discard uninteresting patterns, i.e. those that are very infrequently present in a D database. This amounts to extracting all frequent patterns from D : all patterns whose support is greater than or equal to the *minsup* threshold. The *support* of a sequence S_1 in a database D , denoted

7. A word is defined as a space-separated token in this paper.

$sup_D(S_1)$, is the number of tuples containing S_1 in the database D . For example, the pattern $S_1 = \langle(a)(a)\rangle$ in database D has support $sup_D(S_1) = 2$: sequences 1 and 2 contain an itemset with a followed by an itemset with a . However, the extracted patterns can be very numerous and redundant. To avoid this, we have used a condensed representation without loss of information: *closed sequential patterns* (Yan *et al.*, 2003). A frequent pattern S is closed when there is no superset S' of S that is frequent and shares the same support as S .

To select occurring closed sequential patterns representing the casual, standard, and formal registers, we set *minsup* to 1%. This low value was intended to preserve closed patterns with low frequencies. Since we focused on not missing any interesting patterns, we obtained a large number of results. In the end, we obtained 2,341,661 closed patterns for the casual, 2,735,775 for the standard, and 8,895,962 for the formal. From these three sets of closed patterns representing each language register, we searched for patterns distinguishing one register from another, comparing the sets with each other using emergent sequential pattern mining.

Selection of emergent sequential patterns

Emergent sequential patterns are sequences that exhibit a substantial increase in support (frequency) between two datasets. The growth rate (*GR*) of such a pattern is calculated by dividing its support in a target dataset by its support in a source dataset. If the source dataset support is zero, the growth rate is considered infinite. A pattern is classified as emergent when its *GR* exceeds a user-defined threshold ρ . In our work, we fixed ρ at 1 to get all patterns, even weakly emerging ones. In all, six sets of patterns are discovered (casual vs. standard, casual vs. formal, standard vs. casual, standard vs. formal, formal vs. casual, formal vs. standard). The smallest set has 61,121 emerging sequential patterns, while the largest has 2,356,624. Generally speaking, we can see a discrepancy between the sets of emergent sequential patterns in the formal register and those in the casual and standard registers: on average, there are nine times as many. To explain this difference, we have assumed that there is a greater contrast between the linguistic forms of the formal register and those of the casual and standard registers. In contrast, sequences in the casual and standard registers are more similar to each other.

The main outcome of this section is the successful scaling of the closed and emergent sequential pattern mining algorithms, which validates our methodology. The large number of patterns discovered highlights the necessity of automated processing to produce a more manageable and interpretable set of patterns, as our goal is to obtain a collection that can be manually analyzed.

3.3. Automatic reduction of discovered patterns

We aim to identify linguistic features that differentiate one register from another through pattern mining. To ensure that the results are analyzable and interpretable, the number of patterns must remain manageable. Therefore, our approach to reducing the

set of emerging sequential patterns is guided by a double constraint: minimizing the number of patterns while preserving those that are meaningful. In this study, a pattern is considered meaningful if it effectively distinguishes one register from another.

To achieve this, we reduced the set of emerging sequential patterns by clustering them based on similarity, ensuring that each group remains distinct from the others. Each cluster corresponds to a specific linguistic feature, allowing us to select a single representative pattern from each group. This set of representative patterns forms a smaller, less redundant subset of the original emergent sequential patterns. In this section, we describe the two-step process for obtaining this subset: first, grouping the patterns by similarity, and second, selecting one representative pattern from each group.

3.3.1. Methodology for grouping patterns according to similarity

RGMSE (Regrouping Emerging Sequential Patterns) is a scalable clustering method designed for large sequential pattern datasets. It adapts k-means principles but improves efficiency by avoiding replacement during pattern assignment and automating cluster count determination. This balances intra-cluster cohesion and inter-cluster distinction, making it suitable for high-volume data analysis. RGMSE is described in Algorithm 1.

The user must specify three parameters to RGMSE: *minSim* a minimum similarity threshold between two individuals; *maxSize* a maximum cluster size threshold; *nbrIter* setting the number of iterations that repeat steps 2 and 3. RGMSE takes as input a list of patterns M and returns a set of groups of patterns G . This clustering takes place in four main stages:

1) **The clustering of objects by similarity according to *minSim*:** RGMSE initializes by considering each pattern m_i in the list M as a group g_i on its own. Starting with the first pattern m_1 , corresponding to the first group of patterns g_1 , RGMSE selects the other patterns in M in order of appearance. If the similarity between m_1 and the pattern selected is greater than or equal to *minSim*, then the pattern is added to the g_1 group and removed from M . When RGMSE has finished going through M , it moves on to the next pattern, which is considered the next group (the first one not grouped with m_1), and so on.

2) **The search for medoids for each cluster:** for each G cluster, RGMSE searches for its medoid using two approaches:

- if a group g has a number of patterns $|g|$ greater than or equal to *maxSize*; then the group is divided into $\lceil \frac{|g|}{maxSize} \rceil$ subgroups. For each of these subgroups, RGMSE calculates their medoid. The set of medoids obtained from the subgroups is considered as the set of g patterns from which the final medoid is calculated;

- otherwise, RGMSE searches directly the final medoid from g 's objects.

3) **The redistribution of objects among clusters according to their maximum similarity to all medoids:** RGMSE runs through M starting with m_1 , for which it

calculates its similarity to all medoids in G 's clusters. m_1 joins the cluster of the medoid with which it has maximum similarity. RGMSE then moves on to m_2 , then m_3 , and so on.

4) **The repetition of steps 2 and 3 $nbrIter$ times:** either the user has fixed the value of $nbrIter$; or RGMSE repeats steps 2 and 3 until it converges, i.e. until the distribution of objects in the clusters no longer moves.

RGMSE reduces its search time by setting the number of clusters in step (1). Steps (2) and (3) redistribute the patterns into the clusters with which they are most similar. This redistribution counterbalances the distribution of step (1), which depends on the order in which the patterns are selected.

Algorithm 1 RGMSE

Require: M (patterns), $minSim$, $maxSize$, $nbrIter$

Ensure: G (clusters)

```

1:  $G \leftarrow \emptyset$ ,  $R \leftarrow M$ 
2: while  $R \neq \emptyset$  do
3:    $c \leftarrow \{R[0]\}$ ,  $R \leftarrow R \setminus \{R[0]\}$ 
4:    $c \leftarrow c \cup \{p \in R : sim(R[0], p) \geq minSim\}$ 
5:    $R \leftarrow R \setminus c$ ,  $G \leftarrow G \cup \{c\}$ 
6: end while
7: for  $i = 1$  to  $nbrIter$  do
8:   for  $g \in G$  do
9:     if  $|g| \geq maxSize$  then
10:       $med[g] \leftarrow medoid(medoids(divide(g, maxSize)))$ 
11:     else
12:       $med[g] \leftarrow medoid(g)$ 
13:     end if
14:   end for
15:   for  $m \in M$  do
16:     Assign  $m$  to the cluster  $\arg \max_{g \in G} sim(m, med[g])$ 
17:   end for
18: end for
19: return  $G$ 

```

3.3.2. Clustering patterns by similarity

To determine the criteria on which similarity between two patterns is calculated, we explored the scientific literature on the subject to select a measure adapted to patterns and corresponding to our criteria. We present here the similarity measure used with RGMSE and the result we get to measure the cluster cohesion.

Similarity measure. The S^2MP measure for Similarity Measure for Sequential Patterns (Saneifar *et al.*, 2008) was selected for two key reasons: it enables direct comparison of itemset content without requiring conversion into list formats, and it incorporates both the sequential order of itemsets and the relative distances between similar itemsets into its evaluation. S^2MP is based on two scores to calculate the similarity of two patterns: the correspondence and the order score. The correspondence

score measures the similarity of two sequences based on shared items; while the order score measures the similarity of two sequences based on the order and positions of itemsets in the sequences.

Cluster cohesion. To calculate cluster cohesion, we fixed *minSim* to 0.50 to balance strictness and flexibility in similarity thresholds, *maxSize* to 500 to optimize the size of the subgroup (limiting medoid calculations while reducing computation time) and *nbrIter* to 2 to ensure the quality of distribution without excessive iteration overhead (more iterations didn't bring significant improvements and were very costly). The parameters of this experimental protocol seek to find a compromise between the need to reduce the algorithmic complexity and the quality of the clustering of patterns.

We conducted six experiments, each corresponding to a pair of registers, and evaluated the quality of the resulting partitions using two metrics: the silhouette coefficient (Rousseeuw, 1987) and the Davies-Bouldin index (Davies and Bouldin, 1979). The silhouette coefficients ranged from 0.23 to 0.31, indicating good cohesion and separation within the clusters. Similarly, the Davies-Bouldin index values were consistently low, between 0.40 and 0.53, further confirming the quality of the partitions. These results demonstrate that the partitions obtained for the six register pairs are of high quality.

3.3.3. Automatic selection of representative patterns

For each group of similar patterns, a so-called *representative pattern* is selected. To do this, we assume that for a G group, a good S representative pattern is one with items that are very frequent within a G group, as well as characteristic of the target register. For this reason, the *ItemFreqGR* measure (Equation 2) weights the frequency of a pattern's S items within the same group (given by Equation 1) by its growth rate.

$$ItemFreq(S, G) = \sum_{k=1}^{|S|} freq_G(i_k) \quad (\text{Eq. 1})$$

$$ItemFreqGR(S, G) = ItemFreq(S, G) \times GR(S) \quad (\text{Eq. 2})$$

The final subsets of representative patterns have a minimum of 740 patterns and a maximum of 3,475. Table 1 details the number of patterns, both for the complete pattern set and the representative pattern set: a significant reduction can be seen for the six register pairs. This reduction resulted in the elimination of around 99% of the patterns for the six register pairs.

ID	Register 1	Register 2	Set of patterns	Set of repr. patterns	Reduction rate
1	Casual	Standard	326,552	1,734	99.47 %
2		Formal	226,938	1,338	99.41 %
3	Standard	Casual	416,554	1,753	99.58 %
4		Formal	61,121	740	98.79 %
5	Formal	Casual	2,330,679	3,290	99.86 %
6		Standard	2,356,624	3,475	99.85 %

Table 1. *Quantitative details on the reduction of the complete set of patterns into a subset composed of representative patterns.*

3.4. Evaluation of the final pattern set

To evaluate these final subsets of representative patterns, two independent but complementary evaluations have been implemented: a qualitative and perceptual evaluation, and a quantitative and automatic evaluation.

Perceptual evaluation

This first perceptual evaluation is based on human judgment. It aims to check whether the patterns represented can indeed be used to characterize a language register. To do this, we asked an examiner to select from some tweets the one that most closely belonged to the target register. Only one of the two tweets contained a representative pattern characteristic of the target language register. Three tasks were asked in succession: select the most casual tweet, select the most standard tweet, and select the most formal tweet. The selected tweet was not intended to be considered familiar, casual, or formal in absolute terms, but rather confronted to the other tweet.

The experimental protocol aimed to set up an evaluation task where, between two tweets, the examiner had to select the most casual, the most standard, and the most formal tweet. To this end, three pairs of registers were considered: standard target register, opposite casual source register; casual target register, opposite formal source register; and formal target register, opposite standard source register. To make the evaluation task reasonable in terms of time, each examiner had to decide between 30 pairs of tweets (about 20 minutes). A website dedicated to this evaluation task was set up. To recruit reviewers, we circulated the link to our review platform on various mailing lists of the scientific community (linguistics and NLP). A total of 28 people carried out the evaluation task. When examiners were asked to select the tweet they found the most casual (target register A), they selected the target tweet T_A in 96% of cases. In only three cases did the examiners select the T_B tweet that was not in the target register. For these three cases, the representative pattern contained in the T_A tweet had a low growth rate: 1.30, 1.09, and 1.42. This means that when the representative pattern is weakly characteristic of the target register, it doesn't make it perceptible. In this respect, the growth rate is confirmed as relevant, since its low value refers to the weaker manifestation of the target register in the tweet. The success

rate was lower when the examiners had to select the tweet that seemed most standard to them: in only 68% of cases did they select the tweet T_A actually from the target register A (the standard register). When the examiner selected the T_B tweet, in the majority of cases (i.e. 86% of cases), the tweets in the standard register included patterns with low growth rates: between 2 and 1. Finally, the evaluation task was more successful for the formal register than for the standard one, with 80% of cases passing. In 99% of cases, where the examiner selected the tweet T_B , the MSE of the target register A had a low growth rate (i.e. between 1 and 2). For example, the following pattern $\langle(\text{lemma:que}), (\text{lemma:le}, \text{morpho:singular})\rangle$ had a growth rate of 2.53. Consequently, the tweet that illustrated this pattern presented a rather low degree of characterization of the formal register target. This may explain why the reviewer didn't select it.

This manual pattern evaluation highlighted the fact that a tweet T_A , which contained a pattern representing the target register A , was largely selected as being from that register. Moreover, the growth rate was confirmed as relevant to our task: in the vast majority of cases where the reviewer selected the tweet T_B , which was not from the target register A , the pattern growth rates were very low. In other words, the higher the growth rate of a pattern, the more it contributed to the characterization of the target register; and inversely, the lower the growth rate of a pattern, the less it contributed to the characterization of the target register. The relevance of the growth rate was confirmed during the automatic evaluation, whose work is shown in the next section.

Automatic evaluation

To evaluate the suitability of a subset of representative patterns, we assumed that it should enable a classifier to distinguish texts from an A register from a B register, i.e. to correctly label a given text by assigning it either the A register or the B register. To carry out this evaluation, we used the Random Forest classification algorithm introduced by Breiman (2001) as a binary classifier. The binary classifier had to predict whether a text belonged to the A register or the B register. The set of training features was built up from two sets of representative patterns: the one representing the A register versus the B register, and the one representing the B register versus the A register. In all, three binary classifiers were implemented, for the following three register pairs: casual vs. standard (0.81 f-measure), standard vs. formal (0.91), and formal vs. casual (0.94). The results show that, overall, the results were good. They confirm our intuition that the formal and casual registers are more contrasted than the casual and standard registers, with a higher f-measure for the formal vs. casual pair.

These two evaluations confirm the quality of the subsets of representative patterns and pave the way for their qualitative analysis. As the number of patterns has become reasonable, we are now able to analyze them linguistically to gain new insights into language registers in tweets.

4. What have we learned about language registers in French?

This section outlines what the emergent sequential patterns confirmed, i.e. which linguistic features characteristic of the language registers identified in the scientific literature were retrieved by the patterns. Then, we detail how the emergent sequential patterns show an integration into the linguistic norm of new linguistic elements.

What the emerging sequential patterns confirmed

ID	Features	Representing Patterns	R_t	TC
Lexical level				
1	Punctuating element	$\langle \{ \text{lemma: tout, subword: _tout_} \} \rangle$	F	1.3
2	Onomatopoeia	$\langle \{ \text{subword: _ou} \} \rangle$	F	$+\infty$
		$\langle \{ \text{subword: h_} \} \rangle$	F	2.2
3	"là" punctuating	$\langle \{ \text{syntax: modifier, subword: _là_} \} \rangle$	F	2.4
5	Discourse planners	$\langle \{ \text{lemma: après, pos: preposition} \}$ $\{ \text{pos: article, syntax: specifier} \} \rangle$	S	1.7
Morphosyntactic level				
6	Contraction of "cela" ("This is") en "ça" ("This's")	$\langle \{ \text{lemma: ça, subword: _ça_}$ $\text{pos: pronoun, number: singular} \}$ $\{ \text{person: 3} \} \rangle$	F	$+\infty$
7	Negation without "ne"	$\langle \{ \text{syntax: subject} \}$ $\{ \text{pos: verb, number: singular} \}$ $\{ \text{lemma: pas, subword: _pas_}$ $\text{pos: adverb} \} \rangle$	F	$+\infty$
8	Subject "on" transposed in "nous"	$\langle \{ \text{lemma: nous, syntax: subject}$ $\text{number: plural, person: 1} \} \rangle$	S	$+\infty$
10	Word ending in "-ouze"	$\langle \{ \text{subword: ze_} \} \rangle$	F	$+\infty$
11	Word ending in "-o"	$\langle \{ \text{subword: o_} \} \rangle$	F	1.3
12	Verb "être" ("be") in the singular before a singular noun phrase	$\langle \{ \text{lemma: être, number: singular} \}$ $\{ \text{pos: article, number: singular, genre: feminine} \}$ $\{ \text{syntax: object, number: singular, pos: common noun} \} \rangle$	F	1.3
13	"ça" + verb	$\langle \{ \text{lemma: ça, subword: _ça_}$ $\text{pos: pronoun, number: singular} \}$ $\{ \text{pos: verb, number: singular} \} \rangle$	F	$+\infty$
Syntactic level				
28	Deletion of the impersonal pronoun "il" ("he")	$\langle \text{lemma: y, pos: subject}$ $\text{pos: punctuation, syntax: punctuation, lemma: avoir} \rangle$	F	$+\infty$

Table 2. Features from Mekki *et al.* (2018) found among representative motifs.

Comparing the representative patterns obtained with the list of linguistic features taken from the scientific literature and grouped in Mekki *et al.* (2018) not only verifies their quality, but also statistically confirms the relevance of the features taken from the scientific literature. Table 2 gives the features from this article that were found among the representative patterns. Each row gives a feature with the same ID as it has in

(Ibid.), the representing pattern that represents it, the target register R_t characterized, and its growth rate GR . The higher its growth rate, the more the pattern was encountered in R_t , the highest value being $+\infty$ expressing the absence of the pattern in R_s . 12/28 features were found, which, given the limits imposed by the automatic tools and the type of text tweets, confirms the quality of the results.

What we have discovered with emerging sequential patterns

In this section, we present examples of emerging sequential patterns discovered, as well as examples of tweets supporting them. Two register pairs are presented: the casual target register versus the standard source register, and then the casual target register versus the formal register source. We chose these register pairs because the first one explores emergent sequential patterns from nearby registers, while the second one looks at emergent sequential patterns from distant registers.

Casual vs. standard

ID	Emerging sequential patterns	GR
1	$\langle \{\text{syntax:modifier}\}, \{\text{subword:}_j\} \rangle$	$+\infty$
2	$\langle \{\text{subword:}_t\}, \{\text{pos:verb}\} \rangle$	$+\infty$
3	$\langle \{\text{subword:}_\text{toi}_j\} \rangle$	$+\infty$
4	$\langle \{\text{pos:adjective}\}, \{\text{lemma:gros}\} \rangle$	$+\infty$
5	$\langle \{\text{pos:proper-name}\}, \{\text{pos:subject-clitic, morpho:3}^{eme}\}, \{\text{pos:verb, morpho:present}\} \rangle$	$+\infty$
6	$\langle \{\text{subword:}_c\} \rangle$	$+\infty$
7	$\langle \{\text{lemma:rajouter}\} \rangle$	$+\infty$
8	$\langle \{\text{subword:sh}_j\} \rangle$	$+\infty$
9	$\langle \{\text{subword:rrr}_j\} \rangle$	$+\infty$
10	$\langle \{\text{syntax:modifier}\}, \{\text{lemma:pictogram}\} \rangle$	1,52

Table 3. Examples of emerging sequential patterns characteristic of casual vs. standard.

If we sort the emerging sequential patterns in descending order of growth rate, all patterns with growth rates $+\infty$ are *ex aequo* in the first place. Table 3 presents 10 examples of these emerging sequential patterns that characterize casual versus standard register. For each of the patterns in table 3, examples of tweets from the TREMoLo-Tweets corpus are presented below. Patterns 1, 2, and 3 highlight the use of the first and second singular personal pronouns. They seem to reinforce the tendency for casual register to be used in tweets whose purpose is conversational. The tweets from (6) to (7) are examples of pattern 1: $\langle \{\text{syntax:modifier}\}, \{\text{subword:}_j\} \rangle$.

- (6) @X super entraineur laurent blanc **gros j'**espère il ira jamais chez vous (@X *great coach laurent blanc buddy I hope he never goes to your place*)
- (7) @X: **Mdrrrr j'**avais encore jamais vu un seul GP fais pas **genre j'**suis un anti Gasly tu sais très bien ce que je pense 🤔 (@X: *Looooo I'd never seen a single GP before, don't pretend I'm anti-Gasly, you know very well what I think.*)

Tweets from (8) to (9) are examples of pattern 2: $\langle \{\text{subword:}_t\}, \{\text{pos:verb}\} \rangle$.

- (8) @X **t as** dit je suis choqué par suarez ... il etait top 3 avc cr7 et messi pdt 5 ans (@X you said I'm shocked by suarez ... he was top 3 avc cr7 and messi for 5 years)
- (9) @X Pierre **t inquiète** les gens sont méchant reste comme tu es (@X Pierre dont worry people are mean stay as you are)

Tweets from (10) to (11) are examples of pattern 3: $\langle \{\text{subword:}_\text{toi}\} \rangle$.

- (10) @X @X @X Il éduque sanson t'es un fou **toi** mdr (He's educating sanson you're a fool lol)
- (11) Westbrook il est en train de faire une rondo 🤔🤔 réveil **toi** bro (Westbrook he's doing a rondo 🤔🤔 wake up bro)

Pattern 4 refers to discourse units called *discourse markers (DM)* by Dostie and Pusch (2007). These are language elements that punctuate exchanges that are usually oral.

Tweets from (12) to (13) are examples of pattern 4: $\langle \{\text{pos:adjective}\}, \{\text{lemma:gros}\} \rangle$

- (12) @X @X @X en **gros** hier sur fall guys coro a dit que la K corp perdrait aujourd'hui et qu'il avait lancer une malédiction mdr il a continuer le troll même pendant les games de KCorp voila (@X @X @X basically yesterday on fall guys coro said K corp would lose today and that he had cast a curse lol he continued the troll even during KCorp games here you go)
- (13) @X @X Maroua doit des dettes à Kihou che pas quoi et Maroua paye pas son loyer et traître Kihou voilà en **gros** (@X @X Maroua owes Kihou I don't know what and Maroua doesn't pay his rent and Kihou is a traitor that's all there is to it.)

Patterns 5 and 6, meanwhile, confirm the relevance of the linguistic features proposed in our annotation guide⁸ (Mekki *et al.*, 2021a) since they refer to the features described on pages 18 and 24 of this guide: *Doubled element* and *Electronic writing* (respectively). Tweets from (14) to (15) are examples of pattern 5: $\langle \{\text{pos:proper-name}\}, \{\text{pos:subject-clitic}, \text{morpho:3}^{eme}\}, \{\text{pos:verb}, \text{morpho:present}\} \rangle$.

- (14) **Westbrook** il est comme sa lebron le poster il le block mais vreument url_path (Westbrook he's like his lebron the poster he block but really url_path)
- (15) RT @X: Ptdr non **tootatis elle abuse** du bail là (RT @X: Lol no tootatis she's abusing here)

Tweets from (16) to (17) are examples of pattern 6: $\langle \{\text{subword:}_c\} \rangle$.

- (16) jamais tranquille **c** un truc de malade même contre brest (never easy it's a sick thing even against brest)
- (17) Putain lakers rockets la ils sont en mode precision 3 pts max **c** est une dinguerie. (Fucking lakers rockets they are in precision mode 3 pts max it's a madness.)

8. The annotation guide is available on HAL: <https://hal.archives-ouvertes.fr/hal-03218217>.

Pattern 7, for its part, can refer to new digital uses linked to adding users to a circle of online friends or conversation groups. Tweets from (18) to (19) are examples of pattern 7: $\langle \{ \text{lemma:rajouter} \} \rangle$.

- (18) Rt si tu veux que j'te **rajoute** url_path (*Rt if you want me to add you url_path*)
 (19) @X Mdr desac pas stv jte **rajoute** ds un grp le sang (@X Mdr not desactivates if you want I add you in a group bro)

Patterns 8 and 9 illustrate the value of using *subwords* as linguistic features to describe each *word*. These extracted emergent sequential patterns show that certain morphological endings are specific to the casual register. Tweets from (20) to (21) are examples of pattern 8: $\langle \{ \text{subword:sh_} \} \rangle$.

- (20) P T D R. Vous suiez face à Brest wesh **wesh** sois réaliste url_path (*L O L. You're sweating against Brest wesh be realistic.*)
 (21) @X askip les arohas sont restés sur un site pour voter expres pour faire **crash** le truc pour qu'astro gagne et ça a marché (@X heard the arohas stayed on a site to vote on purpose to wreck the thing so astro would win and it worked.)

Tweets from (22) to (23) are examples of pattern 9: $\langle \{ \text{subword:rrr_} \} \rangle$.

- (22) dans 1 semaine chuis en Suède mdr**rrrr** rien n'est prêt c'est la panik (*in 1 week i'm in sweden loool nothing is ready it's panic*)
 (23) je suis mort Djoko disqualifié parce qu'il a mis une tête à un juge pt-drrrrrr**rrrr** (*I'm dead Djoko disqualified because he headbutted a judge looooooooooooooooooool*)

Finally, pattern 10 shows that pictograms are used as punctuation that can be repeated to mark the intensity of the speaker's modalization of his discourse. Tweets from (24) to (25) are examples of pattern 10: $\langle \{ \text{syntax:modifier} \}, \{ \text{lemma:pictogram} \}, \{ \text{lemma:pictogram} \} \rangle$.

- (24) RT @X: Franchement les 2 sont magnifiques mais Silhouette» ❤️❤️❤️
 (*RT @X: Frankly the 2 are beautiful but Silhouette» ❤️❤️❤️*)
 (25) Kaaris x Bosh ils ont tout plié en Deux Deux 🔥🔥🔥 url_path (*Kaaris x Bosh they folded everything quickly 🔥🔥🔥 url_path*)

These few examples of emergent sequential patterns have demonstrated the robustness of the extraction, bearing in mind that casual and standard registers are two closely related registers whose boundaries can be difficult to delineate. We were able to measure the quality of emergent sequential patterns based on : (1) the fact that some patterns referred directly to linguistic features taken either from the literature on the subject or from linguistic exploration of the corpus, some of which were included in our annotation guide; (2) the fact that the patterns extracted showed linguistic patterns characteristic of the casual at the scale of discriminative endings.

Formal vs. casual

10 examples of emerging sequential patterns characteristic of formal register versus casual source register are shown in Table 4. In contrast to casual vs. standard register pair, characterization of formal register vs. casual is based on two strongly



ID	Pattern	GR
1	$\langle \{\text{subword:vous_}\} \rangle$	$+\infty$
2	$\langle \{\text{subword:_Pour_}\} \rangle$	$+\infty$
3	$\langle \{\text{lemma:alors, syntax:modifier, pos:adverb}\} \rangle$	$+\infty$
4	$\langle \{\text{pos:common name, subword:_\#}, \{\text{pos:punctuation}\} \rangle$	$+\infty$
5	$\langle \{\text{lemma:de}, \{\text{subword:_\#}\} \rangle$	$+\infty$
6	$\langle \{\text{subword:_entre}\} \{\text{subword:_\#}\} \rangle$	$+\infty$
7	$\langle \{\text{syntax:subject}\} \{\text{lemma:pictogram}\} \rangle$	$+\infty$
8	$\langle \{\text{subword:lance_}\} \rangle$	$+\infty$
9	$\langle \{\text{subword:hui_}\} \rangle$	$+\infty$
10	$\langle \{\text{subword:ez_}\} \rangle$	1,76

Table 4. *Examples of emerging sequential patterns characteristic of formal vs. casual.*

contrasting registers. Expected linguistic traits, characteristic of the formal register were found among the extracted emerging sequential patterns. For example, pattern 1, which refers to the use of the formal form of address in French: $\langle \{\text{subword:vous_}\} \rangle$. Tweets from (26) to (27) are examples of pattern 1.

- (26) @X Si **vous** permettez pour une diffusion élargie, chez Plenel, "Tiers État" signifie le Peuple par opposition à la noblesse (soit disant supprimée) et au clergé (religieux soit disant exclu du champ sociétal, mais omniprésent ces 20 dernières années). (@X *If you allow for a wider dissemination, in Plenel, "Third State" means the People as opposed to the nobility (supposedly suppressed) and the clergy (religious supposedly excluded from the societal field, but omnipresent in the last 20 years).*)
- (27) @X @X @X René Bousquet était préfet de la République. Je crois que **vous** n'avez pas compris le sens du tweet. Dire "ministre de la République", c'est pas un totem d'immunité. On n'est pas à chat perché. @X (@X @X @X *René Bousquet was a French prefect. I think you missed the point of the tweet. Saying "Minister of the Republic" isn't a totem of immunity. It isn't off-ground tag. @X*)

Patterns 2 and 3 correspond to a preposition (*pour/for*) and an adverb (*alors/then*), both of which help to structure the text. Tweets from (28) to (29) are examples of pattern 2 which identifies the position of the preposition at the beginning of the sentence thanks to the capitalized sub-word: $\langle \{\text{subword:_Pour_}\} \rangle$.

- (28) | #FranceRelance Les jeunes ont souvent été les premières victimes de la crise économique que nous vivons. **Pour** qu'ils puissent s'insérer rapidement et durablement sur le marché du travail, le @X s'engage :  [url_path](#) (*| #FranceRelance Young people have often been the first victims of the current economic crisis. To help them integrate quickly and sustainably into the job market, @X is committed to :  [url_path](#)*)
- (29) @X @X @X @X **Pour** vous, le masque arrête-t-il le virus? (@X @X @X @X *For you, does the mask stop the virus?*)

The adverb *alors/then* contained in pattern 3, $\langle \{\text{lemma:alors, syntax:modifier, pos:adverb}\} \rangle$, is used to structure tweets from (30) to (31).

- (30) @X Si moi, petit écrivillon de banlieue, je savais dès 1986 à quoi m'en tenir sur les "amours" de #Matzneff, la défense à géométrie variable de Girard ne tient pas 1 seconde "je ne savais pas, j'étais aux USA", **alors** même que les écrits de GB éclairent son rôle de factotum de Berg (@X *If I, a little writer from the suburbs, knew back in 1986 where I stood on the "loves" of #Matzneff, Girard's variable-geometry defense doesn't hold up for 1 second: "I didn't know, I was in the USA", even though GB's writings shed light on his role as Berg's factotum.*)
- (31) Si l'on réfléchit, **alors** on ne peut cautionner ce que fait Plenel: c'est haineux, ignoble et indéfendable 🙄 url_path (*If you think about it, then you can't support what Plenel is doing: it's hateful, despicable and indefensible* 🙄 url_path)




While emergent sequential patterns 1, 2, and 3 are classic linguistic features, emergent sequential patterns 4, 5, and 6 incorporate linguistic elements specific to digital writing. These patterns show that hashtags are integrated into the grammatical norm. Tweets from (32) to (33) are examples of pattern 4: ⟨{pos:common name, subword:_#}, {pos:punctuation}⟩.

- (32) Le module russe Zarya est le premier élément de l'**#iss**. Il est lancé le 20 novembre 1998 avant d'être rejoint 2 semaines plus tard par le module américain Unity. Aujourd'hui il sert principalement d'espace de stockage. CREDIT : NASA #espace #astronomie url_path (*The Russian Zarya module is the first element of the #iss. It was launched on November 20, 1998, and was joined 2 weeks later by the American Unity module. Today it serves mainly as storage space. CREDIT: NASA #space #astronomy url_path*)
- (33) @X Pour l'instant, #CharlieHebdo respecte scrupuleusement les limites définies par les tabous et la **#censure**. Quand @X voudra vraiment tester les limites de la tolérance en France, ce genre de caricature sera publié: url_path (@X *For the moment, #CharlieHebdo scrupulously respects the limits defined by taboos and #censorship. When @X really wants to test the limits of tolerance in France, this kind of cartoon will be published: url_path*)















Tweets from (34) to (35) are examples of pattern 5, which also illustrates the integration of hashtags into the linguistic norm: ⟨{lemma:de}, {subword:_#}⟩.

- (34) Tranquillement, l'équipe de **#Trump** ment et manipule des propos **de #Biden**. Twitter a marqué la publication comme "manipulée". C'est de la désinformation pure et simple. url_path (*Quietly, #Trump's team lies and manipulates comments from #Biden. Twitter marked the publication as "manipulated". This is disinformation, pure and simple. url_path*)
- (35) "Laval Agglomération est connectée et ouverte sur le monde... Nous ne pouvons pas rester insensibles en matière **de #solidarité**..." #Liban | url_path via @X en #Mayenne url_path (*"Laval Agglomération is connected and open to the world... We cannot remain insensitive when it comes to solidarity..." #Liban | url_path by @X in #Mayenne url_path*)













Finally, the emerging sequential pattern 6 also highlights the use of hashtags as classic words: ⟨{subword:_entre} {subword:_#}⟩. Tweets from (36) to (37) are examples of pattern 6.

- (36) Notre note politique sur l'Assemblée de Bretagne a mis en évidence l'absurdité actuelle de la distribution des compétences **entre** . L'avenir est à l'intégration des politiques publiques  #ODD #agenda2030. Il nous faut des Régions et Villes au pouvoir plus intégré. url_path url_path (Our policy brief on the Brittany Assembly highlighted the current absurdity of the distribution of competences between **collter** and FR. The future lies in the integration of public policies  #ODD #agenda2030. We need more integrated regions and cities into political power. url_path url_path)
- (37) #Darmanin se dit à "100.000 lieues" de faire "le lien **entre** #immigration et #insécurité" et invoque ses origines familiales url_path (#Darmanin says he is "100,000 leagues" away from making "the link between #immigration and #insecurity" and cites his family origins url_path)

Another type of linguistic element specific to CMCs (Computer-Mediated Communications) is integrated into the grammatical norm: pictograms. The emerging sequential pattern 7 shows that pictograms can be used in the same way as traditional syntactically integrated lexicon: {syntax:subject}{lemma:pictogram}.

- (38) La destruction de l'Amazonie empire. La  **est** complice de ce désastre en raison de ses importations, @X l'a reconnu l'été dernier. Depuis ? Uniquement des paroles, 0 acte concret. url_path #JeudiPhoto #Marseille  url_path (The destruction of Amazonia is getting worse. The  is complicit in this disaster through its imports, as @X acknowledged last summer. Since then? Only words, 0 concrete actions. url_path #ThursdayPicture #Marseille  url_path)
- (39) #FranceRelance  c'est aussi #EuropeRelance : l'  **sera** présente dans chacun des projets de ce plan, il ne faut pas avoir l'  honteuse, ni l'  invisible  @X @X @X url_path' (#FranceRelance  is also #EuropeRelance : l'  will be present in each of the projects of this plan, we must not have the  ashamed, nor the  invisible  @X @X @X url_path')

Note that the pictograms indicate a rather political and/or institutional communication. Finally, the last patterns, emerging sequential patterns 8, 9, and 10, highlight the contribution of subwords to the characterization of language registers, with three endings characteristic of the formal register. The first one, pattern 8 ({subword:lance_}) is illustrated by tweets from (40) to (41).

- (40)    Comme l'école de la confiance et la bienveillance n'est pas l'école de la transparence il est indispensable de partager et relayer le travail des #stylosrouges qui recensent les cas #Covid_19 dans les établissements scolaires !!!    (   As the school of trust and benevolence is not the school of transparency, it is essential to share and relay the work of #stylosrouges who identify cases #Covid_19 in schools !!!!   )
- (41) Écologie : Réorganiser les moyens de surveillance de l'état environnemental de la France ainsi que les principaux organismes de gestion des sols, des forêts et des eaux. #UPR #FRANÇOISASSELINEAU #ecologie url_path (Ecology: Reorganize France's environmental monitoring resources and key soil, forest and water management bodies. #UPR #FRANÇOISASSELIN-EAU #ecology url_path)

The emerging sequential pattern 9, $\langle \{\text{subword:hui_}\} \rangle$, refers to the ending of the word *aujourd'hui*/today and, like pattern 3 (*alors/then*), provides a temporal structure to the tweet.

- (42) Donc le maire de Stains @X se bat pour le maintien de la fresque d'un violeur. #AdamaVioleur Monsieur le maire, maintenant que les faits sont aujourd'**hui** avérés, allez vous continuer votre combat pour défendre un violeur ? url_path (So Stains mayor @X fights to keep a rapist's mural up. #AdamaVioleur Mr. Mayor, now that the facts are in, will you continue your fight to defend a rapist? url_path)
- (43) Aujourd'**hui** c'est Soral... Demain c'est Dieudonné. "Quand les bandits sont au pouvoir, la place d'un honnête homme est en prison" (Michel Chartrand) url_path (Today it's Soral... Tomorrow it's Dieudonné. "When bandits are in power, an honest man's place is in prison" (Michel Chartrand) url_path)

Lastly, the emerging sequential pattern 10 joins pattern 1, as it indicates the presence of the formal salutation in French: $\langle \{\text{subword:ez_}\} \rangle$.

- (44) @X Bonjour. Le placement automatique "un siège sur deux" n'est plus appliqué à bord des trains depuis début juin. Toutes les places peuvent désormais être occupées. Il est donc possible que vous soy**ez** assis à côté d'une personne que vous ne connais**sez** pas. 1/2 (@X Hello. The automatic "every other seat" placement is no longer applied on trains since the beginning of June. All seats can now be occupied. It is therefore possible that you will be seated next to someone you do not know. 1/2)
- (45) @X Vous ave**z** la réponse à votre question : le VPCE n'est pas président de la section du contentieux. (@X You have the answer to your question: the VPCE is not chairman of the Litigation Division.)

Various emerging sequential patterns characteristic of formal versus casual could be considered of quality because: (1) they include salient features characteristic of formal speech traditionally associated in the scientific literature (such as the use of the *vouvoiement* and the discourse elements that structure it logically and temporally); (2) they correspond to linguistic features mentioned in the annotation guide (syntactic integration of elements specific to CMCs); (3) they confirmed the contribution of subwords to pattern mining by highlighting endings characteristic of formal.

Emerging sequential patterns have confirmed the phenomenon of integrating elements specific to digital discourse into the grammatical norm: hashtags or pictograms are no longer reserved for the casual register, but are instead used for institutional communications.

Conclusion and perspectives

In this paper we introduced a complete pipeline for characterizing language registers from a corpus of French tweets (TREMolo-Tweets). The primary objective of this study was to characterize language registers by extracting emergent sequential patterns from a corpus of tweets. The secondary objective was to propose a pipeline that could be used for other use cases. To meet the first objective, we proposed a processing chain aimed at obtaining a set of patterns distinguishing one language register

from another. In developing this pipeline, we were careful not to make any linguistic *a priori* to be able to apply it to other linguistic phenomena represented with contrasting data. Our results show that we have succeeded in fulfilling our initial aim of characterizing language registers using a robust methodology. They also show that our second goal has been achieved with a very unconstrained pipeline that can be applied to other use cases. From a linguistic point of view, this study has confirmed the possibility of characterizing language registers comparatively by considering several levels of language analysis. It also revealed the integration into the linguistic standard of new linguistic elements specific to tweets (such as hashtags and pictograms, which are used in both standard and formal registers). From a computational point of view, we have addressed several emerging sequential pattern mining issues, enabling our approach to be adapted to characterize all types of linguistic variation illustrated by a corpus of texts made up of contrasting sub-corpora.

Several perspectives can be considered to extend the work presented in this paper. Firstly, when manually annotating part of the TREMoLo-Tweets corpus, annotators had to justify their choice of registers by selecting at least one linguistic feature present in the tweet. A set of linguistic features was thus annotated in proportion to language registers. Future work could explore this set to discover whether features were systematically selected together. Next, in our pipeline, closed sequential patterns were used. A shortcoming of this type of pattern is that a closed pattern $M_1 = \langle (a, b, c), (b, c) \rangle$ with a support of 6 will not be able to contain $M_2 = \langle (a, b, c), (b) \rangle$ whose support is 5. To bring M_1 and M_2 together, we could use the δ -patterns introduced by Holat *et al.* (2014). The δ -patterns bring patterns with neighboring supports together. Their use would therefore make it possible not to differentiate between close patterns, and would perhaps lead to a smaller set of emerging sequential patterns than the one obtained with closed patterns. Finally, several ways of selecting representative patterns could be tested. For example, we could weigh the types of features contained in the patterns. Our work has shown that subwords carry interesting information, such as the position of the word in the sentence, with the presence of a capital letter. We could therefore give more weight to a pattern that includes subwords. This would give rise to more explicit emerging sequential patterns, and perhaps even more interpretable patterns characteristic of language registers.

5. References

- Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R. J., "Sentiment analysis of twitter data", *Proceedings of the workshop on language in social media (LSM 2011)*, p. 30-38, 2011.
- Biber D., *Variation across speech and writing*, Cambridge University Press, 1991.
- Biber D., Conrad S., *Register, genre, and style*, Cambridge University Press, 2019.
- Branca-Rosoff S., "Des innovations et des fonctionnements de langue rapportés à des genres", *Langage & société*, vol. 87, n° 1, p. 115-129, 1999.
- Breiman L., "Random forests", *Machine learning*, vol. 45, n° 1, p. 5-32, 2001.

- Davies D. L., Bouldin D. W., "A cluster separation measure", *IEEE transactions on pattern analysis and machine intelligence*, n° 2, p. 224-227, 1979.
- Dong G., Li J., "Efficient mining of emerging patterns: Discovering trends and differences", *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 43-52, 1999.
- Dostie G., Pusch C. D., "Présentation. Les marqueurs discursifs. Sens et variation", *Langue française*, n° 2, p. 3-12, 2007.
- Egbert J., Biber D., Gray B., *Designing and evaluating language corpora: A practical framework for corpus representativeness*, Cambridge University Press, 2022.
- Ferguson C. A., "Simplified registers and linguistic theory", *Exceptional language and linguistics*, p. 49-66, 1982.
- Fournier-Viger P., Lin J. C.-W., Kiran R. U., Koh Y. S., Thomas R., "A survey of sequential pattern mining", *Data Science and Pattern Recognition*, vol. 1, n° 1, p. 54-77, 2017.
- Frei H., *La grammaire des fautes : introduction à la linguistique fonctionnelle, assimilation et différenciation, brièveté et invariabilité, expressivité*, vol. 1, Slatkine, 1971.
- Gadet F., "Niveaux de langue et variation intrinsèque", *Palimpsestes. Revue de traduction*, n° 10, p. 17-40, 1996.
- Gadet F., "La variation, plus qu'une écume", *Langue française*, p. 5-18, 1997.
- Gadet F., *La variation sociale en français*, Editions Ophrys, 2007.
- Go A., Bhayani R., Huang L., "Twitter sentiment classification using distant supervision", *CS224N project report, Stanford*, vol. 1, n° 12, p. 2009, 2009.
- Halliday M. A. K., Hasan R., "Language, context, and text: Aspects of language in a social-semiotic perspective", (*No Title*), 1989.
- Heller M., Alby S., Brohy C., Candelier M., Castellotti V., Gajo L., Ghimenton A., Ledegen G., Légise I., Matthey M. et al., *Sociolinguistique du contact : dictionnaire des termes et concepts*, ENS éditions, 2013.
- Holat P., Plantevit M., Raïssi C., Tomeh N., Charnois T., Crémilleux B., "Sequence classification based on delta-free sequential patterns", *2014 IEEE International Conference on Data Mining*, IEEE, p. 170-179, 2014.
- Labov W., "The judicial testing of linguistic theory", *Language in Context: Connecting Observation and Understanding*, Norwood, Ablex, 1988.
- Lecorvé G., Ayats H., Fournier B., Mekki J., Chevelu J., Battistelli D., Béchet N., "Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français", *Traitement automatique du langage naturel (TALN)*, 2018.
- Martin L., Muller B., Suárez P. J. O., Dupont Y., Romary L., de La Clergerie É. V., Seddah D., Sagot B., "CamemBERT: a tasty French language model", *arXiv preprint arXiv:1911.03894*, 2019.
- Mekki J., Battistelli D., Lecorvé G., Béchet N., "Identification de descripteurs pour la caractérisation de registres", *Rencontre des jeunes chercheurs en traitement automatique du langage naturel et recherche d'information (CORIA-TALN-RJC)*, 2018.
- Mekki J., Battistelli D., Lecorvé G., Béchet N., "TREMolo-Tweets corpus: guide d'annotation pour un corpus annoté en registres de langue pour le français", 2021a.

- Mekki J., Béchet N., Battistelli D., Lecorvé G., “Caractérisation de registres de langue par extraction de motifs séquentiels émergents”, *JADT 2020: 15èmes Journées Internationales d'Analyse statistique des Données Textuelles*, 2020.
- Mekki J., Lecorvé G., Battistelli D., Béchet N., “TREMolo-Tweets: a Multi-Label Corpus of French Tweets for Language Register Characterization”, *RANLP 2021-Recent Advances in Natural Language Processing*, 2021b.
- Pak A., Paroubek P., “Twitter as a corpus for sentiment analysis and opinion mining.”, *LREc*, vol. 10, p. 1320-1326, 2010.
- Paveau M.-A., Rosier L., *La langue française. Passions et polémiques*, 2008.
- Poudat C., Landragin F., *Explorer un corpus textuel: Méthodes-pratiques-outils*, De Boeck Supérieur, 2017.
- Rousseeuw P. J., “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”, *Journal of computational and applied mathematics*, vol. 20, p. 53-65, 1987.
- Saneifar H., Bringay S., Laurent A., Teisseire M., “S2mp: Similarity measure for sequential patterns”, *AusDM: Australasian Data Mining*, vol. 87, ACS, p. 095-104, 2008.
- Ure J., “Introduction: approaches to the study of register range”, *International Journal of the Sociology of Language*, vol. 1982, n° 35, p. 5-24, 1982.
- URIELI A., “Talismane: construction d’un analyseur syntaxique probabiliste”, 2012.
- Yan X., Han J., Afshar R., “Clospan: Mining: Closed sequential patterns in large datasets”, *Proceedings of the 2003 SIAM international conference on data mining*, SIAM, p. 166-177, 2003.