Traitement automatique des langues

Varia

sous la direction de
Maxime Amblard
Marie Candito
Cécile Fabre
Benoît Favre
Aurélie Névéol
Sophie Rosset

Vol. 66 - n°1 / 2025

ATALA

Varia

ATALA

Revue de l'Association pour le Traitement Automatique des Langues

TAL
Vol.
66

n°1
2025

Varia

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS.

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Traitement automatique des langues

Volume 66 – n°1 / 2025

VARIA

Table des matières

Préface

Les préfaces des numéros non thématiques de la revue *TAL* permettent de faire chaque année le point sur la vie de la revue et de commenter les statistiques que nous présentons traditionnellement pour les numéros des trois dernières années.

La revue a connu un renouvellement important de son comité de rédaction : après le départ en 2023 de Pascale Sébillot, Sophie Rosset a terminé son mandat avec le numéro 64-2, Emmanuel Morin avec le numéro 65-1, et Cécile Fabre avec ce numéro 66-1. Le quatuor s'est reconstitué autour de Maxime Amblard et Benoit Favre, rejoints en janvier 2025 par Marie Candito et Aurélie Névéol. Les rédacteurs en chef poursuivent la réflexion sur les évolutions requises pour assurer une meilleure visibilité à la revue. Après le passage au fil de l'eau des Varia et l'attribution d'un DOI à chaque article qui va prendre effet à partir du volume 64, la question des modalités de diffusion de la revue est en cours d'examen. Nous explorons plusieurs pistes permettant d'héberger la revue sur une plateforme d'édition en libre accès (type OpenEdition ou Mersenne).

Les statistiques du tableau 1 considèrent les dix derniers numéros sur les trois dernières années, du début de 2022 jusqu'à ce numéro *Varia* de 2025 inclus. Ce tableau donne les taux de sélection par numéro et par volume. La ligne du total synthétise ces chiffres sur l'ensemble des dix numéros considérés. L'attractivité des numéros (entre 2 et 8 propositions) reste assez faible, avec un nombre moyen d'articles soumis un peu plus élevé pour les numéros thématiques que pour les varia (6 *vs* 4,5). Le comité de rédaction de la revue est très attaché à sélectionner les articles sur le seul critère de leur qualité, indépendamment du nombre d'articles soumis, et n'hésite pas, comme cela a été le cas pour deux des derniers *Varia*, à préserver cette exigence et à ne retenir qu'un article. Le taux de sélection est de fait très variable, allant de 25 à 75 %. Le choix des deux thématiques annuelles est fait chaque année en comité de rédaction. Un appel à proposition de thématiques est lancé sur la liste *In* en début d'année.

Les statistiques que nous donnons, dans le tableau 2, sur l'origine des articles, prennent en compte le pays du premier auteur, France ou hors de France. Nous considérons également la langue de la soumission, le français ou l'anglais. Les chiffres sont fournis pour la même période de temps que pour le tableau 1. Globalement, un cinquième des premiers auteurs sont des chercheurs hors de France, et près d'un tiers des articles sont en anglais. Néanmoins, ces chiffres cachent des disparités fortes selon les

Intitulé	**Vol.**	**N°**	**Année**	**Soumis**	**Acceptés**	**% acceptés**
- *Varia*	63	1	2022	5	3	60,0 %
- Intermodalité et multimodalité en TAL	63	2	2022	4	3	75,0 %
- États de l'art en TAL	63	3	2022	8	3	37,5 %
Sous-total	**63**		**2022**	**17**	**9**	**52,9%**
- *Varia*	64	1	2023	4	1	25,0 %
- Robustesse et limites des modèles de TAL	64	2	2023	4	2	50,0 %
- Explicabilité des modèles de TAL	64	3	2023	5	5	100,0 %
Sous-total	**64**		**2023**	**13**	**8**	**61,5 %**
- *Varia*	65	1	2024	1	1	100,0 %
- Scholarly Document Processing	65	2	2024	5	1	20,0 %
- Discours de haine	65	3	2024	8	3	37,5%
Sous-total	**65**		**2024**	**14**	**5**	**35,7 %**
Varia	66	1	2025	7	2	28,5 %
Total	**Dix derniers n^{os}**			**51**	**24**	**47,0 %**

Tableau 1. *Taux de sélection aux appels de la revue TAL sur les dix derniers numéros de la période 2017-2020*

Intitulé	**Vol.**	**N°**	**Année**	**% 1er auteur hors France**	**% en anglais**
- *Varia*	63	1	2022	0,0 %	33,3 %
- Intermodalité et multimodalité en TAL	63	2	2022	0,0 %	0,0 %
- États de l'art en TAL	63	3	2022	66,6 %	66,6 %
Pourcentages par volume	**63**		**2022**	**22,2 %**	**33,3 %**
- *Varia*	64	1	2023	0,0 %	0,0 %
- Robustesse et limites des modèles de TAL	64	2	2023	0,0 %	0,0 %
Explicabilité des modèles de TAL	64	3	2023	20,0 %	20,0 %
Pourcentages par volume	**64**		**2023**	**12,5%**	**12,5%**
Varia	65	1	2024	0,0%	0,0%
Traitement automatique de documents scientifiques	65	2	2024	0%	0%
Discours de haine	65	3	2024	66%	66%
Pourcentages par volume	**65**		**2024**	**40%**	**40 %**
Varia	66	1	2025	0%	50%
Pourcentages totaux	**Dix derniers n^{os}**			**20,8 %**	**29,1 %**

Tableau 2. *Proportion des articles publiés d'un premier auteur hors de France et proportion des articles publiés rédigés en anglais sur les dix derniers numéros de la période 2022-2025*

numéros : si certains numéros thématiques ont attiré plus de contributions internationales (c'est le cas des numéros sur les états de l'art et sur les discours de haine), les Varia ont intégré exclusivement des articles rédigés par des chercheurs français. Rappelons que, depuis 2024, les articles peuvent être rédigés indifféremment en anglais ou en français, alors que jusqu'à cette date les chercheurs francophones devaient opter pour le français. La dimension internationale de la revue se manifeste également à travers la composition du comité de rédaction de la revue, la constitution d'un comité scientifique international pour les numéros thématiques, la présence d'un chercheur étranger parmi les rédacteurs invités en charge de ces numéros.

Le présent numéro contient deux articles retenus lors de l'appel non thématique lancé en janvier 2024. Cet appel portait, comme c'est l'usage, sur tous les aspects du traitement automatique des langues. Sept articles ont été soumis entre septembre et décembre 2024. Deux articles ont été retenus à l'issue du processus de sélection habituel à deux tours.

Automatic characterization of French language registers: Illustration on tweets, Jade Mekki (U. Rennes, CNRS, IRISA) — Delphine Battistelli (U. Paris Nanterre, CNRS, MoDyCo), Nicolas Béchet (U. Rennes, CNRS, IRISA), Gwénolé Lecorvé (Orange Labs / U. de Bretagne Sud, CNRS, IRISA)

L'article propose une méthode pour détecter et pour analyser les variations de registre (familier, standard, formel) dans un large corpus de tweets en français. L'annotation d'un échantillon du corpus permet d'entraîner un classifieur pour généraliser l'annotation, afin d'explorer les caractéristiques linguistiques qui distinguent chaque paire de registre. La méthode fait appel à une technique d'extraction de motifs qui rend possible la caractérisation des registres de langue à différents niveaux d'analyse linguistique.

De l'annotation intégrale à l'analyse des réseaux de personnages : un modèle pour la REN dans les textes littéraires en français, Motasem Alrahabi (ObTIC, Sorbonne U.) — Arthur Amalvy (LIA, Avignon U.) — Vincent Labatut (LIA, Avignon U.) — Perrine Maurel (ObTIC, Sorbonne U.)

Cet article présente une contribution du TAL au champ des humanités numériques, et plus particulièrement à l'analyse littéraire. L'étude porte sur la reconnaissance d'entités nommées (EN) dans des textes littéraires en français. Elle se fonde sur l'annotation en EN d'un corpus de sept oeuvres du XIX^{e} siècle, librement disponible, et sur l'utilisation d'un système de reconnaissance d'EN à base d'un modèle CamemBERT préentrainé puis fine-tuné sur ce corpus littéraire. L'article montre comment ces résultats peuvent produire des réseaux de personnages favorisant l'analyse des dynamiques littéraires.

Merci aux membres du comité de rédaction de la revue qui ont participé aux différentes étapes d'élaboration de ce numéro et à la décision collégiale, à l'appui des relectures reçues, qui a abouti à l'acceptation ou au rejet des articles soumis. Merci en

particulier à celles et ceux qui ont pris en charge des relectures (voir la composition du comité sur le site de la revue : http://www.atala.org/content/comité-de-rédaction-0). Merci aux relecteurs spécifiques de ce numéro : Timothée Bernard (LLF), Mathieu Lafourcade (LIRMM), Philippe Langlais (U. de Montréal), Gaël Lejeune (STIH, CERES), Filip Miletic (U. de Stuttgart), Thierry Poibeau (LATTICE), Céline Poudat (BCL).

Nous adressons nos remerciements à la Délégation générale à la langue française et aux langues de France (DGLFLF) pour son soutien financier à la revue.

Maxime Amblard
LORIA, Université de Lorraine

Cécile Fabre
CLLE, Université Toulouse - Jean Jaurès

Benoit Favre
LIS, Aix-Marseille Université

Sophie Rosset
Université Paris-Saclay, CNRS, LISN

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.

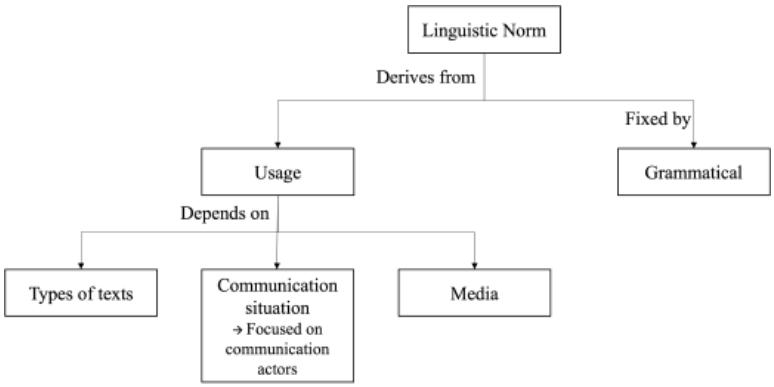


Figure 1. *Linguistic norms.*

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: *politic****ard***), 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 Hallidays' 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. élective 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 community [3]. 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 rank [4], it must be justified by the presence of at least one feature from the list presented in our annotation guide [5]. 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 pretrained 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/`.

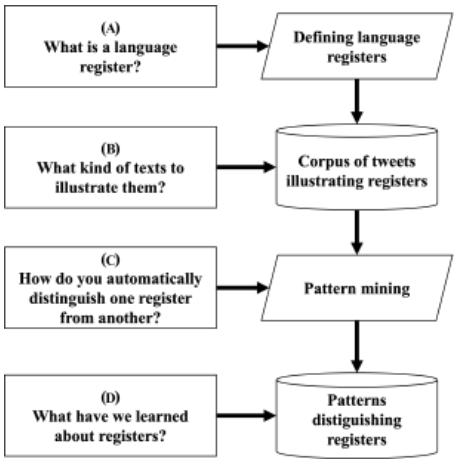


Figure 2. *Main steps for characterizing French language registers.*

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 paper [6], 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 word [7] 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 emerging sequential patterns

Emerging 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 $[\frac{|g|}{maxSize}]$ 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)
$G \leftarrow \emptyset$, $R \leftarrow M$
while $R \neq \emptyset$ **do**
  $c \leftarrow \{R[0]\}$, $R \leftarrow R \setminus \{R[0]\}$
  $c \leftarrow c \cup \{p \in R : sim(R[0], p) \geq minSim\}$
  $R \leftarrow R \setminus c$, $G \leftarrow G \cup \{c\}$
end while
for $i = 1$ **to** $nbrIter$ **do**
  for $g \in G$ **do**
    if $|g| \geq maxSize$ **then**
      $med[g] \leftarrow medoid(medoids(divide(g, maxSize)))$
    else
      $med[g] \leftarrow medoid(g)$
    end if
  end for
  for $m \in M$ **do**
    Assign m to the cluster $\arg\max_{g \in G} sim(m, med[g])$
  end for
end for
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 ⟨(lemma:que), (lemma:le, morpho:singular)⟩ 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	⟨{lemma:tout, subword:_tout_}⟩	F	1.3
2	Onomatopoeia	⟨{subword:_ou}⟩	F	$+\infty$
		⟨{subword:h_}⟩	F	2.2
3	"là" punctuating	⟨{syntax:modifier, subword:_là_}⟩	F	2.4
5	Discourse planners	⟨{lemma:après, pos:preposition} {pos:article, syntax:specifier}⟩	S	1.7
Morphosyntactic level				
6	Contraction of "cela" ("This is") en "ça" ("This's")	⟨{lemma:ça, subword:_ça_, pos:pronoun, number:singular} {person:3}⟩	F	$+\infty$
7	Negation without "ne"	⟨{syntax:subject} {pos:verb, number:singular} {lemma:pas, subword:_pas_, pos:adverb}⟩	F	$+\infty$
8	Subject "on" transposed in "nous"	⟨{lemma:nous, syntax:subject, number:plural, person:1}⟩	S	$+\infty$
10	Word ending in "-ouze"	⟨{subword:ze_}⟩	F	$+\infty$
11	Word ending in "-o"	⟨{subword:o_}⟩	F	1.3
12	Verb "être" ("be") in the singular before a singular noun phrase	⟨{lemma:être, number:singular} {pos:article, number:singular, genre:feminine} {syntax:object, number:singular, pos:common noun}⟩	F	1.3
13	"ça" + verb	⟨{lemma:ça, subword:_ça_, pos:pronoun, number:singular} {pos:verb, number:singular}⟩	F	$+\infty$
Syntaxic level				
28	Deletion of the impersonal pronoun "il" ("he")	⟨lemma:y, pos:subject pos:punctuation, syntax:punctuation, lemma:avoir⟩	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	⟨{syntax:modifier}, {subword:_j}⟩	$+\infty$
2	⟨{subword:_t}, {pos:verb}⟩	$+\infty$
3	⟨{subword:_toi_}⟩	$+\infty$
4	⟨{pos:adjective}, {lemma:gros}⟩	$+\infty$
5	⟨{pos:proper-name}, {pos:subject-clitic, morpho:3^{eme}}, {pos:verb, morpho:present}⟩	$+\infty$
6	⟨{subword:_c_}⟩	$+\infty$
7	⟨{lemma:rajouter}⟩	$+\infty$
8	⟨{subword:sh_}⟩	$+\infty$
9	⟨{subword:rrr_}⟩	$+\infty$
10	⟨{syntax:modifier}, {lemma:pictogram} {lemma:pictogram}⟩	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: ⟨{syntax:modifier}, {subword:_j}⟩.

(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: **Mdrrr 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: Looool 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: ⟨{subword:_t}, {pos:verb}⟩.

(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: ⟨{subword:_toi_}⟩.

(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: ⟨{pos:adjective}, {lemma:gros}⟩

(12) @X @X @X en **gros** hier sur fall guys coro a dit que la K corp perderait 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 guide [8] (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: ⟨{pos:proper-name} , {pos:subject-clitic, morpho:3eme}, {pos:verb, morpho:present}⟩.

(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: ⟨{subword:_c_}⟩.

(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: ⟨{lemma:rajouter}⟩.

(18) Rt si tu veux que j'te **rajoute** url_path (*Rt if you want me to add you url_path*)

(19) @X Mdrr desac pas stv jte **rajoute** ds un grp le sang (*@X Mdrr not desactives 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: ⟨{subword:sh_}⟩.

(20) P T D R. Vous suez face à Brest we**sh** 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 cra**sh** 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 : ⟨{subword:rrr_}⟩.

(22) dans 1 semaine chuis en Suède mdr**rrr** rien n'est prêt c'est la panik (*in 1 week i'm in sweden looool nothing is ready it's panic*)

(23) je suis mort Djoko disqualifié parce qu'il a mis une tête à un juge pt-drrrrrr**rrr** (*I'm dead Djoko disqualified because he headbutted a judge looooooooooooool*)

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: ⟨{syntax:modifier}, {lemma:pictogram}, {lemma:pictogram}⟩.

(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	⟨{subword:vous_}⟩	$+\infty$
2	⟨{subword:_Pour_}⟩	$+\infty$
3	⟨{lemma:alors, syntax:modifier, pos:adverb}⟩	$+\infty$
4	⟨{pos:common name, subword:_#}, {pos:ponctuation}⟩	$+\infty$
5	⟨{lemma:de}, {subword:_#}⟩	$+\infty$
6	⟨{subword:_entre} {subword:_#}⟩	$+\infty$
7	⟨{syntax:subject} {lemma:pictogram}⟩	$+\infty$
8	⟨{subword:lance_}⟩	$+\infty$
9	⟨{subword:hui_}⟩	$+\infty$
10	⟨{subword:ez_}⟩	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: ⟨{subword:vous_} ⟩. 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: ⟨{subword:_Pour_}⟩.

(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, ⟨{lemma:alors, syntax:modifier, pos:adverb}⟩, is used to structure tweets from (30) to (31).

(30) @X Si moi, petit écrivaillon 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:ponctuation}⟩.

(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'#AssembleedeBretagne a mis en évidence l'absurdite actuelle de la distribution des #competences **entre** #collterr 🇫🇷 . L'avenir est à l'#intégration des politiques publiques ➡️ #ODD #agenda2030. Il nous faut des #Regions et #Villes au pouvoir plus intégré. url_path url_path (*Our policy brief on the #BrittanyAssembly highlighted the current absurdity of the distribution of competences between #collterr 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 bienveil**lance** 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 surveil**lance** 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 #FRANÇOISASSELINEAU #ecology url_path*)

The emerging sequential pattern 9, ⟨{subword:hui_}⟩, 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. #AdamaViolеur 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. #AdamaViolеur 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: ⟨{subword:ez_} ⟩.

(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 connaiss**ez** 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 av**ez** 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., Léglise 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 Superieur, 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.

De l'annotation intégrale à l'analyse des réseaux de personnages : un modèle pour la REN dans les textes littéraires en français

Motasem Alrahabi*— Arthur Amalvy— Vincent Labatut**— Perrine Maurel***

* *Sorbonne Université, ObTIC*
** *Avignon Université, Laboratoire Informatique d'Avignon – UPR 4128*

RÉSUMÉ. *Les corpus textuels sont au cœur des humanités numériques, mais leur exploitation à grande échelle reste limitée par l'absence d'outils adaptés, notamment pour la reconnaissance d'entités nommées. Pour y remédier, nous présentons un corpus de sept romans français du* XIX^{e} *siècle, annotés intégralement. Cette ressource permet d'entraîner et d'évaluer des outils robustes aux spécificités littéraires. Nous proposons un modèle de REN basé sur CamemBERT, performant sur ce type de textes, et montrons son intérêt via l'extraction de réseaux de personnages, ouvrant de nouvelles perspectives sur les dynamiques narratives. Les ressources sont librement accessibles en ligne.*

MOTS-CLÉS : *reconnaissance d'entités nommées, corpus littéraires annotés, réseaux de personnages, apprentissage supervisé.*

TITLE. *From complete annotation to character network analysis: A NER model for literary texts in French*

ABSTRACT. *Textual corpora are central to Digital Humanities, but their large-scale use remains limited by the lack of suitable tools, particularly for Named Entity Recognition. To address this, we present a fully annotated corpus of seven 19th-century French novels. This resource enables the training and evaluation of tools robust to literary specificities. We propose a CamemBERT-based NER model, effective on such texts, and demonstrate its relevance through character network extraction, opening new perspectives on narrative dynamics. All resources are freely available online.*

KEYWORDS: *Named entity recognition, annotated literary corpora, Network analysis, supervised learning.*

1. Introduction

La constitution de corpus textuels dans le domaine des humanités numériques (HN) est en plein essor (Ehrmann *et al.*, 2023). À mesure que ces ressources gagnent en volume et en complexité, leur exploration nécessite des instruments d'analyse performants, conçus pour accomplir une variété de tâches en traitement automatique des langues (TAL). Parmi celles-ci, la reconnaissance d'entités nommées (REN), associée à des tâches comme la résolution de coréférence, joue un rôle clé dans l'interprétation et la valorisation des corpus dans différents contextes. En littérature, des problématiques fondamentales liées aux personnages, aux lieux et aux dimensions temporelles occupent une place centrale et ouvrent la voie à de nombreuses applications : identification des noms de lieux (Berragan *et al.*, 2023), analyse des réseaux de personnages (Labatut et Bost, 2019), recherche documentaire guidée par entités (Guo *et al.*, 2009), enrichissement sémantique par lien aux bases de connaissance (Labusch et Neudecker, 2022), ou encore reconstruction de biographies à partir de mentions dispersées dans les corpus (Fokkens *et al.*, 2018). Il faut cependant noter que les corpus textuels issus des HN pourraient être davantage valorisés grâce à des outils d'analyse adaptés (Egloff et Picca, 2020). Ceux-ci, étant généralement issus de l'application de méthodes d'apprentissage supervisé, souffrent d'un défaut de données annotées. Si certaines tâches bénéficient déjà de corpus ou de modèles adaptés pour le français — comme le corpus Democrat pour la résolution de coréférences (Landragin, 2021), ou celui de Durandard *et al.* (2023) pour la détection du discours direct — d'autres, en revanche, restent moins couvertes. C'est notamment le cas de la REN et de la résolution d'alias, lacunes que nous nous proposons de combler dans cet article.

L'une des principales difficultés de la REN réside dans la variabilité des performances des modèles selon la langue, l'époque ou le domaine des textes analysés (Liu *et al.*, 2021), une difficulté accentuée dans les corpus littéraires et historiques, qui se caractérisent par une grande diversité de types de texte, un bruit important lié à l'OCR, des variations diachroniques des conventions d'écriture, et un manque de ressources annotées adaptées. Les meilleures approches existantes (Ehrmann *et al.*, 2023 ; Yamada *et al.*, 2020) obtiennent d'excellents résultats sur des documents contemporains courts, comme les textes journalistiques ou encyclopédiques, qui constituent l'essentiel de leurs données d'apprentissage. Cependant, cette efficacité diminue sensiblement dans des contextes plus complexes, tels que les textes littéraires, notamment narratifs, qui sont marqués par une grande richesse et diversité stylistique (Silva et Moro, 2024). Cette difficulté est accentuée par la variation diachronique, un aspect souvent négligé qui exige des outils capables de traiter les particularités des textes historiques (Ehrmann *et al.*, 2023). Par ailleurs, les défis sont encore plus importants pour les langues peu documentées ou éloignées de l'anglais, pour lesquelles les ressources et les outils sont largement insuffisants. Pour surmonter ces obstacles, il est crucial de concevoir des corpus et des modèles adaptés, permettant ainsi d'améliorer les performances des systèmes REN tout en répondant à des besoins plus spécifiques et diversifiés en TAL.

Notre étude, qui vise à combler ce manque, apporte une contribution majeure en proposant un nouveau corpus de référence (ou *gold standard*) en français, annoté en entités nommées (EN), ainsi qu'un nouveau modèle de REN spécifiquement adapté aux documents littéraires. Notre corpus se compose de romans français du XIX^{e} siècle, annotés dans leur intégralité, sans échantillonnage. Nous adaptons un guide d'annotation existant pour créer ces ressources, ce qui nous permet d'entraîner un modèle de REN et d'évaluer la pertinence des architectures neuronales sur ce type de documents. En termes d'application, nous démontrons l'intérêt de ce modèle en l'utilisant pour extraire des réseaux de personnages de romans (section 4.3). Pour les besoins de cette analyse ainsi que pour remédier au manque de jeux de données disponibles, nous ajoutons à notre corpus une couche d'annotation ciblant la tâche de résolution d'alias, c.-à-d. le fait d'identifier les différentes expressions faisant référence à un même personnage.

L'article est organisé en quatre parties. Nous faisons tout d'abord le point sur les principales contributions concernant la REN dans le domaine du TAL (section 2). Nous présentons ensuite notre corpus annoté en EN, en décrivant la méthode suivie et les difficultés rencontrées (section 3). La partie suivante porte sur l'entraînement de notre modèle de REN et sur son application à l'extraction de réseaux de personnages (section 4). Nous concluons avec une discussion des apports du modèle proposé, mais aussi de ses limites (section 5), et en suggérant des pistes d'amélioration pour les travaux futurs.

2. REN et résolution d'alias : un bref état de l'art

Dans cette section, nous abordons dans un premier temps les méthodes existantes conçues pour reconnaître les EN dans des textes littéraires (section 2.1), avant de fournir une brève description des corpus disponibles pour entraîner et pour évaluer des méthodes automatiques à effectuer cette tâche (section 2.2). Nous traitons également du cas de la résolution d'alias dans la section 2.3.

2.1. *Approches et méthodes de REN*

La REN dans les textes littéraires est complexifiée par un certain nombre de facteurs spécifiques qui les distinguent suffisamment d'autres domaines pour avoir un effet mesurable sur les performances. On peut notamment mentionner la diversité linguistique et stylistique (archaïsmes, néologismes, figures de style), les références implicites (périphrases, titres, surnoms), les contextes narratifs complexes (variations selon les narrateurs ou les dialogues), l'ambiguïté (noms communs utilisés comme noms propres), ainsi que des entités enrichies ou spécifiques influencées par le multilinguisme et les références culturelles (références mythologiques, objets symboliques, entités fictives) (van Dalen-Oskam *et al.*, 2014 ; Dekker *et al.*, 2019 ; Labatut et Bost, 2019).

Plusieurs méthodes ont donc été développées pour la REN dans les corpus littéraires, allant des approches à base de règles et d'expressions régulières aux modèles d'apprentissage automatique supervisés et non supervisés. Les méthodes à base de règles, qui reposent sur des répertoires géographiques (*gazetteers*) et sur des règles spécifiques au domaine, permettent de reconnaître les entités (Moncla *et al.*, 2017), mais leur manque de généralisation constitue une limite importante. Les approches d'apprentissage automatique traditionnel ont établi des bases solides pour la REN (Kogkitsidou et Gambette, 2020). Plus récemment, les méthodes d'apprentissage profond, qui intègrent les plongements lexicaux contextuels et l'apprentissage par transfert, surpassent nettement les autres (Sprugnoli, 2018), mais elles nécessitent en contrepartie de grands corpus annotés pour leur entraînement. L'arrivée des *Transformers* préentraînés, notamment avec des modèles comme BERT et GPT, a marqué une avancée majeure, offrant une meilleure prise en compte du contexte et des spécificités syntaxiques des textes littéraires (Devlin *et al.*, 2019). En français, des modèles comme CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020) sont adaptés aux particularités de la langue, tandis qu'en anglais, des modèles préentraînés tels que BERT et RoBERTa (Liu *et al.*, 2019) sont souvent affinés sur des corpus littéraires.

L'évolution de ces méthodes, des règles aux modèles *Transformers* préentraînés, illustre des avancées majeures, mais les textes littéraires continuent à poser des défis à surmonter tels que la longueur des documents (Amalvy *et al.*, 2023), la difficulté à reconnaître certains noms particuliers (Dekker *et al.*, 2019), la richesse stylistique (Silva et Moro, 2024), le manque de données ou encore l'ambiguïté de certaines entités. Pour répondre à ces enjeux, Ehrmann *et al.* (2023) proposent des stratégies telles que l'apprentissage par transfert, l'apprentissage actif, l'utilisation de modèles de langage historiques, ainsi que la normalisation et l'entraînement sur des données adaptées au contexte.

2.2. *Corpus annotés en REN*

Un petit nombre de travaux s'intéressent à l'annotation de données en REN dans les textes littéraires. Bamman *et al.* (2019) introduisent LitBank, reposant sur des extraits de 100 œuvres du domaine public en anglais (environ 211 000 tokens au total), et couvrant six catégories d'entité définies par ACE 2005 (Linguistic Data Consortium, 2005). De même, Dekker *et al.* (2019) ont créé OWTO, une ressource basée sur 40 chapitres extraits d'autant de romans en anglais (300 phrases par roman), et axée sur les personnages. Dans Frontini *et al.* (2020), des extraits de 400 mots ont été annotés dans chacune des neuf langues du corpus ELTeC (European Literary Text Collection) (Schöch *et al.*, 2021). Les étiquettes comprenaient des catégories telles que les gentilés, les professions, les œuvres d'art, les personnes, les lieux, les événements et les organisations, mais elles ont été simplifiées par la suite pour éviter les annotations imbriquées et les chevauchements (Frontini *et al.*, 2020). Alrahabi *et al.* (2024) ont proposé un corpus annoté d'extraits de trois romans français du XIX^{e} siècle,

couvrant trois catégories d'entité (personnage, lieu et divers) avec une granularité fine pour chaque catégorie. Enfin, le corpus Travel Writings (Sprugnoli, 2018) couvre 38 récits de voyage en anglais (de 1850 à 1940) avec 100 000 tokens annotés, et se concentre sur les entités géographiques.

En complément des corpus annotés manuellement, certains projets utilisent des annotations automatiques, souvent sur des volumes plus importants. Par exemple, BDCamões (Grilo *et al.*, 2020) est un vaste corpus portugais OCRisé (4 millions de mots), couvrant 14 genres littéraires sur cinq siècles, avec des annotations automatiques pour les lieux, organisations, œuvres, événements et entités diverses. En français, GeoNER (Kogkitsidou et Gambette, 2020) se concentre sur trois textes littéraires des XVI^{e} et XVII^{e} siècles, explorant l'impact de la normalisation historique sur la REN géographique.

Du point de vue méthodologique, il est à noter que les guides d'annotation proposés concernent toujours des textes courts ou limités. Jørgensen *et al.* (2020) forment leur guide à partir de l'annotation d'un corpus d'environ 600 000 tokens (soit moitié moins que notre corpus), composé principalement de textes journalistiques, mais aussi de rapports gouvernementaux et de blogs. Le guide ACE08 (Li, 2008) concerne deux corpus linguistiques faits d'extraits de blog, d'actualité, de conversation à la radio ou au téléphone, qui sont des formats courts par nature. Ainsi, le corpus anglophone est composé de six sous-corpus d'une longueur moyenne de 44 166 mots, et le corpus arabophone est composé de trois sous-corpus d'une longueur moyenne de 35 000 mots.

Ces choix méthodologiques illustrent une tendance générale : malgré la diversité des travaux existants, une limitation commune réside dans l'échantillonnage partiel des œuvres littéraires, où seules des portions — parfois très courtes — sont annotées. Cela restreint leur exploitation en termes d'évaluation, car les modèles existants peuvent généralement prendre en entrée l'entièreté de ces extraits. Cependant, les limites de taille de contexte des modèles ou le manque de ressources de calcul mènent souvent à l'application d'une fenêtre glissante quand il s'agit de longs documents. Cette pratique est préjudiciable à la performance de la tâche de REN car les modèles n'ont pas toujours à disposition suffisamment de contexte à l'inférence (Amalvy *et al.*, 2023), ce que ne permettent pas de mesurer des corpus composés de courts documents. Afin de surmonter cette limitation, nous constituons un corpus de romans français entièrement annotés en EN, particulièrement utile pour l'analyse des textes littéraires longs.

2.3. *Résolution d'alias*

La tâche de résolution d'alias consiste à extraire, pour un personnage, tous les alias qui y font référence dans le texte. Dans cet article, nous considérons comme alias d'un personnage toute mention qui serait annotée lors de la phase d'annotation de REN : nous excluons donc les mentions génériques comme les pronoms. Cette

tâche peut être considérée comme une version de la résolution de coréférences limitée aux personnages. Dans le cas de romans complets, cette tâche est considérée comme très difficile, et les approches neuronales récentes se heurtent au coût computationnel du traitement de textes aussi longs (Guo *et al.*, 2023 ; Gupta *et al.*, 2024). La résolution d'alias hérite de ces difficultés, et plusieurs approches se basent donc plutôt sur un ensemble de règles (Vala *et al.*, 2015 ; Cuesta-Lazaro *et al.*, 2022), utilisant une première passe de résolution de coréférences comme une entrée du système. Il s'agit d'une tâche moins étudiée que la REN, et plusieurs auteurs ne s'y intéressent que pour résoudre des problèmes de plus haut niveau, comme l'attribution du locuteur (Cuesta-Lazaro *et al.*, 2022) ou la classification de personnages (Bamman *et al.*, 2014).

La résolution d'alias constitue donc une tâche de bas niveau importante qui précède l'exploitation de textes littéraires lorsqu'on s'intéresse à leurs personnages. Au vu de son importance, nous enrichissons les annotations de notre corpus de REN en relevant également les alias des personnages, ce qui constitue une première pour des romans complets en français. Ces annotations permettront de développer et d'évaluer des systèmes de résolution d'alias à une échelle réaliste.

3. Constitution du corpus

Cette section décrit les méthodes que nous mettons en œuvre pour élaborer notre corpus : sélection des textes (section 3.1), règles appliquées (section 3.2) et déroulement du processus d'annotation (section 3.3). Elle fournit également une brève évaluation du corpus annoté (section 3.4) et revient sur les difficultés rencontrées (section 3.5).

3.1. *Romans sélectionnés*

Notre étude porte principalement sur les romans français du XIX^{e} siècle, un genre qui s'est imposé comme modèle en raison de sa structure narrative distincte, articulée autour des figures de l'auteur, du narrateur et des personnages. Ces romans, caractérisés par une narration souvent au passé et à la troisième personne, conjuguent récit et description. Cette alternance permet aux auteurs de faire avancer l'intrigue tout en marquant des pauses pour installer le décor, ancrant ainsi l'action dans un environnement qui reflète la société de l'époque. Un autre avantage concret de ce choix est que les œuvres produites à cette période sont libres de droits, ce qui facilite le partage de romans entiers sans soulever de problème juridique.

Notre corpus se compose de sept romans, sélectionnés pour leur richesse en personnages et en intrigues, ainsi que pour leur représentativité des divers courants littéraires. Ces œuvres suivent les trois grandes étapes du processus de canonisation décrites par Evans (2000) : elles sont sélectionnées par les critiques pour leur mérite esthétique, validées par les enseignants et universitaires pour l'enseignement,

puis pérennisées par les éditeurs à travers des rééditions successives. Les textes sélectionnés sont listés et décrits dans le tableau 1.

Roman	**Auteur**	**Publication**	**Tokens**
Les Trois Mousquetaires	Alexandre Dumas	1849	294 989
Le Rouge et le Noir	Stendhal	1854	216 445
Eugénie Grandet	Honoré de Balzac	1855	80 659
Germinal	Émile Zola	1885	220 273
Bel-Ami	Guy de Maupassant	1901	138 156
Notre-Dame de Paris	Victor Hugo	1904	221 351
Madame Bovary	Gustave Flaubert	1910	148 861

TABLEAU 1. *Liste des romans constituant notre corpus avec, pour chacun, l'auteur, l'année de publication, et le nombre de tokens*

Aucun prétraitement ni normalisation préalable n'a été appliqué aux documents collectés. Totalisant 1 320 734 tokens, l'ensemble des œuvres provient du site Wikisource[1]. Les URL correspondantes sont répertoriées en annexe (tableau 9).

3.2. *Règles d'annotation*

Le travail décrit par Sims et Bamman (2020) illustre la complexité accrue de l'annotation des EN dans les longs documents, en montrant que l'attribution de locuteurs et la résolution de coréférences posent des défis similaires. La fréquence élevée des mentions et la complexité narrative renforcent la nécessité de modèles adaptés pour les textes littéraires, avec des performances élevées pour garantir la cohérence dans ces textes. Ce constat est confirmé dans d'autres études telles que (Amalvy *et al.*, 2023 ; Tay *et al.*, 2021 ; Labatut et Bost, 2019). La dernière mentionne que la prose littéraire, plus complexe que la prose journalistique sur laquelle la plupart des modèles sont habituellement entraînés, affecte la performance des méthodes génériques pour des tâches NLP variées, comme la modélisation de l'intrigue ou la détection de personnages. Les auteurs soulignent également que les œuvres de fiction possèdent des caractéristiques spécifiques rendant inefficaces certains outils standards de traitement, ce qui renforce la nécessité de méthodes adaptées pour garantir la cohérence et la précision dans l'annotation de textes littéraires longs.

Pour l'annotation de notre corpus en EN, nous choisissons d'utiliser un guide récent, proposé dans le cadre du projet Universal NER[2] (Mayhew *et al.*, 2024). Nous l'avons sélectionné en raison de de son caractère très généraliste, afin de nous permettre de faire évoluer progressivement nos directives d'annotation et de les

1. https://fr.wikisource.org/wiki/Wikisource:Accueil
2. https://www.universalner.org/guidelines/

adapter aux besoins spécifiques de notre corpus. Universal NER, qui est lui-même basé sur le projet Universal Dependencies (UD)[3], propose un guide d'annotation multilingue et un ensemble de données standardisées, visant à établir des bases cohérentes pour la REN dans une dizaine de langues typologiquement variées. Dans sa version actuelle, il couvre notamment l'anglais, le chinois, le russe, le danois et le serbe, mais il n'inclut pas encore le français. Selon le guide d'Universal NER, tous les noms de personnes (réelles ou fictives), d'organisations et de lieux doivent être annotés, à condition que le mot soit ou contienne un nom propre et qu'il ait une référence unique pointant vers une entité spécifique et qui soit constante dans le temps. Le jeu de balises comprend trois types principaux d'entité : les *Personnes* (PERS), les *Lieux* (LOC) et les *Organisations* (ORG), ainsi qu'une catégorie supplémentaire, *Autre* (OTHER) pour les entités jugées pertinentes mais ne correspondant pas à ces types principaux. Nous adaptons ces catégories et ces directives à notre corpus, comme détaillé ci-après.

3.3. *Déroulement de l'annotation*

L'équipe, composée d'un coordinateur de projet et de trois annotateurs, a mené une campagne d'annotation en trois phases reposant sur la plateforme libre d'étiquetage de données Label Studio[4].

Première phase : annotation exploratoire. Nous annotons un échantillon selon le guide, chaque passage étant traité par deux annotateurs. Nous classons dans les catégories PERS, LOC et ORG tout ce qui est explicitement défini et sans ambiguïté. Par exemple, dans le cas de PERS, cela inclut les humains, les animaux (« *— Qu'est-ce que cette Djali ? — C'est la chèvre.* »[5]), les figures mythologiques ou religieuses (« *le tendon d'Achille* », « *Dieu* »), les figures personnifiées (« *Le suisse, alors, se tenait sur le seuil, au milieu du portail à gauche, au-dessous de la Marianne dansant* ») et les figures fictives (« *les accents divins du désespoir de Caroline dans le Matrimonio segreto le firent fondre en larmes* »).

Parmi les premières consignes d'annotation que nous avons jugées importantes d'ajouter au guide, citons :

– annoter la mention continue la plus englobante d'une EN (« *la compagnie des gardes de M. Essart* » [ORG][6]) ;

3. https://universaldependencies.org/
4. https://github.com/HumanSignal/label-studio
5. Dans cet exemple comme dans l'ensemble de cet article, le passage souligné désigne l'extrait pertinent dans son contexte au sein de l'œuvre d'origine.
6. Certaines catégories, comme les dates ou les monnaies, ne sont pas prises en compte dans notre guide. Si la mention la plus englobante appartient à l'une de ces catégories (p. ex. « *soie Louis XVI* »), elle n'est pas annotée. On examine alors la portion la plus restreinte pour déterminer si une annotation est nécessaire.

– ne pas annoter les EN imbriquées (p. ex. « *M. Essart* » dans l'exemple précédent) ;

– annoter les descripteurs qui accompagnent les EN (p. ex. « *la rue Boursault* », « *le prince de Guerche* »).

Deuxième phase : concertation des annotateurs. Cette étape permet à l'équipe d'analyser les difficultés et les désaccords de la première phase, et d'adapter le guide à nos besoins. Parmi ces ajustements, citons notamment :

– annoter les appellations et les titres nobiliaires, car le statut social des différents personnages est souvent central au récit (p. ex. « *Sa Majesté le roi Charles X* [PERS] ») ;

– annoter les déterminants et les prépositions (p. ex. « *le roi* ») car ils permettent parfois de préciser le genre de la personne[7] ;

– nous ajustons l'étiquette (OTHER), définie dans le guide d'Universal NER pour inclure des catégories comme les nationalités et les langues, afin de mieux répondre à nos besoins spécifiques. Nous l'utilisons pour annoter des EN associées aux trois catégories principales, mais qui sont *indéterminées* ou *ambiguës*. Pour nous, la notion d'*indéterminé* concerne les entités qui renvoient à une référence spécifique dans le roman, mais qui, une fois sorties de ce contexte, ne renvoient plus à une personne, à un lieu ou à une organisation précise[8]. Cette catégorie inclut les noms communs représentatifs (« *Tout le monde s'inclina vers le Patron* », « *la bohémienne* », « *les environs de Paris* », « *Français contre Français* », etc.). La notion d'*ambigu*, quant à elle, regroupe les entités que le contexte ne permet pas de classer clairement dans l'une de nos trois catégories. Dans l'exemple « *être responsable par-devant Notre-Dame la Critique* », le contexte manque de précisions pour trancher, et la catégorisation de *Notre-Dame* reste ambiguë : bien qu'il s'agisse de la cathédrale, elle est ici personnifiée.

Troisième phase : annotation finale. Le guide d'annotation, désormais adapté, est appliqué à l'ensemble du corpus. Comme dans la première étape, chaque phrase est traitée par deux annotateurs pour garantir la qualité et la cohérence.

Le tableau 2 présente le nombre total d'entités annotées par roman, ainsi que par catégorie, réalisées par les trois annotateurs. Le corpus produit est publiquement accessible en ligne[9].

7. En français, les particules définies (*le, la, les, l'*) indiquent le genre, le nombre et parfois l'élision, ce qui les rend cruciales pour l'analyse, contrairement à l'anglais où *the* est invariable.
8. L'un des avantages de notre démarche est que les annotateurs disposent d'une vision globale du contexte du roman, contrairement à certains travaux où seuls des échantillons sont annotés.
9. https://github.com/obtic-sorbonne/7-romans

Roman	Nombre de mentions d'entité					Tokens par
	PERS	LOC	ORG	OTHER	**Total**	**mention d'entité**
Les Trois Mousquetaires	8 583	962	45	255	9 845	1,90
Le Rouge et le Noir	4 351	806	39	78	5 274	1,89
Eugénie Grandet	1 389	187	13	34	1 623	1,65
Germinal	3 439	838	195	40	4 512	1,43
Bel-Ami	1 773	400	25	46	2 244	1,77
Notre-Dame de Paris	3 631	1 409	17	94	5 151	2,01
Madame Bovary	2 004	403	11	70	2 488	1,54
Total pour le corpus	25 170	5 005	345	617	31 137	1,80

TABLEAU 2. *Nombre de mentions d'entité par roman et par type, et moyenne du nombre de tokens qui les composent*

3.4. *Accord inter-annotateur*

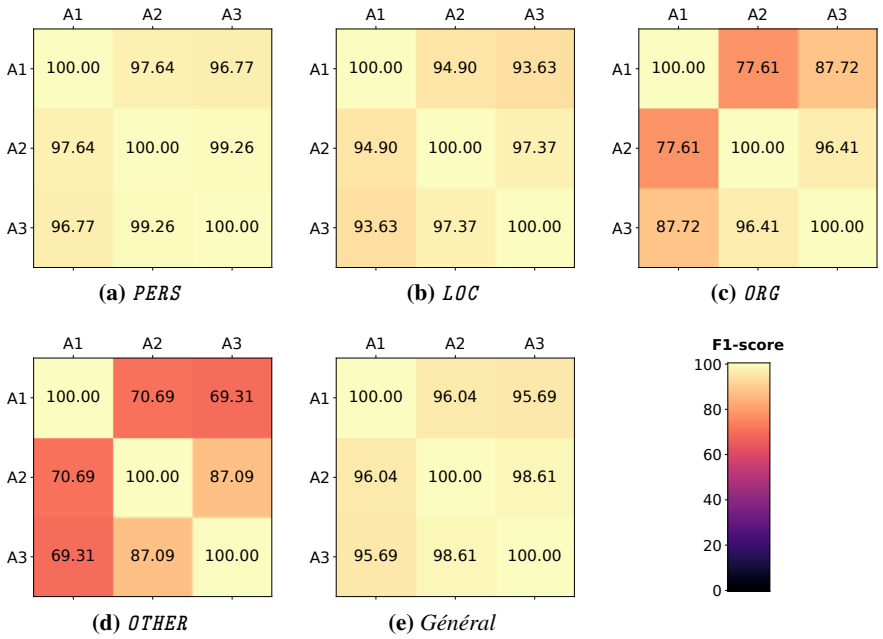
Nous avons fait appel à trois annotateurs aux profils très diversifiés. A1 correspond à un annotateur débutant, au profil mixte entre littérature et informatique. A2 correspond à un annotateur semi-expérimenté, au profil littéraire, qui avait déjà participé à des campagnes d'annotation à plusieurs mains. A3 correspond à un annotateur débutant, au profil informatique.

Nous mesurons l'accord inter-annotateur en calculant le F1-score au niveau des entités, et le coefficient Kappa de Cohen (Cohen, 1960) au niveau des tokens. Nous calculons Kappa en nous limitant aux tokens identifiés comme appartenant à une entité par au moins un annotateur, ignorant ceux qui ne sont annotés par personne. Cette méthode permet d'éviter de surestimer l'accord, en excluant les tokens qui n'ont aucune probabilité d'appartenir à une entité, ces derniers étant majoritaires (Brandsen *et al.*, 2020).

Classe	F1-score	Kappa
PERS	97,89	93,71
LOC	95,30	86,45
ORG	87,25	75,57
OTHER	75,70	47,70
Moyenne	89,04	75,86
Global	96,78	93,35

TABLEAU 3. *Accord inter-annotateur par type d'entité. Les deux dernières lignes indiquent respectivement la moyenne par classe et le score obtenu en considérant toutes les entités sans distinction de classe.*

En considérant toutes les entités sans distinction de classe, nous obtenons un F1-score de 96,78 et un Kappa de Cohen de 93,35, ce qui confirme la cohérence de nos annotations. Mais ces scores agrégés peuvent masquer des disparités, aussi nous les décomposons selon deux dimensions : le type d'entité et l'annotateur. Le tableau 3 détaille l'accord inter-annotateur par classe d'entité : bien que l'accord global soit excellent, la classe `OTHER` semble poser davantage de difficultés, comme nous le détaillons dans la section suivante. La figure 1 donne le détail de l'accord par paire d'annotateurs, et montre que les scores plus faibles pour `ORG` et `OTHER` sont dus à un désaccord plus prononcé du premier annotateur.



(a) *PERS* **(b)** *LOC* **(c)** *ORG*

(d) *OTHER* **(e)** *Général*

FIGURE 1. *Accord inter-annotateur exprimé en F1-score, pour chaque paire d'annotateurs et type d'EN*

3.5. *Difficultés rencontrées*

L'annotation des EN présente plusieurs défis, tenant à la diversité des catégories, des styles littéraires et des contextes d'utilisation. Nous les abordons en harmonisant les annotations et en distinguant deux types principaux de désaccord.

Le premier type regroupe les désaccords faciles à résoudre, qui sont dus à des oublis, à des incertitudes ou à des erreurs d'interprétation. Voici quelques exemples :

– dans la phrase « *Je fais le Sénat au Salut, et, de temps en temps, des chroniques littéraires pour la Planète* », l'entité *La Planète* a été annotée à tort par un annotateur comme `LOC` au lieu d'`ORG` ;

– « *On les nommait à la Chambre la bande à Walter* » : après concertation, nous optons pour `ORG` plutôt que `LOC` ;

– dans la phrase « *Sera à la Saint-Jean gelée,* », le nom propre *Saint-Jean* pourrait porter à confusion, car il s'agit, avec le déterminant, d'un évènement empruntant le nom d'une personne par antonomase : nous annotons la mention la plus restreinte (*Saint-Jean*) en `PERS`.

Le second type concerne les désaccords de fond, plus difficiles à résoudre, car complexes et porteurs d'une certaine ambiguïté. Ils relèvent fréquemment de la catégorie `OTHER`, créée et ajustée lors de la deuxième phase. Voici quelques exemples :

– dans l'énoncé : « *Elle s'empara d'un prie-Dieu et s'agenouilla.* », l'entité nommée est relevée par tous les annotateurs, mais sa catégorisation suscite un débat. En effet, *prie-Dieu* désigne un objet, mais une lecture littérale du terme évoque également une référence au personnage divin (*Dieu*), soulignant l'usage de l'objet pour prier ce dernier ;

– les notions « indéterminé » et « ambigu » de la catégorie `OTHER` peuvent varier en fonction de l'interprétation de l'annotateur, comme dans le cas de « *l'affaire Morel* », qui pourrait être considérée comme `OTHER` ou comme `PERS`, auquel cas seul le mot *Morel* serait annoté ;

– l'inclusion ou l'exclusion de certains syntagmes dans les EN, comme *Français* ou *huguenots* dans l'exemple qui suit, soulève des questionnements, la capitalisation reflétant souvent des conventions typographiques plutôt qu'une intention emphatique de l'auteur : « *où Français devaient combattre contre Français [...] En effet, le sac de La Rochelle, l'assassinat de trois ou quatre mille huguenots [...]* » ;

– les variations stylistiques entre romans et auteurs ont profondément influencé l'évolution du guide d'annotation. Par exemple, Victor Hugo et Alexandre Dumas se distinguent par un usage particulièrement fréquent de noms communs pour désigner certains personnages, contrairement à d'autres auteurs du corpus qui privilégient généralement les noms propres. Ainsi, dans *Notre-Dame de Paris*, une approche d'annotation basée uniquement sur les noms propres ne couvre que 28 % des mentions d'*Esmeralda*, tandis que les désignations par « *l'égyptienne* » et « *la bohémienne* » représentent respectivement 45 % et 26 % de ces mentions.

Ces quelques cas illustrent la complexité de l'annotation des EN dans les textes littéraires, nécessitant une approche flexible et adaptée aux enjeux spécifiques de chaque contexte.

4. Modèle et application proposés pour la REN

Nous mettons à profit notre corpus en entraînant un modèle de REN sur celui-ci (section 4.1). Afin de mobiliser nos annotations dans l'entraînement du modèle, la majorité des désaccords de fond furent levés afin de ne présenter qu'une seule et même annotation au modèle : pour cela, nous avons privilégié l'une des interprétations au profit de l'autre en fonction du consensus atteint au sein de notre équipe. En ce qui concerne la seule exception que nous avons jugée irrésolvable, le modèle a reçu l'une des deux annotations de façon aléatoire. Compte tenu du manque de jeux de données littéraires en français, ce modèle entend avancer l'état de l'art en permettant une bonne performance d'inférence sur ce type de texte (section 4.2). Nous illustrons son utilité sur une tâche d'extraction de réseaux de personnages (section 4.3).

4.1. *Méthode d'entraînement*

Il existe principalement deux modèles encodeurs préentraînés à base de Transformers pour le français : CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020). Leurs architectures, basées sur BERT (Devlin *et al.*, 2019), sont très proches, de même que leurs performances sur différentes tâches.

Nous nous basons sur le modèle CamemBERT-base, de 110 millions de paramètres. Nous ajoutons à celui-ci une couche neuronale de classification, et formalisons la tâche de REN comme un problème de classification de tokens en nous basant sur le format BIO (Ramshaw et Marcus, 1995). Nous effectuons un ajustement fin de nos modèles sur notre corpus entier pendant un maximum de dix cycles d'apprentissage avec arrêt prématuré, avec un taux d'apprentissage de $1 \cdot 10^{-5}$. En pratique, notre stratégie d'arrêt prématuré revient à entraîner le modèle pendant environ trois cycles d'apprentissage. Pour traiter le déséquilibre entre les classes, nous pondérons notre fonction de coût en divisant le nombre d'exemples de la classe ayant le plus d'exemples (`PERS`) par le nombre d'exemples de chaque classe. Pendant l'entraînement et l'inférence, le modèle prédit les entités présentes dans chaque paragraphe sans contexte supplémentaire. Nous distribuons notre modèle librement en ligne[10].

4.2. *Évaluation du modèle*

Afin d'évaluer la performance de notre modèle, nous procédons par validation croisée en 7 blocs : nous produisons 7 modèles différents, chacun entraîné sur un ensemble unique de 5 livres, avec un jeu de développement d'un livre, et évalué sur le dernier livre restant (nous nommons cette configuration *romans complets*). Pour donner une idée de la meilleure performance possible du modèle, nous réalisons

10. http://huggingface.co/compnet-renard/camembert-base-literary-NER-v2

une expérience supplémentaire où nous utilisons un jeu de test correspondant à un septième de chaque roman, un jeu de développement correspondant à un autre septième, et un jeu d'entraînement correspondant à cinq septièmes de chaque roman (configuration *romans échantillonnés*). Cette expérience mesure la performance du modèle lorsque celui-ci a observé des entités du jeu de test au moment de l'apprentissage.

Nous utilisons la librairie `seqeval`[11] dans son mode par défaut, et rapportons nos scores en termes de précision, rappel et F1-score. Le tableau 4 recense les résultats que nous obtenons sur les différents romans, tandis que le tableau 5 détaille les résultats par types d'entité. Nous ne comparons pas notre modèle avec d'autres modèles entraînés sur des textes de domaines différents, car les différences de modalité d'annotation entre corpus les pénaliseraient injustement.

Roman	**F1**	**Précision**	**Rappel**
Les Trois Mousquetaires	71,15	68,49	74,02
Le Rouge et le Noir	88,97	85,23	93,04
Eugénie Grandet	88,56	85,58	91,77
Germinal	89,94	87,54	92,48
Bel-Ami	87,13	82,47	92,34
Notre-Dame de Paris	75,70	73,48	78,04
Madame Bovary	88,25	84,02	92,92
Romans complets	81,21	78,19	84,48
Romans échantillonnés	91,53	88,12	95,21

TABLEAU 4. *Performance de notre modèle de REN évalué par validation croisée. Les résultats par roman sont obtenus dans la configuration* romans complets*.*

Dans l'ensemble, nos résultats sont cohérents avec ceux obtenus précédemment par CamemBERT sur d'autres corpus de domaines différents en français, comme sur le jeu de données journalistique French Treebank (Martin *et al.*, 2020). Nous observons une forte différence entre les configurations *romans échantillonnés* et *romans complets* : le fait de donner accès au modèle à des entités du jeu de test à l'entraînement augmente sensiblement la performance (plus de 10 points de F1-score). Deux romans posent plus de difficultés : *Notre-Dame de Paris* et *Les Trois Mousquetaires*. Dans le cas des *Trois Mousquetaires*, nous analysons manuellement les résultats et observons que le modèle a de grandes difficultés à reconnaître le personnage principal *d'Artagnan*. Ce résultat rejoint les observations de Dekker *et al.* (2019), qui notent que les mentions de personnages qui comportent des caractères spéciaux (en l'occurrence, une apostrophe) sont plus difficiles à détecter. Pour ce qui est de *Notre-Dame de Paris*, les erreurs de précision comme de rappel se concentrent sur les descriptions définies, qui ne contiennent pas de nom propre (*« le roi »*, *« le cardinal »*, etc.). Du côté des résultats par classe, si on observe de bonnes

11. https://github.com/chakki-works/seqeval

Classe	F1	Précision	Rappel
PERS	83,51	81,89	85,20
LOC	81,80	78,69	85,17
ORG	55,74	42,39	81,34
OTHER	34,08	25,15	52,86

TABLEAU 5. *Performance de notre modèle de REN sur les différentes classes*

performances sur les classes PERS et LOC, les classes ORG et OTHER sont très difficiles à détecter. Cela peut s'expliquer par la difficulté inhérente du problème (l'accord inter-annotateur étant plus faible pour ces deux classes comme indiqué dans le tableau 3, mais aussi par le déséquilibre des classes : notre corpus comporte presque 73 fois plus d'entités PERS que d'entités LOC.

Il est à noter que ces résultats sous-estiment probablement les capacités du modèle final que nous mettons à disposition, car celui-ci est entraîné sur l'entièreté des sept romans du corpus.

4.3. *Application à l'extraction de réseaux*

Nous nous tournons maintenant vers une application de notre modèle de REN au domaine de la littérature, à travers la tâche d'extraction de réseaux de personnages (Labatut et Bost, 2019). Par conséquent, elle se concentre sur les EN de type PERS.

Les réseaux de personnages sont des graphes où les sommets représentent des personnages et les arêtes traduisent les relations entre eux. Ils sont particulièrement utiles pour de nombreuses applications. D'une part, ils offrent une visualisation claire des relations entre les personnages d'une œuvre littéraire, permettant ainsi de comprendre intuitivement la dynamique des interactions. D'autre part, ils constituent un outil précieux pour l'analyse littéraire, en apportant une nouvelle perspective sur la structure narrative et les relations interpersonnelles. Enfin, ces réseaux fournissent une modélisation d'un texte narratif long sous la forme compacte d'un graphe d'interactions. Ce type de modèle peut être exploité pour traiter automatiquement différentes tâches, telles que la classification de genre littéraire (Hettinger *et al.*, 2015 ; Holanda *et al.*, 2019), la segmentation d'histoires (Min et Park, 2016) ou l'alignement narratif (Amalvy *et al.*, 2024a).

Extraction des réseaux. L'outil Renard[12] (Amalvy *et al.*, 2024b) est un pipeline modulaire sous forme de librairie Python permettant d'extraire des réseaux de personnages (statiques ou dynamiques) à partir de textes narratifs, et d'analyser

12. https://github.com/CompNet/Renard/

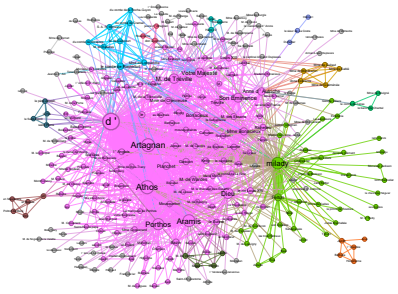
les relations entre personnages, offrant une visualisation de l'évolution des réseaux sociaux au fil du récit. Pour extraire ces réseaux, la librairie résout séquentiellement différentes tâches de TAL. La résolution de la tâche de REN permet d'abord de détecter les mentions des différents personnages dans le texte : il s'agit des entités PERS extraites par le système. Puis, la tâche de *résolution d'alias* permet de rassembler les mentions distinctes faisant référence à un même personnage. Finalement, la librairie extrait les interactions entre les personnages pour déterminer leurs relations. Ces relations peuvent être de différents types : cooccurrences, conversations, rencontres, etc. Compte tenu des annotations à notre disposition, nous nous concentrons sur l'extraction de réseaux de cooccurrences : nous considérons que deux personnages ont une interaction s'ils apparaissent à proximité l'un de l'autre dans le texte, ce qui constitue une approximation de leurs interactions réelles. Dans cette expérience, nous fixons arbitrairement le seuil de proximité à 32 tokens pour tous les romans. À plus long terme, un travail devra être mené pour proposer une méthode plus fondée visant à estimer ce paramètre. Afin de mesurer l'intérêt de notre modèle, nous proposons de l'utiliser lors de la phase de REN de `Renard`, puis de mesurer la qualité des réseaux extraits par rapport à des réseaux de référence.

Pour obtenir ces réseaux de référence, nous utilisons nos annotations en REN pour la classe PERS afin d'obtenir des mentions de référence. De plus, nous annotons manuellement les sept romans à notre disposition en résolution d'alias. Pour ce faire, nous nous inspirons du guide d'annotation du corpus littéraire *Novelties* (Amalvy et Labatut, 2024). Un annotateur (l'un des auteurs de cet article) a adapté les annotations existantes de ce corpus qui portent originellement sur les versions anglaises de nos romans. Nous notons que cette phase d'annotation nous a en outre permis de corriger de rares erreurs restantes dans nos annotations en REN, lorsque les alias à annoter semblaient incorrects. Nous indiquons le nombre de personnages ainsi annotés pour chaque roman dans le tableau 6.

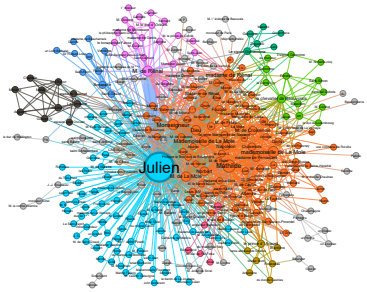
Roman	**Personnages**
Les Trois Mousquetaires	213
Le Rouge et le Noir	318
Eugénie Grandet	107
Germinal	102
Bel-Ami	150
Notre-Dame de Paris	536
Madame Bovary	175

TABLEAU 6. *Nombre total de personnages distincts annotés pour chaque roman*

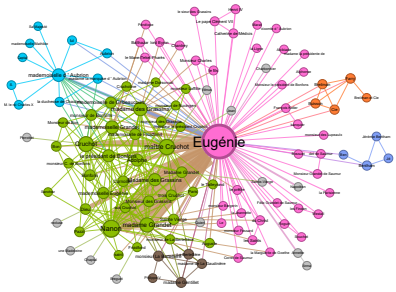
Les réseaux extraits en utilisant notre modèle de REN sont présentés dans la figure 2. Comme pour notre expérience d'évaluation de REN, nous utilisons le modèle dans un schéma de validation croisée pour éviter de surestimer la qualité des réseaux extraits. Le processus étant complètement automatique, on voit apparaître certaines erreurs. Par exemple, dans *Les Trois Mousquetaires*, la présence d'un



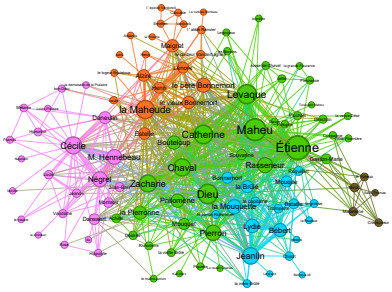
(a) *Les Trois Mousquetaires*



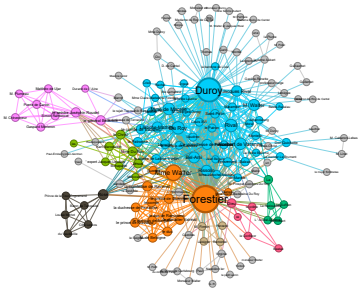
(b) *Le Rouge et le Noir*



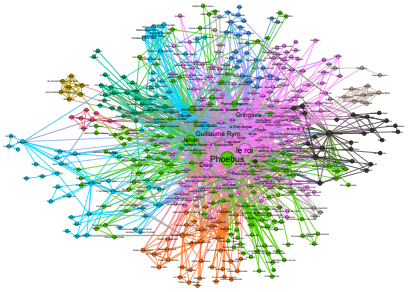
(c) *Eugénie Grandet*



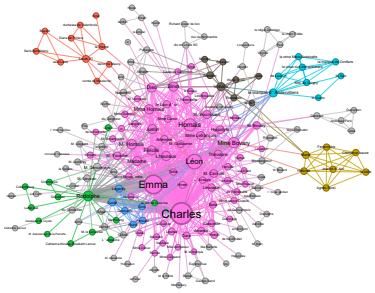
(d) *Germinal*



(e) *Bel-Ami*



(f) *Notre-Dame de Paris*



(g) *Madame Bovary*

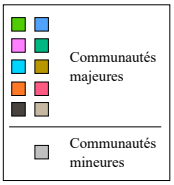


FIGURE 2. *Réseaux extraits par* `Renard` *en utilisant notre modèle de REN sur les différents romans du corpus. Le degré des sommets est indiqué par leur taille, tandis que l'épaisseur des arêtes représente leur poids. La couleur des sommets correspond à leur communauté (gris pour les communautés de trois sommets ou moins).*

sommet incorrect appelé *d'*, ou bien le *Cardinal de Richelieu* représenté par plusieurs sommets (*Son Éminence*, *Richelieu*). Nous revenons sur ce point plus loin, en proposant une évaluation quantitative de ces erreurs. La taille des sommets indique leur degré, leur couleur correspond à leur communauté, et l'épaisseur des arêtes représente leur poids. Les communautés constituent une partition de l'ensemble des sommets V. Nous les identifions au moyen de l'algorithme Girvan–Newman (Girvan et Newman, 2002), une méthode standard consistant à retirer itérativement les arêtes de plus grande intermédiarité. Dans le cadre de l'analyse de réseaux de personnages, les communautés sont souvent interprétées comme des sous-intrigues (Labatut et Bost, 2019). Visuellement, il est possible d'identifier plusieurs types de réseaux dans la figure 2. Certains romans sont clairement construits autour de leur protagoniste, qui centralise de nombreuses relations : c'est notamment le cas de *Le Rouge et le Noir* (Julien Sorel) et d'*Eugénie Grandet* (elle-même). D'autres sont au contraire structurés autour d'une opposition, comme *Les Trois Mousquetaires* (les mousquetaires contre Milady) et *Bel-Ami* (Duroy et Forestier). D'autres encore ont une nature plus chorale : *Germinal* et *Notre Dame de Paris* contiennent de très nombreux personnages, avec des interactions réparties de façon beaucoup plus uniforme. Certains réseaux mettent aussi en évidence l'importance de certaines relations dans l'histoire : le groupe des mousquetaires est très dense dans *Les Trois Mousquetaires*, le lien entre Julien et Mme de Rênal domine dans *Le Rouge et le Noir*, Eugénie et Nanon sont elles aussi fortement connectées dans *Eugénie Grandet*, de même qu'Emma et Charles dans *Madame Bovary*. Ces différentes observations sont cohérentes avec la lecture du roman, et indiquent intuitivement une bonne qualité des réseaux extraits.

Méthode d'évaluation. Afin de mesurer la qualité des réseaux extraits, nous utilisons des métriques qui permettent de mesurer la fidélité du graphe prédit $G_p = (V_p, E_p)$ en termes de sommets et d'arêtes par rapport à un graphe de référence $G_r = (V_r, E_r)$. Pour les sommets, nous utilisons la Précision, le Rappel et le F1 de sommet (Vala *et al.*, 2015) :

$$Pre_V = \max_{f_V} \frac{\sum_{u \in V_p} 1 - \frac{|u - f_V(u)|}{|u|}}{|V_p|} \quad [1]$$

$$Rec_V = \max_{g_V} \frac{\sum_{v \in V_r} [g_V(v) \cap v \neq v_\varnothing]}{|V_r|}, \quad [2]$$

où :

– chaque sommet de V_p et V_r représente l'ensemble des alias d'un personnage ;

– f_V est une fonction associant un sommet de G_p à un sommet de G_r ou au sommet nul $v_\varnothing$ si le personnage prédit n'a pas d'équivalent dans les personnages de référence ;

– similairement, g_V est une fonction associant un sommet de G_r à un sommet de G_p ou au sommet nul ;

– l'expression $[g_N(v) \cap v \neq v_\varnothing]$ vaut 1 si la condition interne est vraie, 0 sinon.

Le F1 de sommet est la moyenne harmonique des deux mesures Pre_V et Rec_V.

Pour ce qui est des arêtes, nous utilisons la Précision, le Rappel et le F1 d'arête (Amalvy *et al.*, 2025) :

$$Pre_E = \max_{f_E} \frac{\left|\{f_E(e) : e \in E_p\} \cap E_r\right|}{|E_p|} \quad [3]$$

$$Rec_E = \max_{g_E} \frac{\left|E_p \cap \{g_E(e) : e \in E_r\}\right|}{|E_r|}, \quad [4]$$

où f_E est une fonction associant une arête de E_p à une arête de E_r ou à l'arête nulle $e_\varnothing$ si une arête n'a pas d'équivalent dans les arêtes de référence, et *vice versa* pour g_E. Enfin, nous utilisons également les versions pondérées de ces mesures d'arête :

$$WPre_E = \max_{f_E} \frac{\sum_{e \in E_p} 1 - \left|w(f_E(e)) - w(e)\right|}{|E_p|} \quad [5]$$

$$WRec_E = \max_{g_E} \frac{\sum_{e \in E_r} 1 - \left|w(e) - w(g_E(e))\right|}{|E_r|}, \quad [6]$$

où $w(e)$ est le poids de l'arête e, normalisé en divisant par le poids maximal du graphe considéré. Ces versions pondérées permettent de prendre en compte la fidélité des poids des arêtes extraites. Par définition, elles ne peuvent atteindre que des scores inférieurs ou égaux à ceux de leurs homologues non pondérées.

Résultats. Nous indiquons la qualité des sommets extraits dans le tableau 7, et la qualité des arêtes dans le tableau 8. La qualité des sommets semble acceptable (de 52,46 à 64,08 $F1_V$), mais les métriques d'arête sont relativement basses (de 20,13 à 48,18 $F1_E$). Nous montrons dans l'annexe B qu'une partie importante des erreurs est due à l'algorithme de résolution d'alias plutôt qu'à la REN. La forte influence de cet algorithme explique, au moins en partie, que nous ne notions pas de corrélation entre la performance de la REN sur un roman et la qualité du réseau extrait de celui-ci. Si *Les Trois Mousquetaires* et *Notre-Dame de Paris* présentaient plus de difficultés en termes de détection d'entités, cela ne se ressent pas fortement sur nos mesures.

De manière générale, les résultats sont un peu en dessous de ceux reportés par Amalvy *et al.* (2025) sur le corpus Litbank (Bamman *et al.*, 2019 ; Bamman *et al.*, 2020), ce qui peut s'expliquer par la différence de longueur entre les textes étudiés. En effet, Litbank ne contient que des extraits d'environ 2 000 tokens, quand les romans de notre corpus peuvent être jusqu'à 150 fois plus longs. Cette longueur implique de plus nombreuses entités à détecter et d'alias à résoudre, ce qui multiplie les risques d'erreur, lors de l'étape de REN comme celle de résolution d'alias.

Il est à noter que, dans cette étude, nous ne prenons en compte que les mentions de personnages annotées dans notre étape de REN. Cela exclut des mentions plus génériques, comme certaines descriptions définies ou les pronoms, qui seraient détectables en utilisant un algorithme de résolution de coréférences. Ces mentions

Roman	$F1_V$	Pre_V	Rec_V
Les Trois Mousquetaires	59,10	47,08	79,34
Le Rouge et le Noir	63,52	57,83	70,44
Eugénie Grandet	52,46	43,37	66,36
Germinal	60,35	47,96	81,37
Bel-Ami	57,25	45,45	77,33
Notre-Dame de Paris	61,81	55,36	69,96
Madame Bovary	64,08	52,47	82,29

TABLEAU 7. *Qualité des sommets des réseaux extraits par un pipeline* `Renard` *utilisant notre modèle de REN*

peuvent avoir une forte influence sur les liens du réseau, comme montré par Amalvy *et al.* (2025).

Roman	$F1_E$	Pre_E	Rec_E	$WF1_E$	$WPre_E$	$WRec_E$
Les Trois Mousquetaires	34,12	27,27	45,56	33,48	26,76	44,72
Le Rouge et le Noir	26,38	25,39	27,45	26,20	25,22	27,26
Eugénie Grandet	20,13	19,69	20,58	19,04	18,63	19,47
Germinal	48,18	43,73	53,65	46,99	42,64	52,32
Bel-Ami	23,91	19,72	30,37	23,52	19,40	29,87
Notre-Dame de Paris	23,45	20,42	27,55	23,24	20,23	27,30
Madame Bovary	29,38	24,79	36,06	29,13	24,58	35,75

TABLEAU 8. *Qualité des arêtes des réseaux extraits par un pipeline* `Renard` *utilisant notre modèle de REN*

Ce travail sur les réseaux de personnages n'est qu'une étude préliminaire, qui appelle plusieurs approfondissements. Il s'agira tout d'abord de mener une analyse descriptive plus poussée, permettant une comparaison avec la littérature existante. Par ailleurs, des expérimentations seront nécessaires afin d'évaluer l'impact des erreurs présentes dans le réseau sur une tâche en aval (Labatut et Bost, 2019), telles que la recommandation ou l'identification des personnages principaux.

5. Conclusion et perspectives

Dans cet article, nous avons présenté un nouveau corpus en français constitué de romans annotés en entités nommées. Notre évaluation a révélé un très bon accord inter-annotateur, ce qui démontre la cohérence de notre processus d'annotation. À la différence de la plupart des jeux de données disponibles en ligne, cette annotation a été menée sur *l'intégralité* de ces romans, et non pas sur des extraits, ce qui en fait

une ressource très précieuse, voire unique, pour le développement et l'évaluation de méthodes de REN sur ce type de textes longs.

Afin d'illustrer cela, nous avons également entraîné un modèle de langage spécifiquement destiné à la REN dans les romans en français. Bien que performant, notre modèle a montré ses limites pour certains romans spécifiques, soulignant les défis posés par les variations stylistiques et narratives, et le besoin de pousser ce travail plus loin. Enfin, nous avons proposé une application de la REN dans les romans, prenant la forme d'une tâche d'extraction de réseaux de personnages. Celle-ci a nécessité de notre part l'ajout d'une nouvelle couche d'annotation à notre corpus, ciblant la tâche de résolution d'alias. Celle-ci est potentiellement plus complexe sur des livres complets, et aucun corpus existant ne proposait jusqu'ici d'annotations de ce type, qui plus est en français. Notre corpus, le code source de nos expériences ainsi que notre modèle sont mis à disposition de la communauté sous licence libre.

Nous identifions trois perspectives directes à notre travail. La première est d'ajuster le guide d'annotation afin d'améliorer encore l'accord inter-annotateur et l'homogénéité des EN. Le guide d'Universal NER a été adapté dans le cadre de notre étude pour intégrer les spécificités linguistiques du français, et pour traiter des cas complexes d'étiquetage des EN (notamment dans la catégorie `OTHER`). Une collaboration avec ce projet pourrait permettre d'affiner ce guide en y intégrant ces particularités tout en préservant sa cohérence multilingue. La seconde piste porte sur la création d'un corpus type *silver* de grande ampleur, qui permettrait de venir enrichir les données disponibles, et potentiellement d'améliorer ainsi l'apprentissage et donc les performances de notre modèle de REN. Une troisième perspective, à plus long terme, est d'appliquer notre méthode d'extraction de réseaux de personnages à un corpus beaucoup plus vaste. Ce serait l'occasion, d'une part, d'analyser les propriétés statistiques des réseaux construits à partir d'un échantillon statistiquement représentatif ; et, d'autre part, de mobiliser ces réseaux pour des tâches de prédiction, telles que la classification ou la régression appliquées aux œuvres littéraires.

6. Bibliographie

Alrahabi M., Brando C., Frontini F., Guide d'annotation manuelle d'entités nommées dans des corpus littéraires, Technical report, Labex OBVIL - L'Observatoire de la vie littéraire, 2024.

Amalvy A., Janickyj M., Mannion S., MacCarron P., Labatut V., « Interconnected Kingdoms : Comparing 'A Song of Ice and Fire' Adaptations Across Media Using Complex Networks », *Social Network Analysis and Mining*, vol. 14, p. 199, 2024a.

Amalvy A., Labatut V., Annotation Guidelines for Corpus Novelties : Part 2 — Alias Resolution, Technical report, Avignon Université, 2024.

Amalvy A., Labatut V., Dufour R., « The Role of Global and Local Context in Named Entity Recognition », *ACL*, p. 714-722, 2023.

Amalvy A., Labatut V., Dufour R., « Renard : A Modular Pipeline for Extracting Character Networks from Narrative Texts », *Journal of Open Source Software*, vol. 9, p. 6574, 2024b.

Amalvy A., Labatut V., Dufour R., « The Role of Natural Language Processing Tasks in Automatic Literary Character Network Construction », *31st International Conference on Computational Linguistics*, p. 8462-8473, 2025.

Bamman D., Lewke O., Mansoor A., « An Annotated Dataset of Coreference in English Literature », *12th Language Resources and Evaluation Conference*, p. 44-54, 2020.

Bamman D., Popat S., Shen S., « An annotated dataset of literary entities », *NAACL*, p. 2138-2144, 2019.

Bamman D., Underwood T., Smith N. A., « A Bayesian Mixed Effects Model of Literary Character », *ACL*, vol. 1, p. 370-379, 2014.

Berragan C., Singleton A., Calafiore A., Morley J., « Transformer based named entity recognition for place name extraction from unstructured text », *International Journal of Geographical Information Science*, vol. 37, nº 4, p. 747-766, 2023.

Brandsen A., Verberne S., Wansleeben M., Lambers K., « Creating a Dataset for Named Entity Recognition in the Archaeology Domain », *Twelfth Language Resources and Evaluation Conference*, p. 4573-4577, 2020.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, p. 37-46, 1960.

Cuesta-Lazaro C., Prasad A., Wood T., « What does the sea say to the shore ? A BERT based DST style approach for speaker to dialogue attribution in novels », *ACL*, 2022.

Dekker N., Kuhn T., van Erp M., « Evaluating named entity recognition tools for extracting social networks from novels », *PeerJ Computer Science*, vol. 5, p. e189, 2019.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *NAACL*, p. 4171-4186, 2019.

Durandard N., Tran V. A., Michel G., Epure E., « Automatic Annotation of Direct Speech in Written French Narratives », *ACL*, vol. 1, p. 7129-7147, 2023.

Egloff M., Picca D., « WeDH - a Friendly Tool for Building Literary Corpora Enriched with Encyclopedic Metadata », *LREC*, p. 813-816, 2020.

Ehrmann M., Hamdi A., Linhares Pontes E., Romanello M., Doucet A., « Named Entity Recognition and Classification in Historical Documents : A Survey », *ACM Computing Surveys*, vol. 56, nº 2, p. 1-47, 2023.

Evans A. B., « Jules Verne and the French Literary Canon », *Jules Verne : Narratives of Modernity*, Liverpool University Press, chapter 2, p. 11-34, 2000.

Fokkens A., ter Braake S., Sluijter R., Arthur P. L., Wandl-Vogt E. (eds), *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*, vol. 2119 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018.

Frontini F., Brando C., Byszuk J., Galleron I., Santos D., Stanković R., « Named Entity Recognition for Distant Reading in ELTeC », *CLARIN Annual Conference*, p. 37-41, 2020.

Girvan M., Newman M. E. J., « Community structure in social and biological networks », *Proceedings of the National Academy of Sciences*, vol. 99, nº 12, p. 7821-7826, 2002.

Grilo S., Bolrinha M., Silva J., Vaz R., Branco A., « The BDCamões Collection of Portuguese Literary Documents : A Research Resource for Digital Humanities and Language Technology », *12th Language Resources and Evaluation Conference*, p. 849-854, 2020.

Guo J., Xu G. X., Cheng X., « Named entity recognition in query », p. 267-274, 2009.

Guo Q., Hu X., Zhang Y., Qiu X., Zhang Z., « Dual Cache for Long Document Neural Coreference Resolution », *ACL*, p. 15272-15285, 2023.

Gupta T., Hatzel H. O., Biemann C., « Coreference in Long Documents using Hierarchical Entity Merging », *LaTeCH-CLfL*, p. 11-17, 2024.

Hettinger L., Becker M., Reger I., Jannidis F., Hotho A., « Genre Classification on German Novels », *DEXA*, p. 249-253, 2015.

Holanda A. J., Matias M., Ferreira S. M. S. P., Benevides G. M. L., Kinouchi O., « Character Networks and Book Genre Classification », *International Journal of Modern Physics*, 2019.

Jørgensen F., Aasmoe T., Ruud Husevåg A.-S., Øvrelid L., Velldal E., « NorNE : Annotating Named Entities for Norwegian », *12th Language Resources and Evaluation Conference*, p. 4547-4556, 2020.

Kogkitsidou E., Gambette P., « Normalisation of 16th and 17th Century Texts in French and Geographical Named Entity Recognition », *4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, p. 28-34, 2020.

Labatut V., Bost X., « Extraction and Analysis of Fictional Character Networks : A Survey », *ACM Computing Surveys*, vol. 52, n° 5, p. 89, 2019.

Labusch K., Neudecker C., « Entity Linking for Cultural Heritage Collections with Wikidata », *DH 2022 : Digital Humanities Conference*, 2022.

Landragin F., « Le corpus DEMOCRAT et son exploitation. Présentation », *Langages*, vol. 224, n° 4, p. 11-24, 2021.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., « FlauBERT : Unsupervised Language Model Pre-training for French », *12th Language Resources and Evaluation Conference*, p. 2479-2490, 2020.

Li F., Automatic Content Extraction 2008 Evaluation Plan (ACE08), Technical report, Linguistic Data Consortium, 2008.

Linguistic Data Consortium, ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Technical report, Linguistic Data Consortium, 2005.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V., « RoBERTa : A Robustly Optimized BERT Pretraining Approach », *arXiv*, 2019.

Liu Z., Xu Y., Yu T., Dai W., Ji Z., Cahyawijaya S., Madotto A., Fung P., « CrossNER : Evaluating Cross-Domain Named Entity Recognition », *AAAI Conference on Artificial Intelligence*, p. 13452-13460, 2021.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie E., Seddah D., Sagot B., « CamemBERT : a Tasty French Language Model », *ACL*, p. 7203-7219, 2020.

Mayhew S., Blevins T., Liu S., Šuppa M., Gonen H., Imperial J. M., Karlsson B. F., Lin P., Ljubešić N., Miranda L. J., Plank B., Riabi A., Pinter Y., « Universal NER : A Gold-Standard Multilingual Named Entity Recognition Benchmark », *NAACL*, 2024.

Min S., Park J., « Network Science and Narratives : Basic Model and Application to Victor Hugo's Les Misérables », *Studies in Computational Intelligence*, vol. 644, p. 257-265, 2016.

Moncla L., Gaio M., Joliveau T., Le Lay Y.-F., « Automated Geoparsing of Paris Street Names in 19th Century Novels », *1st ACM Workshop on Geospatial Humanities*, p. 1-8, 2017.

Ramshaw L., Marcus M., « Text Chunking using Transformation-Based Learning », *3rd Workshop on Very Large Corpora*, 1995.

Schöch C., Patraș R., Erjavec T., Santos D., « Creating the European Literary Text Collection : Challenges and Perspectives », *Modern Languages Open*, vol. 1, n° 25, p. 1-19, 2021.

Silva M. O., Moro M. M., « PPORTAL_ner : An Annotated Corpus of Portuguese Literary Entities », *Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, p. 12927-12937, 2024.

Sims M., Bamman D., « Measuring Information Propagation in Literary Social Networks », *EMNLP*, p. 642-652, 2020.

Sprugnoli R., « Arretium or Arezzo ? A Neural Approach to the Identification of Place Names in Historical Texts », *CLiC-it*, vol. 2253 of *CEUR Workshop Proceedings*, p. 26, 2018.

Tay Y., Dehghani M., Abnar S., Shen Y., Bahri D., Pham P., Rao J., Yang L., Ruder S., Metzler D., « Long Range Arena : A Benchmark for Efficient Transformers », *ICLR*, 2021.

Vala H., Jurgens D., Piper A., Ruths D., « Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized : On The Difficulty of Detecting Characters in Literary Texts », *EMNLP*, p. 769-774, 2015.

van Dalen-Oskam K., de Does J., Marx M., Sijaranamual I., Depuydt K., Verheij B., Geirnaert V., « Named entity recognition and resolution for literary studies », *Computational Linguistics in the Netherlands Journal*, vol. 4, p. 121-136, 2014.

Yamada I., Asai A., Shindo H., Takeda H., Matsumoto Y., « LUKE : Deep Contextualized Entity Representations with Entity-aware Self-attention », *EMNLP*, p. 6442-6454, 2020.

Annexes

A. Précisions sur le corpus

Le tableau 9 indique les URL de chaque roman du corpus proposé. Il s'agit des versions brutes qui ont été utilisées lors du processus d'annotation. La figure 3 montre la distribution des EN pour chaque roman et classe d'entité.

Roman	**URL Wikisource** – `https://fr.wikisource.org/wiki/...`
Les Trois Mousquetaires	`.../Les_Trois_Mousquetaires/Texte_entier`
Le Rouge et le Noir	`.../Le_Rouge_et_le_Noir/Texte_entier`
Eugénie Grandet	`.../Eugénie_Grandet`
Germinal	`.../Germinal/Texte_entier`
Bel-Ami	`.../Bel-Ami/Édition_Ollendorff,_1901/Texte_entier`
Notre-Dame de Paris	`.../Notre-Dame_de_Paris/Texte_entier`
Madame Bovary	`.../Madame_Bovary/Texte_entier`

TABLEAU 9. *URL des versions brutes des romans constituant notre corpus*

B. Impact de la REN sur la qualité des réseaux

Afin de vérifier l'impact de la REN sur la qualité des réseaux de personnages, nous réalisons une extraction en injectant à l'entrée du pipeline `Renard` les annotations en

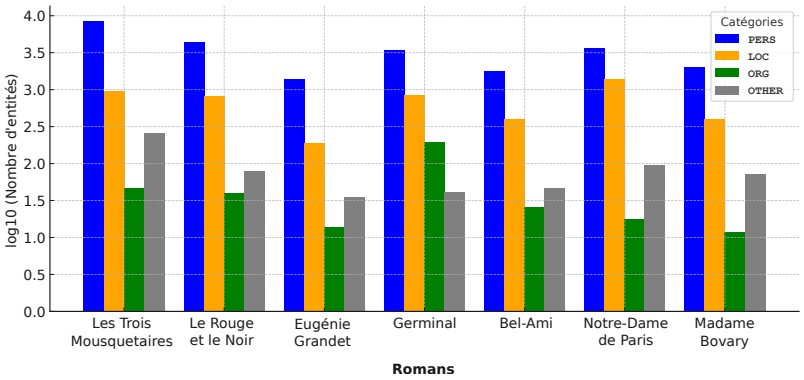


FIGURE 3. *Distribution des EN par roman et par type*

REN de notre jeu de données. Cela nous permet d'observer l'influence de l'algorithme de résolution d'alias sur les résultats, et de donner une borne haute de la performance de l'algorithme d'extraction lorsque l'étape de REN est parfaitement exécutée. Le tableau 10 donne les résultats de cette expérience en termes de mesures de sommets, tandis que le tableau 11 donne les mesures en termes d'arêtes.

Roman	$Prec_V$	Rec_V	$F1_V$
Les Trois Mousquetaires	68,76	92,02	78,71
Le Rouge et le Noir	73,55	83,96	78,41
Eugénie Grandet	66,60	84,11	74,34
Germinal	80,04	85,29	82,58
Bel-Ami	71,44	88,67	79,13
Notre-Dame de Paris	68,04	76,68	72,10
Madame Bovary	71,52	88,00	78,91

TABLEAU 10. *Qualité des sommets des réseaux extraits par un pipeline* `Renard` *en utilisant nos annotations en REN*

Si la REN influence fortement la qualité des sommets extraits (jusqu'à 16,79 $F1_V$ pour *Germinal*), son impact sur les arêtes reste plus limité (8,24 $F1_E$ pour *Madame Bovary*). L'algorithme de résolution d'alias apparaît ainsi comme le principal facteur affectant la qualité, encore imparfaite, des réseaux extraits.

Roman	Pre_E	Rec_E	$F1_E$	$WPre_E$	$WRec_E$	$WF1_E$
Les Trois Mousquetaires	26,56	34,32	29,94	26,32	34,01	29,67
Le Rouge et le Noir	23,39	24,52	23,94	23,31	24,44	23,86
Eugénie Grandet	28,66	30,23	29,42	27,28	28,77	28,00
Germinal	60,04	52,10	55,79	58,44	50,72	54,31
Bel-Ami	27,95	31,89	29,79	27,50	31,37	29,31
Notre-Dame de Paris	21,25	27,00	23,79	20,81	26,45	23,29
Madame Bovary	44,71	48,01	46,30	44,16	47,42	45,73

TABLEAU 11. *Qualité des arêtes des réseaux extraits par un pipeline* `Renard` *en utilisant nos annotations de REN*