Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Du Genre au Continuum : Analyse Ascendante de la Variation Linguistique dans le Corpus NEM.fr

Marina Seghier
LIASD, Université Paris 8, 2 rue de la Liberté, 93526 Saint-Denis France
ms@up8.edu

RÉSUMÉ

Les outils de TAL sont sensibles aux variations linguistiques, fait souvent masqué par des évaluations sur corpus homogènes. Ce travail interroge la pertinence des classifications externes traditionnelles (domaines, genres) face aux caractéristiques linguistiques internes pour appréhender la variation textuelle. Nous présentons le corpus varié NEM.FR, annoté en entités nommées et morphosyntaxe. Avec une approche ascendante non supervisée (ACP et classification floue) sur 33 traits linguistiques, nous démontrons que la variation se structure davantage comme un continuum que comme un ensemble de catégories discrètes. Nos résultats révèlent des dimensions et des méta-catégories transversales, inaccessibles aux approches descendantes classiques. Cette étude confirme que des observables intrinsèques offrent une modélisation plus fine et fiable de la variation que les étiquettes externes. À terme, cette méthodologie vise à corréler configurations linguistiques et performances des modèles pour orienter le TAL vers des stratégies d'évaluation et d'adaptation plus robustes.

ABSTRACT

From Genres to Continuum : Bottom-up Analysis of Linguistic Variation in the NEM.fr Corpus

NLP tools are sensitive to linguistic variation, a fact often masked by evaluations on homogeneous corpora. This work questions the relevance of traditional external classifications (domains, genres), faced to internal linguistic characteristics for understanding textual diversity. We present the varied NEM.FR corpus, annotated with named entities and morphosyntax. Using an unsupervised bottom-up approach (PCA and fuzzy classification) on 33 linguistic features, we demonstrate that variation is structured more like a continuum than as a set of discrete categories. Our results reveal cross-cutting dimensions and meta-categories inaccessible to classic top-down approaches. This study confirms that intrinsics observables offer a more refined and reliable modeling of variation than external labels. Ultimately, this methodology aims to correlate linguistic configurations and model performance to guide NLP towards more robust evaluation and adaptation strategies.

MOTS-CLÉS : variation linguistique, conception de corpus, annotation multi-couches, évaluation.

KEYWORDS: linguistic variation, corpus design, multi-layered annotation, evaluation.

Les outils de TAL sont fortement sensibles aux variations linguistiques, leurs performances fluctuant selon le genre ou le domaine traité (Ponti *et al.*, 2019; Fu *et al.*, 2020; Ziems *et al.*, 2023; Millour *et al.*, 2024). Pourtant, pour des raisons de coûts de production, les évaluations reposent encore sur des jeux de données homogènes, peu représentatifs de la diversité réelle. Pour tendre vers des évaluations plus robustes et fiables, nous faisons l'hypothèse que les baisses de performances observées s'expliquent par des facteurs linguistiques intrinsèques aux différents textes. De fait, confronter un système à des jeux de données variés permet de corréler résultats et variations (Fu *et al.*, 2020) ce qui nécessite un corpus de test représentatif. En TAL, la représentativité des corpus est souvent abordée selon

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 2–13.

une perspective externe (Sinclair, 1996) classant les textes par domaine (ex : droit, médecine) ou modalité (ex : langue parlée vs. écrite), sans considérer leurs divergences linguistiques réelles. Or, cette approche n'est pas la seule façon d'aborder la variation linguistique : des études linguistiques descriptives antérieures menées sur l'anglais (Biber, 1988; Passonneau *et al.*, 2014) ont fourni les outils nécessaires pour démontrer que des textes provenant de différentes sources ou classes partagent des traits linguistiques communs, souvent masqués par les catégorisations humaines. Ce travail s'inscrit dans le contexte de la Reconnaissance d'Entités Nommées (REN) en français, pour lequel nous présentons la conception et l'enrichissement du corpus NEM.FR, une ressource de référence représentative pour l'évaluation de la tâche de REN, et conçu pour l'étude de la variation selon des perspectives tant internes qu'externes.

Après un état de l'art sur l'étude de la variation linguistique, nous détaillons la création, l'annotation et l'extraction des caractéristiques de NEM.FR. Enfin, nous présentons notre méthode d'analyse de la variation linguistique par ACP et classification floue, puis nous discutons des dimensions linguistiques qui émergent de ces regroupements.

1 Variation linguistique et caractéristiques fines

Dans son étude fondamentale, Biber (1988) introduit la notion de variation non seulement d'un point de vue situationnel et fonctionnel (soit externe au texte), mais également en tant qu'observable linguistique (interne au texte) qui peut être déduit par l'étude des cooccurrences de caractéristiques linguistiques. Basée sur une étude empirique de deux grands corpus anglais (écrit et parlé) et de 67 caractéristiques linguistiques couvrant le lexique, la morphosyntaxe et certaines structures syntaxiques, son étude révèle 6 dimensions continues le long desquelles se produisent des variations. Les dimensions sont étiquetées comme suit (1) "Informationnel *versus* Productions impliquées", (2) "Narration *versus* Description", (3) "Référence explicite *versus* Référence dépendante de la situation", (4) "Degrés d'expression manifeste de persuasion", (5) "Abstrait *versus* Non-abstrait" et (6) "Élaboration d'informations en ligne (spontanée)".

Certaines des caractéristiques utilisées dans les travaux de Biber ont été reprises dans une étude plus récente menée par Passonneau *et al.* (2014), où la plupart des caractéristiques linguistiques considérées ont été sélectionnées dans le riche ensemble d'annotations validées manuellement du corpus MASC (*Manually Annotated Sub-Corpora*). Comme dans les travaux de Biber, l'analyse factorielle de Passonneau *et al.* (2014) met en évidence que la variation du texte n'est pas distribuée selon des catégories discrètes (telles que "oral" ou "écrit"). Elle confirme au contraire l'idée générale de variation multidimensionnelle et graduelle. Cette étude identifie des dimensions de variation influencées par des facteurs pragmatiques et cognitifs, tels que le degré d'interactivité, la densité de l'information ou les contraintes contextuelles de production, ce qui correspond aux dimensions de Biber. Travailler avec des caractéristiques linguistiques a ouvert la voie à l'analyse multidimensionnelle (MDA) introduite dans Biber (1992), et à son utilisation pour l'identification automatique de genre (AGI), en particulier dans les corpus extraits du Web (Santini *et al.*, 2011; Laippala *et al.*, 2021; Rao *et al.*, 2021).

2 Conception et enrichissement de corpus

2.1 Échantillonnage de corpus

La présente étude se concentre sur l'observation de la variation textuelle à partir d'un corpus aussi varié qu'équilibré. Cette perspective fait écho à Biber (1993), qui soutient que la conception d'un corpus doit trouver un équilibre entre représentativité et contrôle, garantissant que la variabilité interne des données observées, reflète véritablement la gamme des contextes de communication trouvés dans l'utilisation réelle du langage.

Pour mener notre étude, nous avons choisi de nous appuyer sur une ressource d'évaluation existante de 15 000 tokens pour la tâche du REN français, le corpus FENEC (Millour *et al.*, 2022) annoté manuellement en entités nommées. Il est composé de 15 documents appartenant à six catégories textuelles différentes : POÉSIE, PROSE, SPOKEN, ENCYCLOPÉDIE, INFORMATIONS, MULTISOURCES).

Nous avons rééquilibré ces catégories en enrichissant le corpus avec de nouveaux documents issus des genres existants du corpus FENEC et étendons encore sa couverture avec de nouvelles catégories : POLITIQUE, JURIDIQUE, BIOMÉDICAL, TWEETS, MAILS. Une attention particulière a été accordée à la parité entre les sexes parmi les auteurs de la catégorie PROSE, l'analyse de FENEC ayant révélé une surreprésentation des auteurs masculins. Les autres textes proviennent de sources institutionnelles et open source : discours préparés de présidents français, décisions de justice des tribunaux administratifs, corpus spécialisés (MORFITT (Labrak *et al.*, 2023), POPCORN (Giordano *et al.*, 2024), ESLO (Abouda & Baude, 2005), TREMOLO (Mekki *et al.*, 2021), WIKINER-FR-GOLD (Cao *et al.*, 2024)), et de mails anonymisés. Les échantillons ont été équilibrés en tokens pour assurer une répartition homogène entre les genres. Le tableau 1 donne un aperçu du corpus NEM.FR.

Nous avons également réalisé des *data statements*, cadre proposé par Bender & Friedman (2018). Cette approche encourage les chercheurs à décrire clairement comment leurs corpus ont été construits, y compris des informations sur les sources de données, les conditions de collecte et les profils des locuteurs et des annotateurs. En documentant ces aspects, nous rendons plus visibles nos choix méthodologiques et nos biais potentiels. Ce faisant, le cadre soutient à la fois la transparence et la reproductibilité des ressources utilisées, tout en fournissant une base claire pour relier les modèles de variation textuelle, tels qu'observés dans le cadre de notre corpus.

2.2 Annotation multi-couches

Nous avons annoté notre corpus en entités nommées à l'aide de la plateforme collaborative INCEpTION [1], selon le guide d'annotation des EN français QUAERO [2] et le jeu d'étiquettes (PERS, ORG, LOC, EVENT, TIME, PROD) [3]. Cinq annotateurs [4] ont travaillé pendant environ trois mois, deux réalisant l'annotation et un troisième assurant la curation, chacun alternant dans ces rôles. Les données ont ensuite été exportées et converties au format adapté à notre étude.

Pour la couche morpho-syntaxique (POS) du corpus NEM.FR, nous avons utilisé le modèle POET (*A*

1. https://inception-project.github.io/
2. Disponible sur : http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf.
3. Nous avons exclu l'étiquette AMOUNT qui est rare.
4. 3 chercheurs permanents, 1 doctorant, 1 étudiant en master.

Source (identifiant)	Période	Genre	# tokens
**Traité sur la Tolérance*, Voltaire (prose01)	18th	prose	1 029
**Le Ventre de Paris*, Émile Zola (prose02)	19th		973
**L'Homme qui plantait des arbres*, Jean Giono (prose03)	20th		961
Chéri, Colette (prose04)	20th		1 284
Les Sévriennes, Gabrielle Reval (prose05)	20th		1 357
**Les Fleurs du Mal*, Baudelaire (poetry01)	19th	poetry	1 014
**Œuvres d'Arthur Rimbaud - Vers et proses* (poetry02)	19th		1 027
Anatomie du Mouvement, Huguette Bertrand (poetry03)	20th		1 111
Fleurs Sauvages, Léonise Vallois (poetry04)	19-20th		1 237
Les Vivants et les Morts, Anna de Noailles (poetry05)	19-20th		1 242
*UD French GSD (multi01)	21st	multisources	969
*Sequoia (Candito & Seddah, 2012) (multi02)	21st		954
*French Question Bank (Seddah & Candito, 2016) (multi03)	21st		961
*APIL (office du tourisme Othe-Armance) (information01)	21st	information	935
*Wikinews (information02)	21st		1 022
**L'Est Républicain* (ATILF & CLLE, 2020) (information03)	21st		928
*French WikiNER (encyclopedia01)	21st	encyclopedia	1 010
WikiNER-FR gold (Cao *et al.*, 2024) (encyclopedia02-06, 5 échantillons)	21st		4 667
*Rhapsodie (Lacheret *et al.*, 2014) (spoken01-03, 3 échantillons)	21st	spoken	2 967
ESLO (Abouda & Baude, 2005) (spoken04-05, 2 échantillons)	21st		2 242
Discours au Parlement, Macron (politique01)	21st	political	1 122
Discours aux français, Macron (politique02)	21st		1 109
Appel du 18 juin 1940, De Gaulle (politique03)	20th		443
Discours aux algériens, De Gaulle (politique04)	20th		634
Cour d'appel administrative, Anonymisé (juridique01)	21st	legal	1 147
Tribunaux d'appels, Anonymisé (juridique02)	21st		1 165
Conseil d'état, Anonymisé (juridique03)	21st		959
Cour d'appel administrative, Anonymisé (juridique04)	21st		1 112
Spécialité Chimie (Labrak *et al.*, 2023) (biomedical01)	21st	biomedical	1 108
Spécialité Génétique (Labrak *et al.*, 2023) (biomedical02)	21st		1 113
Spécialité Chirurgie (Labrak *et al.*, 2023) (biomedical03)	21st		1 104
Spécialité Psychologie (Labrak *et al.*, 2023) (biomedical04)	21st		1 121
Tweets dans un registre courant (Mekki *et al.*, 2021) (tweets01)	21st	tweets	1 100
Tweets dans un registre familier (Mekki *et al.*, 2021) (tweets02)	21st		1 100
Tweets dans un registre soutenu (Mekki *et al.*, 2021) (tweets03)	21st		1 103
Extrait du corpus d'entraînement (Giordano *et al.*, 2024) (defense01-03)	21st	defense	2 218
Extrait du corpus de test (Giordano *et al.*, 2024) (defense02-04)	21st		2 197
Emails professionnels et privés anonymisés (mail01-26)	21st	mail	3 735
		12 genres	51 480 tokens

TABLE 1 – Aperçu du corpus NEM.FR annoté manuellement. Les sections préexistantes au corpus NEM.FR sont précédées d'un astérisque (*).

French Extended Part-of-Speech Tagger [5], basé sur les plongements lexicaux `Flair` et `CamemBERT`, au sein du framework `Flair`). Afin de disposer d'un corpus de référence de haute qualité pour une analyse fiable et robuste, les étiquettes ont été corrigées manuellement. Nous avons également choisi de réduire le jeu d'étiquettes fines initialement fournit par `POET` au jeu d'étiquettes UPOS (*Universal Dependencies*), ceci dans le but d'améliorer la lisibilité des données pour les différents genres de NEM.FR.

Nous avons tout de même réalisé une évaluation rapide de cet outil sur un échantillon de tokens de

5. Disponible ici : https://huggingface.co/qanastek/pos-french-camembert-flair

différents genres. Le modèle a globalement bien fonctionné, avec de meilleurs résultats sur les textes écrits (jusqu'à 94,6% de précision sur les textes encyclopédiques) que sur la parole transcrite (78,4%).

La précision relativement faible de l'étiquetage en POS observée sur la catégorie SPOKEN (78,43%) peut s'expliquer par les fortes différences entre la syntaxe orale et écrite. L'extrait suivant illustre les erreurs d'annotations typiques produites par le modèle `POET` (`Flair`) :

> *et euh et ces rails du tram eh ben je vais je vais les plus long au rond-point c'est tout droit direction Saint-Jean-de-Maurienne et donc euh voilà ou je vais me promener.*

Sur 37 tokens, 11 ont mal été étiquetés. La plupart des erreurs se situent sur des disfluences et des interjections telles que "euh", "eh" et "ben", qui sont étiquetées comme verbes (`VERB`) ou noms propres (`PROPN`) au lieu de `INTJ`. D'autres erreurs proviennent de confusions contextuelles dans des constructions ambiguës ("tout droit" étiqueté `DET + NMS` plutôt que `ADV + ADJ`). Ces erreurs révèlent les limites des modèles entraînés sur un certain type de texte lorsqu'ils sont appliqués à un nouveau type. Les documents de notre corpus étiquetés SPOKEN contiennent des hésitations, des répétitions et une syntaxe incomplète qui perturbent les régularités linguistiques capturées dans les données écrites standards. La plus faible précision observée (78,43%) ne doit donc pas être interprétée comme une mauvaise performance du modèle, mais plutôt comme la preuve d'un problème plus large : à titre d'illustration ici, `POET` (Flair) — entraîné principalement sur du français écrit *canonique* [6] — peine à gérer des phénomènes oraux absents de ses données d'entraînement.

2.3 Caractéristiques Linguistiques

Nous avons utilisé un ensemble de 33 caractéristiques pour vectoriser nos documents : **(i)** 19 relatifs aux étiquettes POS : `ADJ`, `ADP`, `SCONJ`, `CCONJ`, `ADV`, `PROPN`, `NUM`, `AUX`, `VERB`, `DET`, `NOUN`, `PRON`, `PPER1S`, `PPER2S`, `PPER3`, `INTJ`, `SYM`, `PUNCT`, `X` ; **(ii)** 7 liées aux EN : `LOC`, `ORG`, `PERS`, `PROD`, `EVENT`, `TIME`, `TOTAL_EN` ; **(iii)** 1 caractéristique verbale : `PART_PASSE`, **(iii)** 2 mesures textuelles : `LONGUEUR_MOTS`, `TTR` (ratio type/token) ; **(v)** 4 signes de ponctuation distincts : `,` `.` `!` `?`

Ce jeu de caractéristiques présente plusieurs similitudes avec celui utilisé par Biber (1988). Premièrement, les caractéristiques morpho-syntaxiques obtenues à partir du *POS-tagging* se superposent directement à plusieurs caractéristiques de Biber : les pronoms personnels (`PRON`, `PPER1S`, `PPER2S`, `PPER3`), les verbes (`VERB`, `AUX`, `PART_PASSE`), les noms et leur caractérisation (`NOUN`, `ADJ`), et les conjonctions de subordination et de coordination (`SCONJ`, `CCONJ`). De même, le ratio type-token (`TTR`) et la longueur moyenne des mots (`LONGUEUR_MOTS`) correspondent à ses mesures de spécificité lexicale. Enfin, les caractéristiques liées aux étiquettes et au nombre total d'entités nommées, bien qu'absentes de celles de Biber, étendent son cadre pour mieux rendre compte des variations linguistiques dans les corpus contemporains utilisés en TAL.

6. Au sens "conforme à des règles, à un standard".

3 Expériences et résultats

3.1 Analyse factorielle et classification floue

Pour explorer la structure sous-jacente de notre corpus et identifier les principales dimensions de la variation, nous avons appliqué une Analyse en Composantes Principales (ACP) à nos données vectorisées (voir Figure 1). L'ACP permet de réduire la dimensionnalité tout en conservant autant que possible la variance d'origine. Contrairement aux techniques de visualisation non linéaires telles que l'algorithme t-SNE (*t-Distributed Stochastic Neighbor Embedding*) ou UMAP (*Uniform Manifold Approximation and Projection*), l'ACP fournit des axes interprétables et permet une analyse qualitative directe des caractéristiques linguistiques extraites précédemment.

Chaque document du corpus [7] a été échantillonné et transformé en un vecteur des 33 caractéristiques linguistiques présentées dans la section précédente. Les quatre premières composantes principales capturent 50% de la variance totale du corpus. Si nous concentrons notre analyse détaillée sur les deux premiers axes (32,1%), qui synthétisent les facteurs de variation les plus structurants, ces résultats globaux confirment que notre réduction de dimensionnalité saisit une part significative de la variation linguistique du corpus.

Comme le souligne l'enquête de Kuzman & Ljubešić (2025) sur l'identification automatique des genres, l'un des défis majeurs de l'annotation automatique et manuelle réside dans "l'existence de documents hybrides et de documents multitextes dans des ensembles de données Web". Pour résoudre ce problème et en suivant l'approche de Lee & Jiang (2014), nous avons appliqué une méthode de classification floue (*fuzzy clustering*) à nos deux premières composantes principales. Plutôt que d'attribuer chaque document à une seule catégorie, cet algorithme calcule des degrés d'appartenance pour chaque échantillon de texte — ainsi un document peut être, par exemple, à mi-chemin entre le discours narratif et le discours informatif. Cette approche nous aide à visualiser les zones de transition entre les genres et à identifier des sous-groupes cohérents au sein du corpus.

3.2 Dimensions linguistiques de NEM.FR

La Figure 1 présente la projection des documents (par catégorie externe) et des caractéristiques linguistiques sur les deux premières composantes principales, qui expliquent respectivement 18,8% et 13,3% de la variance totale. Les flèches représentent les contributions des variables, tandis que les points correspondent aux documents, colorés selon leur genre d'origine.

La première composante principale (PC1) oppose clairement deux configurations linguistiques. Du côté positif de PC1 (droite), on observe une forte association avec les pronoms de manière générale et à la première personne du singulier (`PRON`, `PPER1S`), les verbes (`VERB`), les adverbes (`ADV`) et les interjections (`INTJ`). Ces caractéristiques sont typiques de productions impliquées, où le locuteur est présent et s'exprime directement. La forte proportion de verbes et d'adverbes, la présence d'interjections et de pronoms personnels, renvoient à des usages plus interactionnels, nuancés et expressifs. Les échantillons des genres SPOKEN, TWEETS, MAIL, ainsi qu'une partie de PROSE et **textsc**, se projettent majoritairement dans cette région.

À l'inverse, du côté négatif de PC1, les documents sont associés à une plus forte proportion de nombres

7. Nous avons exclu les documents MULTISOURCES en raison de leur étiquette ambiguë.

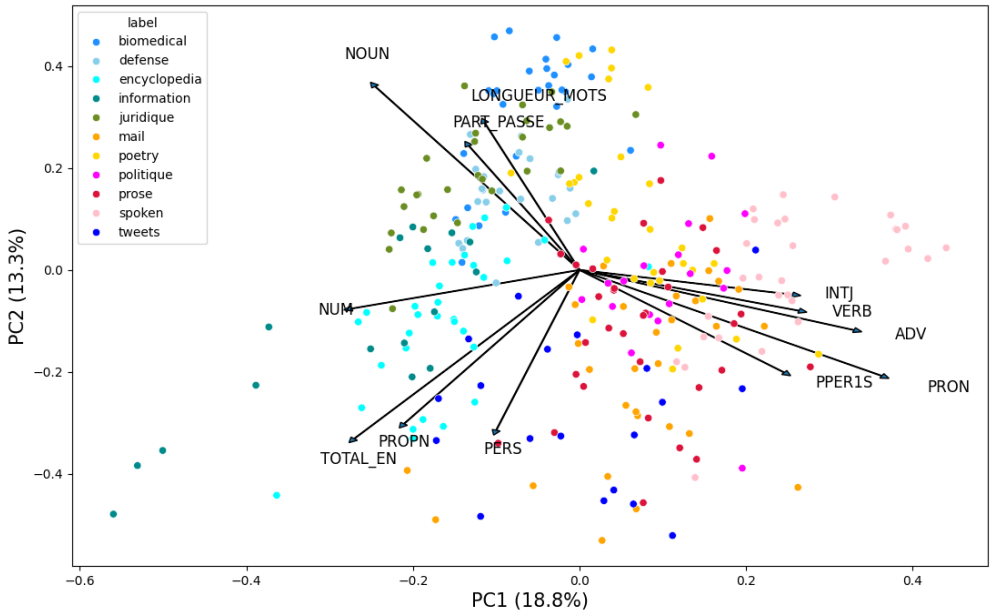

FIGURE 1 – Première et deuxième composante de l'ACP sur les 33 caractéristiques extraites des documents de NEM.FR.

(NUM) : en effet, cette caractéristique se retrouve principalement dans les genres ENCYCLOPEDIA et INFORMATION. Cette distribution reflète la présence fréquente de données chiffrées, typiques des textes à visée informative ou descriptive.

Dans la partie supérieure du plan (valeurs positives de PC2), les documents se caractérisent par une proportion plus élevée de noms communs (NOUN), une plus grande longueur moyenne des mots (LONGUEUR_MOTS), ainsi que de participes passés fréquents (PART_PASSE). Ces propriétés semblent particulièrement présentes dans les genres BIOMEDICAL, DEFENSE et JURIDIQUE. Elles correspondent à des textes à forte densité lexicale, mobilisant un vocabulaire spécialisé et des constructions syntaxiques complexes, souvent associées à des descriptions techniques, des définitions ou des formulations institutionnelles.

À l'inverse, la partie inférieure du plan (valeurs négatives de PC2) est associée à une plus forte proportion de noms propres (PROPN), ainsi qu'à une fréquence plus élevée d'entités nommées en général (TOTAL_EN) mais surtout de mentions de personnes (PERS). Cette zone regroupe principalement des documents issus des genres TWEETS, MAIL, SPOKEN et, dans une moindre mesure, POLITIQUE et PROSE. Cette configuration reflète des textes davantage ancrés avec des référents spécifiques, où les individus, les organisations ou les lieux sont explicitement mentionnés, ce qui correspond à des usages plus contextualisés.

Le diagramme en bâtons présenté dans la figure 2 illustre la distribution des scores des échantillons du corpus NEM.FR le long de la première composante principale (PC1), qui capture la plus grande part de la variance. La partie droite du graphique (valeurs positives) regroupe majoritairement les échantillons issus des genres SPOKEN, POLITIQUE, MAIL, PROSE et TWEETS. Linguistiquement, ce pôle est fortement corrélé aux marques de l'énonciation (PPER1S) et à une dynamique verbale (VERB, ADV, INTJ). À l'opposé, la partie gauche (valeurs négatives) montre une forte concentration

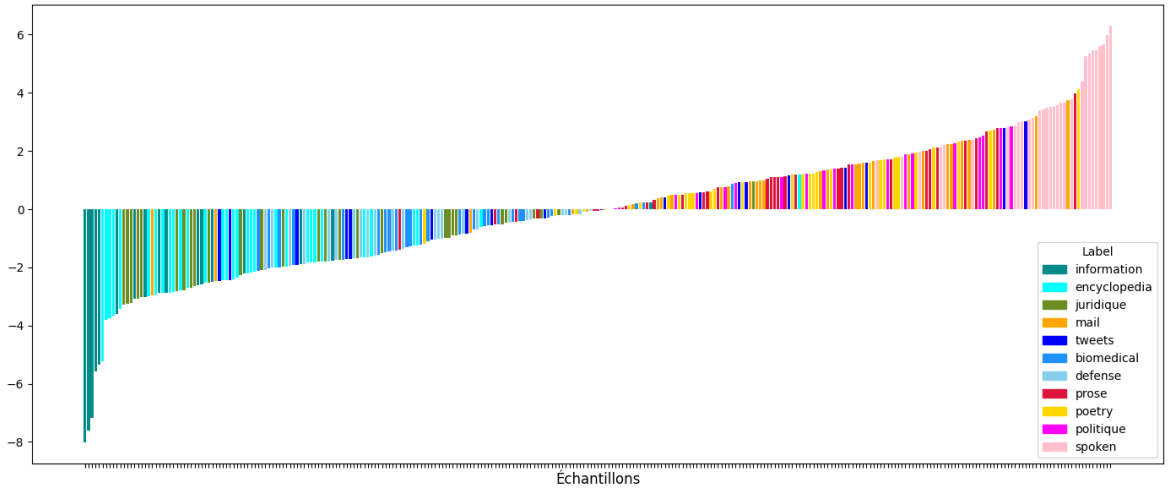

FIGURE 2 – Distribution des scores des échantillons du corpus sur notre première dimension (PC1).

	k = 2	
label	**cluster 0**	**cluster 1**
biomedical	3,70	96,30
defense	0,00	100,00
encyclopedia	3,13	96,88
information	0,00	100,00
juridique	0,00	100,00
mail	90,32	9,68
poetry	51,28	48,72
politique	94,74	5,26
prose	87,50	12,50
spoken	100,00	0,00
tweets	68,42	31,58

FIGURE 3 – Répartition (%) des échantillons par genre pour $k = 2$

des genres INFORMATION, JURIDIQUE, ENCYCLOPEDIA, DEFENSE et du BIOMEDICAL. Ce pôle est défini par des caractéristiques liées à la densité nominale (NOUN, LONGUEUR_MOTS) et à la précision factuelle (PROPN, TOTAL_EN, PERS, NUM). Enfin, la partie centrale du diagramme (scores proches de zéro) montre une zone de transition où les échantillons de différentes catégories se chevauchent. Nous pouvons également noter que certains échantillons d'un genre qui est majoritairement d'un côté, peuvent apparaître de l'autre (ex : TWEETS, POETRY, PROSE, MAIL).

3.3 Vers une opposition stylistique

Afin de compléter notre lecture et d'identifier des structures linguistiques latentes au sein de notre corpus, nous avons appliqué l'algorithme de classification floue *Fuzzy C-Means* à nos caractéristiques linguistiques. Cette méthode est particulièrement pertinente pour notre étude, où la frontière entre les genres textuels est relativement poreuse.

Pour déterminer le nombre optimal de clusters (k), nous avons été guidés par trois métriques de

performance classiques, dont les résultats convergent de manière univoque vers une partition en deux clusters (k = 2) : le *Fuzzy Partition Coefficient* (FPC), la *Partition Entropy* (PE), et la *Silhouette Score*.

L'étude de la répartition des échantillons au sein des deux clusters (voir Tableau 3) confirme que l'algorithme a capturé une structure sous-jacente liée à la formalité et à l'élaboration du discours, plutôt qu'à une simple proximité thématique.

On observe une polarisation quasi parfaite du corpus. **(i)** Le cluster 0 (pôle expressif/spontané) : ce pôle rassemble les genres marqués par l'immédiateté ou l'oralité. Le score de 100% pour le genre SPOKEN est ici un indicateur fort. La présence massive des échantillons des genres POLITIQUE (94,7%), MAIL (90,3%) et PROSE (87,5%) suggère que ce cluster capte les traits liés à l'interaction, comme l'utilisation de pronoms personnels et de marques de l'oralité. **(ii)** Le cluster 1 (pôle référentiel/technique) : Ce cluster regroupe l'intégralité ou la quasi-totalité des échantillons des genres institutionnels et techniques (DEFENSE, JURIDIQUE, INFORMATION, BIOMEDICAL, ENCYCLOPEDIA). Ces textes se caractérisent par une syntaxe élaborée, une forte densité de noms propres et de termes techniques, et une absence de marques de subjectivité. Toutefois, certains genres de notre corpus présentent une distribution plus hybride.

Le genre POETRY est le seul genre de notre corpus à présenter une répartition quasi équilibrée (51% en cluster 0, 49% en cluster 1). L'examen des échantillons révèle que cette division ne repose pas sur le thème, mais sur la structure syntaxique et la densité des entités. **(i)** La poésie du cluster 1 (pôle référentiel/technique) concerne les textes de Rimbaud et d'Huguette Bertrand et rejoignent le pôle des textes techniques (ex : JURIDIQUE, BIOMEDICAL) pour des raisons précises. Rimbaud multiplie les syntagmes nominaux et les noms propres géographiques (ex : "l'Épire et le Péloponèse, ou que la grande île du Japon, ou que l'Arabi"). Cette accumulation de noms propres et de noms crée une signature statistique proche du texte informatif ou encyclopédique. Le texte d'Huguette Bertrand, bien que très métaphorique, utilise des noms concrets et anatomiques (ex : "ventres", "corps", "respir", "main"). L'absence de pronoms personnels (je/tu) le détache du pôle spontané pour le placer dans une forme d'observation objective du mouvement. **(ii)** La poésie du cluster 0 (pôle expressif/spontané) concerne les œuvres de Baudelaire, Vallois et Noailles sont classées aux côtés du SPOKEN et des MAIL en raison de leur nature intrinsèquement interactive et déictique. Chez Baudelaire, l'utilisation d'interjections, de tirets et autres signes de ponctuation (ex : "–Cependant tu vas gueusant") mime une forme de spontanéité ou de rupture syntaxique propre au langage parlé : la poésie devient ici une parole plutôt qu'une description. Quant aux extraits de Vallois et Noailles, ils contiennent de nombreux pronoms personnels de première et deuxième personnes (ex : "Je ne sais pas", "Tu me feras", "Nos cœurs voudraient vous retenir"). La forte fréquence de ces pronoms place ces textes dans la catégorie du discours impliqué, caractéristique du cluster 0.

Les échantillons du genre TWEETS se répartissent de manière asymétrique : 68% dans le cluster 0 (pôle expressif/spontané) et 32% dans le cluster 1 (pôle référentiel/technique). Cette dualité reflète les deux fonctions primaires de la plateforme, à savoir l'interaction et la diffusion d'actualités. **(i)** Dans le groupe du cluster 1, on retrouve la totalité des tweets au registre **soutenu** (5) et un seul tweet **courant**. Dans le groupe du cluster 1, on retrouve la totalité des tweets de registre soutenu (5) et un seul tweet courant. Ils affichent une densité élevée de noms propres et d'entités nommées liées aux lieux, organisations ou personnalités politiques. La structure syntaxique est proche de celle de la dépêche ou de la synthèse informative (ex : "Dépakine : le laboratoire français Sanofi mis en examen pour homicides involontaires"). La longueur des mots et la densité des phrases les éloignent de la spontanéité et les rapprochent de l'écrit élaboré. **(ii)** Le cluster 0 capte la totalité des tweets au registre **familier** (6), quasi-totalité du **courant** (5) et deux seuls au **soutenu**. Ces tweets sont caractérisés par

une interactivité maximale (réponses parfois courtes, sans contexte, ex : "@XXX @XXX Chut faut pas le dire à Camille"), traitent de sujets "intenses" (sport, gaming, vlogs) et de réactions à chaud (insultes, enthousiasme). Ils sont saturés en interjections, verbes au présent, et par l'omniprésence de la première personne du singulier. L'usage massif de l'argot (ex : "wesh", "mif", "bougs"), d'émojis, et de marqueurs de sentiment (ex : "j'ai le seum", "bichette") entre en contraste avec les tweets "froids" et dépersonnalisés de l'autre cluster.

Cette répartition démontre que la classification floue n'oppose pas simplement des genres (ex : biomédical vs Twitter), mais des postures énonciatives. Un tweet peut être **informationnel** par sa forme s'il relaie une mise en examen de laboratoire (registre courant/soutenu), ou **impliqué** s'il exprime une opinion virulente sur un médicament (familier).

4 Conclusion

Cette étude visait à explorer si la variation linguistique au sein d'un corpus pouvait être mieux appréhendée par des caractéristiques linguistiques internes au texte plutôt que par des catégories prédéfinies externes, telles que le domaine ou le genre. Nos résultats suggèrent que les méthodes ascendantes non supervisées basées sur les caractéristiques linguistiques – ici, l'ACP avec classification floue – permettent de révéler les dimensions latentes de la variation, souvent masquées par les classifications descendantes traditionnelles. La structure interne révélée par ces méthodes montre que la variation textuelle est mieux représentée comme un continuum plutôt que comme un ensemble de catégories discrètes. Certaines catégories traditionnellement considérées comme distinctes (par exemple, le langage parlé, politique et par correspondance) présentent des caractéristiques linguistiques communes, tandis que d'autres (par exemple, le langage informationnel et encyclopédique) forment des groupes plus stables et homogènes. Nous avons montré qu'avec seulement 33 caractéristiques, l'ACP combinée à la classification floue offre des perspectives intéressantes, révélant des méta-catégories ainsi que des catégories subdivisées. Ce résultat ouvre la voie à de nouvelles expérimentations avec des caractéristiques supplémentaires — comme les dépendances syntaxiques ou les plongements sémantiques — afin d'affiner l'analyse. Sans aucune hypothèse préalable concernant les catégories de notre corpus, l'ensemble de caractéristiques utilisées s'est avéré suffisant pour en discriminer certaines de manière pertinente. Ce résultat confirme également notre hypothèse initiale selon laquelle des documents d'un même genre peuvent, en pratique, appartenir à plusieurs classes, en fonction de leur profil linguistique. Ces résultats confirment l'idée que la variation linguistique peut être décrite avec plus de précision et de fiabilité par des observables linguistiques que par des étiquettes externes. NEM.FR vise ainsi à fournir un cadre pour la caractérisation linguistique des genres textuels et l'influence de la variation textuelle sur le comportement d'un modèle. Il a d'ailleurs déjà été mis à profit dans des travaux récents (Xu *et al.*, 2026) visant à évaluer la robustesse face à la variation textuelle des grands modèles de langage (LLMs) pour la tâche de REN. Dans de futurs travaux, nous avons l'intention d'évaluer les performances des systèmes de TAL au sein des différentes classes linguistiques identifiées par regroupement. Cela nous permettra de mieux comprendre si ce sont des configurations linguistiques spécifiques — plutôt que des catégories textuelles — qui sont responsables des variations de performance, et, en fin de compte, d'orienter la conception des corpus et les stratégies d'adaptation des modèles vers des pratiques d'évaluation plus fiables et fondées sur des critères linguistiques solides.

Références

ABOUDA L. & BAUDE O. (2005). Du Français Fondamental aux ESLO. In *Cahiers de linguistique*, volume 33 de *Cahiers de linguistique*, p. 131–146, Lyon, France. HAL : halshs-01162533.

ATILF & CLLE (2020). Corpus journalistique issu de l'est républicain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BENDER E. M. & FRIEDMAN B. (2018). Data Statements for Natural Language Processing : Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, **6**, 587–604. DOI : 10.1162/tacl_a_00041.

BIBER D. (1988). *Variation across Speech and Writing*. Cambridge University Press. DOI : 10.1017/CBO9780511621024.

BIBER D. (1992). The multi-dimensional approach to linguistic analyses of genre variation : An overview of methodology and findings. *Computers and the Humanities*, **26**(5/6), 331–345.

BIBER D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, **8**(4), 243–257.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.

CAO D., BÉCHET N. & MARTEAU P.-F. (2024). WikiNER-fr-gold : A Gold-Standard NER Corpus. arXiv :2411.00030 [cs], DOI : 10.48550/arXiv.2411.00030.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FU J., LIU P. & NEUBIG G. (2020). Interpretable multi-dataset evaluation for named entity recognition. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6058–6069, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.489.

GIORDANO B., PRIEUR M., VUTH N., VERDY S., COUSOT K., SÉRASSET G., GADEK G., SCHWAB D. & LOPEZ C. (2024). POPCORN : Fictional and Synthetic Intelligence Reports for Named Entity Recognition and Relation Extraction Tasks. *Procedia Computer Science*, **246**, 1170–1180. DOI : 10.1016/j.procs.2024.09.542.

KUZMAN T. & LJUBEŠIĆ N. (2025). Automatic genre identification : a survey. *Language Resources and Evaluation*, **59**(1), 537–570.

LABRAK Y., ROUVIER M. & DUFOUR R. (2023). MORFITT : Un corpus multi-labels d'articles scientifiques français dans le domaine biomédical.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *4e Congrès Mondial de Linguistique Française*, volume 8, p. 2675–2689, Berlin, Allemagne. DOI : 10.1051/shsconf/20140801305, HAL : halshs-01061368.

LAIPPALA V., EGBERT J., BIBER D. & KYRÖLÄINEN A.-J. (2021). Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language Resources and Evaluation*, **55**, 757–788.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEE S.-J. & JIANG J.-Y. (2014). Multilabel text categorization based on fuzzy relevance clustering. *IEEE Transactions on Fuzzy Systems*, **22**(6), 1457–1471. DOI : 10.1109/TFUZZ.2013.2294355.

MEKKI J., BATTISTELLI D., LECORVÉ G. & BÉCHET N. (2021). TREMoLo-Tweets corpus : guide d'annotation pour un corpus annoté en registres de langue pour le français.

MILLOUR A., DUPONT Y., FORT K. & DUIGNAN L. (2024). Unveiling Strengths and Weaknesses of NLP Systems Based on a Rich Evaluation Corpus : the Case of NER in French. In *LREC-COLING 2024*, Turin, Italy. HAL : hal-04534593.

MILLOUR A., DUPONT Y., JOUGLAR A. & FORT K. (2022). FENEC : un corpus équilibré pour l'évaluation des entités nommées en français (FENEC : a balanced sample corpus for French named entity recognition). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 82–94, Avignon, France : ATALA.

PASSONNEAU R. J., IDE N., SU S. & STUART J. (2014). Biber redux : Reconsidering dimensions of variation in American English. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 565–576, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

PONTI E. M., O'HORAN H., BERZAK Y., VULIĆ I., REICHART R., POIBEAU T., SHUTOVA E. & KORHONEN A. (2019). Modeling language variation and universals : A survey on typological linguistics for natural language processing. *Computational Linguistics*, **45**(3), 559–601. DOI : 10.1162/coli_a_00357.

RAO M. S., KALYAN O. P., KUMAR N. N., TABASSUM M. T. & SRIHARI B. (2021). Automatic music genre classification based on linguistic frequencies using machine learning. In *2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*, p. 1–5 : IEEE.

SANTINI M., MEHLER A. & SHAROFF S. (2011). *Riding the Rough Waves of Genre on the Web*, p. 3–30. DOI : 10.1007/978-90-481-9178-9_1.

SEDDAH D. & CANDITO M. (2016). Hard time parsing questions : Building a QuestionBank for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2366–2370, Portorož, Slovenia : European Language Resources Association (ELRA).

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SINCLAIR J. (1996). *Preliminary Recommendations on Corpus Typology*.

XU Z., SEGHIER M., MILLOUR A., GONZALEZ-GAILLARDO C.-E. & ANTOINE J.-Y. (2026). Evaluating the adaptability of large language models to linguistic variation. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC-2026)* : ELRA Language Resources Association.

ZIEMS C., HELD W., YANG J., DHAMALA J., GUPTA R. & YANG D. (2023). Multi-value : A framework for cross-dialectal english nlp.