Global Scientific JOURNALS

GSJ: Volume 11, Issue 5, May 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

ANALYSE DES REGISTRES DE LANGUE ET DE LA VARIATION SOCIALE DANS UN CORPUS LITTÉRAIRE : UNE APPROCHE BASÉE SUR LE TRAITEMENT AUTOMATIQUE DU LANGAGE

ASSOULI Zoubida, KHATTABI Redouane.

ASSOULI Zoubida, Professeur encadrant Jamila Bellamqaddam, doctorant en Sciences du Langage et Communication, Laboratoire Langage et Société CNRST-URAC56, Faculté des Lettres, Langues et Arts, Université Ibn Tofaïl, Kénitra-Maroc, +212672816532, zoubida.assouli@uit.ac.ma
KHATTABI Redouane, Professeur encadrant Jamila Bellamqaddam, doctorant en Sciences du Langage et Communication, Laboratoire Langage et Société CNRST-URAC56, Faculté des Lettres, Langues et Arts, Université Ibn Tofaïl, Kénitra-Maroc, +212661891287, redouane.khattabi@uit.ac.ma

KeyWords
Language register, Natural Language Processing (NLP), Literary corpus, social variation, social status, linguistic variable, relative lexical frequency.

ABSTRACT

In this study, we focus on the perception of language registers in social variation and their connection to the social status of speakers. While oral corpora have been extensively studied for the analysis of language registers, written corpora, particularly literary works, have been overlooked. We question the existence of a diversity of French language registers in a written corpus, specifically in Émile Zola's novel Germinal. Our research aims to highlight variations in language registers based on indicators of social status present in the work. The adopted methodology is based on corpus preparation, including formatting, cleaning, and labeling of the text. We then select indicators of social status, contextual situations, and linguistic variables for our analysis. The automatic processing of the corpus will allow us to identify occurrences and relative frequencies of linguistic variables in relation to the discourse. We posit that lexical diversity is a key indicator of variation in language registers. Thus, we begin by identifying indicators of social status among the characters in our corpus and select a sample for our study. We then choose relevant contextual situations. The third step involves defining the linguistic variables or descriptors that will be automatically processed using natural language processing (NLP) software. Our automatic exploration of the corpus will proceed in several stages. Firstly, we will calculate the relative lexical frequencies of language registers in the discourse of each character in our sample, in order to classify them within our corpus. Next, we will identify the variables to search for and analyze their occurrences in the characters' discourse. Finally, we will compare the results obtained with the relative lexical frequencies, seeking to establish connections between the language registers used and the indicators of social status of the characters. This semi-automatic and statistical approach to exploring a literary corpus will allow us to highlight language registers based on the social status of speakers. The findings of this study will provide new insights into social variation in language and contribute to a better understanding of the specificities of linguistic expression in different contexts.

Dans l'étude de la variation sociale, la perception des registres de langue revêt une importance cruciale pour comprendre les spécificités de l'expression linguistique des locuteurs. Cette évaluation, qui implique un jugement de valeur accompagné d'une hiérarchie sociale, permet de situer chaque production linguistique dans un registre donné. Ainsi, les locuteurs sont souvent attribués à un espace social déterminé (Labov, 1972 ; Milroy, 1987). Bien que les corpus oraux présentent une alternance de registres de langue en fonction de l'usage et de l'usager, les corpus écrits ou littéraires sont généralement associés à l'utilisation d'un registre soutenu (Biber et Conrad, 2009). Cependant, les œuvres littéraires, comprenant des dialogues, reflétant des échanges verbaux entre des personnages attribués d'indicateurs de statut social varié. En traitement automatique des langues (TAL), les registres de langue sont peu étudiés en raison du manque de logiciels spécifiques à cette exploration. Cet article présente une approche d'exploration semi-automatique et statistique d'un corpus textuel littéraire en vue de mettre en évidence les registres de langue des locuteurs en fonction de leur statut social. En effet, la perception et l'évaluation des registres de langue constituants des éléments fondamentaux pour appréhender la variation sociale dans la langue.

Toutefois, les études antérieures se sont principalement concentrées sur les corpus oraux, négligeant ainsi les corpus écrits, notamment les œuvres littéraires. Par conséquent, il est nécessaire de s'interroger sur l'existence d'une diversité de registres de langue française dans un corpus écrit, plus précisément dans l'œuvre germinal d'Émile Zola. Ainsi, la problématique de cette recherche se formule comme suit : dans quelle mesure l'analyse des structures discursives de Germinal permet-elle de mettre en évidence une variation des registres de langue en fonction des indicateurs du statut social, à la lumière des travaux de Françoise Gadet et Blanche-Benveniste (Gadet, 2007 ; Blanche-Benveniste, 2010) ? La méthode adoptée dans notre recherche repose en premier sur la préparation du corpus : formatage, nettoyage et étiquetage de la matière textuelle. Ensuite, le choix des indicateurs du statut social, des situations contextuelles et des variables linguistiques. Le traitement automatique du corpus tentera de mettre en évidence les occurrences et les fréquences relatives des variables linguistiques en fonction des discours. Dans cet article, nous aborderons tout d'abord notre positionnement théorique en mettant en avant les concepts clés nécessaires à la compréhension de notre recherche. Par la suite, nous fournirons un aperçu des travaux de recherche existants dans le domaine afin de contextualiser notre étude. En ce qui concerne la méthodologie, nous détaillerons les différentes étapes et procédures que nous avons suivies pour mener notre recherche. Nous présenterons également le corpus d'étude que nous avons utilisé, ainsi que les critères qui ont guidé notre sélection des indicateurs du statut social, des situations contextuelles et des variables linguistiques. Enfin, nous exposerons les résultats de notre recherche, en mettant en relation les résultats obtenus avec notre positionnement théorique et en les discutant à la lumière de l'état de l'art.

1. Positionnement théorique et état de l'art

Cet article se situe à la croisée de deux champs disciplinaires : la sociolinguistique, domaine d'étude qui explore la relation entre la langue et la société, qui examine comment les facteurs sociaux tels que la classe sociale, l'ethnie, le genre et l'âge influencent l'utilisation et la variation linguistique. Le deuxième domaine est le Traitement Automatique du Langage Naturel (TALN) qui se rapporte à la technologie qui permet aux ordinateurs de comprendre, d'analyser et de générer le langage humain de manière automatique. Le principe du registre de langue se rapporte aux méthodologies d'évaluation et de catégorisation des productions langagières des sujets parlants au sein d'un groupe linguistique spécifique, comme dans les travaux de (Bourquin, 1965) de (Halliday, 1978) et de (Offord, 1990). On peut aussi différencier divers registres basés sur plusieurs traits distinctifs tels que la complexité des termes, l'ordre des mots, le temps des verbes, la longueur des phrases, et plus encore, ces distinctions sont généralement conceptualisées sur un continuum varié qui inclut des niveaux de langue tels que le registre soutenu, courant, familier, etc. L'organisation catégorielle peut manifester une diversité de portée dépendant de la définition prescrite du concept de "registre" - un terme qui, en lui-même, est sujet à controverses académiques - et a le potentiel d'exposer des degrés variables d'études nuancées comme dans les travaux de recherches menées par Gadet (1996), et celles de Biber et Finnegan (1994). L'objet d'analyse peut également s'étendre à l'incidence des médias de communication révélée par les travaux de Charaudeau (1997) ou sur le discours (Moirand, 2007). Dans le cadre de notre recherche, nous optons pour une méthodologie plus traditionnelle, classifiant les registres de langue en trois catégories distinctes : le registre familier, le registre courant et le registre soutenu. Selon nos recherches bibliographiques, les recherches axées spécifiquement sur les registres de langue sont rares, bien que de nombreux travaux sur le traitement du style fournissent une méthodologie et des outils théoriques. Toutefois, l'exploration des registres de langue partage certaines caractéristiques communes avec les domaines d'étude consacrés aux méthodes qui analysent les choix de mots, la syntaxe, et d'autres éléments stylistiques pour identifier l'auteur d'un texte comme les travaux de Stamatatos (2009) et (Swain et al., 2017) d'autres recherches ont été menées sur l'analyse des nouveaux médias (Ouertatani et al., 2018) (Bellot et al., 2019). Des corpus de référence destinés à ces divers médias ont été publiés en nombre significatif. Les méthodes employées dans l'analyse automatique du langage reposent sur l'extraction d'un ensemble de descripteurs significatifs prélevés des textes soumis à l'analyse. Les études en attribution d'auteur, en raison de leur portée historique, ont permis l'identification d'un vaste spectre de ces descripteurs. Stamatatos (2009) souligne que les préférences et les décisions stylistiques d'un auteur se manifestent à divers niveaux linguistiques. Le niveau lexical est l'aspect le plus manifeste et le plus étudié de l'analyse linguistique, illustré notamment par l'examen de la longueur des mots et des phrases, la variété lexicale, ainsi que les fréquences d'apparition des mots (Dubois, et Dubois-Charlier, 2002), (Melka, 2014) et (. Chollet et Guryev, 2016). Dans le domaine de la syntaxe, l'emploi de marqueurs ou variables issus d'investigations morphosyntaxiques et syntaxiques est fréquemment répertorié comme moyen de caractérisation du style, comme l'indiquent diverses études (Rainer et Poudat, 2010 ; Danlos et Namer, 2015 ; Mazière et Prévost, 2016). En somme, un certain nombre de travaux se sont intéressés au domaine de l'information graphémique, en considérant des n-grammes de caractères, les types de graphèmes ou encore en se penchant sur les mesures de compression de l'information (Ribalet, 2012;. Petitjean et Brun, 2013 ; Schang, 2015).

2. Méthodologie

Dans le cadre de notre étude, nous postulons que la diversité lexicale se manifeste généralement en tant que principal indicateur de la variation du registre linguistique adopté par les locuteurs. Ainsi, la première étape de notre approche consiste à mettre en évidence les indicateurs du statut social des personnages de notre corpus et d'en choisir un échantillon pour notre étude. La deuxième étape propose le choix des situations contextuelles. La troisième étape consiste à délimiter les variables linguistiques ou descripteurs qui seront traités automatiquement à l'aide de logiciels TALN.

L'exploration automatique du corpus sera comme suit : tout d'abord, nous allons calculer les fréquences lexicales relatives des registres de langue dans les discours de chaque personnage de notre échantillon afin de mettre en place un classement des registres de langue dans notre corpus. Puis, nous allons déterminer des variables à rechercher dont nous allons calculer les occurrences dans les discours des personnages et les analyser. Enfin, nous allons procéder à la comparaison des résultats obtenus (occurrences des variables à rechercher) avec le résultat de la fréquence lexicale relative. Et nous allons essayer de mettre la lumière sur le lien entre le registre de langue utilisé et les indicateurs du statut social des personnages.

2.1. Le corpus

Germinal", une œuvre littéraire écrite par Émile Zola publiée en 1885. Ce roman est le treizième de la collection intitulée « Les Rougon-Macquart », une série volumineuse comprenant vingt romans qui offrent une représentation exhaustive de la société française durant la période du Second Empire. "Germinal" se concentre sur la vie des mineurs dans le nord de la France, décrivant les conditions de vie difficiles, le travail épuisant et l'exploitation des travailleurs par les propriétaires des mines. Le choix de "Germinal" comme corpus de notre étude est motivé par plusieurs facteurs. Premièrement, le roman est riche en dialogues qui mettent en lumière les différents registres de langue, faisant de lui un terrain fertile pour une étude sociolinguistique. De plus, les personnages de "Germinal" proviennent de diverses couches de la société, offrant une variété de registres à explorer. Le traitement automatique des registres de langue dans "Germinal" offre une occasion unique d'étudier les variations linguistiques dans leur contexte social. Les progrès substantiels réalisés dans le domaine de l'ingénierie linguistique, notamment le traitement automatique du langage naturel (TALN), ont rendu possible l'examen minutieux des registres de langue, permettant ainsi une compréhension plus approfondie de la façon dont le langage reflète la stratification sociale dans le roman.

2.2. Le choix des indicateurs du statut social

Le tableau suivant donne un aperçu de la diversité des classes sociales représentées dans "Germinal", et de la manière dont ces différences sont illustrées par les variables sociales des personnages.

Personnage	Âge (approximatif)	Sexe	Métier	Salaire (indicatif)	Niveau d'étude
Étienne Lantier	26 ans	M	Mineur	Bas	Faible
Toussaint Maheu	42 ans	M	Mineur	Bas	Faible
Constance Maheude	39 ans	F	Femme au foyer/Mineuse	Très bas	Très faible
Catherine	15 ans	F	Mineuse	Bas	Très faible
Chaval	25 ans	M	Mineur	Bas	Faible
Victor Deneulin	50 ans	M	Propriétaire de mine	Élevé	Élevé
M. Hennebeau	55 ans	M	Directeur de la Compagnie des	Très élevé	Très élevé
Mme Hennebeau	50 ans	F	Femme au foyer	N/A	Élevé

Tableau 1 : les indicateurs du statut social

Notre choix s'est porté sur deux échantillons issus de deux couches sociales diamétralement opposées. D'une part, les capitalistes (bourgeois) très instruits : Victor Deneulin et M. Philipe Hennebeau. Et d'autre part, les prolétaires (mineurs) vivant dans la misère et peu ou pas instruits : Étienne Lantier, Toussaint Maheu et Constance Maheude.

2.3. Le choix des situations contextuelles

D'après les travaux de F. Gadet et W. Labov, une corrélation significative est identifiée entre le statut socio-économique des interlocuteurs et la qualité de leur utilisation de la langue, ce qui implique que l'analyse des comportements langagiers peut être assimilée à une évaluation

sociale basée sur l'usage linguistique. La conceptualisation du niveau de langue ne peut être séparée de celle du statut socio-économique, entraînant une association implicite entre les comportements linguistiques et les comportements sociaux.

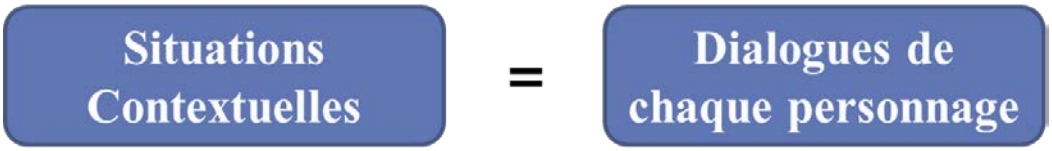


Figure 1 : Les situations contextuelles

2.4. Les variables

Dans le cadre de notre analyse, plusieurs variables linguistiques ont été choisies pour évaluer les registres de langue utilisés par les personnages de "Germinal". La variable de référence est le lexique, avec une attention particulière accordée à la fréquence relative des différents registres utilisés.

En plus de cette variable de référence, d'autres aspects du langage seront examinés, y compris la fréquence des adjectifs et des adverbes, une caractéristique souvent associée à un langage plus formel (Gadet, 2007), ainsi que l'utilisation du présent indicatif.

Nous analyserons également la surprésence des onomatopées, un indicateur de langage informel et expressif (Ilmola, 2012), ainsi que l'utilisation fréquente du mot "là" (Gadet, 1997), et l'effacement du pronom sujet "il " dans les constructions impersonnelles, où "il" est remplacé par "y" (Favart, 1966-2006).

Enfin, nous étudierons la fréquence des éléments "ponctuant" (Gadet, 2003), et la répétition des signes de ponctuation (Branca-Rosoff, 1999), deux caractéristiques qui peuvent influencer le ton et le rythme du discours.

L'analyse de ces variables permettra une vue d'ensemble, mais non exhaustive des registres de langue utilisés par les personnages, et de la manière dont ils modifient leur position sociale.

1. La variable de référence : Le lexique (la fréquence relative des registres).

2. Les variables à rechercher :

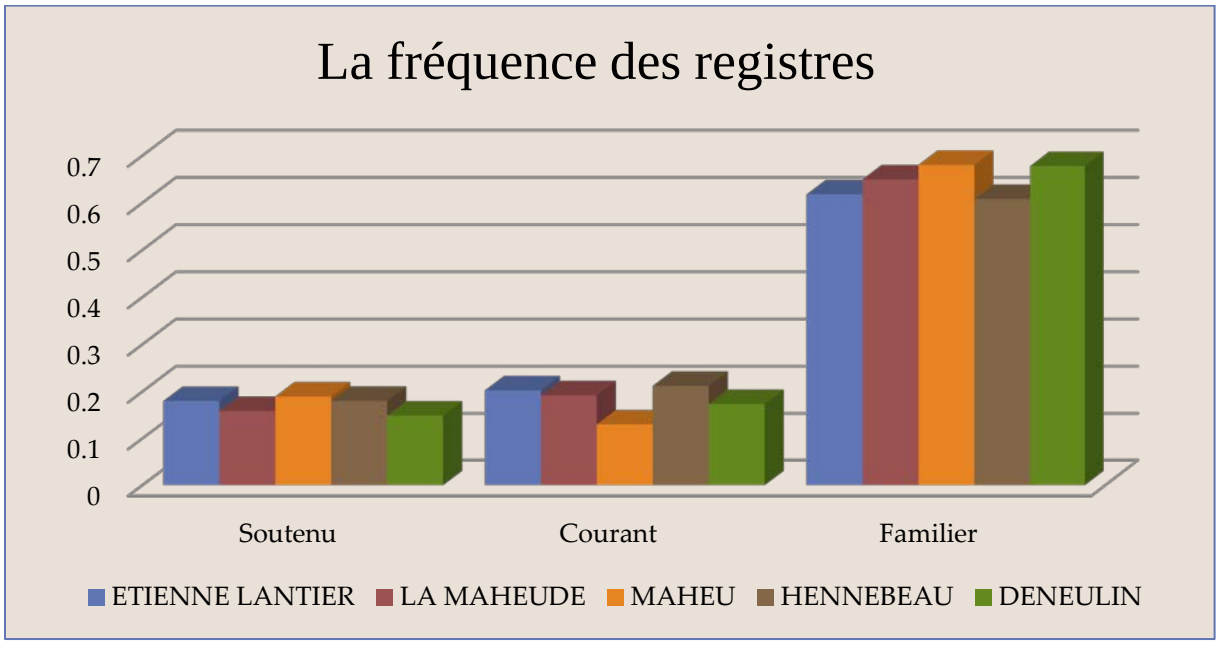
- La fréquence des Adjectifs et des Adverbes.
- Surutilisation du présent indicatif
- Sur présence des Onomatopées
- Sur présence de "là"
- Effacement du pronom sujet "il" dans les constructions impersonnelles - il = y
- La fréquence des éléments "ponctuant"
- Répétition des signes de ponctuation

3. Résultats

3.1. Le lexique : la fréquence des registres

	Étienne Lantier	La Maheude	Maheu	Hennebeau	Deneulin
Registre Soutenu	17,99 %	15,86 %	18,94 %	18,02 %	14,89 %
Registre Courant	20,25 %	19,19 %	13,02 %	21,19 %	17,33 %
Registre Familier	61,76 %	64,95 %	68,03 %	60,79 %	67,78 %

Tableau 2 : La fréquence des registres de langue en fonction du lexique



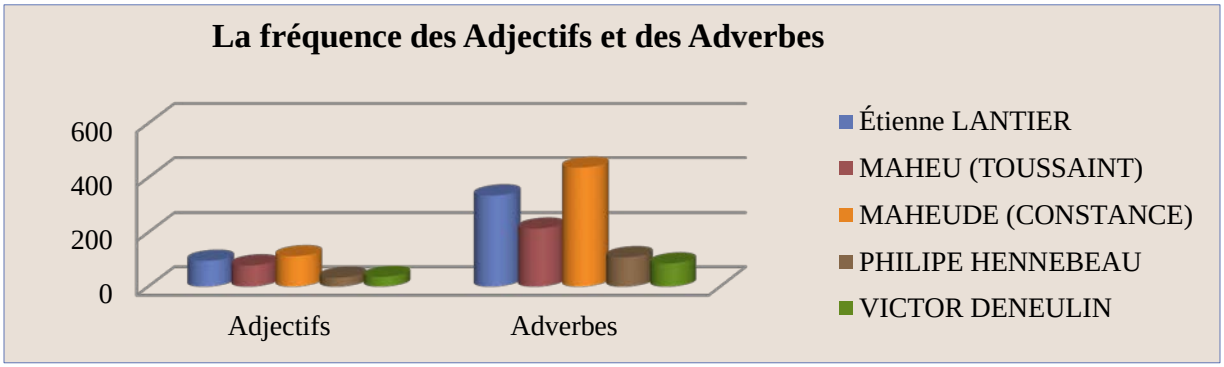
Graphe 1 : La fréquence des registres de langue en fonction du lexique

L'analyse du graphe montre que tous les personnages ont presque les mêmes fréquences d'utilisation des registres dans leurs discours. Les discours des personnages sont dominés par un registre familier (avec une moyenne de 64,66 %), suivi d'un registre courant (18,33 %), et enfin un registre soutenu avec (17,16 %). Nous pouvons déduire que le registre de langue caractéristique des discours des personnages est le registre familier.

3.2. La fréquence des Adjectifs et des Adverbes

	Étienne Lantier	Maheu (Toussaint)	Maheude (Constance)	Philipe Hennebeau	Victor Deneulin
Adjectifs	95	78	112	33	36
Adverbes	334	214	437	108	85

Tableau 3 : La fréquence des Adjectifs et des Adverbes



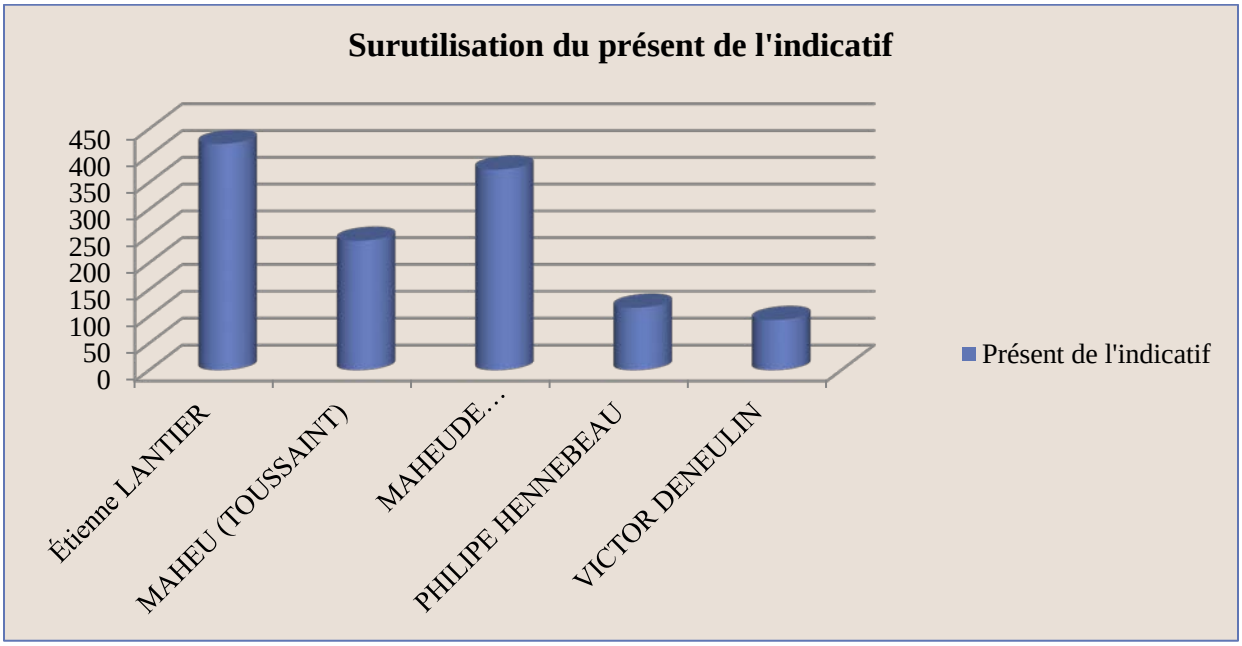
Graphe 2 : La fréquence des Adjectifs et des Adverbes

Nous constatons que la fréquence d'utilisation des adjectifs et des adverbes par les mineurs est deux fois à quatre fois supérieures à celle des bourgeois. Certains mineurs comme Maheude (Constance) présentent une surutilisation des adjectifs et des adverbes dans son discours.

3.2. La surutilisation du présent de l'indicatif

	Présent de l'indicatif
Étienne LANTIER	422
Maheu (Toussaint)	241
Maheude (Constance)	374
Philipe Hennebeau	116
Victor Deneulin	93

Tableau 4 : La surutilisation du présent de l'indicatif



Graphe 3 : La surutilisation du présent de l'indicatif

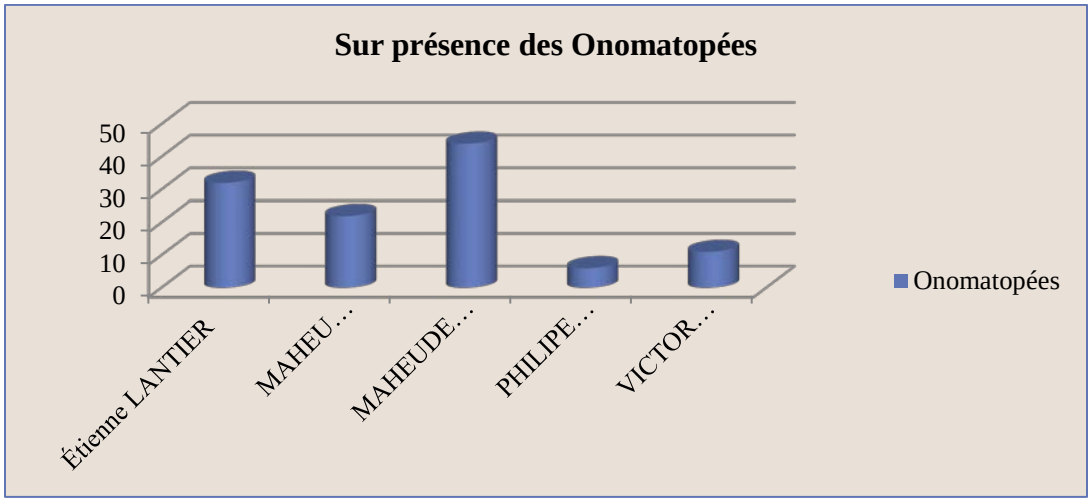
Nous remarquons que la fréquence du présent de l'indicatif chez les mineurs est nettement supérieure à celle chez les bourgeois.

3.3. La sur présence des Onomatopées

	Sur présence des Onomatopées
Étienne Lantier	32
Maheu (Toussaint)	22
Maheude (Constance)	44

Philipe Hennebeau	6
Victor Deneulin	11

Tableau 5 : La sur présence des Onomatopées



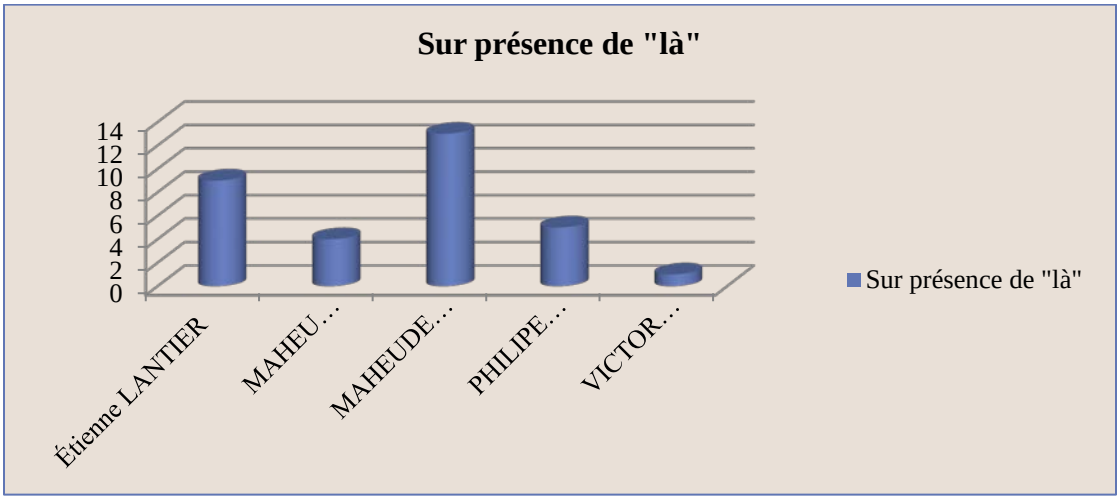
Graphe 4 : La sur présence des Onomatopées

Il se dégage clairement du graphe 4 une plus grande proportion d'onomatopées dans le discours des mineurs.

3.4. La sur présence de "là"

	Sur présence de "là"
Étienne Lantier	9
Maheu (Toussaint)	4
Maheude (Constance)	13
Philipe Hennebeau	5
Victor Deneulin	1

Tableau 6 : La sur présence de "là"



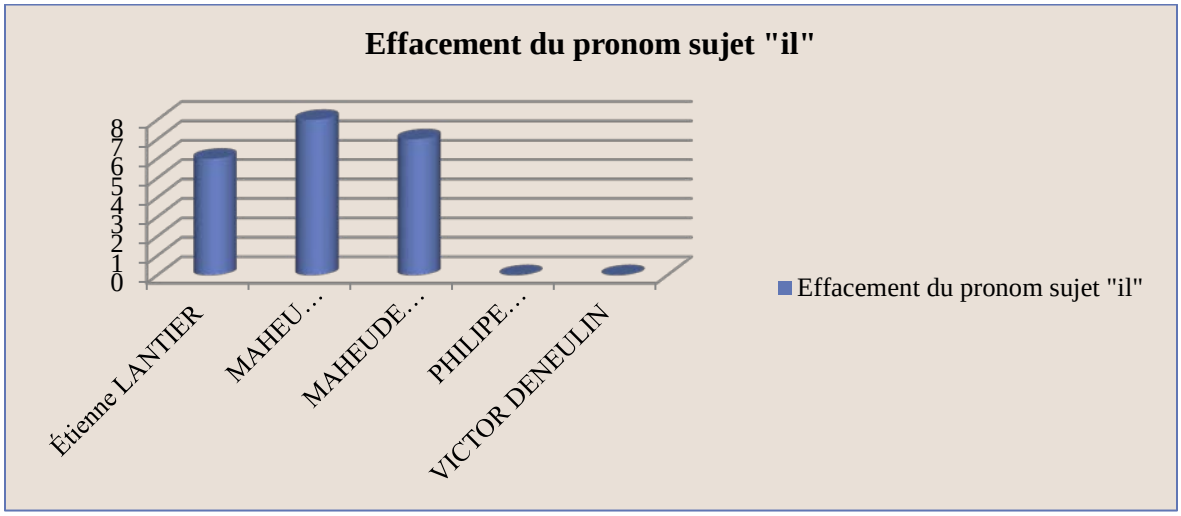
Graphe 5 : La sur présence de "là"

Nous remarquons que sur présence de "là" dans les discours des mineurs sont nettement supérieurs à celle des bourgeois.

3.5. L'effacement du pronom sujet "il"

	Effacement du pronom sujet "il"
Étienne Lantier	6
Maheu (Toussaint)	8
Maheude (Constance)	7
Philipe Hennebeau	0
Victor Deneulin	0

Tableau 7 : L'effacement du pronom sujet "il"



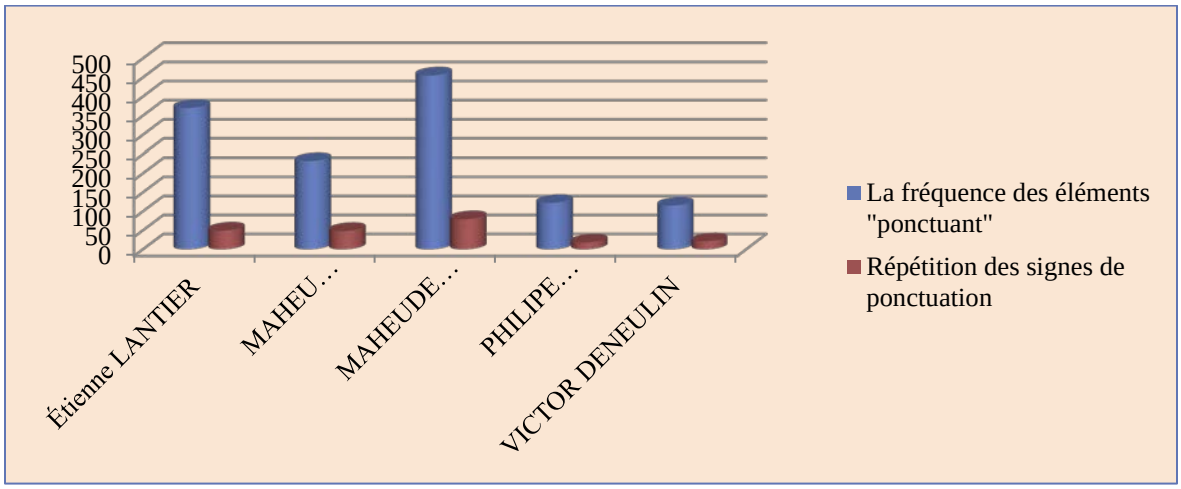
Graphe 6 : L'effacement du pronom sujet "il"

La variable est nettement confirmée puisque les deux bourgeois (Hennebeau & Deneulin) n'ont commis aucun effacement du pronom sujet « il » dans leurs discours.

3.6. La fréquence des éléments "ponctuant"

	La fréquence des éléments "ponctuant"	Répétition des signes de ponctuation
Étienne Lantier	370	49
Maheu (Toussaint)	232	48
Maheude (Constance)	455	79
Philipe Hennebeau	122	18
Victor Deneulin	115	21

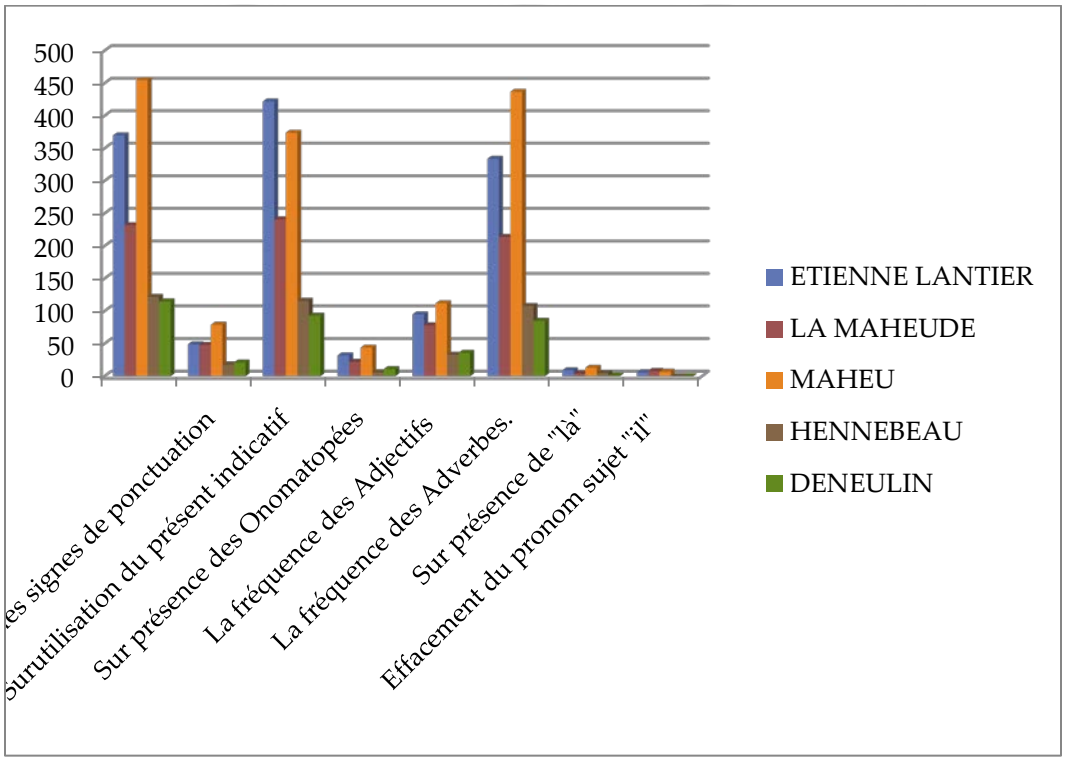
Tableau 8 : La fréquence des éléments "ponctuant"



Graphe 7 : La fréquence des éléments "ponctuant"

Les deux variables sont nettement confirmées dans le diagramme. Le discours des mineurs présente une forte fréquence d'éléments ponctuant et de répétition des signes de ponctuation (!!,…).

3.7. Synthèse des résultats



Graphe 8 : Synthèse des résultats

4. Discussion

Les variables testées ont une fréquence élevée dans les discours des trois mineurs, on peut en déduire que :

- ✓ Les mineurs présentent une pauvreté du vocabulaire. En effet, ils sont incapables de trouver les substantifs et les verbes adéquats pour communiquer convenablement, ce qui les pousse à user des adjectifs et des adverbes pour donner plus de sens à leurs phrases.

- ✓ La simplicité de conjugaison au présent de l'indicatif donne une aisance d'utilisation.
- ✓ La pauvreté du lexique des mineurs constitue une entrave communicationnelle, ils se trouvent obligés à utiliser des onomatopées pour combler ce manque.
- ✓ La fréquence d'éléments ponctuant est un indicateur d'utilisation de la parataxe dans la construction phrastique des mineurs, due à une méconnaissance de la syntaxe.
- ✓ La répétition des signes de ponctuation montre que le discours des mineurs présente des phrases incomplètes ou concises, ce qui s'impacte négativement sur la qualité de communication.
- ✓ Un relâchement et une économie dans le discours des mineurs.

Conclusion

L'étude de la fréquence lexicale relative révèle une homogénéité quant à l'utilisation des registres linguistiques par les protagonistes, malgré les différences discernables dans leurs indicateurs de statut social. Cependant, un examen des occurrences des variables linguistiques révèle une divergence notable du registre de langue en fonction desdits indicateurs. Cela soulève une interrogation notable : quel est le facteur sous-jacent à cette discordance dans les résultats obtenus ?

En effet, l'identification précise des limites entre les registres linguistiques représente un défi pour les lexicographes, comme le souligne Françoise Gadet (Gadet F., 1997). De plus, l'emploi du registre soutenu par les mineurs peut s'expliquer par la nécessité d'engager des négociations avec les propriétaires de mines, imposant ainsi un changement de registre pour répondre aux exigences formelles de la situation. Par ailleurs, l'utilisation du registre familier par les bourgeois peut être attribuée à leurs tentatives de négocier la cessation des grèves des mineurs, nécessitant un langage plus accessible, voire simplifié, pour assurer une compréhension optimale. Même en dehors des négociations, il semble nécessaire pour les bourgeois de maintenir ce registre familier lorsqu'ils s'adressent aux mineurs. Ainsi, les cinq protagonistes ont démontré une capacité à adapter leur production linguistique en fonction des différents contextes d'énonciation.

Nous préconisons que toute recherche future dans ce domaine prenne en considération la situation communicationnelle spécifique de chaque protagoniste, plutôt que de traiter les discours comme une entité autonome.

References

Labov, W. (1972). Modèles sociolinguistiques. Philadelphie : University of Pennsylvania Press.
Milroy, L. (1987). Langage et réseaux sociaux. Oxford : Basil Blackwell.
Biber, D., & Conrad, S. (2009). Registre, Genre et Style. Cambridge : Cambridge University Press.
Gadet, F. (2007). Variation sociale en français. Paris : ArmandColin.
Blanche-Benveniste, C. (2010). Approches de la langue parlée en français. Paris : Ophrys.
Bell, A. (1984). Le style de langage comme conception d'audience. Langue dans la société, 13 (2), 145-204.
Eckert, P., & McConnell-Ginet, S. (1992). Penser concrètement et regarder localement : Langue et genre en tant que pratique communautaire. Revue annuelle d'anthropologie, 21, 461-488.
Gadet, F. (1989). Le français ordinaire. Armand Colin.
Hudson, RA (1996). Sociolinguistique. La presse de l'Universite de Cambridge.
Labov, W. (1966). La stratification sociale de l'anglais à New York. Centre de linguistique appliquée.
ZOLA Émile, Germinal, Éditions Gallimard, Paris, 1978
Favart, Françoise (2010). La représentation de l'oralité populaire dans quelques romans du second XXe siècle (1966-2006). Thèse de doctorat. Lille : Atelier national de reproduction des thèses.
Gadet F. (1997) : Le français ordinaire. Éd. Armond Colin, P. 18.).
Patrice Bellot, Lerch Soëlie, Bruno Emmanuel, Elisabeth Murisasco. Influence des lexiques d'émotions et de sentiments sur l'analyse de sentiments - Application à des critiques de livres. Conférence en Recherche d'Informations et Applications - CORIA 2019, 16th French Information Retrieval Conference, Mar 2019, Lyon, France.
Swain, S., Mishra, G., & Sindhu, C. (2017, April). Recent approaches on authorship attribution techniques—An overview. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 1, pp. 557-566). IEEE.
Ouertatani, A., Gasmi, G., & Latiri, C. (2018). Détection d'opinion argumentée à partir de Twitter. In CORIA.
Patrick Paroubek, Cyril Grouin, Patrice Bellot, Vincent Claveau, Iris Eshkol-Taravella, et al.. DEFT2018 : Recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France. DEFT 2018 - 14ème atelier Défi Fouille de Texte, May 2018, Rennes, France. pp.1-11
Chollet, M., & Guryev, A. (2016). Diversité lexicale et effet de domaine : approche statistique. Revue TAL, 57 (2), 55-82.
Melka, F. (2014). La mesure de la diversité lexicale : méthodes et applications. Peter Lang.
Dubois, D., & Dubois-Charlier, F. (2002). Corpus, méthodes, statistiques et informatique textuelle. Armand Colin.
Rainer, F., & Poudat, C. (2010). La variation stylistique. Armand Colin.
Danlos, L., & Namer, F. (2015). Syntaxe du français contemporain : de la phrase au texte. Armand Colin.
Mazière, F., & Prévost, S. (2016). Stylistique et esthétique du discours littéraire : mélanges offerts à Jean-Louis Aroui. Classiques Garnier.
Ribalet, F. (2012). Étude de la langue française par les graphèmes : modèles et application à la détection de plagiat. Thèse de doctorat, Université de

Toulouse.
Petitjean, F., & Brun, L. (2013). Utilisation de n-grammes de caractères pour l'identification d'expressions figées. Revue TAL, 54 (1), 119-139.
Schang, E. (2015). Systèmes de signes. dans Nouveaux regards sur la linguistique. Lambert-Lucas.