



## **Comparer le discours des enseignants en formation et en exercice : une approche par les méthodes du Traitement Automatique du Langage**

**Haute école pédagogique du canton de Vaud**

**Master of Science HEP Vaud pour le degré secondaire II**

**Discipline Informatique**

**Année académique : 2024–2025**

**Auteur : Gerber Nino**

**n° HEP : p60693**

**Directeur de mémoire : Parriaux Gabriel**

**Membre du jury : Pellet Jean-Philippe**

**Date de soutenance : 23.06.2025**

## **Résumé :**

Ce mémoire propose une analyse comparative du discours d'enseignants en formation (pre-service) et en exercice (in-service) à l'aide d'outils de Traitement Automatique du Langage (TAL). Le corpus, constitué de transcriptions de discours d'enseignants francophones, est analysé au moyen d'un pipeline incluant prétraitement, embeddings (Sentence-BERT), réduction de dimension (PCA), modélisation thématique (BERTopic) et analyse de sentiments.

L'analyse thématique met en évidence des différences marquées entre les deux groupes : les enseignants en formation ont un discours plus spontané et orienté vers l'action, tandis que les enseignants en exercice utilisent un langage plus structuré, normatif et abstrait. Ces tendances sont confirmées par la projection PCA, dont les dimensions révèlent des oppositions sémantiques cohérentes avec les thématiques extraites. En revanche, l'analyse de sentiments ne met pas en évidence de différences significatives entre les deux groupes.

**Mots-clés :** Traitement Automatique du Langage, enseignants, discours, embeddings, TPACK, préservice, in-service, PCA, clustering, sentiment analysis

## Table des matières

<b>Résumé :</b> .....	<b>2</b>
<b>Revue de la littérature</b> .....	<b>4</b>
A Glimpse Into Primary and Secondary Teachers' Knowledge to Teach Informatics (Parriaux, Jean-Philippe, & Vassilis , 2024) .....	4
Is it in their words? Teachers' enthusiasm and their natural language in class-A sentiment analysis approach (Frenzel, 2025) .....	4
Exploring teacher discourse patterns: Comparative insights from novice and expert teachers in junior high school EFL contexts (Tong, 2024) .....	5
Un angle encore peu traité dans la littérature actuelle .....	5
<b>Explication des données</b> .....	<b>6</b>
<b>Méthodes utilisées</b> .....	<b>7</b>
Prétraitement : .....	7
Vectorisation sémantique par embeddings .....	7
BERTopic et approche statistique .....	7
Réduction de dimension (PCA) .....	8
Analyse de sentiments et d'affects .....	8
<b>Résultats</b> .....	<b>9</b>
Analyse affective (sentiment) : .....	9
Les thématiques les plus discriminants .....	10
Visualisation de l'espace sémantique (PCA & UMAP) .....	12
<b>Conclusion</b> .....	<b>14</b>
<b>Perspectives de recherche</b> .....	<b>14</b>
<b>Bibliographie</b> .....	<b>15</b>
<b>Annexe</b> .....	<b>16</b>
Github .....	16

## **Revue de la littérature**

### **A Glimpse Into Primary and Secondary Teachers' Knowledge to Teach Informatics (Parriaux, Jean-Philippe, & Vassilis , 2024)**

Cet article propose une analyse des connaissances des enseignants pour enseigner l'informatique, à travers une auto-évaluation basée sur les sept dimensions du modèle Technological Pedagogical Content Knowledge TPACK. L'étude, menée auprès de 64 enseignants de Suisse romande issus des niveaux primaire, secondaire I, secondaire II et supérieur, met en évidence des profils différenciés selon le degré d'enseignement. Les enseignants du secondaire II déclarent des compétences solides tant sur le plan disciplinaire que didactique, tandis que ceux du primaire et du secondaire I se sentent moins à l'aise sur les contenus d'informatique mais plus compétents sur les aspects technologiques.

Ce mémoire prolonge ces travaux à travers une analyse du discours effectif des enseignants, en comparant ceux en formation (pré-service) et ceux en exercice (in-service). À l'aide de méthodes avancées de Traitement Automatique du Langage (TAL) — embeddings sémantiques, réduction de dimension, clustering et classification — il explore comment l'expérience professionnelle se reflète dans le langage utilisé. Cette approche vise à démontrer le potentiel des outils de TAL pour analyser des corpus issus du monde de l'enseignement et à explorer des pistes concrètes d'application de ces méthodes dans le domaine de la formation et du développement professionnel des enseignants.

### **Is it in their words? Teachers' enthusiasm and their natural language in class-A sentiment analysis approach (Frenzel, 2025)**

L'étude propose une approche innovante pour mesurer l'enthousiasme des enseignants à travers l'analyse de sentiment de leur discours en classe. Les auteurs montrent que les enseignants qui se déclarent enthousiastes utilisent davantage de mots à connotation positive, et que ce lien est particulièrement fort pour l'enthousiasme in-situ. Ils rappellent également que l'enthousiasme des enseignants a un effet positif documenté sur la motivation et la performance des élèves.

Plus précisément, les enseignants enthousiastes n'utilisent pas nécessairement un langage explicitement émotionnel comme begeistert ou freuen, mais manifestent plutôt leur enthousiasme par l'usage fréquent de mots modérément ou contextuellement positifs, en particulier ceux qui expriment la valeur ou la clarté, comme wichtig, klar ou super.

Ce travail s'inscrit directement dans la logique de ce mémoire, qui vise à comparer le discours d'enseignants en formation et en exercice à l'aide des méthodes du Traitement Automatique du Langage (TAL). Tandis que Frenzel et al. explorent le lien entre affect et langage dans des situations d'enseignement, mon étude examine comment l'expérience professionnelle se reflète dans le style discursif, les contenus pédagogiques et la tonalité émotionnelle.

### **Exploring teacher discourse patterns: Comparative insights from novice and expert teachers in junior high school EFL contexts (Tong, 2024)**

Tong et al. ont analysé les différences de discours entre enseignants novices et expérimentés dans des classes d'anglais langue étrangère en Chine. À partir des transcriptions de six cours, ils ont mis en évidence des structures conversationnelles dominantes, en particulier le schéma question-réponse-feedback. Leur étude montre que les enseignants expérimentés mobilisent davantage de variantes interactives susceptibles de favoriser l'engagement des élèves, tandis que les novices ont recours à des formats plus rigides et directifs. Adossée à une méthodologie mixte combinant analyse qualitative (entretiens) et analyse quantitative (tests du Chi carré). Les auteurs ont comparé la fréquence des différents types de structures discursives (comme IRF, IRFR, IR0, etc.) entre enseignants novices et expérimentés. Cette recherche met en lumière l'impact de l'expérience sur la complexité et la souplesse du discours pédagogique. Les auteurs plaident ainsi pour une formation explicite des enseignants débutants à des formes discursives plus variées et participatives.

Dans le cadre de ce mémoire, nous cherchons à retrouver de manière automatique certaines de ces distinctions discursives à l'aide d'outils issus du Traitement Automatique du Langage (TAL).

### **Un angle encore peu traité dans la littérature actuelle**

Nous nous efforçons de relier ces travaux à des méthodes issues du traitement automatique des langues, telles que la visualisation d'embeddings, le clustering ou d'autres approches classiques, mais nous n'avons pas identifié d'études traitant directement de ces questions. La majorité des publications existantes se concentrent plutôt sur l'intégration des grands modèles de langage (LLM) dans les contextes scolaires — à l'image d'Extance (2023), qui analyse comment ChatGPT peut transformer les pratiques pédagogiques. À rebours de cette orientation, notre travail cherche à explorer comment les représentations vectorielles extraites

de textes éducatifs peuvent être projetées et regroupées afin de révéler des structures sémantiques latentes, d'évaluer la cohérence de clusters pédagogiques et d'enrichir l'analyse du discours enseignant et apprenant. Notre objectif est ainsi de développer un cadre méthodologique combinant visualisation d'embeddings et techniques de clustering pour approfondir la compréhension des dynamiques discursives en contexte éducatif.

## **Explication des données**

Les données proviennent d'une recherche menée dans le cadre du modèle TPACK, visant à explorer les connaissances didactiques et disciplinaires des enseignants en informatique à travers l'analyse de leur discours (Parriaux, Jean-Philippe, & Vassilis, 2024). Il contient des transcriptions textuelles (`text`) de segments de discours extraits de séances d'enseignement, annotés selon leur position temporelle dans le déroulement du cours (`time_slice`), permettant ainsi de situer chaque intervention dans le fil de l'action pédagogique (début, milieu ou fin de la séance). Chaque segment est également associé à une information clé sur le statut professionnel de l'intervenant (`pre_or_in_service`), distinguant les enseignants en formation initiale de ceux en exercice. Cette structuration rend possible l'étude comparative de styles discursifs et de stratégies pédagogiques selon l'expérience, tout en ouvrant la voie à des analyses automatisées par traitement du langage naturel.

## Méthodes utilisées

L'étude s'appuie sur un pipeline TAL complet allant du prétraitement des données textuelles jusqu'à l'analyse interprétative. Les étapes clés de ce pipeline sont détaillées ci-dessous :

**Prétraitement** : Chaque corpus (regroupant d'une part les textes d'enseignants en formation, d'autre part ceux d'enseignants en exercice) a été nettoyé et normalisé. Concrètement, cela inclut la tokenisation (segmentation en mots), la lemmatisation (réduction des mots à leur forme de base) et l'élimination des mots vides (articles, conjonctions fréquentes, etc.). Le prétraitement garantit que les comparaisons ne soient pas biaisées par des variations superficielles (majuscules, pluriels, etc.) et que l'analyse porte sur le contenu linguistique significatif.

**Vectorisation sémantique par embeddings** : Pour comparer finement le contenu des discours, chaque phrase d'enseignant a été convertie en vecteur numérique reflétant son sens global. Nous avons pour cela utilisé un modèle d'embeddings contextualisés de la famille BERT (Bidirectional Encoder Representations from Transformers). BERT est un modèle neuronal avancé, introduit par Google, qui produit des représentations numériques de phrases en tenant compte du contexte bidirectionnel (gauche et droite) (Jacob Devlin, 2019). Contrairement aux anciennes approches (Word2Vec, GloVe), qui attribuaient à chaque mot un vecteur fixe indépendamment du contexte, BERT fournit des vecteurs dynamiques : par exemple, le mot *classe* n'aura pas la même représentation selon qu'il désigne un groupe d'élèves ou une catégorie abstraite. Nous avons utilisé un modèle BERT pré-entraîné adapté à la langue des données (multilingue) (Reimers, 2019) fin d'obtenir, pour chaque phrase, un embedding de dimension élevée (384 composantes) représentant sa signature sémantique.

**BERTopic et approche statistique** : L'algorithme BERTopic (Grootendorst, 2022) est une méthode de modélisation de sujets (topics) qui combine des représentations sémantiques issues de modèles de type transformers avec des techniques classiques de clustering. Son fonctionnement repose sur les étapes suivantes :

**Vectorisation sémantique** : chaque texte (ou segment de texte) est converti en un vecteur dense à l'aide d'un modèle multilingue basé sur BERT (ou similaire), capable de capturer les relations contextuelles entre les mots.

**Réduction de dimension** : les vecteurs sont ensuite projetés dans un espace de plus faible dimension (généralement par UMAP) afin de faciliter le regroupement.

**Clustering** : l'algorithme HDBSCAN regroupe les vecteurs proches les uns des autres en topics. Chaque cluster est supposé représenter un thème récurrent dans le corpus.

**Extraction de mots-clés** : pour chaque topic, BERTopic identifie les mots les plus représentatifs à l'aide d'une pondération TF-IDF modifiée (appelée c-TF-IDF).

Afin d'évaluer les différences de fréquence des topics entre les deux groupes d'enseignants (en formation vs en exercice), nous avons calculé un *z-score* pour chaque topic. Ce score mesure l'écart entre les proportions observées dans les deux groupes, rapporté à la variance attendue sous l'hypothèse nulle d'égalité des proportions. Plus la valeur absolue du *z* est élevée, plus la différence est susceptible d'être significative sur le plan statistique.

Pour tenir compte des comparaisons multiples et limiter le risque de faux positifs, une correction de multitest a été appliquée.

**Réduction de dimension (PCA)** : Les vecteurs produits étant de grande dimension, nous leur avons appliqué des techniques de réduction dimensionnelle pour pouvoir les visualiser et en extraire les tendances principales. Une Analyse en Composantes Principales (ACP/PCA) a permis de projeter les données dans un espace à 2 ou 3 dimensions en maximisant la variance expliquée. Concrètement, nous avons obtenu des nuages de points où chaque point représente un texte d'enseignant, et où la distance entre deux points reflète la différence de contenu discursif. Ces visualisations offrent un premier aperçu des similarités ou différences entre les deux groupes.

**Analyse de sentiments et d'affects** : Nous appliquons une analyse de sentiment aux textes afin d'identifier le ton émotionnel et affectif des discours. Pour cela, nous adoptons une méthode lexicale qui attribue à chaque texte un score global de sentiment (positif, négatif ou neutre). Cette démarche s'inscrit dans la lignée de recherches récentes qui utilisent l'analyse des sentiments pour quantifier le ton émotionnel dans le discours enseignant. Par exemple, Frenzel et al. (2025) mesurent l'enthousiasme des professeurs à partir de la polarité de leurs interventions en classe. De notre côté, nous cherchons à déterminer si les enseignants novices expriment davantage d'émotions (positives ou négatives) que leurs homologues expérimentés. Nous attribuons à chaque texte un score de sentiment moyen, puis comparons les distributions entre les deux groupes.



## Résultats

Les analyses révèlent plusieurs différences et similarités notables dans le discours des enseignants novices et expérimentés. Nous présentons ci-dessous les principaux résultats, structurés par type d'analyse, que nous illustrons et discutons.

### Analyse affective (sentiment) :

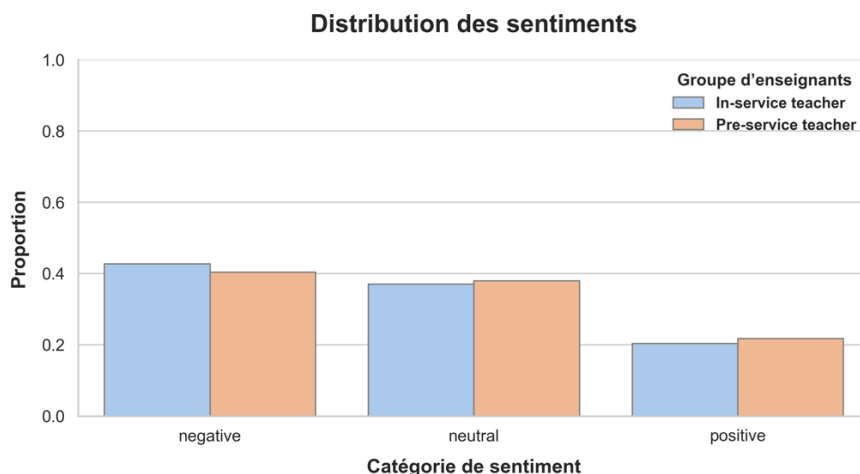


Figure 1: Distribution proportionnelle des sentiments par groupe d'enseignants

La Figure 1 présente la distribution des sentiments pour chaque groupe d'enseignants selon trois catégories : négatif, neutre et positif. Nous normalisons les proportions au sein de chaque groupe afin de faciliter la comparaison visuelle.

Dans l'étude *In Their Own Words* (Frenzel, 2025), les auteurs montrent que les enseignants les plus enthousiastes utilisent plus fréquemment des mots à connotation positive, même modérée ou contextuelle. Ils interprètent cette tendance comme une manifestation implicite de l'enthousiasme pédagogique, qui influence positivement l'apprentissage des élèves.

En nous appuyant sur ce cadre, nous explorons la proportion de sentiments positifs dans les discours des enseignants en formation (pre-service) et en exercice (in-service), à l'aide d'une analyse de sentiment multilingue.

Notre première observation : le graphique ne révèle pas de différence visuelle marquée entre les groupes. Pour vérifier si cette différence est significative, nous réalisons un test z de comparaison de proportions, avec les hypothèses suivantes :

**Hypothèse nulle ( $H_0$ ) :** La proportion de sentiments positifs est égale dans les deux groupes d'enseignants.

**Hypothèse alternative ( $H_1$ ) :** La proportion de sentiments positifs est différente entre les deux groupes d'enseignants.

Le résultat du test z est : Z-statistique =  $-1.569$  et valeur-p =  $0.1166$

Ces résultats ne permettent pas de rejeter l'hypothèse nulle à un seuil de signification de 5 %. Autrement dit, nous ne pouvons pas conclure que les enseignants en exercice utilisent significativement moins de formulations positives que ceux en formation. Cela dit, le graphique suggère tout de même que les enseignants en exercice adoptent un discours légèrement plus négatif que les enseignants novices. Cette tendance contredit partiellement les observations de Frenzel et al. (2025), qui associent l'expérience enseignante à un langage plus positif, reflet d'un enthousiasme pédagogique plus affirmé.

### **Les thématiques les plus discriminantes**

L'analyse des thématiques les plus discriminantes, que nous obtenons à l'aide de BERTopic et d'un test z de proportion, met en évidence des contrastes discursifs nets entre les enseignants en formation (pre-service) et ceux en exercice (in-service). Le z-score indique dans quelle mesure un thème est statistiquement surreprésenté dans l'un ou l'autre groupe : un z positif signale une prévalence du thème chez les enseignants en exercice, tandis qu'un z négatif reflète une occurrence plus fréquente chez les enseignants en formation.

Nous présentons dans la Figure 2 les vingt thèmes affichant les différences les plus significatives. Leur interprétation met en lumière des styles discursifs et des focalisations pédagogiques distincts selon l'expérience professionnelle.

Chez les enseignants en exercice ( $z > 0$ ), plusieurs thèmes révèlent un rapport plus institutionnalisé et structuré à l'enseignement. Le thème « Références élèves » (Topic 0,  $z \approx 26.5$ ), particulièrement saillant, renvoie à une forte présence de noms d'élèves ou de rappels de situations concrètes, traduisant une mémoire professionnelle ancrée dans les interactions pédagogiques. Le thème « Couleurs et créativité » (Topic 15,  $z \approx 5.5$ ) contient des termes comme changer ou amuser, qui suggèrent une attention portée à l'esthétique ou à l'engagement ludique. D'autres thèmes, tels que « Configuration technique » (Topic 73,  $z \approx 5.2$ ) ou « Grandeurs et tailles » (Topic 50,  $z \approx 4.9$ ), reflètent un usage plus technique ou spécialisé du langage, du langage, possiblement lié à des pratiques didactiques stabilisées. Les thèmes « Références enseignants » (Topic 72,  $z \approx 3.9$ ) et « Profils élèves » (Topic 120,  $z$

Top 20 Topics les plus discriminants (z-test)



Figure 2: Top 20 Topics les plus discriminant selon le z-test.

$\approx 3.7$ ) confirment cette orientation vers une représentation plus systématisée du métier : références à des collègues, catégorisations implicites des élèves, etc.

À l'inverse, chez les enseignants en formation ( $z < 0$ ), les thématiques dominantes reflètent un style discursif plus spontané, expressif et situé dans le vécu immédiat. Le thème « Interjections » (Topic 19,  $z \approx -4.6$ ), qui regroupe des expressions orales informelles comme ouai, wow, raaaaah, est révélateur d'une oralité marquée. Le thème « Excuses / retards » (Topic 58,  $z \approx -4.4$ ) contient des mots comme récré ou excuse, qui peuvent traduire une focalisation sur les imprévus du quotidien scolaire. Le thème « Expressions orales » (Topic 63,  $z \approx -4.2$ ), dominé par des marques d'hésitation telles que hm, ph, ahhh, illustre également la dimension spontanée et moins formalisée du discours. Enfin, des thèmes liés aux tâches

scolaires concrètes, comme « Exercices et menus » (Topic 52,  $z \approx -4.6$ ) ou « Commandes Tortue » (Topic 1,  $z \approx -3.3$ ), indiquent une centration sur la tâche ou le déroulement immédiat de l'activité.

Ainsi, les résultats obtenus mettent en lumière une tendance générale : les enseignants en exercice mobilisent davantage un discours structuré, orienté vers des objets professionnels stabilisés (élèves, collègues, concepts techniques), tandis que les enseignants en formation expriment plus fréquemment des affects, des réactions situées ou des éléments du vécu quotidien en classe.

### Visualisation de l'espace sémantique (PCA & UMAP)

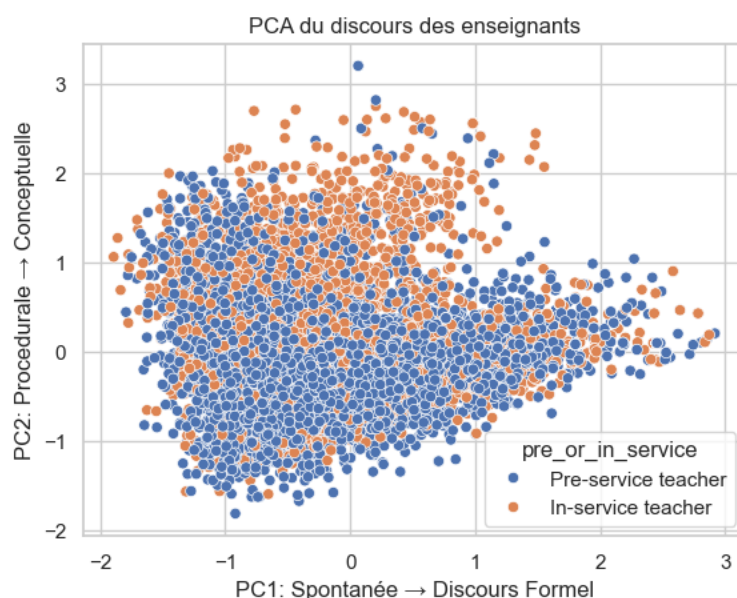


Figure 3: PCA du discours des enseignants

L'analyse en composantes principales (PCA) présentée ci-dessus projette les segments de discours des enseignants dans un espace bidimensionnel construit à partir des principales dimensions de variance sémantique du corpus. Chaque point représente un extrait de discours, coloré selon le statut professionnel de l'enseignant : les enseignants en formation sont représentés en bleu, tandis que les enseignants en exercice apparaissent en orange. La distribution spatiale révèle une tendance différenciée : les discours des enseignants en formation se concentrent davantage dans la partie inférieure gauche du graphique, tandis que ceux des enseignants en exercice se retrouvent relativement plus fréquemment dans la partie supérieure droite, bien que les deux groupes se recouvrent partiellement.

Les deux axes de la projection ont été interprétés à partir des mots les plus corrélés à chacune des composantes principales. L'axe horizontal (PC1) oppose un style discursif spontané, informel et expressif à gauche, marqué par des interjections, des hésitations et un registre oral, à un style plus formel, institutionnalisé et structuré à droite, caractérisé par un vocabulaire normatif, lié à l'encadrement et à la planification pédagogique. L'axe vertical (PC2) distingue un discours à orientation procédurale, tourné vers l'action immédiate ou l'exécution de consignes en bas, d'un discours à portée plus conceptuelle ou réflexive en haut, mettant en jeu des entités abstraites, des références à d'autres acteurs (élèves ou enseignants) et des verbes cognitifs ou évaluatifs.

La signification attribuée à ces axes repose sur une inspection manuelle des vecteurs de projection, c'est-à-dire sur l'analyse des mots les plus fortement corrélés à chacune des composantes principales issues de l'analyse en composantes principales (PCA). Cette démarche exploratoire a ensuite été confrontée aux résultats de la modélisation thématique réalisée avec BERTopic. Il apparaît que les dimensions identifiées par PCA convergent largement avec les topics les plus discriminants entre enseignants en formation et enseignants en exercice. L'axe horizontal (PC1) oppose ainsi un discours spontané, expressif et informel, illustré par des topics tels que *Interjections* (ouai, wow, raaaaah) ou *Expressions orales* (hm, ph, ahhh), à un discours plus normatif et structuré, représenté par des topics comme *Commandes Tortue* (tourner, devoir, avancer) ou *Accord / Présentation* (présentation, limite, oui). L'axe vertical (PC2) distingue quant à lui un style davantage procédural, tourné vers l'action immédiate et l'exécution de tâches (*Exercices & menus, Vitesse / Mouvement*), d'un style plus conceptuel ou réflexif, mobilisant des références à des acteurs et des structures institutionnelles (*Références élèves, Métier informatique*).

## Conclusion

Cette étude visait à comparer le discours d'enseignants en formation et en exercice à l'aide de plusieurs approches issues du Traitement Automatique du Langage. L'analyse de sentiment, inspirée de travaux antérieurs, n'a pas permis de mettre en évidence de différence statistiquement significative entre les deux groupes quant à l'usage de formulations positives. En revanche, la modélisation thématique par BERTopic a révélé des axes discursifs saillants : un langage spontané, expressif et centré sur l'expérience immédiate chez les enseignants en formation, et un discours plus normatif, structuré et réflexif chez les enseignants en exercice. Ces contrastes ont été confirmés par l'analyse en composantes principales (PCA), dont les dimensions sémantiques recoupent les grandes orientations identifiées par BERTopic. La convergence entre ces méthodes souligne la robustesse des résultats, qui s'alignent avec les tendances décrites dans la littérature : les enseignants expérimentés favorisent un discours engageant et structurant, tandis que les novices adoptent des formats plus rigides et directifs.

## Perspectives de recherche

Dans le prolongement de cette étude, il serait pertinent d'établir un lien explicite entre les profils discursifs identifiés par les méthodes de TAL et les scores TPACK des enseignants, tels qu'ils ont été mesurés dans l'étude de Parriaux, Pellet et Komis (2024). Ce travail propose une évaluation des connaissances pour enseigner l'informatique fondée sur les sept dimensions du cadre TPACK, en distinguant notamment les différences de perception selon les niveaux d'enseignement. L'intégration des scores TPACK permettrait de valider empiriquement certaines hypothèses formulées à partir des analyses textuelles : par exemple, vérifier si les enseignants exprimant un discours plus structuré ou plus réflexif présentent également des scores plus élevés dans les dimensions didactiques ou technologiques du modèle.

Une telle mise en relation ouvrirait des perspectives intéressantes pour le développement professionnel : en croisant indicateurs qualitatifs (discours) et quantitatifs (auto-évaluations TPACK), il deviendrait possible de construire des profils typologiques d'enseignants, de mieux cibler les besoins de formation, et d'enrichir les outils de diagnostic utilisés dans la formation initiale et continue. En ce sens, cette approche hybride contribuerait à renforcer la complémentarité entre les analyses issues des sciences de l'éducation et celles fondées sur des techniques computationnelles.

## Bibliographie

- Frenzel, A. e. (2025). Is it in their words? Teachers' enthusiasm and their natural language in class – A sentiment analysis approach. *British Journal of Educational Psychology*.  
*British Journal of Educational Psychology*.
- Tong, Z. e. (2024). Exploring teacher discourse patterns: Comparative insights from novice and expert teachers in junior high school EFL contexts. *Heliyon*, e36435.
- Jacob Devlin, M.-W. C. (2019, 10 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arxiv.org*.
- Sentiment analysis of preservice teachers' reflections using a large language model. (2024).  
*arxiv* .
- Extance, A. (2023). ChatGPT has entered the classroom: how LLMs could transform education. *Nature*.
- Parriaux, G., Jean-Philippe, P., & Vassilis, K. (2024). A Glimpse Into Primary and Secondary Teachers' Knowledge to Teach Informatics. *Pluhár, Z., Gaál, B. (eds) Informatics in Schools. Innovative Approaches to Computer Science Teaching and Learning.*, 125-138.
- Nils Reimers, I. G. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv*.
- Reimers, N. a. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Association for Computational Linguistics*.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794*.

## **Annexe**

### **Github**

Tout le code du projet se trouve sur GitHub sous le lien : [https://github.com/nino-gerb/nlp\\_classroom\\_observations](https://github.com/nino-gerb/nlp_classroom_observations)