Les attendus des formations sur Parcoursup sont-ils socialement discriminants ?

Gabriel AMMOUR

Étudiant en $2^{ème}$ année d'Économétrie Appliquée, Nantes Université

15 mai 2025

Résumé

Ce travail s'intéresse aux attendus publiés sur la plateforme Parcoursup et à la manière dont leur formulation peut, parfois, refléter des attentes implicites susceptibles de désavantager certains lycéen·es. En mobilisant des outils de traitement automatique du langage (NLP), nous avons analysé l'ensemble des attendus associés aux formations en licence, selon plusieurs dimensions : lexicale, thématique, syntaxique et affective. Grâce à une méthode de classification non supervisée, les formations ont été regroupées en deux grands ensembles : d'un côté les Sciences Naturelles et Technologiques (SNT), de l'autre les Sciences Humaines et Sociales (SHS). Nous observons que les attendus des SNT utilisent un langage plus normatif et centré sur la maîtrise de compétences académiques, avec des termes comme *rigueur* ou *méthode*. À l'inverse, les attendus des SHS mettent davantage en avant des capacités comme *argumenter*, *raisonner* ou *travailler en autonomie*, compétences potentiellement plus difficiles à mobiliser pour des élèves ayant un capital culturel, linguistique ou cognitif moins élevé. Bien que le ton général des attendus reste globalement positif, le type de compétences valorisées posent question quant au rôle que peut jouer la formulation des attendus dans la reproduction des inégalités scolaires.

Mots clés : Inégalités sociales, Education, Microéconométrie appliquée, NLP

1 Introduction

Il y a 40 ans, le gouvernement français, par l'intermédiaire de son ministre de l'éducation national Jean-Pierre Chevènement, fixait comme objectif de mener 80 % d'une génération au niveau baccalauréat. Aujourd'hui, cet objectif est largement dépassé puisqu'en 2024, le taux de réussite au bac s'élevait à 90,9 %. Cette hausse du nombre de bachelier·ère s'est méchaniquement traduit par une hausse du nombre d'étudiants inscrits dans le supérieur. Entre 1990 et 2018, le nombre d'étudiants inscrits dans l'enseignement supérieur est passé 1 717 000 à 2 678 700 [1]. Cette démocratisation, quoique incontestable, a largement été nuancée par la recherche académique. En effet, celle-ci concernerait d'avantages les cycles courts comme les licences plutôt que les master ou les doctorats (Albouy et Tavan, 2007). Du reste, le budget de l'enseignement supérieur et de la recherche par étudiant chute continuellement depuis 2010 [2], et, de nombreuses inégalités d'accès au supérieur subsistent encore (Van Zanten, 2015).

L'accès aux études supérieures à au cours de son histoire pris plusieurs formes. Conformément à la dématérialisation des usages et des procédures administratives, l'accès aux études supérieures se dématérialise en 2009 avec la création de plateforme Admission Post-Bac (APB). Souhaitant mettre un terme au système de tirage au sort de la plateforme, le ministère de l'éducation national décide de clore la plateforme prématurément en 2017. Malgré de vifs débats et polémiques à son sujet, c'est la plateforme en ligne Parcoursup qui va venir remplacer APB et s'imposer en 2018 comme le lieu de référence servant à recueillir et traiter les vœux d'affectation des futurs étudiants de l'enseignement supérieur. Pour émettre leurs voeux dans les meilleurs conditions possibles, les lycéen·es ont la possiblité de consulter les attendus de chaque licence universitaire sur le site de l'ONISEP. Connaître les attendus d'une licence peut permettre à un·e lycéen·e de mieux cibler

1. INSEE, Tableau de bord de l'économie française, 2020
2. Lucas Chancel, La chute du budget par étudiant en France, 2008-2021

son choix d'orientation en fonction de ses qualités et/ou capacités personnelles [3]. Si l'étudiant·e dispose d'autres moyens pour affiner son choix d'orientation (enseignants, proches, forum, journées lycéennnes, etc.), les attendus de formations représentent une ressources commune à la disposition de tous·tes lycéen·es doté d'une connexion internet [4].

Cependant, nous savons que les logiques des plateformes tendent à maintenir voir aggraver les inégalités sociales (Lemêtre et Orange, 2017). Le système Parcoursup, en formulant des attendus spécifiques pour chaque formation, peut renforcer les inégalités. En effet, les élèves disposant d'un capital culturel élevé sont plus à même de comprendre et de répondre à ces attendus, tandis que les autres peuvent se sentir exclus ou mal préparés. De plus, le manque de capital culturel peut également engendrer un sentiment d'illégitimité chez les élèves des classes populaires, les conduisant à douter de leurs capacités et à se sentir intimidés par les attendus implicites des formations sélectives. Ce manque de confiance en soi peut freiner leurs ambitions et les dissuader de postuler à certaines filières. À ce titre, le concept d'autocensure scolaire souligne le fait que certains stéréotypes sociaux intégrés par les élèves de catégories sociales plus populaires auraient tendance à créer une asymétrie d'information et/ou un manque d'ambition quant aux perspectives scolaires (Huillery et Guyon, 2014).

Dans ce contexte, nous pouvons légitimement nous demander si il n'y aurait pas, dès la phase des vœux d'orientation, des inégalités quant aux informations fournies par les attendus des formations. Plus précisément, les attendus des formations sur Parcoursup sont-ils socialement discriminants ? Pour fournir quelques éléments de réponses, nous ferons usage du *Natural Language Processing* (NLP). Le NLP regroupe l'ensemble des méthodes issues de l'intelligence artificielle et de la linguistique informatique, qui permettent aux machines de traiter, comprendre et analyser automatiquement des données textuelles. Dans le cadre de cette étude, le NLP constitue un outil pertinent pour conduire une analyse textuelle des attendus, en identifiant des similitudes entre formations, des thèmes dominants ou encore des variations de ton ou de vocabulaire susceptibles de refléter certaines normes ou certains attendus spécifiques.

2 Données

Comme mentionnée plus tôt, les attendus de formations sont disponibles sur le site de l'ONISEP. Une façon de récolter les données aurait été de faire du web scraping, mais cela aurait été long et fastidieux. Toutefois, le ministère de l'éductation nationale a mis à disposition un fichier `pdf` de 57 pages se nommant *Elements de cadrage national des attendus pour les mentions de licence* qui liste tous les attendus des Licences ouvertes aux lycéen·es. Même si le document a été publié en 2020, il y a de grandes chances que les attendues soient les mêmes qu'aujourd'hui. Ce choix de travailler sur les formations universitaires (Licence) est motivé d'une part, comme nous venons de le voir, par la disponibilité des données, et, d'une autre part, par le fait qu'en 2018, 61% des élèves inscrits dans l'enseignement supérieur l'étaient à l'université.

Le document brut a tout d'abord été importé puis lu via la libraire python `PyMuPDF`. Une fois importé, il a fallu lui apporter quelques modifications pour le rendre exploitable. Dans un premiere temps, nous avons procéder à l'élimination des mots vides ou *stopwords*. Ces mots vides sont des mots sans significations ou réels intérêts pour l'analyse textuelle et la compréhension d'un texte. En français, on va penser aux articles définis comme *le*, *la* ou bien aux articles indéfinis comme *un*, *une*, ou encore *des*. Il est important de les retirer pour ne pas fausser l'analyse avec des mots qui n'apportent pas d'informations. Ensuite, dans un second temps, nous avons procéder à une lemmatisation. La lemmatisation est une technique utilisée dans le traitement automatique du langage naturel qui permet de conserver l'unité lexicale d'un mot. Par exemple, si le texte est composé des mots suivants : *mange*, *manges*, *mangea*, *manger*, la lemmatisation les ramènera tous à leur forme de base, qui est *manger*. Toutes les modifications ont été réalisés à l'aide de la librairie python `spacy` et du modèle `fr_core_news_sm`.

Enfin, une fois les textes nettoyés et lemmatisés, il a été nécessaire de les vectoriser pour les rendre exploitables. Pour ce faire, nous avons utilisé la méthode *TF-IDF* (*Term Frequency – Inverse Document Frequency*), qui pondère chaque terme en fonction de sa fréquence dans un document. La

3. Du moins celles que l'on aura identifiées lors de son parcours scolaire.
4. On notera à ce titre que le simple fait de devoir disposer d'une connexion internet peut être un frein en soi.

vectorisation a été réalisée à l'aide de la classe `TfidfVectorizer` issue de la librairie `scikit-learn`.

3 Méthodologie

L'essor récent des méthodes de traitement automatique du langage naturel a profondément renouvelé les pratiques de recherche en sciences sociales. Ces outils sont désormais utilisés aussi bien en sociologie qu'en économie pour explorer les biais, normes et représentations véhiculés par les textes. Le NLP offre un cadre méthodologique pertinent pour interroger la manière dont les attendus Parcoursup peuvent, par leur forme ou leur contenu, refléter des exigences implicites ou des biais sociaux susceptibles de désavantager certains publics. Combinée à des méthodes de classification et modélisation statistique, le NLP constitue un levier important pour mettre en évidence d'éventuelles inégalités dès la phase d'orientation des lycéen·e·s.

3.1 Clusters et classification

Afin d'explorer la diversité des formulations présentes dans les attendus Parcoursup, une méthode de classification non supervisée a été utilisée. L'objectif est de regrouper les formations selon des similarités linguistiques observées dans les attendus (sans présupposer des catégories prédéfinies). Ce type d'approche permet de faire émerger des regroupements thématiques qui peuvent refléter des différences dans les exigences formulées selon les types de filières. En d'autres termes, nous cherchons à obtenir des groupes aux traits lexicaux communs.

La méthode de clustering retenue repose sur l'algorithme des *k-means*, une technique largement utilisée en apprentissage non supervisé. Cet algorithme attribue chaque document à l'un des k groupes en minimisant la distance entre le document vectorisé (selon son contenu lexical) et le centre du cluster. Le choix du nombre optimal de clusters k est un paramètre central de l'analyse. Il a été déterminé à l'aide de la méthode du coude (*elbow method*), qui consiste à comparer la variance intra-cluster pour différents niveaux de k et à identifier le point au-delà duquel l'ajout de nouveaux clusters n'apporte plus de réduction significative de variance.

Les différents clusters servent de socle à l'ensemble des analyses menées dans la suite du travail. Ils permettent d'organiser les attendus selon des thématiques communes, ce qui facilite la mise en évidence de régularités ou, au contraire, de particularités dans la formulation des exigences. Surtout, cette classification permet de simplifier l'analyse : plutôt que de comparer individuellement les 45 mentions de Licence, nous travaillons à partir de quelques groupes seulement. À partir de ces clusters, nous pourrons ainsi étudier plus finement les différences lexicales, les tonalités employées, ainsi que les éventuelles normes implicites utilisées dans les différentes formations.

3.2 L'analyse textuelle

Une fois les clusters identifiés, nous avons cherché à comprendre ce qui caractérise chacun d'eux. Pour cela, nous avons utilisé différentes méthodes d'analyse textuelle visant à identifier les mots les plus représentatifs ou spécifiques à chaque groupe. En effet, si certains mots apparaissent beaucoup plus fréquemment dans un cluster que dans les autres, ils peuvent nous aider à mieux comprendre les thématiques ou les attentes dominantes dans ce groupe.

Deux approches ont été utilisées, la première repose sur le score de spécificité, qui compare la fréquence d'un mot dans un cluster donné à sa fréquence dans le reste du corpus. Cette méthode met en évidence les termes véritablement discriminants, c'est-à-dire ceux dont l'usage est quasi exclusif à un groupe. La seconde approche consiste à extraire, pour chaque cluster, les mots les plus représentatifs, aussi appelés *topwords*. Cette méthode s'appuie sur les scores moyens de pondération *TF-IDF*, qui mesurent l'importance d'un mot en fonction de sa fréquence dans un document et de sa rareté à l'échelle du corpus. Les mots les mieux classés permettent de dégager les registres de langage dominants dans chaque groupe et d'en identifier plus facilement les thématiques. Ces methodes nous offrent ainsi des indices sur ce qui est valorisé - ou au contraire peu mis en avant - dans certains parcours, ce qui peut contribuer à renforcer des inégalités d'interprétation ou d'accès pour les lycéen·nes.

En complément, une analyse de sentiment a été menée. Elle permet de mesurer la tonalité du langage utilisé dans les attendus. On cherche à savoir si certains textes sont plus accueillants ou si d'autres sont plus formels ou plus exigeants. Dans le cadre de notre question de recherche, cela peut être révélateur de différences dans la manière dont les formations s'adressent aux lycéen·nes. Si certains groupes expriment des attentes de façon plus froide ou abstraite, cela peut décourager certains profils, notamment ceux qui ont moins confiance en eux ou qui ne possèdent pas les bons codes. Pour réaliser cette analyse de sentiment, nous avons utilisé un modèle de *nlptown*[5], accessible via la librairie `transformers`. Ce modèle, entraîné sur un grand nombre de textes en plusieurs langues, permet d'attribuer une note de sentiment allant de 1 à 5 à chaque attendu, 1 correspondant à un ton très négatif et 5 à un ton très positif. Ces scores sont ensuite comparés entre clusters afin d'évaluer s'il existe des différences dans le ton employé d'un groupe à l'autre.

3.3 Régression logistique

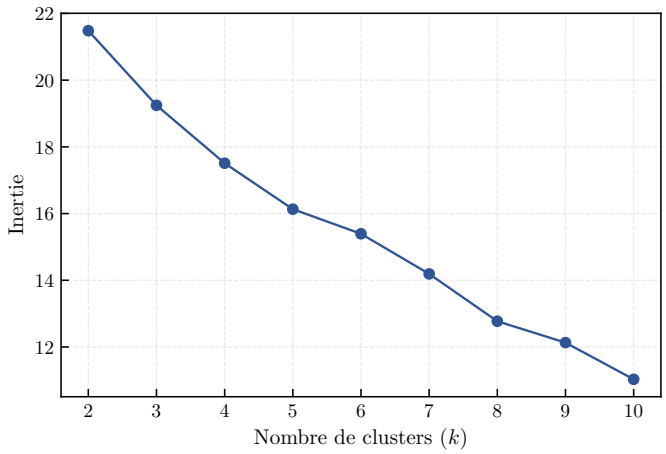
Enfin, une régression logistique a été mise en place pour essayer de prédire, à partir du contenu textuel d'un attendu, à quel cluster il appartient. Dans notre cas, les variables explicatives sont les mots contenus dans les attendus, qui ont été vectorisés à l'aide de la méthode *TF-IDF*. L'objectif est de voir si le langage utilisé dans un attendu permet d'identifier de manière fiable le cluster auquel il est associé. Si c'est le cas, cela signifie que les groupes formés sont bien différenciés sur le plan lexical, ce qui renforce la validité de notre classification.

4 Résultats

4.1 Choix du nombre de cluster

Choisir le nombre de clusters est une étape importante de la classification non supervisée. La Figure 1 nous montre que la diminution de l'inertie n'est pas nette entre les différents cluster, ne laissant pas l'impression d'un coude. Dès lors, nous conservons 2 clusters pour le reste de l'analyse, au-delà, l'inertie ne baisse pas significativement.

FIGURE 1 – Évolution de l'inertie en fonction du nombre de clusters k pour l'algorithme k-means



Nous avons donc choisi de nous concentrer sur deux clusters, qui, à eux deux, regroupent toutes les mentions présentes (Table I). Le premier cluster, que l'on nommera Sciences Naturelles et Technologiques (SNT), regroupe 15 mentions. Le second, les Sciences Humaines et Sociales (SHS), regroupe 29 mentions. Ce déséquilibre entre le nombre de mentions consitue une limite dont on reparlera plus tard. Ces deux appellations permettent de donner une première lecture des regroupements observés, mais ne sont pas parfaites. En effet, certaines mentions, notamment artistiques ou interdisciplinaires, s'intègrent difficilement dans des catégories strictes.

5. `bert-base-multilingual-uncased-sentiment`

TABLE I – Licences par clusters

Cluster 1 : Sciences Naturelles & Technologiques	Cluster 2 : Sciences Humaines & Sociales
— **Mathématiques et Sciences Fondamentales :** Mathématiques, Physique, Chimie — **Sciences de la Vie et de la Terre :** Sciences de la Vie, Sciences de la Terre — **Sciences pour l'Ingénieur :** Électronique, Énergie Électrique, Automatique, Mécanique, Génie Civil — **Informatique et Applications :** Informatique, Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales	— **Sciences Juridiques et Politiques :** Droit, Administration Publique, Science Politique — **Sciences Économiques :** Économie, Gestion, Administration Économique et Sociale — **Sciences Humaines :** Histoire, Géographie et Aménagement, Philosophie, Sciences de l'Éducation, Histoire de l'Art et Archéologie — **Sciences Sociales :** Sociologie, Anthropologie, Ethnologie, Psychologie, Sciences Sanitaires et Sociales — **Lettres et Arts :** Langues, Littératures et Civilisations Étrangères et Régionales, Sciences du Langage, Lettres, Arts Plastiques, Arts du Spectacle, Musicologie — **Autres :** Information Communication, Sciences Techniques des Activités Physiques et Sportives

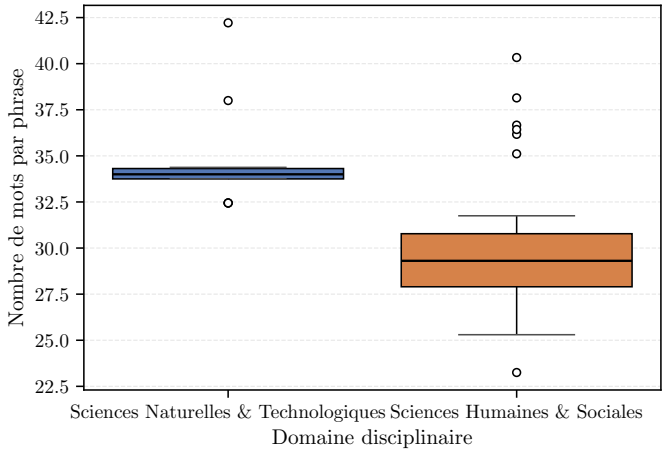
La clusterisation nous aura ainsi permis d'établir une dichotomie entre deux grands domaines disciplinaires. Cette distinction rappelle en partie la séparation classique entre les sciences dites *dures* (mathématiques, physique, biologie, etc.) et les sciences dites *molles* (sciences sociales, lettres, arts, etc.), une opposition ancienne et familière qui continue de structurer certaines représentations sociales.

4.2 Analyse textuelle

4.2.1 Complexité syntaxique

Avant d'entrer dans l'analyse du contenu lexical, il est intéressant d'observer certaines caractéristiques plus formelles des textes. La Figure 2 présente la distribution du nombre moyen de mots par phrase dans les attendus, selon les deux clusters.

FIGURE 2 – Complexité syntaxique (mots par phrase) par cluster

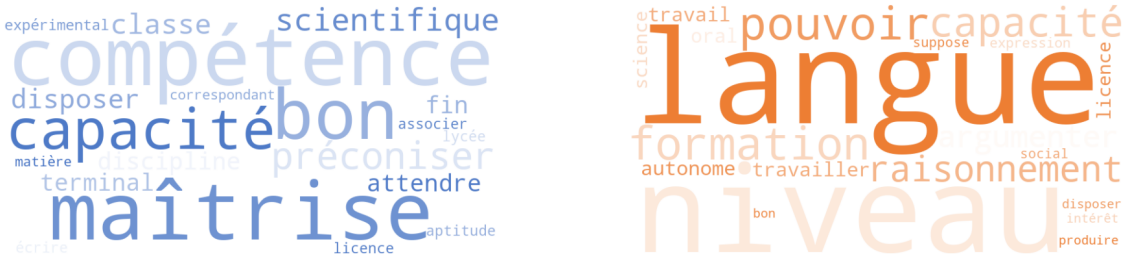


On observe que les attendus des formations en SNT présentent en moyenne des phrases légèrement plus longues, avec environ 34 mots par phrase, contre 29 mots en moyenne pour les SHS, soit un écart d'environ 5 mots. Cette différence laisse suggérer une structure syntaxique un peu plus dense dans les formations scientifiques. La distribution des longueurs de phrases est également plus homogène pour les formations SNT. L'écart-type est faible, et la majorité des attendus se situent autour de la moyenne. À l'inverse, la distribution est plus étalée pour les formations SHS, avec des phrases pouvant descendre à 25 mots ou monter jusqu'à 32 mots en moyenne selon les mentions. Du côté des valeurs atypiques, les attendus les plus longs sont ceux des Sciences de l'ingénieur, avec une moyenne de 42 mots par phrase, tandis que ceux des Lettres et Langues dans le groupe SHS atteignent en moyenne 40 mots. La mention STAPS présente des phrases significativement plus courtes, avec 23 mots en moyenne. Ces résultats semblent cohérents avec la nature même des formations.

4.2.2 Topwords

La Figure 3 présente les nuages de mots générés pour chacun des deux clusters. Ces nuages donnent une représentation visuelle des mots les plus fréquents et les plus caractéristiques de chaque groupe, avec une taille proportionnelle à leur importance.

FIGURE 3 – Wordcloud des mots les plus fréquents par cluster



(a) Sciences Naturelles & Technologiques

(b) Sciences Humaines & Sociales

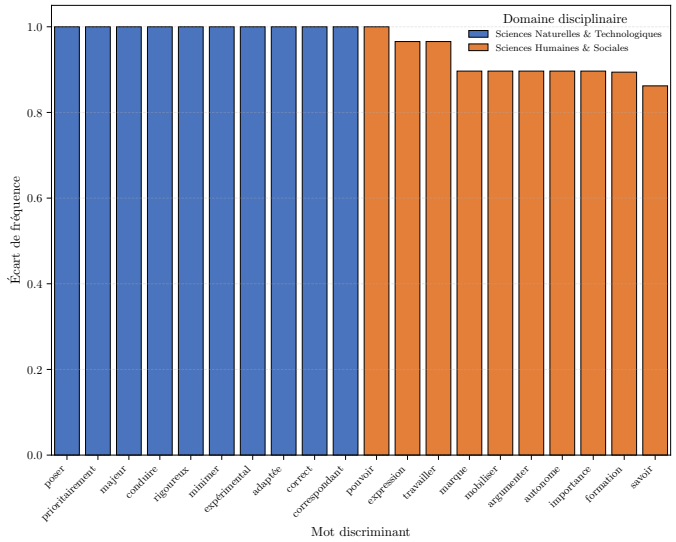
Le cluster 1, celui des SNT, est marqué par la présence de mots liés aux compétences scolaires et académiques attendues. On y retrouve des termes tels que *compétence*, *maîtrise*, *capacité* ou encore *discipline*, qui renvoient à une logique d'évaluation des savoirs et savoir-faire techniques. Le mot *scientifique* occupe également une place importante. On observe également des termes comme *préconiser* ou *attendre*, qui introduisent une certaine normativité dans le discours, en indiquant ce qu'il est recommandé ou requis de posséder. Enfin, des mots comme *bon* ou *classe* peuvent faire référence au niveau académique et au parcours scolaire antérieur, notamment en terminale. Dans l'ensemble, le vocabulaire de ce cluster semble plus être centré sur des connaissances scolaires formelles, avec un lexique plutôt orienté vers l'exigence et l'évaluation.

Le cluster des SHS se distingue par un vocabulaire plus orienté vers les capacités cognitives. On y retrouve des termes comme *raisonnement*, *argumenter*, *autonome*, ou encore *travailler*, qui renvoient davantage à des compétences transversales (*soft skills*) ou intellectuelles, souvent mobilisées dans les disciplines des SHS. Le mot *langue* apparait également de façon notable renvoyant probablement aux compétences linguistiques en lien avec les filières littéraires et de communication. On note la présence du verbe *pouvoir*, qui introduit une certaine ouverture ou flexibilité dans la formulation, contrastant avec le ton plus normatif du cluster précédent. De même, *formation* et *travail* sont des termes plus généraux, qui orientent le discours autour du parcours de l'étudiant plutôt que des critères de sélection stricts. Nous noterons toutefois la présence du mot *niveau*, suggérant une attention portée au niveau académique (ou peut-être dans une langue) général requis à l'entrée dans la formation.

4.2.3 Mots discriminants par clusters

La Figure 4 présente les 10 mots les plus discriminants pour chacun des deux clusters, c'est-à-dire ceux dont la fréquence d'apparition varie fortement d'un groupe à l'autre. Plus précisément, l'axe vertical représente l'écart de fréquence du mot entre un cluster et l'ensemble du corpus. Si l'écart de fréquence est égal à 1, cela signifie que le mot est exclusivement utilisé dans le cluster, et donc fortement spécifique à celui-ci.

FIGURE 4 – Mots les plus discriminants par cluster



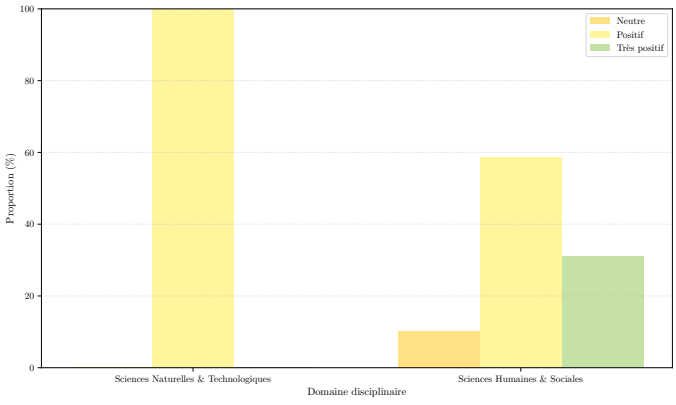
Du côté des SNT, on retrouve des termes tels que *poser*, *majeur*, *conduire*, *rigoureux* ou encore *expérimental*. Ces mots, uniquement présents dans ce cluster, évoquent des notions de rigueur, de méthode. D'autres mots comme *adaptée* ou *correspondant* peuvent renvoyer à l'idée d'une adéquation ou d'un niveau requis. Ces observations sont cohérentes avec l'analyse des *topwords* du cluster 1, où dominaient des termes comme *compétence*, *maîtrise* ou *scientifique*, qui renforcent l'idée d'un registre assez normatif, centré sur l'acquisition de savoirs techniques et évaluables.

Dans le cluster associé aux SHS, les mots discriminants sont d'un tout autre registre lexical. On y retrouve des verbes ou noms d'action comme *pouvoir*, *travailler*, *mobiliser*, *argumenter* ou encore *expression*, qui peuvent traduisent une certaine valorisation de la capacité à mettre en œuvre ses connaissances dans un raisonnement ou une démarche plus intellectuelle. La présence du mot *autonome* renvoie à une attente en termes d'engagement personnel et de responsabilité dans le travail universitaire. Il ne s'agit pas simplement de maîtriser des connaissances, mais d'être capable de s'organiser, de réfléchir par soi-même et de construire un parcours universitaire et intellectuelle sans encadrement constant. Le mot *savoir* est également significatif. Il peut suggérer une orientation vers des connaissances moins strictement codifiées, peut-être parfois implicites, dont la maîtrise peut dépendre fortement des connaissances personnelles de l'élèves. Ces éléments prolongent l'analyse du *topwords* du cluster 2, où figuraient déjà *langue*, *raisonnement* ou encore *formation*, et confirment l'existence d'un registre plus ouvert, moins centré sur des critères scolaires formels.

4.2.4 Analyse des sentiments

La Figure 5 présente la répartition des niveaux de sentiment observés dans les attendus, pour chacun des deux clusters.

FIGURE 5 – Répartition des sentiments par cluster



On observe d'abord que les attendus associés aux formations en SNT présentent une distribution

très homogène. En effet, tous les textes sont classés comme ayant un ton positif, sans variation. Le score moyen dans ce cluster est de 4, ce qui correspond à un ton positif. À l'inverse, les attendus rattachés aux formations de SHS montrent une plus grande diversité de tonalités. Si une majorité de textes restent positifs, environ un tiers sont classés comme très positifs, et une petite proportion comme neutres. Le score moyen pour ce cluster est légèrement plus élevé, à 4,2. Malgré ces légers écarts, l'ensemble des attendus analysés sont globalement associés à un ton positif.

4.2.5 Analyse thématique

Afin de compléter l'analyse lexicale, une modélisation thématique a été réalisée pour identifier les grands sujets abordés dans les attendus, à l'intérieur de chaque cluster. Pour cela, nous avons utilisé une méthode de type *topic modeling* (modélisation de sujets), qui permet de regrouper les mots les plus fréquemment associés dans les textes, et ainsi de dégager automatiquement des thématiques récurrentes.

Table II – Thèmes émergents par cluster

Cluster / Thème	**Lexique caractéristique**
Cluster 1 - Sciences Naturelles & Technologiques	
Thème 1 : Compétences scientifiques	maîtrise, compétence, bon, capacité, préconiser, scientifique, classe, disposer, discipline, terminal
Thème 2 : Parcours et orientation	compétence, capacité, pace, filière, visent, bon, disposer, maîtrise, connaissance, national
Cluster 2 - Sciences Humaines & Sociales	
Thème 1 : Capacités cognitives et autonomie	langue, niveau, pouvoir, capacité, formation, raisonnement, argumenter, travailler, autonome, oral
Thème 2 : Contenu disciplinaire	économie, gestion, filière, mathématique, capable, compétence, sociétal, étudiant, licence, structurer

Note. Les termes sont présentés par ordre décroissant de significativité statistique au sein de chaque thème identifié.

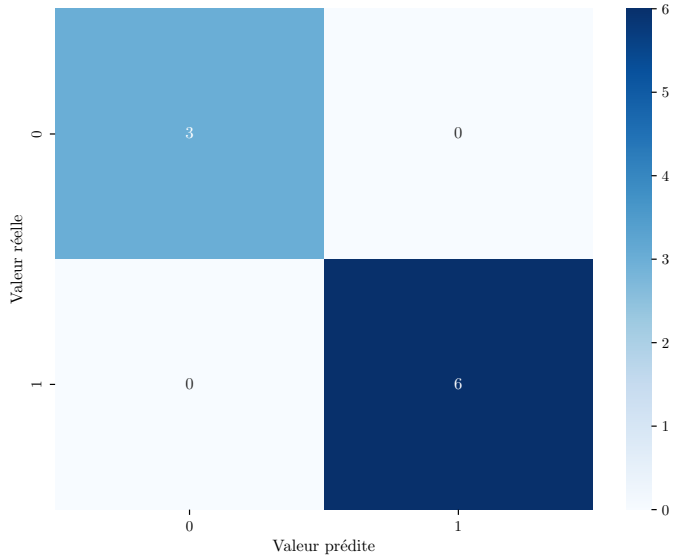
Dans le premier cluster, deux grands thèmes se dégagent. Le premier est centré sur des notions de maîtrise des compétences scientifiques, avec des mots comme *maîtrise*, *compétence*, *scientifique*, *discipline*, ou encore *terminal*, qui renvoient peut-être au niveau attendu en fin de Lycée. On retrouve également *préconiser* et *disposer*, qui évoquent des formulations normatives sur les acquis à posséder. Le second thème du cluster 1 se concentre davantage sur la structure des parcours et l'orientation, avec des termes tels que *filière*, *PACES* (première année commune aux études de santé), *visent*, *connaissance* et *national*.

Dans le second cluster, le premier thème semble identifier un ensemble de compétences cognitives, avec des mots comme *langue*, *niveau*, *pouvoir*, *argumenter*, *oral*, *travailler* ou encore *autonome*. Ce groupe de termes met en évidence une attente tournée vers la capacité à s'exprimer, raisonner et travailler en autonomie dans un cadre universitaire. Le second thème du cluster 2 porte sur le contenu disciplinaire voir interdisciplinaire, avec les mots *économie*, *gestion*, *mathématique*, *sociétal*, *licence* ou *structurer*. Ce thème suggère une certaine ouverture aux enjeux de société, comme le montre la présence du mot *sociétal*.

4.3 Régression logistique

La Figure 6 présente la matrice de confusion obtenue à l'issue de la régression logistique menée pour prédire l'appartenance des attendus à l'un des deux clusters identifiés. Le modèle parvient à classer correctement l'ensemble des observations du jeu de test, les prédictions sont parfaitement alignées avec les valeurs réelles.

Figure 6 – Matrice de confusion



Les scores de performance sont également parfaits, avec une accuracy et un F1-score égaux à 1. Ces résultats indiquent que, dans notre échantillon, le contenu lexical des attendus permet de distinguer de manière parfaite les deux groupes disciplinaires. Cependant, il est important de souligner que cette performance doit être interprétée avec prudence. Elle s'explique en grande partie par la structure très marquée des données (dichotomie forte entre les clusters) et par le faible volume de textes disponibles pour l'entraînement du modèle.

5 Discussion

Les résultats de cette étude offrent plusieurs pistes de réflexion sur la façon dont les attendus Parcoursup peuvent contenir des attentes implicites, parfois discriminantes. L'analyse textuelle montre des différences claires entre les deux grands groupes de formations identifiés. Ces différences ne relèvent pas seulement du style, elles traduisent aussi des visions différentes des qualités d'un·e futur·e étudiant·e. La première différence observée concerne le registre de langage utilisé dans chaque cluster. Du côté des SNT, les textes sont marqués par un vocabulaire plus normatif et scolaire. On y retrouve des notions de maîtrise, de compétences ou encore de rigueur qui y sont centrales. À l'inverse, les attendus des SHS mettent davantage l'accent sur des capacités cognitives comme le fait d'argumenter, de raisonner ou encore de mobiliser des connaissances déjà acquises. Ces compétences - par exemple, l'aisance à l'oral, ou la capacité à argumenter - sont probablement plus facilement mobilisées par des individus issus de milieux socialement favorisés, en raison de leur socialisation scolaire et familiale. Dès lors, les résultats obtenus dans le cluster des SHS, où ces compétences sont fortement mises en avant, peuvent suggérer une forme d'exigence plus difficile à décrypter ou à atteindre pour des publics moins dotés en capital culturel. L'analyse de la complexité syntaxique ne vient pas forcément renforcer cette lecture. Si les formations des SHS présentent une plus grande hétérogénéité dans la longueur des phrases, et certaines mentions, comme les lettres ou les langues, se distinguent par des formulations longues et denses, elles restent légèrement inférieurs à celles des SNT. Cependant, cette analyse doit être nuancée par le déséquilibre d'effectifs et de diversité disciplinaire entre les deux clusters.

Sur le plan affectif, l'analyse des sentiments montre des attendus globalement positifs dans l'ensemble des formations, ce qui peut être vu comme un signe d'accueil et de bienveillance. Cependant, un tier des mentions SHS affichent un ton très positif, tandis que les SNT restent dans un registre plus constant, mais modérément positif. Si cette variation est faible, elle pourrait toutefois renforcer des écarts de perception pour certains lycéen·es, notamment celles et ceux qui doutent de leur légitimité à postuler dans certaines filières plus techniques ou plus prestigieuses. Toutefois, cette analyse dépends fortement de la methode utilisé pour répertorier les sentiments. En effet, ici la methode semble assez limité puisque la distinction entre les différentes connotations reste très floue. Par exemple, le mot *économie* a une connotation très positive lorsque le mot *socioéconomique* a une connotation neutre. En ce qui concerne les résultats de la régression logistique, ils confirment la

clarté de la séparation entre les deux clusters. Cette forte dichotomie, observable dès les premières étapes de l'analyse, reflète une différenciation très nette entre les deux groupes. Toutefois, cela peut aussi être une conséquence directe de la manière dont les formations sont rédigées et catégorisées administrativement.

Plusieurs limites doivent être soulignées dans ce travail. D'abord, l'importance des attendus dans le processus d'orientation reste incertain. En effet, il semblerait que l'attendu d'une formation ne soit pas la variable d'ajustement avec le plus d'intérêt lorsque l'on s'intérèsse à la phase d'orientation. De plus, il semblerait que pour la grande majorité des lycéen·es, les attendus ne soient pas une source d'inquiétude (Bretton-Wilk et Haute, 2021). Par ailleurs, il peut même être difficile de croire que les attendus face l'objet d'une réel lecture critique de la part des élèves. Ensuite, le choix du nombre de clusters et la taille très restreinte de l'échantillon doivent être pris en compte. La séparation binaire des clusters simplifie l'analyse mais peut masquer des nuances propres à certains groupes de formations. De plus, comme mentionné plus haut, le déséquilibre entre le nombre de mentions par cluster - avec deux fois plus de formations en SHS qu'en SNT - peut avoir influencé certaines analyses, notamment celle des sentiments. Enfin, le fait de travailler sur un corpus français ne facilite pas l'analyse textuelle puisqu'un certain nombre de méthodes sont bien plus adaptées aux dictionnaires anglophones.

6 Conclusion

Ce travail avait pour objectif d'examiner dans quelle mesure les attendus publiés sur la plateforme Parcoursup pouvaient refléter des attentes implicites potentiellement discriminantes. À travers une combinaison d'analyses lexicale, syntaxique, thématique et affective, nous avons mis en lumière une dichotomie nette des attendus selon les disciplines, plus précisément, un différence marquée entre formations identifiées comme appartenant aux Sciences Naturelles et Technologiques (SNT) et celles relevant aux Sciences Humaines et Sociales (SHS). Dans l'ensemble, les attendus adoptent un ton majoritairement positif, suggérant peut-être une volonté de rendre les formations accueillantes pour les lycéen·nes. Toutefois, au-delà de cette tonalité générale, des différences apparaissent dans le langage mobilisé. Tandis que les formations appartenant aux SNT privilégient un vocabulaire normatif comme la *rigueur* ou la *méthode*, les formations associées aux SHS font davantage appel à des compétences intellectuelles telles que l'argumentation, la mobilisation de savoirs ou le raisonnement. Or, plusieurs travaux ont montré que ces compétences sont souvent plus aisément mobilisées par des élèves issus de milieux favorisés disposant d'un capital culturel, linguistique ou cognitif plus important en raison de leur socialisation et de leur familiarité avec les codes de l'école (Draelants et Ballatore, 2014). En ce sens, les attendus peuvent participer à entretenir des inégalités sociales d'accès ou de compréhension, en valorisant certaines formes de savoir et d'expression plus accessibles à certaines classes.

Ces résultats s'inscrivent dans une réflexion plus générale sur le rôle de l'école comme instance de reproduction des inégalités sociales, notamment à travers la valorisation d'une culture légitime dont la maîtrise varie selon les classes sociales. Ils invitent à interroger plus systématiquement la formulation des attendus dans les documents d'orientation institutionnelle (pas seulement à l'école) et à réfléchir à leur accessibilité réelle pour l'ensemble des publics visés. Plusieurs pistes pourraient être explorées dans la suite de ce travail. Il serait tout d'abord pertinent de comparer les attendus des formations universitaires avec ceux des classes préparatoires ou des grandes écoles, pour analyser si les différences de niveau d'exigence s'accompagnent également de différences dans la formulation des attendus. Par ailleurs, une approche plus fine pourrait s'intéresser aux dimensions de genre. En effet, il semblerait que les jeunes filles ont souvent une confiance en elles moindre que les garçons à niveau égal, ce qui pourrait influencer leur perception des attendus, en particulier dans les formations perçues comme exigeantes ou codées. Enfin, un approfondissement de la méthodologie - avec un corpus plus étendu et des outils mieux adaptés au français - permettrait peut-être d'affiner les résultats. En conclusion, cette étude montre que le langage utilisé dans les documents d'orientation n'est pas neutre. Il peut véhiculer des représentations, des normes et des attentes qui méritent d'être examinées de manière critique, pour garantir une meilleure égalité d'accès à l'enseignement supérieur.

Références

Albouy, V., & Tavan, C. (2007). Accès à l'enseignement supérieur en France : une démocratisation réelle mais de faible ampleur. *Économie et statistique*, *410*(1), 3-22.

Van Zanten, A. (2015). 5. Les inégalités d'accès à l'enseignement supérieur. *Regards croisés sur l'économie*, *16*(1), 80-92.

Lemêtre, C., & Orange, S. (2017). Les bacheliers professionnels face à Admission Post-Bac (APB) :≪logique commune≫ versus ≪logique formelle≫ de l'orientation. *Revue française de pédagogie. Recherches en éducation*, (198), 49-60.

Huillery, E., & Guyon, N. (2014). *Choix d'orientation et origine sociale : mesurer et comprendre l'autocensure scolaire* [thèse de doct., Sciences Po-Institut d'études politiques de Paris ; Laboratory for ...].

Bretton-Wilk, R., & Haute, T. (2021). ≪En attente≫ : les logiques plurielles du sentiment d'injustice face à Parcoursup. *Sélections, du système éducatif au marché du travail*, 373-388.

Draelants, H., & Ballatore, M. (2014). Capital culturel et reproduction scolaire. Un bilan critique. *Revue française de pédagogie. Recherches en éducation*, (186), 115-142.