Analyse de questions d'apprenants et de profils associés dans des environnements en ligne

Fatima Harrak

Thèse de doctorat

Pour obtenir le grade de Docteur de

SORBONNE UNIVERSITÉ

École doctorale Informatique, Télécommunications et Électronique (Paris)

UMR 7606 Laboratoire d'Informatique de Paris 6, Équipe Modèle et Outils en Ingénierie des Connaissances pour l'Apprentissage Humain (MOCAH)

Analyse de questions d'apprenants et de profils associés dans des environnements en ligne

Spécialité : **Informatique**

Par **Fatima HARRAK**

Dirigée par **Vanda LUENGO**

Encadrée par **François BOUCHET**

JURY

Agathe MERCERON. Professeur. Beuth University of Applied Sciences, Rapporteure.
Anne-Laure LIGOZAT. Maîtresse de conférences, HDR. ENSIIE, LIMSI, Rapporteure.

Bénédicte LE GRAND. Professeur. Université Paris 1, Centre de Recherche Informatique, Examinatrice.
Fabrice POPINEAU. Professeur. Centrale Supélec, Examinateur.
Christophe MARSALA. Professeur. Sorbonne Université, Faculté des Sciences, Examinateur.

Vanda LUENGO. Professeur. Sorbonne Université, LIP6, Directrice de thèse.
François BOUCHET. Maître de Conférences, Sorbonne Université, LIP6, Encadrant.

Remerciements

Ce manuscrit est le résultat de nombreuses recherches que j'ai pu mener au cours de ces années de thèse. Je tiens à remercier toutes les personnes avec qui j'ai pu interagir au cours de cette période.

Je remercie les rapporteures, Agathe Merceron et Anne-Laure Ligozat, pour leurs rapports et commentaires avisés m'ayant permis d'améliorer ce manuscrit. Je remercie également Christophe Marsala et Fabrice Popineau d'avoir accepté de participer au jury de cette thèse. Un grand merci à Bénédicte Le Grand, qui fut bien plus qu'un membre de ce jury, pour m'avoir accompagné depuis mon master 2, par ses précieux conseils et encouragements toutes ces années.

Ce travail n'aurait pas été possible sans le soutien de Vanda Luengo, ma directrice de thèse, et François Bouchet mon encadrant de thèse. Ils m'ont apporté un grand soutien scientifique et morale. Ils m'ont accompagné avec patience et dévouement tout au long de ce doctorat. Travailler avec une personne passionnée par la recherche et très chaleureuse comme Vanda, est un véritable plaisir. François de par son savoir-faire et sa rigueur m'a appris tant sur les aspects scientifiques que méthodologiques pour mener à bien mes recherches tout au long de ces années. Nos discussions et échanges, menant souvent bien au-delà de mon sujet de thèse, sont à la fois enrichissantes et motivantes.

Je remercie l'équipe MOCAH dans son ensemble : Jean-Marc, Amel, Thibault, Iryna, Mathieu, Odette et Hélène pour leurs conseils et remarques concernant mes présentations. Je tiens à remercier particulièrement Monique d'avoir accepté d'évaluer une partie de ce travail. Je remercie bien évidemment mon collègue Mathieu G. pour l'aide qu'il m'a apportée durant toutes ces années ainsi que mes autres collègues : Mathieu V., Guy, Gorgoumack, Thomas...

Je tiens également à remercier ma famille pour son soutien et son encouragement : ma mère, mon père, ma sœur et mes frères ainsi que mes beaux-parents et mes amis.

Je remercie mon mari Zouhair, pour sa patience, son écoute et son soutien permanent : je n'y serai pas arrivée au bout sans lui ! Mon enfant chéri Abderahmane, que j'ai accouché pendant ce doctorat et qui me donne à chaque fois l'énergie et la motivation d'aller jusqu'au bout malgré les difficultés.

Enfin, une pensée à mon futur bébé qui viendra prochainement au monde !

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

Les questions des élèves jouent un rôle important dans le processus d'apprentissage, car elles sont une ressource potentielle pour l'apprentissage et les sciences d'éducation [Chin & Osborne, 2008]. Les questions des élèves indiquent qu'ils ont réfléchi aux idées présentées et qu'ils ont essayé de les relier à d'autres choses qu'ils connaissent. Graesser & Person [1994] avançaient l'hypothèse que l'un des mécanismes qui suscitent des questions vient de la nécessité de corriger les déficits de connaissances déclaratives. Par ailleurs, les questions peuvent provenir de la curiosité à l'égard du monde qui nous entoure ainsi que d'événements et d'interactions avec des enjeux du monde réel [Chin & Brown, 2002].

Pour les élèves, poser leurs propres questions est une première étape pour combler leurs lacunes en matière de connaissances et résoudre leurs problèmes. Le processus de poser des questions leur permet d'améliorer leur compréhension actuelle d'un sujet, d'établir des liens avec d'autres idées, de prendre conscience de ce qu'ils savent ou ne savent pas et aussi d'acquérir de nouvelles connaissances [Chin & Brown, 2000]. Lire ou entendre les questions d'autres apprenants peut aussi stimuler les étudiants à considérer différentes perspectives d'une question, améliorant ainsi la qualité de leur apprentissage. Les questions des élèves peuvent refléter l'apprentissage actif, la construction des connaissances, la curiosité et la profondeur du processus d'apprentissage [Graesser *et al.* , 2010].

Bien que les questions des élèves soient utiles pour les apprenants, elles servent également aux enseignants pour susciter la réflexion et l'engagement des élèves. Les questions des élèves aident non seulement les élèves à mieux apprendre, mais servent aussi la pédagogie de l'enseignant, lui permettant de déterminer ce qui a été compris et donc d'adapter le contenu et le rythme de son cours en conséquence [White & Gunstone, 1992; Etkina & Harper, 2002]. Par exemple, le fait de connaître les difficultés auxquelles les élèves font face aide l'enseignant à fournir des analogies, des précisions, des exemples et des questions qui aident les élèves à comprendre le contenu du cours [Harper *et al.* , 2003; Colbert *et al.* , 2007].

L'analyse de ces questions peut également indiquer le type des questions (cette notion est rarement définie par les auteurs, dans notre cas par exemple : le type d'une question

peut être une demande de ré-explication ou de vérification d'un concept) posées par les élèves qui ont réussi ou échoué. Par conséquent, les questions des élèves ne fournissent pas seulement une rétroaction à l'enseignant sur la compréhension des élèves, mais selon ces questions, l'enseignant a souvent une "intuition" du niveau et du profil des élèves dans la classe (par exemple : s'ils redoublent l'année ou suivent régulièrement le cours).

Malgré la capacité des questions des élèves à améliorer l'apprentissage, une grande partie de ce potentiel reste encore inexploitée. Les questions posées par les élèves en classe ne représentent qu'une petite proportion de toutes les questions posées pendant l'enseignement. En effet, Graesser & Person [1994] ont constaté que les élèves posaient peu de questions en classe, et encore moins lorsqu'ils étaient à la recherche de la connaissance. Cela se produit probablement parce que les élèves ne veulent pas attirer l'attention sur eux ou parce que les enseignants n'encouragent pas souvent les élèves à poser des questions. Les environnements en ligne et EIAH peuvent éliminer de nombreux obstacles qui empêchent les élèves à poser des questions en classe [Otero & Graesser, 2001]. Les forums de discussions et MOOCs sont considérés également comme une source riche pour extraire des informations et analyser les questions des élèves [Elgort *et al.* , 2018]. Cependant, dans un contexte universitaire, le traitement des questions posées en ligne par les étudiants peut être problématique du point de vue de l'enseignant. L'un des principaux problèmes est le volume de questions posées, ce qui empêche les enseignants de répondre à chaque question en classe ou en ligne.

Un autre problème typique est la difficulté pour l'enseignant d'établir un lien entre le profil des élèves et les questions posées, ce qui l'amène à choisir empiriquement les questions auxquelles il doit répondre en priorité, au lieu de pouvoir utiliser une stratégie plus avancée et adaptée (par exemple, ne pas répondre uniquement aux questions des bons élèves, ne pas considérer de la même façon une question provenant d'un élève qui ne suit pas habituellement le cours ou qui redouble l'année).

Nous nous intéressons plus particulièrement dans cette thèse à savoir si le type de questions posées par les étudiants sur une plate-forme en ligne peut être lié à leur performance et à leur comportement d'apprentissage global. Nous étudions principalement cette question dans le contexte d'une formation hybride (comme [Liu *et al.* , 2016a]), dans lequel chaque semaine les étudiants doivent poser des questions concernant le matériel en ligne qu'ils ont étudié à la maison (selon une approche de classe inversée), avant le cours, pour aider les enseignants à préparer leur séances de questions-réponses en présentiel. Dans notre contexte, les enseignants reçoivent actuellement par courriel, une semaine avant leur cours, une liste de questions posées par les élèves triées uniquement en fonction du nombre de votes associés à chaque question. Compte tenu du volume de questions posées, le courriel reçu est parfois long et difficile à lire (certains enseignants parlent d'un "mur de questions"). De plus, les enseignants n'ont souvent pas assez de temps pour répondre à chaque question et doivent donc sélectionner celles auxquelles ils vont répondre. Pour aborder cette problématique, les enseignants encouragent les étudiants à voter sur les questions déjà posées avant d'en poser de nouvelles pour limiter le nombre de questions et les aider dans ce choix. D'un point de vue pédagogique, cela suppose que les étudiants lisent les questions des autres, ce qui peut également avoir un impact positif en les forçant à s'interroger sur leur propre compré-

hension des points abordés par leurs camarades. Mais nous pouvons aussi penser qu'un vote n'est pas exactement équivalent à une question. En effet, dans le cadre théorique Active-Constructive-Interactive (ACI) proposé par Chi [2009], le fait de poser une question est une activité de nature « Constructive » (mobilisant des processus cognitifs tels que la recherche de lacunes dans ses connaissances et la restructuration de celles-ci), plus à même d'aider l'apprentissage qu'une activité « Active » comme le fait de voter (mettant uniquement en jeu une recherche dans ses connaissances pour savoir si on saurait ou non répondre à cette question). Par conséquent, nous avons conduit des analyses pour savoir comment les questions sont liées au comportement des étudiants (assiduité, redoublement, nombre de votes, etc.) et explorer plus particulièrement la valeur du vote.

Notre travail se situe dans la communauté de LA (Learning analytics), puisque le but principal est d'outiller l'enseignant en l'occurrence pour qu'il détermine les types de questions posées par les différents groupes d'apprenants (par exemple : lui proposer des organisations de questions qui l'aide à préparer ses sessions de questions-réponses) et non de lui proposer des solutions dans lesquelles le système intervient lui-même auprès des groupes d'apprenants. Notre objectif global est de fournir aux enseignants des informations supplémentaires pour les aider à choisir les questions pendant leur séances de questions-réponses, en fonction de la stratégie qu'ils jugent appropriée, ce qui leur permettra de "briser le mur des questions". Notre objectif ici est donc d'identifier la relation entre les questions des étudiants et leurs profils, à partir des données existantes, et d'utiliser ce travail pour aller vers la prédiction du profil des futurs étudiants dans les cours continus, en appliquant les analyses précédentes aux prochaines itérations des cours.

Pour valider la généralisabilité du travail mené, nous nous sommes également intéressé à la possibilité d'appliquer certaines étapes du processus suivi dans le contexte de cette formation hybride à un contexte différent dans le cadre d'un MOOC.

1.1 Questions de recherche

Nous avons donc été amenés tout au long de cette thèse à répondre à différentes questions de recherche :

QR1. Peut-on définir un schéma de codage pertinent pour analyser les questions des élèves en termes d'intentions ?

QR2. Comment automatiser l'annotation des questions des étudiants selon ce schéma de codage ?

QR3. Une fois les questions annotées selon le schéma de codage défini, peut-on utiliser les questions posées par les étudiants pour identifier leur profils, substantiellement différents en termes de performance et divers aspects du comportement du cours (assiduité, nombre de questions et votes) ?

QR4. Peut-on utiliser un modèle entraîné sur des sessions antécédentes de cours afin de prédire les profils des étudiants dans les prochaines sessions de ce cours ?

QR5. Peut-on proposer à l'enseignant des organisations pertinentes de questions pour l'aider dans la préparation de ses séances questions-réponses ?

QR6. Peut-on répliquer notre processus dans un autre contexte d'étude, tel un MOOC, pour vérifier la généralisabilité de notre approche ?

La QR3 sera la principale question de recherche dans ce manuscrit

1.2 Plan

Cette thèse est structurée en huit chapitres. Outre ce premier chapitre spécifiant la problématique, les objectifs et les questions de recherche abordées, nous présentons dans le **chapitre 2** un état de l'art s'articulant autour les différentes typologies de questions et les recherches qui ont été menées sur la relation entre les questions des élèves et leur comportement. Nous introduirons ensuite, dans le **chapitre 3**, le contexte de la formation hybride (PACES) selon une approche de classe inversée et les données utilisées. Nous présenterons par la suite dans le **chapitre 4** la démarche suivie pour définir un schéma de codage adapté aux questions des étudiants, puis le processus d'annotation automatique de questions associé à ce schéma de codage pour concevoir et évaluer différents systèmes d'annotation automatique. Dans le **chapitre 5**, nous étudierons le lien entre la nature de questions posées par les étudiants et leur comportements en utilisant une approche de clustering, puis nous explorerons la valeur du vote et son lien avec le comportement de questionnement et la performance. L'évaluation de la possibilité d'utiliser un modèle entraîné sur des données passées pour prédire le profil de nouveaux élèves dans de nouvelles itérations de cours fera l'objet du **chapitre 6**. Ensuite, nous proposerons des organisations de questions aux enseignants et présenterons comment elles ont été évaluées par les enseignants ayant répondu à un questionnaire sur celles-ci dans le **chapitre 7**. Nous introduirons dans **le chapitre 8** une nouvelle étude de cas pour répliquer notre processus dans le contexte d'un MOOC. Nous montrerons comment nous avons étendu le schéma de codage défini dans le chapitre 4 pour annoter les messages des étudiants dans les forums de discussions du MOOC et concevoir un nouveau modèle d'annotation automatique en cascade permettant d'annoter l'ensemble du corpus de questions. Nous étudierons par la suite le lien entre d'une part la nature des questions posées par les étudiants et d'autre part leur performance et leur capacités d'autorégulation. Ce travail permettra donc de vérifier la réplicabilité de notre processus d'annotation et d'analyse et la généralisabilité de notre approche. Enfin, nous conclurons ce travail par un bilan, des perspectives et pistes de travail à venir.

Chapitre 2

Etat de l'art

Les chercheurs ont étudié le comportement de questionnement des élèves dans divers contextes éducatifs, tels que la classe [Chin & Kayalvizhi, 2002], le tutorat [Graesser & Person, 1994] et les environnements d'apprentissage en ligne [Li *et al.* , 2014]. De nombreuses études se sont d'ailleurs intéressées à la définition de typologies de questions, et dans certains cas, ont analysé le comportement des élèves en fonction de leur profil de questions afin d'améliorer l'efficacité de leur apprentissage [Chin & Osborne, 2008; Graesser & Person, 1994]. Nous nous intéressons donc dans ce chapitre à l'étude de la nature de questions des élèves et son lien avec leur comportement.

Nous présentons dans ce chapitre les typologies de questions définies dans différents contextes éducatifs et comment elles étaient évaluées, en particulier pour les questions posées dans les environnements d'apprentissage en ligne (MOOCs, plateformes d'e-learning et forums de discussions). Nous identifions les caractéristiques des apprenants et spécifiquement comment le vote a été exploré dans les environnements en ligne pour analyser le comportement des élèves.

2.1 Nature et type de questions posées

Dans le cadre des travaux menés qui explorent la nature des questions des étudiants, Chin & Osborne [2008] présentent un état de l'art sur le rôle et l'importance des questions des étudiants dans le processus d'apprentissage à la fois sous l'angle de l'apprenant et de l'enseignant, et a par conséquent constitué l'une de nos principales sources pour cette section.

Avant d'évoquer les travaux existants sur la nature et type de questions posées, il nous semble important de définir la notion de taxonomie des questions et montrer la distinction entre une *taxonomie* et un *schéma de codage*.

En informatique, le terme taxonomie désigne une méthode de classification des informations dans une architecture structurée de manière évolutive. Elle peut désigner une

représentation hiérarchique de concepts, d'objets ou encore de disciplines (Wikipedia). A titre d'exemple : la taxonomie de Bloom, appelée Taxonomie des Objectifs d'Éducation, est un modèle pédagogique proposant une classification des niveaux d'acquisition des connaissances du processus cognitif [Bloom & Engelhart, 1956]. Elle peut aider les enseignants à poser des questions qui permettent de situer le niveau de compréhension des élèves, comme l'illustre la figure 2.1

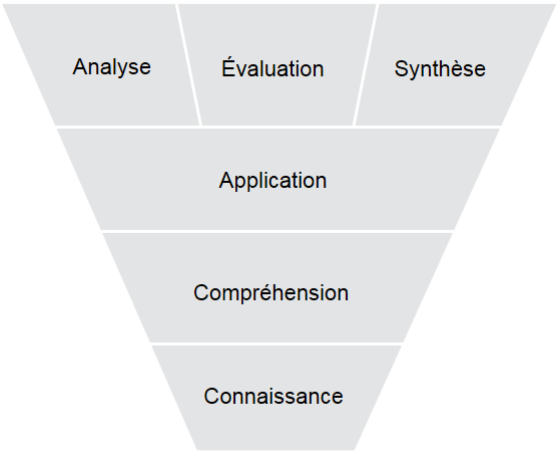


Figure 2.1 – Schéma de taxonomie de Bloom

La distinction entre un schéma de codage et une taxonomie est que la taxonomie a une structure hiérarchique avec une relation de subsomption entre ses éléments, alors que dans un schéma de codage (ou catégorisation) il n'y a pas forcément cette notion de relation entre concepts. Les catégories définies peuvent être complètement indépendantes les unes des autres. Dans les travaux présentés ici, le schéma de codage proposé (*cf.* chapitre 4) n'ayant pas de structure hiérarchique, il ne s'agit donc pas d'une taxonomie.

Dans la suite de ce travail, lorsque nous souhaitons parler indistinctement de schéma de codage/catégorisation et de taxonomie, c'est-à-dire de toute manière de distinguer des questions les unes entre les autres, nous emploierons le terme de *typologie de questions*.

2.1.1 Typologies de questions

Scardamalia & Bereiter [1992] ont mené trois études pour explorer la capacité des élèves de primaire à poser et reconnaître des questions productives basées sur les connaissances. Les auteurs font la distinction entre les *questions fondées sur le texte* et les *questions fondées sur le savoir*. Les questions fondées sur le texte font référence aux questions posées par les élèves après la lecture d'un texte dont les réponses peuvent être trouvées dans le texte donné, tandis que les questions fondées sur le savoir (ou les connaissances) sont des questions posées spontanément qui portent sur un intérêt profond ou un effort pour donner un sens à ce qui est étudié et approfondir un concept. Les questions fondées sur les connaissances formulées avant d'étudier le sujet se révèlent être d'un ordre plus élevé que les questions fondées sur le texte générées en réponse aux documents textuels. Ceci a mené les auteurs à conclure que les questions fondées sur les connaissances ont

un plus grand potentiel éducatif (explications et causes) que celles produites dans le contexte d'un questionnement fondé sur le texte (faits). Scardamalia & Bereiter [1992] ont également analysé la nature des questions posées par les élèves sur un sujet familier comparé à un autre dont ils possédaient peu de connaissances préalables. Ils ont constaté que le manque de connaissance du domaine ne semblait pas empêcher les élèves de poser des questions, le nombre de questions posées étant presque identique dans les deux cas. Cependant, il y avait une différence dans les types de questions posées. Les élèves ont principalement posé des *questions "information de base"* sur le sujet moins familier, mais plus de *questions "d'émerveillement"* (reflétant curiosité, perplexité, scepticisme) pour le sujet plus familier. Ce résultat indique que le manque de connaissance spécifiques à un domaine peut influer la nature de questions posées, mais pas nécessairement la quantité. Les chercheurs ont conclu que les questions de type "émerveillement" ont un potentiel éducatif plus élevé que celles de type "information de base". Les résultats de ces études montrent que les différents types de questions peuvent diriger le processus d'apprentissage différemment.

Chin & Kayalvizhi [2002] ont proposé une typologie de questions dans un contexte de classe en primaire pour étudier les sciences, qui distingue les *questions pouvant faire l'objet d'une investigation* de celles qui ne le permettent pas afin d'aider les élèves à générer des questions qu'ils peuvent élucider eux-mêmes. Les questions pouvant faire l'objet d'une investigation sont celles auxquelles les élèves peuvent trouver des réponses en concevant et en effectuant eux-mêmes des enquêtes pratiques, qui comprennent la comparaison, la cause-effet, la prédiction, la conception, l'exploration, la description et validation des questions du modèle mental. A titre d'exemples : "quel type de matériel est le meilleur pour garder l'eau chaude ?" (Comparaison), "comment la concentration affecte-t-elle le taux de dissolution du sel dans l'eau ?" (Cause et effet) et "quels types d'insectes vivent dans notre jardin ?" (description). D'un autre côté, les questions qui ne peuvent pas faire l'objet d'une enquête ne sont pas des questions d'ordre pratique ; elles comprennent des questions d'information de base ou complexes (simple dont la réponse peut être trouvée dans un livre, ou complexe par une explication d'un théorème scientifique).

Les questions des élèves peuvent aussi être classées selon le niveau de réflexion (ou des processus cognitifs) requis pour y répondre. Graesser & Person [1994] ont ainsi développé un schéma de catégorisation de questions posées pendant les séances de tutorat pour améliorer non seulement les questions posées par les tuteurs, mais aussi les questions que les étudiants produisent sur demande des tuteurs. Leur schéma de catégorisation est composé de 4 catégories principales : *questions suscitant réponses courtes*, *questions suscitant réponses longues*, *assertion* et *demande/consigne*). Les questions suscitant une réponse courte est un ensemble de catégories de questions de vérification, élaboration de concept, spécification des fonctions, *etc.* . Par ailleurs, les questions suscitant une réponses longue incluent les demandes de définition, d'exemple, de comparaison et d'autres questions de raisonnement approfondi. Ils ont décrit des questions de haut niveau comme celles qui impliquent des inférences, un raisonnement en plusieurs étapes, l'application d'une idée à un nouveau domaine de connaissances, la synthèse d'une nouvelle idée à partir de plusieurs sources d'information ou l'évaluation d'une nouvelle demande. Bien que leur schéma de catégorisation puisse être pertinent pour notre

travail, certaines catégories comprenaient des questions de raisonnement approfondi de haute qualité et qui sont associées à des modèles de raisonnement difficiles à identifier automatiquement (en particulier, des questions telles que celles décrites comme "antécédents", "conséquences", "orientation vers un objectif" ou "instruments/procédures et habilitation").

Une autre classification de questions des élèves a été proposée par Pedrosa de Jesus *et al.* [2003] pour distinguer les *questions de "confirmation"* des *questions de "transformation"*. Les auteurs ont placé les questions sur un continuum allant des questions de confirmation d'un côté aux questions de transformation de l'autre, plutôt que de les définir sur différents niveaux hiérarchiques. Les questions de confirmation cherchent à clarifier une information et un détail, tenter de faire la différence entre un fait et une spéculation et demander un exemple et/ou une définition. Ces questions de confirmation visent à déterminer quelle information est pertinente et à vérifier sur quelle base elle peut être incluse dans un contexte particulier. Les questions de transformation, d'un autre côté, impliquent une restructuration ou une réorganisation de la compréhension des élèves. Ces questions ont tendance à être associées à un raisonnement hypothético-déductif, à chercher des extensions dans la connaissance, à explorer les étapes d'argumentation, à identifier les omissions, à examiner les structures de la pensée et à remettre en question les raisonnements acceptés. Selon les auteurs, les deux catégories sont nécessaires et se complètent, et le type approprié de questions dépend de la nature de la situation. Suivant une idée similaire, Watts *et al.* [1997] ont classé les questions en trois catégories selon les périodes du processus de changement conceptuel : (1) *questions de consolidation*, où les élèves tentaient de confirmer des explications et de consolider la compréhension de nouvelles idées ; (2) *questions d'exploration*, visant à élargir les connaissances et à tester des concepts et (3) *questions d'élaboration* permettent aux élèves d'examiner des réclamations, de réconcilier différentes compréhensions et de résoudre des conflits. Bien que cette classification reflète une progression du développement de la pensée des élèves, elle dépend aussi du contexte pour qu'elle ait un sens : il faut savoir quand la question a été posée au cours du processus de développement conceptuel.

Marbach-Ad & Sokolove [2000] ont proposé une taxonomie de questions pour aider les étudiants universitaire en biologie à reconnaître les *bonnes questions* (c'est-à-dire celles considérées comme originales et perspicaces dans un contexte scientifique) et poser de plus en plus de questions de meilleure qualité. La taxonomie a été développée empiriquement après avoir examiné plus de 150 questions posées en classe par les étudiants et regroupé des types de questions similaires. La taxonomie propose quatre catégories principales de questions : (a) celles fondées sur un malentendu ou une idée fausse ou n'ont pas de sens logique ou grammaticales ; (b) celles concernant une définition simple ou complexe ; (c) celles qui nécessitent des informations en dehors du manuel scolaire (cela inclut des questions morales, philosophiques ou socio-politiques basées sur des motivations et intentions, ainsi que celles cherchant une explication fonctionnelle ou évolutive) ; et (d) celles qui impliquent l'utilisation de capacités de réflexion de haut niveau de la part des élèves (cela inclut des questions provenant d'une réflexion approfondie et d'une synthèse de connaissances et d'informations antérieures, ainsi que des questions hypothétiques et de recherche).

Une autre taxonomie déjà mentionnée et bien connue en matière de questions et d'éducation car largement utilisée par les enseignants est celle de Bloom & Engelhart [1956], ce que l'on appelle aussi "taxonomie des objectifs". Cette taxonomie permet à l'enseignant de formuler des questions dans le cadre des objectifs d'enseignement ou d'évaluation, et selon Chin & Osborne [2008] peut aussi s'appliquer aux questions des élèves. En effet, elle peut permettre de concevoir des activités pédagogiques, de présenter les informations selon les niveaux de pensée et de construire une progression pédagogique. Elle inclut une hiérarchie de niveaux allant de la connaissance à la compréhension, l'application, l'analyse, la synthèse et l'évaluation (*cf.* Figure 2.1). Cette taxonomie a été revue par Anderson *et al.* [2001] pour tenir compte des différents niveaux du processus cognitif, qui sont regroupés en six grandes catégories : se souvenir, comprendre, appliquer, analyser, évaluer et créer. Des recherches récentes [Supraja *et al.* , 2017] ont utilisé une version réduite de la taxonomie de Bloom pour établir un lien entre la rétroaction pratique et la performance de l'apprenant en matière d'évaluation. Cette taxonomie, en raison de son origine, tend à être plus appropriée aux questions de l'enseignant (par exemple : l'évaluation) qu'à celles des élèves.

Pizzini & Shepardson [1991] ont mis au point un autre schéma de codage qui catégorise les questions des élèves selon les niveaux cognitifs comme *entrée*, *traitement* et *sortie*. Les questions d'entrée exigent aux élèves de se rappeler de l'information, les questions de traitement amènent les élèves à établir des liens entre les données rappelées et les questions de sortie demandent aux élèves d'aller au-delà des données et de les utiliser d'une nouvelle manière. Le schéma de codage a été développé pour comparer la quantité et la qualité des questions des élèves dans la résolution de problèmes suivant des instructions par rapport à celles posées en laboratoire dirigé par l'enseignant. Cependant, ce schéma à trois niveaux n'est pas suffisamment descriptif des niveaux cognitifs et ne reflète pas l'éventail complet des processus cognitifs associés aux questions comme présentés par Bloom & Engelhart [1956]; Anderson *et al.* [2001].

D'une manière générale, les typologies de questions présentées jusqu'ici ont été définies manuellement par les chercheurs dans le contexte d'une classe ou un tutorat. En élargissant un peu notre centre d'intérêt, au-delà de la classification des questions, les chercheurs se sont intéressés à la détection des questions dans les actes de dialogue comme étant une des catégories principales à identifier. Les chercheurs ont exploré différentes approches permettant de classifier automatiquement des actes de dialogue. En effet, Kim *et al.* [2010b] ont utilisé leurs propres catégories d'actes de dialogue ('question', 'réponse', 'erreur', 'remerciement', 'correction') basés sur leur expérience en interactions des élèves dans les discussions en ligne. Ils ont présenté une approche permettant de classifier automatiquement les discussions des étudiants afin d'identifier celles qui contiennent des questions sans réponse et qui nécessitent l'attention de l'enseignant. Toutefois, la classification des messages avec des données incohérentes et bruitées restait un défi à relever. Bayat *et al.* [2016] ont utilisé et étendu les catégories d'actes de dialogue introduites par Kim *et al.* [2010b] pour inclure la catégorie ('référence') proposée par Merceron [2014] sur les mêmes types de données. Ils ont comparé et évalué les performances du nouveau classificateur sur les messages dans les forums de discussions en allemand.

Contrairement à Kim *et al.* [2010b], Qadir & Riloff [2011] se sont appuyés sur la taxonomie originale de Searle [1969] pour les actes de dialogue ('commissif', 'directif', 'expressif' et représentatif') et n'ont pas créé de catégories spécifiques à un domaine. Ce modèle permet de retrouver les phrases des actes de dialogue dans n'importe quel domaine.

Rus *et al.* [2012] ont proposé une approche dirigée par les données pour découvrir automatiquement les catégories d'actes de dialogue ('question', 'demande', 'réaction', 'déclaration', 'méta-déclaration', 'remerciement', 'évaluation expressive', 'autre') en regroupant les énoncés prononcés par les participants dans des jeux éducatifs. Gautam *et al.* [2018] ont utilisé les mêmes catégories de Rus *et al.* [2012] pour classifier automatiquement des actes de dialogue dans un environnement d'apprentissage virtuel.

Bouchet [2009] a proposé de son côté une catégorisation des demandes d'assistance faites en français à un agent conversationnel adjoint, qui faisait la distinction entre les demandes d'assistance directe et indirecte. Les demandes ont été classées selon quatre activités principales : contrôle, assistance directe, assistance indirecte et chat.

2.1.2 Limites des typologies de questions existantes pour notre contexte

Dans la plupart des travaux cités jusqu'à présent, les questions des étudiants ont été posées dans le contexte de l'enseignement formel en classe et les typologies proposées dépendaient principalement de ce contexte et ne correspondaient pas vraiment à ce qu'on veut annoter ici (analyser les questions des élèves en termes d'intentions et ne pas leur recommander un type de questions à poser). La typologie proposée par Scardamalia & Bereiter [1992] nécessite une transcription des cours (diapositives et vidéos dans notre contexte) avec lesquels les étudiants ont interagi avant de poser leurs questions en ligne. Bien que leur typologie nous semble pertinente, il était donc difficile d'identifier automatiquement si une question était textuelle ou non sans accéder aux transcriptions. Une autre limite de cette étude est la difficulté de reprendre la démarche suivie en dehors du cadre de l'expérimentation (manuelle et non automatisée) ou même de répliquer les analyses dans un autre contexte.

La typologie de Chin & Kayalvizhi [2002] nécessite l'aide de l'enseignant pour déterminer si les questions posées en ligne font l'objet d'une enquête pratique ou pas. Bien que notre but ne soit pas de prescrire aux élèves de poser un certain type de questions, nous ne pouvons pas considérer que les questions posées par les élèves se prêtent toutes à des enquêtes pratiques mais seulement les apprendre à faire la distinction entre des questions investigables et non-investigables.

Le principal problème de la taxonomie proposée dans [Marbach-Ad & Sokolove, 2000] est la difficulté d'automatiser l'identification des questions décrites dans les catégories, ce qui est essentiel dans notre contexte.

La taxonomie de Bloom & Engelhart [1956] inclut des catégories abstraites et com-

plexes de processus cognitifs, qui ne sont pas adaptées aux questions des élèves dans notre contexte.

Bien que la typologie proposée par Bouchet [2009] est l'une des rares ressources disponibles en français, elle semblait trop large pour nos besoins ici (la dimension d'assistance directe/indirecte aurait été intéressante à considérer mais pas suffisante).

Nous avons décidé qu'il serait plus pertinent pour notre objectif final de définir un nouveau schéma de codage utilisant une approche ascendante axée sur les données, où les mots clés et les expressions sont identifiés au fur et à mesure que le système est défini, et de le comparer a posteriori à certains des travaux mentionnés ci-dessus (voir le Tableau de synthèse 4.5).

2.2 Lien entre type de questions et comportement d'apprenant

2.2.1 Lien entre questions et une variable caractéristique de l'apprenant

Les chercheurs ont étudié le comportement de questionnement des élèves dans divers contextes éducatifs, en classe, en tutorat et dans des environnements d'apprentissage en ligne [Li *et al.* , 2014]. Certaines études récentes ont également porté sur la détection de questions audio [Cook *et al.* , 2018] plutôt que sur le texte. L'analyse des questions d'apprentissage a été utilisée à des fins très diverses afin d'améliorer l'enseignement et l'apprentissage des élèves. Dans cette section, nous allons présenter quelques recherches qui avaient étudiées la relation entre les questions des élèves et diverses variables, tel que le niveau de réussite, les approches d'apprentissage, les styles d'apprentissage [1], la nature de l'enseignement et la qualité de connaissances.

2.2.1.1 Réussite

Harper *et al.* [2003] ont étudié la relation entre les types de questions posées par les élèves de collège en physique et les notions qu'ils avaient comprises dans différents sujets. L'un des aspects de la réussite scolaire d'un élève est la compréhension conceptuelle du contenu de la matière. Les élèves ont rédigé un rapport hebdomadaire dans lequel ils ont répondu à trois questions portant sur ce qu'ils avaient appris et comment, quelles questions restaient obscures et quelles questions ils poseraient à leurs élèves s'ils étaient l'enseignant. Les chercheurs ont trouvé qu'il n'y avait pas de corrélation significative entre le nombre de questions posées et la réussite. Toutefois, les élèves qui ont posé des questions de haut niveau (similaire à la taxonomie développée par [Marbach-Ad & Sokolove, 2000]) ont obtenu de meilleurs résultats au test de performance conceptuelle que ceux qui n'ont posé que des questions simples, ce qui indique une relation directe entre la profondeur des questions et les connaissances conceptuelles antérieures. Les

1. Même si leur intérêt pédagogique est ardemment discuté [Pashler *et al.* , 2008]

élèves qui ont posé principalement des questions sur les équations n'ont pas aussi bien réussi que ceux qui ont posé des questions sur la cohérence, suggérant que les élèves ayant une base conceptuelle assez solide ont posé des questions qui les aident à relier divers éléments de leurs connaissances. De plus, les élèves qui manquaient de connaissances conceptuelles et qui posaient des questions pour combler les lacunes ont obtenu de meilleurs résultats aux tests conceptuels subséquents que ceux qui n'en avaient pas. Ces résultats indiquent que le simple fait d'encourager les élèves à poser des questions ne se traduit pas nécessairement par un meilleur apprentissage. Les questions de haut niveau concernant la cohérence et les limites étaient plutôt liées à une meilleure compréhension conceptuelle. Graesser & Person [1994] ont également trouvé que la réussite est positivement corrélée à la qualité des questions posées par les élèves qui ont acquis une certaine expérience en tutorat, tandis que la fréquence des questions n'a pas été corrélée à la réussite. Les élèves ont partiellement auto-régulé leur apprentissage en identifiant les déficits de connaissances et les comblent en posant des questions, mais ils ont besoin de formation et d'entraînement pour améliorer ces compétences.

2.2.1.2 Approches d'apprentissage

Chin & Brown [2002] se sont focalisés sur la relation entre les questions des élèves, la nature de leur réflexion et les actions adoptées durant le processus de construction des connaissances en classe. Ils ont montré que les types de questions posées par les élèves dépendent de la façon dont ils abordent leurs tâches d'apprentissage. En effet, les questions des élèves de type "information de base" qui portaient sur des faits et des procédures (et qui sont typiques d'une approche d'apprentissage superficielle) ont suscité peu de discussions productives. En revanche, les "questions d'étonnement" axées sur la compréhension, la prédiction, la détection des anomalies, l'application et la planification (et qui caractérisent une approche d'apprentissage approfondie) ont amené les élèves à s'engager dans des idées de réflexion et des discussions de groupe. Ces résultats suggèrent que les enseignants ont besoin de reconnaître les approches d'apprentissage (c'est-à-dire profonde par rapport à la superficielle, qui reflète la maîtrise par rapport à la performance) adoptées par leurs élèves et être sensibles et réactifs au type et à la profondeur des questions posées. Par ailleurs, l'étude menée par Li *et al.* [2014] consiste à examiner la fréquence des questions posées lors de la résolution de problèmes en groupes et leurs ampleurs en termes de difficulté des tâches et de phase de jeu. Les auteurs avaient utilisé la taxonomie proposée par Graesser & Person [1994] pour catégoriser les questions des élèves de collège et lycée qui ont participé au jeu sérieux "Land Science". Le jeu est composé de 14 étapes sur trois niveaux de difficultés, facile, moyen et difficile en fonction de la familiarité et de la complexité de la tâche effectuée. Ils ont constaté que les joueurs posaient plus de questions superficielles que profondes pour apprendre dans un environnement de jeux sérieux en ligne.

2.2.1.3 Styles d'apprentissages

Pedrosa de Jesus *et al.* [2004] ont étudié les relations entre les questions des étudiants et les différents styles d'apprentissage (divergent, convergent, assimilateur, accommoda-

teur) basés sur la théorie de Kolb [1984, 1985] en utilisant quatre études de cas dans un cours universitaire en chimie. Les résultats ont montré que bien qu'un élève peut avoir une nette préférence pour un style d'apprentissage particulier, certains élèves peuvent déployer tous les styles d'apprentissage et utiliser divers types de questions. Cependant, les étudiants à un niveau de développement des connaissances moins avancé n'avaient pas le niveau nécessaire pour poser une variété de questions, ce qui a mené les auteurs à conclure que le type de questions posées par les élèves est non seulement influencé par leur style d'apprentissage, mais qu'il dépend aussi (et peut-être surtout) de leur niveau de développement des connaissances. Par conséquent, les chercheurs n'ont pas réussi à identifier l'influence des styles d'apprentissage considérés sur le type de questions posées par les élèves.

2.2.1.4 Nature d'enseignement

L'étude réalisée par Marbach-Ad & Sokolove [2000] consiste à comparer l'apprentissage actif et l'apprentissage traditionnel. Les chercheurs avaient catégorisé les questions des élèves selon une taxonomie préalablement définie [Marbach-Ad & Sokolove, 2000] et examiné la différence de répartition de ces questions sur les différentes catégories en utilisant des tests inférentiels statistiques (Wilcoxon). Ils ont constaté que les étudiants du premier cycle en biologie qui ont suivi une approche d'apprentissage active étaient capables de poser des questions de meilleure qualité et de plus haut niveau après avoir lu des chapitres du manuel que celles enseignées dans un format traditionnel. La classe d'apprentissage actif employait des approches pédagogiques axées sur l'élève, constructives et interactives favorisait le travail en groupe et la discussion de problèmes et encourageait les élèves à poser des questions. D'un autre côté, la classe traditionnelle a été enseignée sous forme de cours magistraux avec peu de temps alloué à la discussion ouverte et aux questions. Au fil du temps, les questions du groupe d'apprentissage actif sont devenues plus perspicaces, réfléchies, liées au contenu et axées sur la recherche, et difficile d'y répondre en consultant le manuel scolaire ou une autre source facilement accessible. En revanche, la qualité des questions posées par les élèves de la classe traditionnelle est demeurée pratiquement inchangée. Par conséquent, ces résultats montrent que les questions des élèves dépendaient de la nature de l'enseignement.

2.2.1.5 Qualité de connaissances

Lai & Law [2013] ont exploré le lien entre la qualité de questions posées et la qualité de connaissances construites de deux niveaux d'étudiants (6ème et 2nde) dans des environnements d'apprentissage collaboratif assisté par ordinateur. Ils ont utilisé les catégories de questions prédéfinies "recherche d'explication" et "recherche de faits", largement utilisées pour évaluer le niveau d'engagement dans la construction de connaissances Hakkarainen [2003]; Zhang *et al.* [2007], et une troisième catégorie "simple clarification" a été incluse à la typologie et également utilisée par van Aalst [2009] dans la construction de connaissances. Bien que les corrélations observées varient d'une année d'étude à l'autre, le niveau moyen d'explication n'est pas corrélé au nombre de questions posées quelque soit la classe étudiée. D'autre part, une corrélation significative a été constatée entre le

niveau moyen d'explication et la qualité de questions posées par les élèves (cependant cette étude n'a pris en compte que les questions de nature épistémologique).

2.2.1.6 Autres variables

D'autres travaux réalisés par Cao *et al.* [2017] visent à analyser le comportement de questionnement des élèves dans les forums de discussion au sein de MOOCs. Les auteurs ont utilisé la quantité de questions posées, la classification et les variations dans le temps, mais ils n'ont pas été en mesure d'établir des relations claires entre les caractéristiques qu'ils ont examinées et les catégories de sujets.

Dans l'ensemble, ces travaux de recherche tendent à tenir compte du comportement ou des caractéristiques d'un élève donné (style d'apprentissage, niveau de compréhension, etc.) et à rechercher les différences dans les questions posées par les différents profils des élèves. Bien que nous avons suivi une approche similaire, le processus d'annotation et d'analyse de questions adopté dans ces travaux est généralement issu d'un processus manuel (car il s'agit d'une étude ponctuelle), et font rarement apparaître un souci d'automatisation. Nous avons développé un outil automatique pour annoter tout le corpus de questions, puis recherché une corrélation entre les questions posées par les élèves et d'autres caractéristiques (niveau, assiduité, etc.). Dans la section suivante, nous passons donc en revue quelques travaux du domaine s'intéressant à catégoriser de manière automatique des messages, questions, etc.

2.2.2 Applications de fouille de texte et fouille de données

La fouille de texte est une spécialisation de fouille de données qui désigne un ensemble de traitements informatiques consistant à rechercher une information, extraire une connaissance, annotation et classification de texte. Les techniques de fouille de texte peuvent être utilisées dans différents types d'applications nécessitant un traitement informatique de données textuelles au préalable.

Séguéla [2017] a distingué les applications ayant pour objectif l'automatisation d'une tâche (classification, étiquetage de documents, etc.) de celles ayant pour objectif la compréhension approfondie d'un corpus de textes à travers l'analyse de son vocabulaire.

Parmi le premier type d'applications ayant pour finalité l'automatisation d'une tâche, on peut noter :

- La catégorisation de messages, qui consiste à affecter automatiquement un message à une ou plusieurs catégories connues au préalable à partir de l'analyse du texte et d'un algorithme de classification supervisée. La détection de spam [Cormack, 2008] et la catégorisation selon le thème [Kessler *et al.* , 2006] sont ainsi deux applications très répandues.
- Le clustering de messages, par opposition à la catégorisation, vise à extraire des classes de messages homogènes du point de vue du contenu textuel à l'aide d'un algorithme de classification non supervisé [Azzag *et al.* , 2006].

- La recherche de documents et plus généralement la recherche d'information [Manning *et al.* , 2010]. Ce domaine de l'informatique englobe les travaux dont le but est d'obtenir un classement des documents relativement à leur pertinence vis-à-vis d'une requête utilisateur.
- Le résumé automatique de texte.
- L'extraction automatique de mots-clés pour représenter un texte.

Parmi le deuxième type d'applications visant à obtenir des conclusions sur un corpus de texte spécifique, on distingue notamment :

- La lexicométrie (étude statistique de l'usage des mots) et la stylométrie (mesure du style). Ces disciplines de l'analyse textuelle sont par exemple utilisées pour la comparaison de discours, la caractérisation lexicale d'un corpus de conversations, ou encore l'attribution d'auteur.
- Le traitement statistique des questions ouvertes dans les enquêtes [Lebart, 2001].

Dans ce type d'application, les pré-traitements appliqués aux données sont généralement peu nombreux afin de ne pas modifier la sémantique du texte. Nous avons utilisé dans ce travail des techniques de fouilles de texte visant à automatiser la classification des questions des élèves à partir des méthodes de classification supervisées selon des catégories de questions définies au préalable.

Comme vu précédemment, l'apprentissage automatique a pour principal objectif d'analyser un ensemble d'observations préalablement recueillies dans le but de construire une procédure permettant de classer, d'estimer, de regrouper ou encore de prédire de nouvelles données Mitchell [1997]. L'apprentissage automatique est un champ de recherche très vaste qui occupe beaucoup de chercheurs dans différents domaines. Nous nous intéressons plus particulièrement aux techniques issues de l'apprentissage automatique supervisé et non supervisé qui s'appliquent à la tâche de classification de texte. Ces techniques ne traitent que des données tabulaires. Elle ne sont donc utilisées pour la classification de textes qu'après transformation de ces derniers en tableaux (via les approches de type "sac de mots"). D'autres approches plus développées, comme Glove [Pennington *et al.* , 2014], Word2Vec [Mikolov *et al.* , 2013; Gautam *et al.* , 2018] et LSA [Dascalu *et al.* , 2013] peuvent également être utilisées pour extraire des caractéristiques sémantiques du texte et le transformer en tableau de vecteurs.

L'apprentissage supervisé a pour principal objectif de rechercher des fonctions permettant de prédire une variable d'intérêt, à partir d'un ensemble d'observations. Le type d'apprentissage supervisé est défini en fonction de la nature de la variable à prédire : classification (si la variable à prédire est qualitative) ou régression (si la variable est quantitative). Ces techniques nécessitent l'utilisation d'un jeu de données étiquetées manuellement pour construire un modèle de classifieur. Ces techniques sont très utilisées pour catégoriser les messages dans des environnements en ligne (*cf.* section 2.2.3).

Certains chercheurs ont utilisé des techniques d'apprentissage non supervisé pour la classification du texte. Ces techniques sont plus complexes puisque le système doit lui même détecter les similarités dans les données, mais restent moins coûteuses que l'apprentissage supervisé parce qu'elles ne nécessitent pas une annotation manuelle des

données. Sindhgatta *et al.* [2017] ont utilisé le regroupement agglomératif hiérarchique (CAH) pour identifier des groupes de questions dans les forums et fournir aux conférenciers des informations sur les questions fréquentes. Leur approche est basée sur la similitude des mots dans la question, en combinant la similitude lexicale et sémantique des questions. L'analyse de clusters peut également être utilisée pour aider les chercheurs à analyser les caractéristiques du comportement d'apprentissage et à élaborer des profils fondés sur les activités des apprenants [Antonenko *et al.* , 2012].

2.2.3 MOOCs, plateformes d'e-learning et forums de discussions

Les forums de discussion sont un élément clé de l'apprentissage en ligne et des MOOCs en particulier [Andresen, 2009], et peuvent être une source riche pour analyser les questions des élèves. Ce n'est pas seulement un lieu de socialisation, c'est aussi un lieu d'apprentissage, où les apprenants échangent des questions, des opinions et des préoccupations, qui sont consultés, votés et répondus par les autres apprenants et/ou le personnel enseignant.

Une revue systématique récente de littérature de forums de discussion dans les MOOCs a été faite par Almatrafi & Johri [2018]. Les chercheurs ont exploré la contribution des participants aux forums et les efforts fournis pour organiser efficacement le contenu des forums de discussions. Hecking *et al.* [2015] ont suivi l'évolution des sujets et ont constaté que les sujets évoluent dans le temps et qu'ils peuvent se diviser en plusieurs sous-sujets ou évoluer vers un sujet plus général. De plus, certaines plateformes de MOOC comme edX [2] demandent aux utilisateurs de spécifier le type de message parmi deux catégories : "discussion" ou "question". Cela permet à l'équipe enseignante d'avoir accès aux questions plus rapidement et d'y répondre de façon appropriée. Les sujets de discussion sont sélectionnés selon une liste prédéfinie de sujets proposés par le MOOC. Bien que cela puisse être utile dans l'organisation des messages du forum de discussion, il y a de fortes chances que des sujets redondants soient créés, ou que trop de sujets soient créés, ce qui peut être difficile à gérer. La participation abondante au forum de discussion est probablement une indication de la motivation, de l'engagement et de l'apprentissage des élèves [Goel & Polepeddi, 2016]. Toutefois, un taux de participation élevé peut être un défi pour l'équipe enseignante qui doit fournir aux élèves des commentaires opportuns, personnalisés et de grande qualité. En effet, la qualité et la rapidité des réponses de l'équipe enseignante aux questions des étudiants est un élément important dans le succès de l'apprentissage et la réussite des étudiants [Kim, 2013]. De plus, étant donné le taux élevé de participation des étudiants au forum de discussion, le personnel n'a pas forcément le temps de répondre à chaque message avec une réponse de qualité en temps opportun [Goel & Polepeddi, 2016]. Il est donc difficile pour les enseignants de revoir tous les messages de questions et commentaires. Par conséquent, des travaux se sont attachés à proposer une représentation alternative des données écrites dans les forums de discussion afin que les enseignants puissent, sans trop d'effort, avoir un aperçu de l'information intégrée dans la discussion (à partir des indicateurs de participation des étudiants) [Dringus & Ellis, 2005] et être informés lorsque de nouveaux messages d'intérêt

2. Organisation Forum edX

sont publiés [Lin *et al.* , 2009].

Une approche un peu différente consiste à automatiser les réponses à l'aide d'un agent conversationnel (chatbot). Ainsi, [Eicher *et al.* , 2017] ont développé un système d'assistance d'enseignement virtuel, appelé *"Jill Watson"* pour répondre de manière autonome aux questions fréquemment posées dans un forum de discussion (Piazza [3]). L'objectif est d'améliorer l'apprentissage et la rétention et augmenter l'interaction entre l'étudiant et l'équipe enseignante. Les enseignants communiquent avec les élèves via le forum de discussion de la classe qui agit comme une salle de classe virtuelle. Les réponses du système virtuel aux nouvelles questions sont basées sur son répertoire de paires questions-réponses construit à partir des occurrences passées du cours et classées par catégories de questions, en utilisant des techniques automatiques d'analyse sémantique basées sur des représentations conceptuelles. Le système répond correctement à quasiment toutes les questions posées par les élèves.

En terme de contenu, une utilisation courante des forums dans les MOOCs est pour les questions/réponses sur le contenu du cours ; une telle utilisation répond aux besoins cognitifs des apprenants (par exemple, corriger des idées fausses ou clarifier des concepts difficiles) [Breslow *et al.* , 2013]. D'autres utilisations (par exemple, établir des liens sociaux ou résoudre des problèmes techniques) peuvent être importantes pour la gestion des cours et le renforcement de la communauté, mais répondent principalement à des besoins non cognitifs. En raison de leur nature profondément différente, de nombreux MOOCs conçoivent des forums séparés pour ces différents types d'utilisations (contenu, social, technique, etc.). Cependant, les utilisateurs envoient souvent des messages inattendus qui ne tiennent pas compte de ces limites. C'est l'un des facteurs qui contribuent au problème de la surcharge d'information dans les discussions en ligne [Peters & Hewitt, 2010]. Dans les systèmes de mise en relation directe entre apprenants à l'aide d'algorithmes de recommandations [Khosravi, 2017], comme par exemple dans [Labarthe *et al.* , 2016; Bouchet *et al.* , 2017], le point d'entrée du système est d'ailleurs l'intention de l'apprenant : contacter d'autres apprenants pour socialiser, ou au contraire pour poser une question précise sur le contenu pédagogique du cours.

Macina *et al.* [2017] ont présenté une approche pour la recommandation de nouvelles questions aux élèves qui sont susceptibles de fournir des réponses. Les auteurs ont également observé que certaines questions du MOOC ne peuvent pas être répondues par des non spécialistes et nécessitent la réponse des instructeurs. La classification a été utilisée pour identifier les dimensions spécifiques des messages sur la base de catégories prédéfinies afin d'étudier les forums de discussion du MOOC et d'identifier les messages liés au contenu ou exprimant un sentiment comme la confusion [Agrawal *et al.* , 2015; Wen *et al.* , 2014; Wise *et al.* , 2017].

Il existe plusieurs méthodologies pour catégoriser et étiqueter les messages du MOOC. Parmi les premières, Stump *et al.* [2013] catégorisent les messages du MOOC en deux dimensions : le sujet du message et le rôle de la personne qui l'envoie. Le sujet du message est divisé en différentes sous-catégories : contenu, social/affectif, plate-forme du cours/technologie, structure/ politique du cours, etc. Le rôle de la personne qui l'envoie,

3. Sankar, P. (2013). Piazza : Our Story. Retrieved from https ://piazza.com/about/story

comprend : demandeur d'aide, apporteur d'aide ou autre. L'évaluation de l'ensemble des catégories du schéma de codage développé a été effectuée manuellement par deux annotateurs humains indépendants non familiers avec les données sur un échantillon de 4500 messages. Dans [Nylén *et al.* , 2015], plusieurs sous-catégories ont été rajoutées au rôle de l'expéditeur, telles que plaignant, fournisseur ou conciliateur. En outre, les auteurs ont proposé deux autres dimensions : la qualité du message et la réciprocité dans la discussion, et ouvrent la perspective d'automatiser l'analyse de données des forums de discussions de MOOC dans les futurs travaux. Les messages ont également été classés selon l'acte de dialogue associé dans les catégories suivantes : question, réponse, clarification, rétroaction positive, rétroaction négative et hors-tâche [Liu *et al.* , 2016b; Arguello & Shaffer, 2015]. À rebours de l'approche très "manuelle" proposée dans [Stump *et al.* , 2013; Nylén *et al.* , 2015], Liu *et al.* [2016b]; Arguello & Shaffer [2015] ont utilisé des méthodes d'apprentissage automatique tels que la régression logistique, modèle bayésien, arbres aléatoires et SVM pour catégoriser et analyser les actes de paroles dans les forums de MOOC.

Un autre modèle s'est attaché à classer les messages selon six dimensions : question, réponse, opinion, confusion, sentiment (positif/négatif) et urgence [Agrawal *et al.* , 2015]. Chaque message prend une valeur sur chaque dimension. Question, réponse et opinion sont des variables binaires, tandis que les autres dimensions peuvent prendre des valeurs discrètes variant de 1 à 7. Agrawal *et al.* [2015] ont établi un modèle de classification basée sur une approche de type sac de mots, des méta-données de messages, et la prédiction du message dans chacune des six dimensions mentionnées précédemment. Le modèle a été entraîné en utilisant une régression logistique. Les résultats obtenus varient selon le domaine d'application (Kappa de 0.62 pour le domaine des sciences humaines et de 0.36 pour le domaine de l'éducation). Bien que la confusion et l'urgence soient corrélées dans les deux domaines testés, la confusion mesure le niveau de clarté que l'élève manifeste dans son message tandis que l'urgence mesure le niveau critique du message pour attirer l'attention et l'intervention des instructeurs.

Une autre classification intéressante est celle de Cui & Wise [2015] visant à organiser les messages dans les forums de discussion selon que le contenu est lié au cours ou non. Les messages liés au cours incluent des questions d'éclaircissements, d'approfondissement ainsi que les discussions sur les ressources partagées. Les messages non liés au contenu du cours étaient variés, mais impliquait généralement des questions d'ordre logistique ou technique sur les évaluations et les certificats, des messages de socialisation et de partage de liens. Wise *et al.* [2017] ont catégorisé les messages dans les forums selon la classification proposée par Cui & Wise [2015] et ils ont appliqué la régression logistique aux mots (unigramme et bigramme) extraits automatiquement selon l'approche sac de mots. La validation du modèle est réalisée sur un ensemble de messages du même domaine, et a été testé ultérieurement sur des différentes disciplines. Bien qu'ils ont trouvé que peu d'étudiants posaient des questions de cours et que l'équipe enseignante répondaient principalement aux questions non liées au cours, leur modèle a montré une fiabilité modérée (Kappa moyen de 0.58) pour la classification des messages.

Almatrafi *et al.* [2018] proposent un modèle de classification permettant d'identifier les messages urgents dans les forums de discussion du MOOC en utilisant des caracté-

ristiques linguistiques indépendantes au sujet des messages. Différentes méthodes d'extraction de mots, ainsi que des techniques d'apprentissage automatique (SVM, Naive Bayes, régression logistique, arbres arbitraires et AdaBoost) ont été utilisées pour annoter les messages urgents (classification binaire). Ces chercheurs ont évalué le modèle en validation croisée et testé sa validité sur un ensemble de test non utilisé auparavant. Les résultats montraient que le modèle est modérément à substantiellement fiable pour identifier les messages urgents (Kappa mesuré entre 0.58 et 0.64).

De nombreux travaux ont tenté d'extraire des informations des messages des élèves. Ainsi, Kim *et al.* [2010a] ont utilisé des techniques de traitement de la langue naturelle (TF-IDF) pour extraire les messages utiles afin d'aider les élèves à participer à des discussions en ligne. Ils ont classifié ensuite les messages des élèves en question et réponse à partir de méthodes d'apprentissage supervisées (SVM) pour analyser la répartition des participants par genre.

2.2.4 Lien entre vote et caractéristiques des questions et des apprenants

Les votes peuvent être utilisés dans les forums de discussions comme un moyen pour faciliter le classement des messages pertinents et aussi un élément important pour caractériser l'activité des utilisateurs [Almatrafi & Johri, 2018]. Les votes des élèves ont donc également été étudiés, d'une manière différente, pour analyser le comportement des élèves dans les forums, où le nombre de votes qu'un message reçoit indique la qualité du message [Longstaff, 2017].

Par exemple Bihani *et al.* [2018] ont utilisé le nombre de votes sur les questions/réponses des étudiants et les réponses de l'enseignant pour révéler les paires questions/réponses pertinentes pour les futurs cours en utilisant les données du forum Piazza. Zeng *et al.* [2017] ont également utilisé un ensemble de caractéristiques y compris le nombre de votes pour détecter les messages de confusion dans un forum de discussion. Ils ont constaté que les messages codant la confusion sont considérés comme importants et étaient associés à un plus grand nombre de votes. Wise *et al.* [2017] ont trouvé que le nombre de vues et de votes ne permet pas d'identifier les messages liés au contenu. Cependant, l'ajout de ces variables améliore la prédiction des messages nécessitant une intervention en urgence, tel que décrit dans Almatrafi *et al.* [2018]. De la même manière, Jenders *et al.* [2016] présentent un modèle de classification pour prédire la réponse correcte aux questions posées dans les forums de discussion. Ils ont constaté que les variables liées au contenu du message n'étaient pas aussi prédictives que d'autres variables telles que le nombre de votes, le nombre de commentaires et le nombre de réponses acceptées.

Atapattu *et al.* [2016] ont effectué une série d'analyses statistiques pour mesurer la corrélation entre les sujets de discussion et d'autres variables telles que les votes. Les résultats ont montré que les sujets les plus discutés ne sont pas toujours ceux qui reçoivent le plus de votes. Klüsener & Fortenbacher [2015] ont étudié le lien entre la réussite et des variables dérivées de l'activité de l'apprenant dans le forum (tel que le

nombre de votes positifs et le nombre de réponses). Ils ont trouvé que le nombre de votes que reçoivent les participants est la variable la plus indicative pour les apprenants qui réussissent. Cependant, les réponses vagues ont été négativement corrélées avec les votes positifs [Reich *et al.* , 2014].

De nombreuses études se sont par ailleurs focalisées sur l'analyse d'un groupe spécial de participants, appelés participants actifs. La description des participants actifs a toutefois été affinée dans chaque étude. Certains les considèrent comme des utilisateurs qui ont participé au moins une fois dans les forums de discussions (publier, commenter, ou voter sur un message) [Mustafaraj & Bu, 2015]. Les résultats ont montré que seulement 30% des participants actifs finissent le cours. D'autres les caractérisent comme des utilisateurs qui participent constamment aux forums pendant au moins 6 semaines (maintien de la participation et non de la quantité). Ils ont été caractérisé sous le terme d'influenceurs (reflété par le nombre élevé de vues et de votes reçus) [Wong *et al.* , 2015]. Dans certains cas, un plus petit groupe de participants actifs a été choisi pour l'analyse, comme les participants réguliers (présents fréquemment dans les forums) Oleksandra & Shane [2016] ou les super-participants (top 5% des participants) [Huang *et al.* , 2014].

Une étude réalisée par Jiang *et al.* [2015] consiste à analyser le comportement et la performance des étudiants considérés comme "influenceurs" (utilisateurs dont les messages génèrent beaucoup de réponses dans les forums d'un MOOC). Ces influenceurs ont des résultats plus faibles et reçoivent moins de votes que les utilisateurs actifs (ceux qui postent régulièrement sur le forum). De même, Wong *et al.* [2015] ont analysé les votes, positifs tant que négatifs, sur les messages et les commentaires des utilisateurs actifs pour évaluer s'ils apportent des contributions positives au forum du MOOC (les messages et commentaires qui reçoivent le plus de votes positifs ont été considérés comme des contributions positives, dont le contenu est utile et bénéfique pour les autres utilisateurs). Contrairement à Jiang *et al.* [2015], ils ont constaté que les utilisateurs actifs sont aussi des utilisateurs influenceurs qui apportent généralement une contribution positive aux discussions du forum.

La plupart des forums de questions et réponses intègrent un mécanisme de vote collaboratif pour apprendre l'expertise des utilisateurs (par exemple : Stack Overflow). À l'aide de ces données, les chercheurs peuvent prédire les meilleurs répondants et les potentiels assistants à partir de votes données aux questions auxquelles ils ont déjà répondues [Ishola & McCalla, 2017b; Tian *et al.* , 2013]. De plus, Ishola & McCalla [2017a] ont utilisé les votes comme facteur d'évaluation de la qualité des réponses, afin de prédire le niveau de performance des utilisateurs à répondre à de nouvelles questions en fonction de leurs contributions précédentes. Dans [Klüsener & Fortenbacher, 2015], les auteurs ont utilisé une méthode d'apprentissage automatique pour classer les élèves comme "vont réussir" ou "à risque" en fonction de leur performance dans les forums de discussion, y compris leur nombre de messages, commentaires, votes, *etc.*

Globalement, dans les études présentées ci-dessus, les votes des élèves ont surtout été utilisés dans le contexte de discussions de forum pour analyser le comportement des élèves, mais nous n'en avons trouvé aucun qui examine le lien entre les votes des élèves et le processus d'apprentissage. Nous nous intéresserons ici plutôt à l'analyse de la

nature des questions votées et de leur lien avec les votes des élèves et des caractéristiques comme la réussite, l'assiduité, *etc.*

2.3 Synthèse

Dans un premier temps, nous avons présenté un éventail de typologies de questions proposées dans différents contextes éducatifs : classe, tutorat et environnements en ligne. Nous avons situé notre contexte de travail par rapport à chacune des typologies présentées dans la littérature et soulevé quelques limites. Globalement, nous avons constaté que la plupart des typologies définies dépendent du contexte, raison pour laquelle nous avons développé notre propre schéma de codage de questions.

L'analyse de questions des élèves a été utilisée à des fins très diverses. Dans un deuxième temps, nous avons étudié le lien entre le type de questions posées et différentes caractéristiques d'apprentissage, telles que la performance/ la réussite, l'engagement *etc.* à travers les travaux existants. Certaines analyses ont été effectuées manuellement, ce qui les rendent difficile à répliquer dans d'autres contextes (tel que le notre). D'autres s'intéressaient également à automatiser l'annotation et l'analyse de questions des élèves pour les rendre plus accessible et applicable dans d'autres contextes. Nous avons donné également un aperçu des techniques de fouille de données utilisées pour la classification de messages.

Ensuite nous avons exploré comment le vote a été associé aux messages dans les forums de discussions pour analyser le comportement des apprenants. Des corrélations positives ont été trouvées entre le comportement des apprenants dans les environnements en ligne et le vote (positif ou négatif). Ce dernier a été également utilisé pour détecter la confusion et l'urgence dans les messages des forums de discussions. En revanche, la nature des questions votées n'a apparemment pas encore été explorée.

Le Tableau 2.1 résume l'ensemble des typologies présentées dans la section 2.1 selon ces différents éléments : la typologie, la technique de validation utilisée, ses usages, l'ensemble des pré-traitements effectués et les méthodes et techniques d'analyse de questions utilisées.

Étude	Validation typologie	Pré-traitement	Usages	Techniques
Scardamalia & Bereiter [1992]	Questions évaluées par deux annotateurs, retient uniquement les cas d'accord, selon score de fiabilité	Filtrage manuel des questions socio-politiques	Encourager les élèves à poser des questions productives (expérimentation sur classes de 5e et 6e)	Pas d'automatisation (entretiens)
Graesser & Person [1994]	Questions annotées et examinées par 6 annotateurs sur 3 dimensions : catégorie de question, mécanisme de génération et degré de spécification, puis mesure de chaque annotation par un score de fiabilité. Priorités attribuées aux catégories en cas de question hybride (appartenant à plusieurs catégories)	-	Améliorer la génération de questions chez tuteurs et aider l'apprenant à auto-réguler son apprentissage	Analyse de variance (ANOVA) de proportion de questions posées par élèves versus tuteur
Anderson *et al.* [2001]	Groupe d'experts en sciences cognitives, programme scolaire, test et évaluation	Mots stemmatisés manuellement	Donner aux éducateurs un cadre conceptuel	-
Pizzini & Shepardson [1991]	Questions annotées par 2 annotateurs entraînés selon les poids attribués à chaque catégorie. Calcul d'un coefficient de corrélation entre les 2 annotateurs	-	Déterminer la quantité et la qualité des questions des élèves dans un modèle pédagogique de résolution de problèmes	Test de Chi-2 pour comparer les questions de différents groupes
Pedrosa de Jesus *et al.* [2003] Teixeira-Dias *et al.* [2005]	Questions annotées par 4 annotateurs (enseignants et chimistes), résultat d'accord mesuré	-	Stimuler l'apprentissage actif et améliorer la participation des élèves à des interactions en classe	
Marbach-Ad & Sokolove [2000] Marbach-Ad & Sokolove [2000]	Développée empiriquement et examinée sur 150 questions (regroupement des questions du même type)	-	Comparer l'apprentissage actif et traditionnel	-

Bayat *et al.* [2016]	Messages annotés manuellement par 2 annotateurs, Kappa inter-annotateur sur chaque catégorie	Pré-traitement manuel, segmentation en phrases, suppresion des phrases courantes non pertinentes, stemmatisation, liste de stop words	Classification automatique des messages	Méthode de classification supervisée (SVM)
Qadir & Riloff [2011]	Messages annotés manuellement par 2 annotateurs, Kappa inter-annotateur calculé sur chaque catégorie	Nettoyage basique, suppression de balises HTML	Classification automatique des messages sur n'importe quel domaine	SVM
Gautam *et al.* [2018]	2000 énoncés annotés manuellement par 3 annotateurs, Kappa inter-annotateur	Lemmatisation des mots et suppression de ponctuation	Classification automatique des actes de dialogue	Sent2vec pour représenter la phrase (sentence embeddings), 3 méthodes de classification testées (Naive Bayes, arbre de décision et réseaux de neurones (RN)), validation croisée appliquée sur le fichier test. Deux stratégies d'entraînement (RN entraîné sur données bruitées et non bruitées)

Tableau 2.1: Tableau de comparaison de taxonomies

Chapitre 3

Environnement PACES

3.1 Contexte d'étude

3.1.1 L'enseignement de PACES

La Première Année des Etudes de Santé, dite PACES, est commune aux études de santé toutes orientations comprises (médecine, pharmacie, odontologie et maïeutique). Les enseignements sont donc communs, et sont composés de neuf unités d'enseignement (UE), dont une spécifique par orientation. Le nombre des UE de PACES peut varier selon chaque faculté de médecine. Les disciplines issues de cette unité d'enseignement ne sont pas obligatoirement suivies par tous les étudiants. Seul l'enseignement des ou de l'orientation(s) choisie(s) par l'étudiant devra être suivi. À la fin de cette première année, suite aux concours, un classement est élaboré visant à sélectionner les étudiants autorisés (les 200 premiers étudiants sur les 1600 inscrits) à s'inscrire en deuxième année sur chaque orientation.

L'enseignement universitaire de PACES est structuré en deux semestres, comportant chacun des unités d'enseignements communes aux quatre filières auxquelles s'ajoutent durant le second semestre une unité d'enseignement dite spécifique à chacune des orientations. Le premier semestre est composé de six disciplines réparties sous quatre unités d'enseignements.

Le second semestre est composé de huit disciplines réparties sous cinq unités d'enseignements, dont une regroupe les disciplines de chaque orientation.

Unités d'enseignement et disciplines du S1	
UE 1	Biochimie (BCE)
UE 2	Histoire et Biologie du Développement (HBD) Biologie Cellulaire (BCE)
UE 3-1	Biophysique (BPH)
UE 4	Bio Statistiques (BSTAT)

Tableau 3.1 – Unités d'enseignements du premier semestre

Unités d'enseignement et disciplines du S2	
UE 3-2	Physiologie (PHS)
UE 5	Anatomie (ANT)
UE 6	Initiation à la Connaissance du Médicament (ICM)
UE 7	Santé, Société, Humanité (SSH)
UE spécialité	Orientation Médecine (MED) Orientation Pharmacie (PHAR) Orientation Maïeutique (MAIEU) Orientation Odontologie (ODON)

Tableau 3.2 – Unités d'enseignements du deuxième semestre

Chaque semestre de l'enseignement universitaire de PACES s'achève par un concours. Ces concours sont des épreuves communes portant sur les contenus pédagogiques étudiés pendant le semestre. Au cours du second semestre, les étudiants sont amenés à choisir une ou plusieurs orientations parmi les quatre suivantes : médecine, pharmacie, maïeutique et odontologie. Chaque étudiant participe au concours correspondant à son ou ses orientation(s) choisie(s).

3.1.2 PACES de Grenoble

La faculté de médecine de Joseph Fourier de Grenoble a mis en place une réforme pédagogique pour la première année (PACES). L'enseignement d'une discipline est structuré par quatre activités d'apprentissage formant une séquence d'apprentissage suivant un système de formation hybride (une partie du travail se fait à distance et l'autre partie en classe) avec une classe inversée (pas de cours magistral, les aspects transmissifs se préparent seul avant le temps passé avec l'enseignant). Chaque discipline est composée de deux à six séquences de quatre semaines, consacrant une semaine pour chaque activité d'apprentissage comme l'illustre la figure 3.1. Chaque discipline est étudiée par un groupe d'étudiants (la promotion est divisée en 8 groupes) pendant une séquence d'apprentissage.

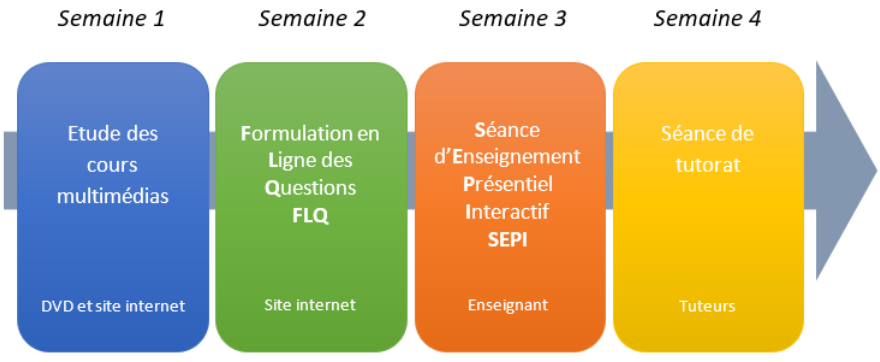


Figure 3.1 – Les quatre activités d'une séquence d'apprentissage sur quatre semaines

Dans chaque matière, une séquence d'apprentissage se répartit sur quatre semaines :

- La première semaine est consacrée à l'étude des cours. Plusieurs disciplines différentes sont alors étudiées sous forme de cours multimédia animés et sonorisés par des commentaires

de l'enseignant sur DVD. Des informations complémentaires sont disponibles sur internet.

- La deuxième semaine est consacrée à la Formulation en Ligne des Questions (FLQ) destinées aux enseignants. Des questions portant exclusivement sur les cours multimédia, étudiés la semaine précédente, sont formulées sur le site internet de la Faculté de Médecine et de Pharmacie de Grenoble. Ces questions s'adressent aux enseignants, et forment la base de l'enseignement de la semaine suivante. Chaque étudiant peut poser autant de questions qu'il le veut, et voter sur les questions posées par les autres étudiants de son groupe. Il ne peut en revanche pas répondre aux questions des autres.

- Cette troisième semaine est dédiée aux Séances d'Enseignement Présentiel Interactif (SEPI), qui est un enseignement explicatif et applicatif en présentiel avec les enseignants. Chaque discipline étudiée au cours de la première semaine donne lieu à une Séance d'Enseignement Présentiel Interactif. Ces séances sont réalisées en groupe d'étudiants. Elles sont assurées par les enseignants en charge des cours. Ces derniers répondent aux questions posées en ligne par le groupe d'étudiants.

- La quatrième semaine est celle des séances de tutorat, c'est-à-dire l'enseignement applicatif. Elles sont supervisées par les enseignants de PACES mais aussi par des étudiants tuteurs. Ces séances permettent de mettre en application les connaissances acquises lors des activités d'apprentissage précédentes par la pratique de questions à choix multiples (QCM). En amont de ces séances de tutorat, les QCM sont rédigées par les tuteurs puis sélectionnées et corrigées par les enseignants. Puis chaque question sélectionnée pour le tutorat est expliquée et commentée par les enseignants à l'ensemble des tuteurs au cours d'une séance de préparation. Lors de la séance de tutorat, les étudiants répondent aux questions durant quarante minutes, puis les tuteurs les corrigent et les expliquent durant une heure et vingt minutes. Le nombre de questions à choix multiples diffère selon la discipline étudiée. Chaque étudiant peut vérifier individuellement ses notes et son classement par rapport à l'ensemble de la promotion, et il est nécessaire d'être connecté à la plateforme de questions pour pouvoir consulter ses notes. Ces résultats de tutorat permettent aux élèves d'évaluer leurs performances.

3.1.3 La pédagogie inversée

Au-delà de la simple inversion des activités traditionnelles typiques, la classe inversée a ceci de spécifique qu'elle apporte une véritable nouvelle perspective pédagogique. Lebrun [2015] l'a définit comme suit

> L'approche des classes inversées est surtout un changement de paradigme, de mentalités dans les rapports que nous construisons avec les termes "savoir", "apprendre" et "enseigner". Le concept porte donc sur une grande variété de pratiques, de méthodes et de techniques à la fois relationnelles et techniques.

En effet, la pédagogie inversée offre la possibilité d'effectuer des activités pédagogiques en dehors de la classe notamment par l'intermédiaire des nouvelles technologies numériques. Des activités de pédagogie classiquement organisées à l'intérieur de la classe et à l'extérieur sont inversées tout en favorisant les interactions entre les enseignants et les apprenants [Nizet & Meyer, 2016]. Les travaux de Mazur [1997] puis d'autres [Bergmann & Sams, 2012], ont remis le concept de classe inversée en avant en proposant des capsules vidéos de leur cours à leurs élèves afin d'apporter un soutien aux élèves qui ne pouvaient pas être présents en classe. Observant l'intérêt de cette pratique à motiver des élèves pour préparer les leçons (à distance,

à domicile) et afin de rendre les cours en présence de l'enseignant (en classe) interactifs, ils ont lancé le concept de "Flipped classroom" [1]. Le cours traditionnel conserve sa place classique dans les étapes des activités chronologiques d'apprentissage mais la forme et l'espace/temps dans lequel il s'inscrit est différent. Il est délivré à distance par l'usage de la technologie ce qui permet à l'apprenant d'avoir les moyens d'assister à ce cours en fonction de ses besoins. En effet, le support enregistré lui permet de le suivre autant de fois que souhaité, de revenir sur les points qui posent problème et d'aller plus loin en ayant la possibilité de préparer des questions. L'applicatif qui est réalisé à distance (devoirs à la maison) se fait en présence de l'enseignant au sein de la classe. Des activités pédagogiques (résolution de problèmes, tâche à effectuer) sont proposées par l'enseignant pour accompagner les étudiants dans l'utilisation des savoirs théoriques. Il s'agit d'un enseignement asynchrone pour un apprentissage synchrone [Villiot-Leclercq *et al.* , 2012; Lebrun, 2015]. Ce sont donc les temps et les espaces d'enseignement qui sont inversés au profit d'une dynamisation de l'enseignement.

Actuellement, il n'y a pas de vision claire concernant les effets de la pédagogie inversée sur la performance des élèves. Les différentes recherches réalisées sont partagées à ce sujet. Mais, ce concept pédagogique n'est pas sans être contraint par différentes conditions de réussite. En effet, Nizet & Meyer [2016] expliquent que l'enseignant doit être volontaire pour modifier ses pratiques classiques d'enseignement. Il doit s'inscrire dans une connaissance du processus d'apprentissage et de la pédagogie active centrée sur l'apprenant. De plus, les temps de préparation des cours sont particulièrement plus importants car l'enseignant doit préparer les supports à distance et proposer des activités en présentiel qui soient riches en apports cognitifs. Par ailleurs, il est nécessaire de prendre en compte que certains étudiants peuvent ne pas faire le travail d'apprentissage à distance, soit parce qu'ils ne seront pas motivés, soit parce qu'ils ne seront pas suffisamment autonomes pour cette réalisation. En effet, cela demande une autonomie pas toujours acquise. Il est utile de bien évaluer les moments où l'apprenant sera apte à aborder seul de nouveaux supports de connaissances. Il s'agit donc d'anticiper les problèmes pédagogiques à résoudre, d'avoir une approche personnalisée en lien avec les spécificités du milieu et du public dans lesquels l'enseignant doit mettre en place la pédagogie inversée. La structure doit d'ailleurs favoriser son introduction en adoptant une certaine souplesse par rapport à la gestion des groupes qui seront plus actifs. Les moyens techniques doivent être adaptés pour l'ensemble des apprenants et l'équipe enseignante capable de fournir un soutien technique. Enfin, capter l'attention du public nécessite des ressources modernes et de bonne qualité.

Dans le contexte particulier de PACES, Gillois *et al.* [2013] expliquent dans leurs travaux de recherche qu'avant 2005, à la faculté de Médecine et de Pharmacie de Grenoble, l'enseignement de la première année de médecine ne satisfaisait ni les étudiants, ni les enseignants. En effet, avec la hausse du numerus clausus [2], le nombre d'étudiants en première année de médecine était en croissance constante. Les moyens pour accueillir les étudiants avaient atteint leur limite au niveau de la taille des amphithéâtres, des horaires et du nombre d'enseignants. Ainsi, selon eux, le concours avait fini par ne plus remplir son rôle de formation et la prise en charge des apprenants devenait difficile. Après la réforme de 2009, le dispositif PACES de la faculté de Grenoble est devenu un espace de formation hybride de type classe inversée au sein d'un environnement numérique de travail nommé Medatice [3].

1. Flipped classroom en anglais, classe inversée en français.
2. Nombre de personnes admises à concourir, à exercer une fonction ou un métier, limité et décidé par une autorité publique
3. paces.medatice-grenoble.fr

BCH	BPH	HBD	BCE	ANT	PHS	SSH	ICM	MAT	Spéc.
19%	17%	15%	11%	10%	9%	8%	6%	3%	1%

BCH = Biochimie, BPH = Biophysique, HBD = Histoire et biologie du développement, BCE = Biologie cellulaire, ANT = Anatomie, PHS = Physiologie, SSH = Santé, société, humanité, ICM = Initiation à la connaissance du médicament, MAT = Mathématique, Spécialité = Médecine, Pharmacie, Odontologie, Maïeutique

Tableau 3.3 – Distribution des questions posées par cours

3.2 Données

Nous avons considéré l'ensemble des questions posées par les étudiants de PACES de l'université Joseph-Fourier de Grenoble en 2012-2013 sur la plateforme en ligne Medatice. 1608 étudiants étaient inscrits cette année-là, bien que tous n'aient pas posé des questions. Par conséquent, pour chacun des 13 cours, nous avons 2 à 6 ensembles de questions (un par séquence) posées par 429 étudiants (6457 questions en tout) pendant la deuxième semaine de chaque séquence (semaine FLQ). La répartition des questions est inégale (*cf* Tableau 3.3), avec plus de questions au 1er semestre, en particulier parce que certains étudiants sont obligés d'arrêter à la fin du 1er semestre, en fonction de leurs résultats à l'examen.

Il est important de noter que, dans l'ensemble, seulement un élève sur quatre a posé au moins une question. Ce phénomène pourrait s'expliquer par le fait que les enseignants encouragent les étudiants à voter au lieu de poser des questions (d'un côté pour les forcer à lire les questions posées par les autres élèves, et d'un autre côté pour réduire le nombre de questions qu'ils reçoivent par courriel avant leur séances de questions-réponses). Nous émettons également l'hypothèse que tous les enseignants n'utilisent pas la plateforme en ligne autant qu'ils le devraient, ce qui pourrait expliquer en partie pourquoi de moins en moins d'étudiants posent des questions avec le temps.

En plus des questions, nous avons également accès à un certain nombre d'informations relatives aux 1608 étudiants inscrits. Les variables dont nous disposons pour chaque étudiant et chaque cours sont fournies dans le Tableau 3.4.

Variable	Description
NotMoy*	note moyenne obtenue sur les QCMs du cours (sur 20)
NotFin*	note finale obtenue au concours à cette matière (sur 20)
AssGlb	proportion de l'assiduité globale sur les deux semestres ((de 0 (jamais là) à 1 (toujours là))
AssCou	proportion de l'assiduité pour chaque cours (de 0 (jamais là) à 1 (toujours là))
NbQst	nombre de questions posées sur chaque cours
NbVotRec	nombre de votes reçus par les autres étudiants sur ses questions dans chaque cours (0 si aucune question n'est posée)
NbVotFait	nombre de votes effectués sur les questions des autres étudiants dans chaque cours
EtuRed	si l'étudiant était redoublant (variable binaire, égale à 1 pour les étudiants dont le rang à l'examen final est inférieur à 200, 0 sinon)
EtuReu	si l'étudiant a réussi ou non l'examen final (égale à 1 pour les étudiants dont le rang à l'examen final est inférieur à 200, 0 sinon)

* Quant un étudiant ne s'est pas présenté à un QCM en fin de séquence ou au concours en fin de matière, la note correspondante est mise à 0, en conformité avec la politique de l'université

Tableau 3.4 – Description des 9 variables disponibles pour chaque étudiant sur chaque cours

Chapitre 4

Schéma de codage de questions

Comme vu dans l'état de l'art, les typologies de questions proposées dépendent principalement du contexte étudié et fournissent rarement un ensemble complet de mots-clés pour permettre une identification automatique de questions, et encore moins des outils dédiés permettant cette classification (même en anglais, ce qui n'est par ailleurs pas notre cas ici). Notre objectif essentiel est de fournir des catégories de questions qui prennent en compte l'intention de l'élève (définie par Garcia-Fernandez [2010] comme *"la réponse à laquelle s'attend un locuteur lorsqu'il pose sa question"*), et nourrir la réflexion de l'enseignant pour lui donner éventuellement une réaction pédagogique différente sur les questions posées. Par conséquent, nous avons décidé de définir notre propre schéma de codage pour identifier le type des questions posées par les étudiants, utilisant une approche ascendante fondée sur les données tout en prenant en compte ces différentes contraintes :

- Avoir une granularité assez fine : les taxonomies existantes sont trop génériques [Pedrosa de Jesus *et al.* , 2003; Chin & Kayalvizhi, 2002] ou trop détaillées [Graesser & Person, 1994]. Notre objectif est de fournir aux enseignants suffisamment d'informations sur la nature de questions posées par les élèves et en même temps construire un schéma de codage assez générique pour analyser leur questions.

- Être indépendant du contexte, l'identification des catégories doit être indépendante des connaissances du domaine (médecine), pour pouvoir l'utiliser sur d'autres contextes (ex : MOOCs) et faciliter la réplication des analyses.

- Être facile à automatiser, à partir des approches simples basées sur l'identification de mots-clés ou autres.

Les résultats présentés dans ce chapitre ont donné lieu à une publication dans la conférence internationale LAK (Learning Analytics and Knowledge) [Harrak *et al.* , 2018] et EDM (Educational Data Mining) [Harrak *et al.* , 2017, 2019b].

4.1 Une démarche exploratoire ascendante

Nous avons travaillé sur trois échantillons de 200 questions [$PACES_{600}$] et un échantillon de 152 questions (voir l'explication plus loin). Le corpus $PACES_{INIT}$ composé des 4 échantillons (*cf.* figure 4.1), est utilisé pour définir le schéma d'annotation et construire le système de classification des questions, représente 12% du corpus total de questions [$PACES_{TOT}$] introduit dans la section 3.2. Les questions extraites aléatoirement de deux cours (BCH et HBD, *cf.* Tableau 3.3), considérés par l'équipe pédagogique comme étant parmi les plus difficiles et ayant suscité le plus de questions (*cf.* 3.3). Chaque échantillon est utilisé dans l'une des 4 étape de catégorisation suivantes : (1) étape de découverte [$PACES_{DEC}$], (2) étape de consolidation [$PACES_{CON}$], (3) étape de validation [$PACES_{VAL}$] et (4) étape d'évaluation [$PACES_{EVAL}$].

(1) L'**étape de découverte** a consisté à regrouper empiriquement des phrases, du premier échantillon $PACES_{DEC}$, ayant des similitudes pour en extraire des concepts significatifs. Bien que l'équipe pédagogique demandait aux élèves de poser des questions simples (c.-à-d. centrées sur un seul sujet, en évitant des questions comme « Pourriez-vous expliquer à nouveau X ? De plus, Y n'était pas clair »), nous avons constaté qu'un sous-ensemble important des questions pouvait être divisé en plusieurs questions indépendantes dans 40 % des cas. Une fois les phrases segmentées en questions dites simples, nous avons regroupé des questions dont la structure (ex : « qu'est-ce que X ? » et « qu'est-ce que Y ? ») et la sémantique (ex : « qu'est-ce que X ? » et « pourriez-vous définir X ? ») semblent similaires. Des groupes de questions ont ensuite reçu des "étiquettes" (par exemple, « définition d'un concept ») pouvant être regroupées en catégories de niveau supérieur. Ensuite, nous avons identifié les exclusions mutuelles entre étiquettes (ex : une question simple ne peut pas être à la fois une vérification et une demande de ré-explication), et inversement, les étiquettes compatibles (ex : une vérification et une ré-explication pourraient être relatives à la correction d'un exercice). Cela nous a conduit à définir le concept de « dimensions », composées d'ensembles d'étiquettes de types de questions similaires mais mutuellement exclusives (dans l'exemple précédent, on ne peut pas en même temps vérifier la validité d'une affirmation et demander une ré-explication d'un concept). Chacune de ces étiquettes individuelles ("vérification", "ré-explication"...) sont des valeurs pouvant être associées à une dimension. En même temps que l'identification d'une valeur dans une dimension, l'annotateur humain identifiait les mots-clés ou expressions idiomatiques indicatifs de cette valeur de dimension (par exemple, dans Dimension1, pour la valeur de dimension "Ré-expliquer", certains des mots-clés identifiés seraient "ré-expliquer", "rappeler", "redéfinir", "refaire", "répéter", "résumer", "revenir", etc.). En résumé, le schéma de codage est constitué de dimensions qui sont un ensemble de valeurs avec une liste de mots-clés associés à chacune de ces valeurs. Chaque question simple peut alors être associée à une annotation dans ce schéma de codage en choisissant, pour chaque dimension, une et une seule valeur. Une annotation associée à une question peut donc être vue comme un vecteur de N valeurs, N étant le nombre de dimensions du schéma de codage (ex : "Pourriez-vous réexpliquer la différence entre un composé ionisable et un composé partiellement ionisable ?" est une demande de ré-explication [Ree] du lien entre deux concepts [Lie], est représentée par le vecteur [Ree,0,Lie,0]), avec aucune valeur identifiée pour la dimension 2 et 4 [0].

(2) L'**étape de consolidation** a consisté à annoter le deuxième échantillon $PACES_{CON}$ pour valider les dimensions et les valeurs précédemment identifiées. Cela a conduit à

divers ajustements des dimensions pour s'assurer qu'elles étaient bien indépendantes les unes des autres (par exemple l'ajout de la valeur "correction" dans Dim2, non identifiée précédemment). Parallèlement, les dimensions identifiées ont été revues et validées par un professeur expert enseignant dans le cadre de PACES, qui a estimé que les catégories étaient potentiellement pertinentes pour analyser les questions des étudiants.

(3) Lors de l'**étape de validation**, nous avons effectué une double annotation pour valider l'ensemble de nos catégories sur le troisième échantillon $PACES_{VAL}$. Premièrement, les 200 phrases ont été segmentées manuellement, fournissant 238 segments. Ensuite, deux annotateurs humains (les chercheurs qui ont défini le schéma de codage et le système d'annotation) ont utilisé comme référence unique le schéma de codage créé à la fin de l'étape précédente pour annoter chacun de ces segments. A l'issue de l'étape précédente, quatre dimensions avaient été identifiées : Dim1 (relative au type de question), Dim2 (relative à la modalité d'explication), Dim4 (facultative, annotée uniquement si la question est une vérification, relative à la nature de ce qui est vérifié) et une autre dimension relative à la nature des fautes dans les phrases (grammaticale, mots manquants, orthographe...). Cette dernière dimension n'est pas abordée ici parce qu'elle a été exclue plus loin en raison de la difficulté à l'automatiser, et de l'intérêt pédagogique éventuellement discutable. La dimension appelée "Dim3" plus loin n'existait pas encore à cette étape. Les annotateurs humains ont fait deux annotations distinctes et indépendantes sur chaque dimension, et leur accord a été évalué à l'aide du Kappa de Cohen (*cf.* section 4.4.4.2) [$\kappa 1$=0.72, $\kappa 2$=0.62 où $\kappa 1$ et $\kappa 2$ correspondent respectivement au Kappa de Dim1 et Dim2]. Pour Dim4, en raison de son caractère facultatif, les deux annotateurs n'ont pas nécessairement annoté les mêmes questions : un annotateur a annoté 82 questions, et l'autre 68, avec un chevauchement de 68 questions. Le kappa calculé sur ces 68 questions valait 0.66. Puis ils se sont rencontrés pour discuter et résoudre les désaccords, essentiellement des cas ambigus. Cela a conduit à un affinement final des catégories (par exemple, séparation des catégories Dim1 et Dim4, ajout de la catégorie Dim3). Finalement, le corpus $PACES_{600}$ a été annoté de nouveau sur 4 dimensions (Dim1 à 4) par un seul annotateur pour tenir compte des changements et fournir une référence à laquelle l'annotation automatique pourrait être comparée. Cette version finale du schéma de codage en quatre dimensions est présentée dans le Tableau 4.4.

(4) Finalement, dans l'**étape d'évaluation** [1], le dernier échantillon $PACES_{EVAL}$ a été annoté manuellement par les deux annotateurs experts (avec un Kappa accru de 0.83 sur Dim1, 0.76 sur Dim2 et 0.76 sur Dim3, *cf.* Tableau 4.1). Comme dans l'étape de validation, une discussion a été faite entre les annotateurs et par ailleurs une convergence plus forte et des valeurs Kappa plus élevées. Cet échantillon, non utilisé pour l'entraînement de l'annotateur automatique, a été utilisé pour son test (*cf.* section 4.4).

Il existe plusieurs façons d'obtenir une estimation du taux d'accord entre les annotateurs pour une tâche d'annotation [Fort, 2012]. Nous avons utilisé essentiellement dans cette thèse le coefficient Kappa de Cohen [Cohen, 1960] pour mesurer l'accord inter-annotateurs (*cf.* section 4.4.4.2). Les accords que nous allons présenter sont valables pour deux annotateurs annotant les mêmes instances.

1. Cette étape est arrivée ultérieurement et on s'est concentré uniquement sur les segments au lieu de questions

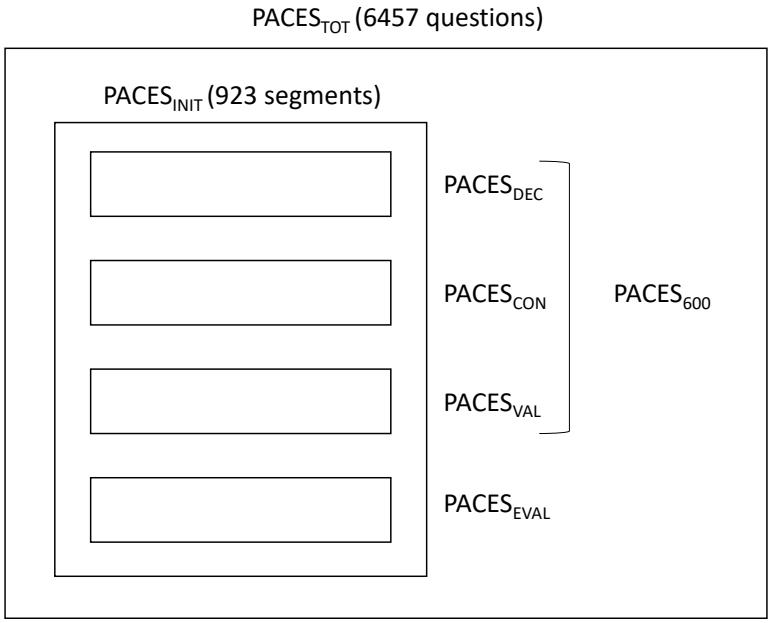


Figure 4.1 – Découpage du corpus PACES

Annotateur	**Dim1**	**Dim2**	**Dim3**	**Dim4**
Kappa entre expert 1 et expert 2	0.83	0.76	0.70	0.85 (56*)

(*) le nombre de chevauchement de segments entre les deux annotateurs

Tableau 4.1 – Les valeurs de Kappa entre les annotations manuelles des deux experts sur $PACES_{EVAL}$ de l'étape d'évaluation

Corpus	$PACES_{TOT}$	$PACES_{INIT}$	$PACES_{600}$	$PACES_{DEC}$	$PACES_{CON}$	$PACES_{VAL}$	$PACES_{EVAL}$
Nb. questions	6457	752	600	200	200	200	152
Nb. segments	8465	923	723	240	245	238	200

Tableau 4.2 – Caractéristiques des corpus PACES utilisés

4.2 Schéma de codage proposé à partir de l'annotation manuelle

Le résultat principal de cette étape était la création d'un schéma de codage des questions présenté dans le Tableau 4.4. Ce schéma de codage est présenté avec une liste non exhaustive de mots-clés pour donner une idée de leur nature (une liste des expressions régulières plus détaillée à partir des mots-clés est donnée en Annexe B). Mais nous avons également établi quelques principes à partir de l'étape d'annotation manuelle qui ont été utiles pour annoter automatiquement le reste des données :

4.2.1 Principes d'annotation

— Toutes les questions doivent être simples (c'est-à-dire les questions combinées doivent être préalablement segmentées en plusieurs questions simples, tableau 4.2).

— L'annotation doit être unique sur chaque dimension (c'est-à-dire qu'une question simple peut correspondre à la catégorie "approfondir un concept" ou "ré-expliquer un concept", mais pas les deux), mais il est possible de ne pas avoir d'annotation sur certaines dimensions (par exemple, une question peut ne pas être sur un exemple, ni un schéma, ni une correction).

— La dimension 4 (vérification) est annotée uniquement pour les questions identifiées comme vérification de la dimension 1. Cette dimension est la plus difficile à identifier automatiquement et pourrait idéalement nécessiter une analyse sémantique, ou au moins une représentation du contenu du cours sous forme de vecteur de mots, pour mieux identifier si des connaissances sont liées au cours ou non (mais cela irait à l'encontre de la généricité recherchée de l'approche).

— En cas de non identification de mots-clés associés à une des différentes catégories possibles dans la question, la dimension correspondante est annotée avec la valeur '0'.

4.2.2 Discussion

Ce qui nous guide dans ce schéma de codage, comme dit auparavant, est l'intention de l'élève et la réaction pédagogique de l'enseignant. Nous avons donc analysé certaines réactions de l'enseignant par exemple une demande de "ré-explication" et "explication" d'un concept. En effet, l'intention de l'élève qui demande une "ré-explication" (réaction attendue par l'enseignant est de ne pas de présenter le concept de la même manière comme il a été déjà vu en cours) est différente de celle qui demande une simple "explication". D'ailleurs, il est toujours explicite qu'un concept a déjà été vu en cours dans les questions de ce corpus, ce qui nous a amené à

séparer les demandes de "ré-explication" et d'"explication" dans la dimension 1 pour faire la différence entre ces deux intentions. D'un autre côté, l'élève pourrait demander à l'enseignant de lui expliquer un concept sous forme d'"exemple" ou "schéma", mais l'enseignant pourrait choisir de répondre à la question selon la modalité d'explication la plus convenable.

4.2.3 Utilisabilité du schéma de codage par un tiers

Afin d'évaluer la compréhensibilité de notre schéma de codage par une personne non impliquée dans la création du schéma, l'échantillon $PACES_{EVAL}$ a été annoté par un annotateur novice (qui n'a pas été impliqué dans cette étude et n'avait pas l'habitude non plus d'annoter des phrases). Les valeurs Kappa obtenues entre chacun des deux annotateurs experts et l'annotateur novice sont données dans le tableau 4.3.

Annotateur	Dim1	Dim2	Dim3	Dim4
Kappa entre expert 1 et novice	0.59	0.38	0.56	0.71 (48*)
Kappa entre expert 2 et novice	0.67	0.38	0.55	0.78 (57*)

(*) le nombre de chevauchement de segments entre les deux annotateurs

Tableau 4.3 – Les valeurs de Kappa entre différentes annotations manuelles sur $PACES_{EVAL}$ de l'étape d'évaluation

Les valeurs Kappa obtenues dans le tableau 4.3 peuvent nous laisser espérer que notre schéma de codage peut être utilisé même par d'autres personnes qui n'ont pas été impliquées dans cette étude.

Dim1	**Type de question**	**Description**	**Mots-clés**
Ree	Ré-expliquer / redéfinir	Demander de revenir sur un concept déjà expliqué en cours	Ré-expliquer, redéfinir, répéter, revenir, rappeler, résumer, reprendre, etc.
App	Approfondir un concept	Approfondir une connaissance, clarifier une ambiguïté ou demander plus de détails pour mieux comprendre	Expliquer, détailler, préciser, développer, décrire, etc.
Ver	Validation / vérification	Vérifier ou valider une hypothèse	Est-il, peut-on, faut-il, etc.
Dim2	**Modalité d'explication**	**Description**	**Mots-clés**
Exe	Exemple	Exemple d'application (cours/exercice)	Exemple, etc.
Sch	Schéma	Schéma d'application ou explication sur ce dernier	Schéma, représentation, etc.
Cor	Correction	Correction d'un exercice en cours/examen	correction, réponse, etc.
Dim3	**Type d'explication**	**Description**	**Mots-clés**
Def	Définir	Définir un concept ou un terme	Définir, signifier, veut dire, etc.
Man	Manière (comment ?)	Demander la manière de procéder	Comment, etc.
Rai	Raison (pourquoi ?)	Demander la raison	Pourquoi, raison, etc.
Rol	Rôles (utilité ?)	Demander l'utilité / fonction	Rôle, utilité, fonction, etc.
Lie	Lien entre concepts	Vérifier le lien entre deux concepts, le définir	Différence entre, correspond, relation, similaire, etc.
Dim4	**Type de vérification (facultatif)**	**Description**	**Mots-clés**
Err	Erreur / contradiction	Détecter une erreur/ contradiction en cours ou dans l'explication de l'enseignant	Erreur, contradiction, comment est-ce possible, etc.
Con	Connaissances liées au cours	Vérifier une connaissance	(*)
Exa	Examen	Vérifier une connaissance attendue dans l'examen	Concours, faut-il savoir, par cœur, etc.

(*) Pas de mots-clés spécifique à cette sous-catégorie car les connaissances en cours nécessitent une analyse sémantique. Par conséquent, cette sous-catégorie n'est annotée que si la question de vérification n'est pas une vérification d'erreur ou d'examen (c'est à dire Err et Exa n'ont pas été marquées comme valeur pour cette question).

Tableau 4.4 – Schéma de codage créé à partir de l'annotation manuelle

4.3 Similarités entre taxonomies

Bien que notre schéma de codage ait été développé indépendamment, nous avons remarqué a posteriori un certain chevauchement entre certaines taxonomies présentées dans l'état de l'art et notre schéma de codage. En particulier, la dimension 1 correspond bien aux catégories de plusieurs taxonomies, comme "Ré-expliquer / redéfinir" (Ree) qui est incluse dans la catégorie des questions pouvant faire l'objet d'une "investigation" [Chin & Kayalvizhi, 2002], la catégorie "Transformation" [Pedrosa de Jesus *et al.* , 2003] et la catégorie "Entrée" [Pizzini & Shepardson, 1991]. Les questions d'approfondissement (App) sont également incluses dans plusieurs catégories étudiées, tels que "Exploration" [Watts *et al.* , 1997], "Transformation" [Pedrosa de Jesus *et al.* , 2003] et "Réflexion approfondie" [Marbach-Ad & Sokolove, 2000]. La modalité d'explication (Dim2) semble moins courante, avec seulement la notion d'"exemple" (Exe) incluse dans la catégorie "Confirmation" de [Pedrosa de Jesus *et al.* , 2003] et "Exemple" de [Graesser & Person, 1994]. De même, seules les catégories "lien entre concepts" (Lie) et "définir" (Def) en dimension 3 sont incluses respectivement dans les catégories de questions pouvant faire l'objet d'une investigation et autres [Chin & Kayalvizhi, 2002]. Notons également que certaines catégories de la dimension 4 se retrouvent dans d'autres taxonomies, comme la catégorie "erreur/contradiction" (Err) qui correspond à la catégorie "Elaboration" [Watts *et al.* , 1997] et "Malentendu / idée fausse" [Marbach-Ad & Sokolove, 2000], les questions visant à vérifier des "connaissance en cours" (Con) sont incluses dans la catégorie "Non investigation" [Chin & Kayalvizhi, 2002] et correspondent partiellement à la catégorie "Sortie" [Pizzini & Shepardson, 1991]. Le Tableau 4.5 résume ces chevauchements.

4.4 Automatisation de taxonomie de questions

Afin d'annoter l'ensemble des questions posées par les étudiants, il était indispensable pour nous d'automatiser la classification des questions. C'est une tâche qui demande beaucoup de temps et d'effort manuellement étant donné le nombre important des questions posées par les étudiants. Nous avons développé dans un premier temps, un outil semi-automatique pour annoter le corpus $PACES_{TOT}$ (et à terme pour l'utiliser en ligne pour analyser les questions collectées). Ensuite, nous avons utilisé d'autres techniques de classification entièrement automatiques et comparé les performances de chacune d'entre elles. Les approches considérées ici sont les suivantes :

1. Annotateur à base de règles d'expert
2. Annotateur à base d'approches statistiques : avec deux variantes, une utilisant uniquement des techniques d'apprentissage automatique et l'autre y associant TF-IDF (Fréquence du terme - Fréquence Inverse de document)
3. Annotateur à base d'ensembles (hybride)

4.4.1 Objectif de l'étude

Notre objectif est de construire un système de classification de chaque question sur chaque dimension, en considérant les dimensions comme indépendantes. Cette tâche étant donc considérée comme quatre tâches de classification multi-classes et peut être effectuée selon différentes approches (règles d'expert, statistiques et ensemblistes). Par conséquent, ceci nous permettra éventuellement de répondre à la deuxième question de recherche introduite en section 1.1.

Dim1	Type de question	[Chin & Kayalvizhi, 2002]	[Watts *et al.* , 1997]	[Pedrosa de Jesus *et al.* , 2003]	[Graesser & Person, 1994]	[Marbach-Ad & Sokolove, 2000]	[Pizzini & Shepardson, 1991]
Ree	Ré-expliquer / redéfinir	Incluse dans Non investigation	-	Incluse dans Transformation	-	-	Incluse dans Entrée
App	Approfondir un concept	-	Incluse dans Exploration	Incluse dans Transformation	-	Inluse dans Réflexion approfondie	-
Ver	Validation / vérification	Investigation (partiellement)	Consolidation (partiellement)	Confirmation (partiellement)	Vérification (partiellement)	-	Sortie (partiellement)
Dim2	Modalité d'explication						
Exe	Exemple	-	-	Incluse dans Confirmation	Incluse dans Exemple	-	-
Sch	Schéma	-	-	-	-	-	-
Cor	Correction	-	-	-	-	-	-
Dim3	Type d'explication						
Def	Définir	Incluse dans Non investigation	-	Incluse dans Confirmation	Incluse dans Définition	Incluse dans Définition	-
Man	Manière (comment)	-	-	-	-	-	-
Rai	Raison (pourquoi)	-	-	-	-	-	-
Rol	Rôles (utilité)	-	-	-	-	-	-
Lie	Lien entre concepts	Incluse dans Investigation	-	-	Comparaison (partiellement)	-	Incluse dans Traitement
Dim4	Type de vérification						
Err	Erreur / contradiction	-	Incluse dans Élaboration	Incluse dans Confirmation	-	Incluse dans Malentendu	-
Con	Connaissances liées au cours	Incluse dans Non investigation	-	-	-	-	Sortie (partiellement)
Exa	Examen	-	-	-	-	-	Sortie (partiellement)

Tableau 4.5 – Résumé des similarités entre taxonomies existantes et le schéma de codage proposé

4.4.2 Protocole d'évaluation

Afin de réaliser l'apprentissage et la classification, le corpus des questions annotées manuellement est découpé en deux ensembles. Le premier est utilisé pour l'apprentissage tandis que le second permet de tester le modèle sur de nouvelles questions. Le choix des ensembles d'apprentissage et de test est crucial dans la procédure d'évaluation des différents systèmes d'annotation proposés. Par exemple, utiliser trop peu de questions pour l'apprentissage ne permet pas de développer un modèle robuste capable de prédire correctement les types de questions utilisées pour les tests. A l'inverse, laisser trop peu de questions pour les tests rend les performances de prédiction peu significatives, notamment dans la mesure où certaines catégories sont moins fréquentes que d'autres. Il faut également tenir compte de la représentation de chaque type de questions dans chacune des deux ensembles. En effet, utiliser uniquement des questions mal segmentées dans l'ensemble d'apprentissage ne permettra pas d'évaluer correctement des questions bien segmentées lors des tests. Enfin, les ensembles d'apprentissage et de test doivent être distincts afin de ne pas surévaluer la performance.

La validation croisée permet d'éviter ce genre de problèmes. L'idée consiste à diviser l'ensemble de questions annotées manuellement en k échantillons distincts, puis à utiliser un échantillon pour les tests et les $k-1$ restants pour l'apprentissage. En répétant cette opération pour les k échantillons, nous obtenons une classification pour chaque question de la base. Pour des valeurs de k élevées (par exemple 10) cela permet d'utiliser 90% de la base pour l'apprentissage, ce qui suffit généralement à établir un modèle pertinent. L'inconvénient de cette approche est la nécessité de générer k modèles, et donc d'effectuer k apprentissages. Toutefois, nos jeux de données étant réduits, nous aurons tendance à privilégier une valeur de k élevée car d'une part le temps d'apprentissage de chaque modèle est suffisamment faible, et d'autre par ce n'est pas vraiment un critère principal pour choisir le modèle ici, puisque nos calculs ne sont pas effectués en ligne.

Il est important de créer ces échantillons aléatoirement, de façon à pouvoir répéter les expériences et observer la reproductibilité des résultats. En effet si les erreurs sont deux fois plus importantes d'un essai à l'autre, cela signifie que le modèle est très fortement sensible au choix des questions et que la capacité de généralisation du modèle est faible. Donc, il faudra bien s'assurer que les résultats sont reproductibles.

Dans les sections 4.4.3 et 4.4.4, nous avons considéré le corpus $PACES_{INIT}$ (*cf.* tableau 4.2) annotés manuellement selon les 4 dimensions du schéma de codage pour l'entraînement et le test des annotateurs automatiques. Dans un premier temps, nous allons distinguer manuellement les ensembles d'apprentissage et de test afin de comparer les classifieurs sur la même base de questions (puisque nous présentons différentes approches non seulement statistique mais aussi à base de règles d'expert). Pour ceci, le corpus $PACES_{600}$ de l'étape de validation a été utilisé pour entraîner les classifieurs, et le corpus $PACES_{EVAL}$ de l'étape d'évaluation (*cf.* section 4.1) a été utilisé pour tester leur performance. Nous n'utilisons donc pas de validation croisée.

Ensuite, dans la section 4.4.5, nous allons utiliser la validation croisée pour entraîner et tester l'annotateur hybride à partir des données de méta-niveau (les étiquettes prédites des 200 segments par chaque modèle).

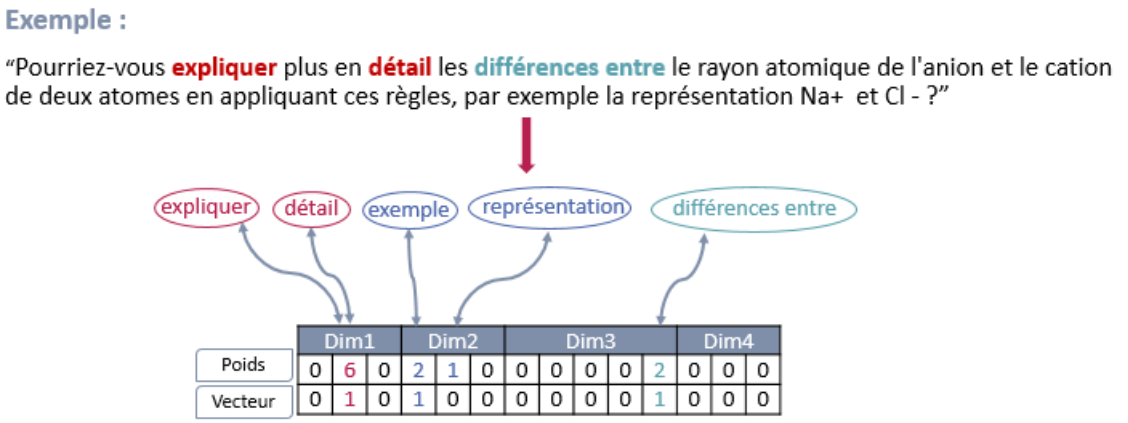


Figure 4.2 – Exemple d'une annotation automatique à base de règles d'une question à l'aide de mots-clés pondérés

4.4.3 Annotation automatique à base de règles d'expert (RE)

4.4.3.1 Méthodes et techniques

Nous avons développé un outil à base de règles, qui permet de segmenter les questions, d'identifier les mots-clés depuis les segments et de leur associer un poids, comme détaillé dans les paragraphes suivants.

Tout d'abord, à partir d'une question d'un étudiant, nous identifions les mots-clés représentatifs (définis manuellement à l'étape de découverte dans la section 4.1) de chaque valeur dans chaque dimension à l'aide d'une correspondance de chaînes de caractères et d'un ensemble d'expressions régulières (*cf.* Figure 4.2) sur le corpus $PACES_{600}$. Les expressions régulières ont été utilisées pour élargir les mots-clés identifiés pendant l'annotation manuelle (*cf.* Tableau en annexe B).

Chaque question est automatiquement segmentée en plusieurs segments (segment peut être une phrase ou un bout de phrase selon la phrase formulée et si elle a été bien ou mal ponctuée) en fonction d'un système de détection de la limite de la phrase intégré dans NLTK [Kiss & Strunk, 2006], qui est l'un des systèmes de traitement automatique de la langue [TAL] fonctionnant en français[2]. Ce système, repose sur une approche non supervisée, a été largement testé sur différentes langues et sur différents genres de textes. Il permet d'obtenir de bons résultats sans autres modifications ou ressources spécifiques à la langue. Bien que les questions de certains élèves puissent être mal rédigées et mal formulées, la méthode de segmentation semble fonctionner assez bien dans ce contexte. Il convient également de noter qu'en pratique, lors de l'annotation manuelle des segments, aucun des experts humains n'a trouvé une situation où il estimait que le segment qui leur était fourni aurait dû être plus segmenté qu'il ne l'était. Cette phase de segmentation est commune à tous les annotateurs, y compris ceux fondés sur les autres approches.

Ensuite, pour chaque segment de la question, pour chaque dimension, nous marquons le segment dans cette dimension en fonction de la valeur à laquelle sont associés le plus

2. Autres outils de TAL par exemple : CoreNLP, Spacy, OpenNLP, etc.

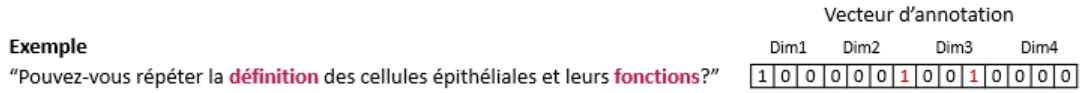


Figure 4.3 – Exemple de l'ambiguïté de l'annotation automatique à base de règles d'une question

de mots-clés (par exemple pour la dimension 1, un segment avec deux mots-clés associés à la valeur "ré-expliquer" et un mot-clé associé à la valeur "validation" serait marqué comme une question de "ré-expliquer"). Au final chaque segment, identifié par les valeurs associées à chaque dimension, est représenté par un vecteur d'annotation binaire (comme le montrent les figures 4.2 et 4.3).

Le problème qui peut survenir dans ce processus est la présence d'une ambiguïté au sein d'une dimension (par exemple, dans l'exemple de la figure 4.3, les deux catégories de dimension 3 "Def" et "Rol" sont annotées dans la même question). Pour résoudre ce problème, l'annotateur automatique utilise un ensemble de poids associés à chaque mot-clé de chaque dimension (par ex. "expliquer" : 7, "quoi/comment" : 3), et définis à l'aide des du corpus $PACES_{600}$ (723 segments). Plus précisément, ces pondérations ont été déterminées en deux étapes : premièrement, les annotateurs experts ont empiriquement associé une pondération entre 1 et 9 à chaque mot-clé, selon ce qu'ils pensaient être très marginales (1), significatives (5) ou très significatives (9) associées à une dimension donnée. Ensuite, en deuxième étape, l'annotateur automatique a été utilisé sur les 723 segments annotés manuellement, et les poids ont été ajustés manuellement (en rajoutant ou en supprimant 1) sur certains mots clés pour les segments dont l'annotation manuelle et automatique étaient différentes, jusqu'à l'obtention d'un accord complet sur presque tous les 723 segments. Dans l'exemple de la figure 4.2, l'introduction de poids aide l'annotateur à choisir entre "exemple" et "représentation" pour la valeur associée à la dimension 2. L'annotation automatique est donc sensible aux variations de poids, néanmoins, après les ajustements de poids, seulement 1% des questions étaient encore considérées comme ambiguës (*cf.* Figure 4.3, les mots "définition" et "fonctions" avaient le même poids dans la dimension 3) et dans ce cas l'annotateur à base de RE choisira une seule valeur aléatoirement. Enfin, nous avons utilisé le corpus $PACES_{EVAL}$ de 200 segments, qui n'ont pas été utilisés dans la phase d'ajustement du poids, pour évaluer l'annotateur à base de règles et calculer les valeurs de Kappa par dimension.

En pratique, nous avons implémenté l'annotateur automatique en utilisant NLTK (Natural Language Toolkit), une bibliothèque majeure en Python permettant de traiter les données en langage naturel. En particulier, nous avons utilisé le module "Punkt Sentence tokenizer" pour diviser chaque phrase en une liste de segments, et chaque segment en une liste de mots. Ce tokenizer, déjà entraîné à la langue française, utilise l'algorithme non supervisé susmentionné pour construire un modèle de mots d'abréviations, de collocations, de mots et de caractères et pour marquer le début et la fin d'une phrase.

4.4.3.2 Résultats

Les valeurs de Kappas par dimension pour les annotations provenant des deux annotateurs experts sont données dans le Tableau 4.6.

Annotateur	Dim1	Dim2	Dim3	Dim4
Kappa entre expert 1 et annotateur à base de RE	0.66	0.50	0.63	0.55
Kappa entre expert 2 et annotateur à base de RE	0.76	0.69	0.70	0.65
Kappa moyen des deux annotateurs	0.71	0.60	0.67	0.60

Tableau 4.6 – Les valeurs de Kappa entre l'annotation automatique et manuelle

Nous considérons que les valeurs de Kappa moyen des deux annotateurs (entre 0.60 et 0.71) sont suffisamment élevées pour appliquer l'annotateur automatique au corpus $PACES_{TOT}$, même si ce type de décision est toujours partiellement arbitraire [Artstein & Poesio, 2008]. De plus, ces valeurs comprennent des phrases non annotées (phrases sur lesquelles aucune expression régulière ne correspondait à la phrase et sont étiquetées '0' par l'annotateur automatique dans toutes les dimensions). Par exemple, sans questions non annotées, les valeurs de Kappa de la dimension 1 sont comprises entre 0.80 et 0.86, pour les deux annotateurs humains.

Nous avons décidé de prendre les valeurs de Kappa obtenues entre l'annotateur à base de RE et l'expert 2 (*cf.* la deuxième ligne du Tableau 4.6), revue par l'expert 1, comme référence pour la suite afin de comparer les performances des différents systèmes d'annotation automatique proposés ici.

4.4.4 Annotateur automatique à base d'approches statistiques

Les approches fondées sur les statistiques ont mené à une révolution dans la façon d'aborder les tâches du TAL. Cependant, pour être efficaces, ces méthodes s'appuient généralement sur des corpus importants, et il n'était pas clair a priori si une technique basée sur une approche statistique pouvait améliorer la performance de l'annotateur à base de règles. Deux étapes ont été effectuées manuellement dans la conception de l'annotateur à base de règles : (a) l'identification des mots-clés, et (b) l'association d'un ensemble de mots-clés à une valeur pour chaque dimension. Nous proposons dans cette section une annotation entièrement automatisée basée sur TF-IDF et des techniques d'apprentissage automatique (AA) pour annoter le corpus des questions.

4.4.4.1 Transformation de questions en vecteurs de mots

Nous allons présenter ici les principales étapes de la préparation des textes pour l'extraction des connaissances (*cf.* figure 4.4). Différentes techniques (lemmatisation, racinisation, filtrage de mots, recherche de synonymes dans WordNet, etc.) sont utilisées afin de réduire de manière importante la taille du vocabulaire avant les analyses.

Tokenisation. La tokenization est un procédé permettant de découper un texte en token (ou phrases). Avant de découper les phrases en tokens, nous avons supprimé d'abord les accents (permet d'avoir moins de variabilité sur les mots extraits sans introduire du bruit ou de problème sémantique dans notre cas), la ponctuation (jugée utile pour la phase de segmentation présentée précédemment) et les nombres qui peuvent rajouter du bruit au corpus et perturber le reste du traitement.

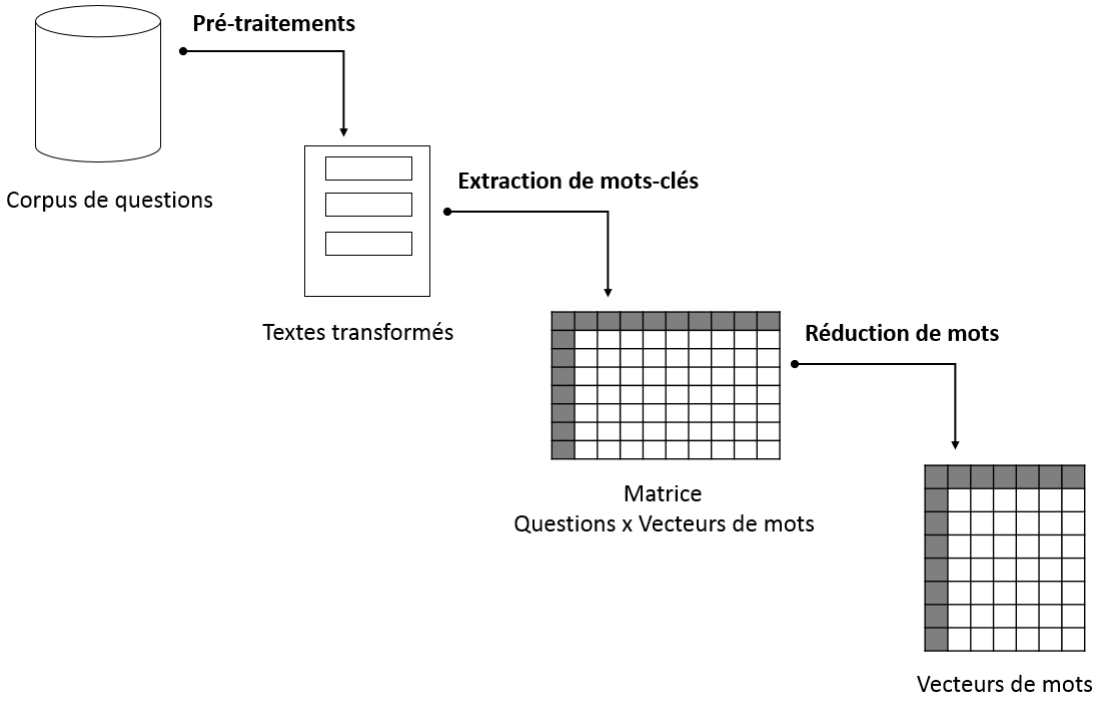


Figure 4.4 – Processus de préparation de questions

Lemmatisation/racinisation. La taille du vocabulaire initial d'un corpus de documents peut parfois dépasser une centaine de milliers de termes distincts. Des pré-traitements permettent une première réduction de la taille du vocabulaire en éliminant des mots jugés non pertinents. L'étiquetage morpho-syntaxique (grammatical) est la tâche qui précède la lemmatisation, et qui consiste à identifier la catégorie grammaticale associée à chaque mot au sein de la phrase (nom, verbe, adjectif, etc.). Cette identification permet ensuite la lemmatisation, procédé qui prend en compte la flexion [3] des mots afin de les ramener au "lemme" ou "forme canonique" (masculin singulier pour un adjectif, infinitif pour un verbe conjugué, etc.). La lemmatisation permet de regrouper toutes les flexions rencontrées dans le texte sous un unique mot. Un des algorithmes les plus utilisés pour la lemmatisation et adapté à la langue française est l'algorithme de Schmid [1994]. Contrairement à la lemmatisation, la racinisation (en anglais "stemming") permet de regrouper les différentes formes d'un mot autour d'une racine commune (sans prendre en compte la flexion des mots). L'algorithme de Porter [1980] est le plus connu pour la racinisation des mots en anglais, d'autres algorithmes ont été développés sur les même principe pour traiter les mots en français [Paternostre *et al.* , 2002]. Les algorithmes de lemmatisation/racinisation développés pour la langue française sont jugés moins performants qu'en anglais. Néanmoins, ceci ne nous a pas empêché d'utiliser la racinisation dans notre travail, qui est moins sensible aux fautes d'orthographes que la lemmatisation et n'a pas besoin du contexte pour fonctionner, afin d'agréger différentes formes en un mot commun (l'algorithme [4] [Porter, 2006] a été intégré dans NLTK sous Python).

3. Variation de la forme d'un mot en fonction de facteurs grammaticaux
4. Voir http://snowball.tartarus.org/

Filtrage de mots. Le filtrage des mots non pertinents ou "stopwords" en anglais, consiste à supprimer des mots vides (ex : "le", "la", "les", "un", "une", "de", "du", etc.). Ces mots sont éliminés car ils ne contiennent pas d'information sémantique. Certains outils linguistiques proposent des listes de mots vides [5]. Une alternative permettant également l'élimination des mots vides consiste à filtrer les mots les plus fréquents. Nous avons créé une liste non exhaustive de mots vides au lieu d'utiliser les filtres automatiques de tous les mots vides et perdre de l'information. En effet, nous avons considéré qu'il pouvait parfois être important de garder certains mots généralement jugés comme vides pour notre analyse (ex : "il", "et", "ou", "mais", "on", etc.). Ces mots peuvent être significatifs en tant qu'unigramme [6] ("ou", "mais") ou bigramme [7] ("faut-il", "peut-on") dans l'annotation des questions selon notre schéma de codage. Nous avons également filtré certains mots jugés sans intérêt pour notre analyse ("merci", "bonjour", "svp", "cordialement", etc.).

WordNet. WordNet est une base de données lexicale, développée par Miller [1995], reliant des concepts sémantiques entre eux dans une ontologie selon une variété de relations sémantiques (tels que synonymes et hyperonymes). Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour. Nous avons utilisé le WordNet libre du français [Sagot & Fišer, 2008] afin de diminuer la diversité lexicale et de renforcer certaines expressions pour le traitement qui suit (l'idée est la même que celle derrière le procédé de lemmatisation ou racinisation). Il existe dans WordNet, plusieurs types de concepts : les termes (issus du lexique) et le sens des termes (appelé synset). Par exemple, le terme «bois», peut désigner l'idée de matière (1), l'idée d'un regroupement d'arbres (2), l'idée d'une famille d'instruments de musique (3) ou encore l'idée d'un accessoire de golf (4). L'objectif est de ramener différentes expressions synonymes à une même expression dans les questions (par exemple pour la dim 3-Rai, les mots synonymes «cause», «raison» et «motif» sont remplacés dans le texte par «pourquoi»). Plutôt que de regrouper les concepts autour de leur forme lexicale, WordNet, à travers les synsets, regroupe les concepts en fonction de leur sens en contexte. Ainsi, WordNet construit deux types de relation :

1. entre un synset et les termes employés pour le dénoter en contexte
2. entre un synset et sa définition en contexte

D'autres types de relation sont alors introduits entre synsets :

— hyponymie : si X est hyponyme de Y alors X spécifie Y (le sens 2 de «bois» est une spécification de la «végétation»)

— hyperonymie : si X est un hyperonyme de Y alors X généralise Y («végetation» est un hyperonyme du sens 2 de «bois»)

— méronymie : si X est un méronyme de Y alors X est composé de Y (le sens 2 de «bois» est un méronyme d'«arbre» dans le sens de la flore)

— holonymie : si X est un holonyme Y alors X est une partie de Y (l' «arbre» dans le sens de la flore est un holonyme de «bois»)

Le second sens du mot bois («Bois2») peut s'exprimer avec les mots «bois» ou «forêt» dans un contexte donné (ils sont synonymes et donc interchangeables en contexte).

Nous avons utilisé l'ensemble de mots-clés extraits manuellement (*cf.* Table 4.4) pour chercher des synonymes et hyperonymes à deux niveaux hiérarchiques (quand c'est disponible) dans

5. Par exemple, se reporter à l'algorithme *Snowball*
6. Une sous-séquence d'un terme
7. Une sous-séquence de deux termes

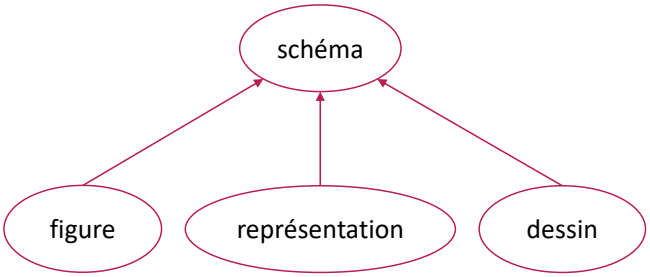


Figure 4.5 – Exemple de synset dans WordNet

WordNet (*cf.* figure 4.5). Les mots extraits par WordNet peuvent ne pas être tous pertinents pour nos dimensions. Nous filtrons donc manuellement les synonymes et hyperonymes non liés à notre contexte (*cf.* Tableau de mots WordNet en annexe C).

Pour mieux comprendre l'ensemble des pré-traitements illustrés sur la figure 4.6, nous présentons un exemple de question sur chacune des étapes des pré-traitements suivantes :

- Exemple de question : "Est-il possible de donner une représentation de liaisons entre atomes ?"

- Après suppression d'accents et ponctuation : "Est il possible de donner une representation de liaisons entre atomes"

- Après Tokenization : 'Est', 'il', 'possible', 'de', 'donner', 'une', 'representation', 'de', 'liaisons', 'entre', 'atomes'

- Après racinisation : 'Est', 'il', 'possibl', 'de', 'don', 'une', 'represent', 'de', 'liaison', 'entre', 'atom'

- Après filtrage de mots : 'Est', 'il', 'possibl', 'don', 'represent', 'liaison', 'entre', 'atom'

- Après WordNet : 'Est', 'il', 'possibl', 'don', '**schema**', 'liaison', 'entre', 'atom'

- Après extraction d'unigrammes et bigrammes : 'Est', 'il', 'possibl', 'don', 'schema', 'liaison', 'entre', 'atom', 'Est il', 'il possibl', 'possibl don', 'don schema', 'liaison entre', 'entre atom'

- Après filtrage d'unigrammes vides : 'Est', 'possibl', 'don', 'schema', 'liaison', 'entre', 'atom', 'Est il', 'il possibl', 'possibl don', 'don schema', 'liaison entre', 'entre atom'

Une fois ces pré-traitements effectués, nous avons extrait les unigrammes et bigrammes sur le corpus $PACES_{600}$ et appliqué un deuxième filtre (*cf.* figure 4.6) pour éliminer les unigrammes vides (ex : "il", "on", *etc.*). Ensuite, les poids des expressions (unigrammes/bigrammes) sont calculés selon deux méthodes différentes : (1) TF-IDF (décrit dans la section suivante), (2) compter les occurrences ('1' si le mot est dans la question, '0' sinon). Chacune des 723 questions est représentée par un vecteur de mots selon (1) ou (2). Nous avons fina-

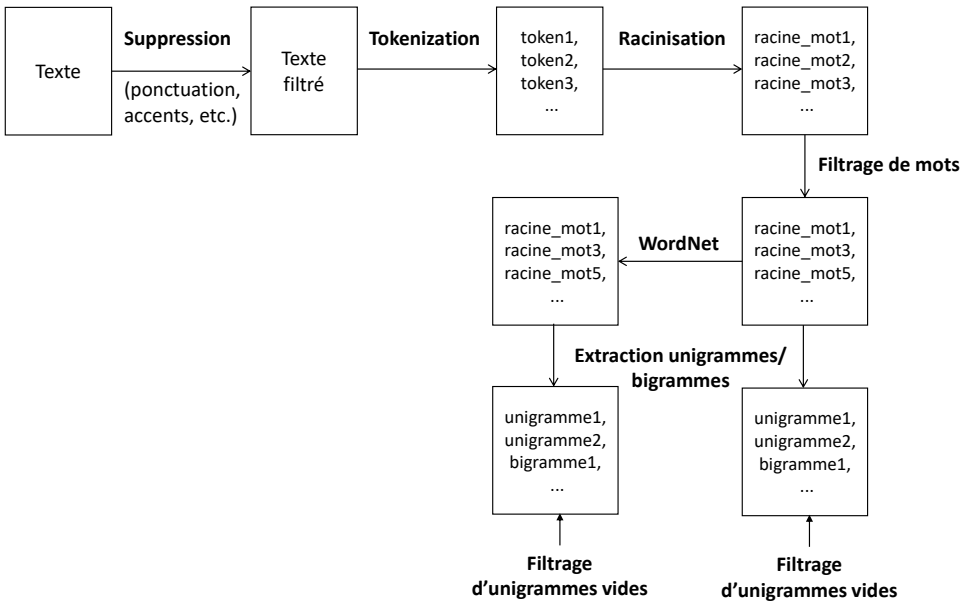


Figure 4.6 – Ensemble des pré-traitements

lement réduit le nombre de mots-clés extraits à l'aide d'une technique de sélection d'attributs (suppression des unigrammes et bigrammes les moins fréquents et corrélés) pour conserver les mots-clés les plus importants et les plus significatifs (comme décrit dans la figure 4.4).

Sélection d'attributs. Des techniques sont utilisées afin de réduire le nombre de termes étudiés tout en conservant l'information pertinente. La réduction du nombre de termes permet de gagner en temps de calcul lorsque les algorithmes sont appliqués à des matrices de très grande dimension comme c'est le cas en fouille de données textuelles. Elle permet parfois également d'améliorer les résultats obtenus grâce à l'élimination des termes constituant du bruit. Nous distinguons différentes mesures (TF-IDF, nombre d'occurrences) permettant d'attribuer des poids aux termes. Les termes sont ensuite ordonnés selon leur poids et éliminés sur la base d'un seuil minimal fixé ou du nombre de termes que l'on souhaite conserver. Ce choix est arbitraire et demeure une tâche délicate car il dépend du type d'application. Nous avons mené plusieurs expérimentations afin de décider du seuil/du nombre de termes à conserver.

4.4.4.2 Annotation à base d'apprentissage automatique (AA)

Notre objectif ici est de faire une classification multi-classe pour chacune des 4 dimensions (à partir des méthodes d'apprentissage automatique) après la transformation de questions en vecteurs de mots (pré-traitements de questions, extraction mots-clés et réduction d'attributs), comme illustré dans la figure 4.7.

Méthode. Pour associer automatiquement des mots-clés aux valeurs associées à chaque dimension (ex : "ré-expliquer"), nous avons essayé différents algorithmes d'apprentissage automatique sur les segments représentés par des vecteurs annotés : (a) avec les mots-clés extraits manuellement (issus des expressions régulières, *cf.* annexe B), (b) avec l'ensemble des mots-clés extraits automatiquement (*cf.* figure 4.7) utilisant WordNet (les mots-clés extraits et réduits

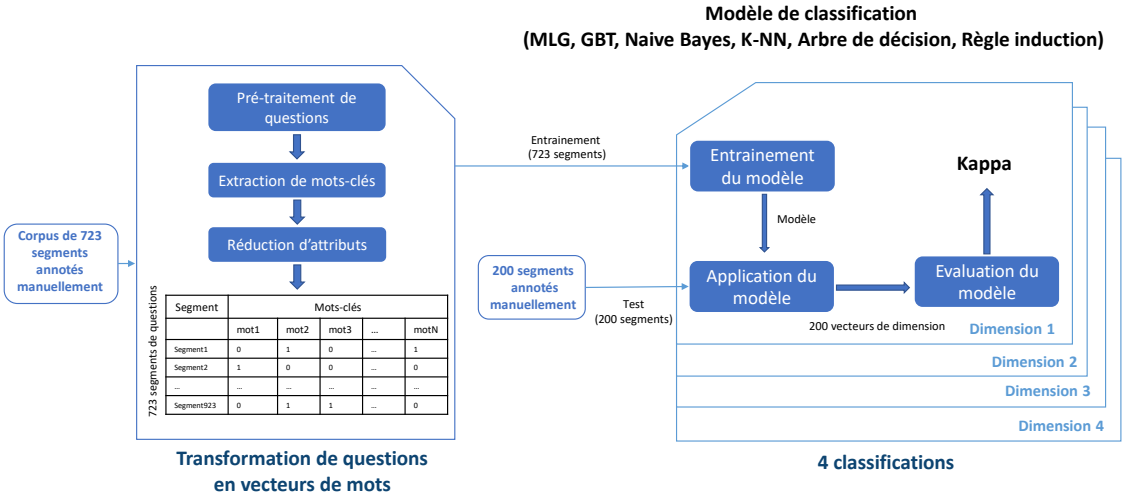


Figure 4.7 – Processus d'annotation à base d'apprentissage automatique

après le traitement WordNet), et (c) avec l'ensemble des mots-clés extraits automatiquement (*cf.* figure 4.7) sans WordNet (utiliser l'ensemble de mots-clés extraits sans traitement WordNet).

Nous avons essayé 6 techniques de classification différentes (Modèle linéaire généralisé [GLM], Gradient Boosted Trees [GBT], Arbre de décision [DT], K voisins les plus proches [KNN], Règle d'induction [RI] et Naive Bayes [NB]) sur chaque dimension séparément (les dimensions étant conçues comme indépendantes). Le principe de chacune de ces méthodes est expliqué plus loin dans cette section. Chaque classifieur est entraîné en prenant en entrée un ensemble de vecteurs de mots représentant les 723 segments de l'ensemble d'entraînement ($PACES_{600}$), et l'étiquette à prédire est la valeur qui a été associée manuellement au segment dans cette dimension. Le modèle est ensuite évalué sur un échantillon indépendant de 200 segments ($PACES_{EVAL}$) sans étiquettes, afin d'assurer une bonne estimation de la performance sur des données non vues. Enfin, nous avons calculé les valeurs Kappa entre les valeurs prédites par le classifieur et les valeurs correspondantes trouvées par l'annotation manuelle (*cf.* Figure 4.7).

Résultats. Les valeurs de Kappa par dimension pour les deux méthodes d'extraction (manuelle et automatique) représentées par les versions (a), (b) et (c) provenant de l'annotation à base d'apprentissage automatique avec les 6 techniques de classification supervisées sont fournies dans la première partie du Tableau 4.7.

Discussion. L'extraction de tous les unigrammes/bigrammes puis faire une sélection d'attributs a donné un ensemble de 270 mots (à comparer à 149 mots identifiés manuellement). Parmi ces mots-clés, 50 mots étaient communs aux deux ensembles de données (par exemple : "rôle", "pourquoi", "pouvons-nous", "faut-il"), mais la majorité d'entre eux se trouvaient uniquement dans la liste des mots-clés automatiques (par exemple : "méthode", "aussi", "enzyme").

Nous avons également comparé les annotateurs à base de techniques d'apprentissage automatique (AA) et de règles d'expert non seulement par rapport aux valeurs de Kappa, mais aussi par rapport au nombre de segments annotés à '0' sur toutes les dimensions des 200 segments de l'échantillon de test. Nous avons trouvé 16 segments annotés à '0' manuellement (c-à-d. ne sont pas des questions), 22 segments ont été annotés à '0' par l'annotateur à base de règles d'expert d'expert et aucun segment par l'annotateur à base d'AA. Il est important de noter que sur les 22 segments non identifiés comme des questions par l'annotateur à base de règles, il y en a 10 qui sont communs avec l'annotation manuelle. Donc dans le cas de l'annotateur à base de règles d'expert, on est capable dans une certaine mesure de détecter les segments qui ne correspondent pas à des questions du schéma de codage, alors qu'avec l'annotateur à base d'AA, tout segment est forcément annoté dans une dimension (c'est à dire pas de question annotée sous forme de vecteur [0,0,0,0], et '0' signifie qu'aucune étiquette du schéma de codage ne correspondait à cette dimension) ce qui augmente le nombre de faux positifs. C'est d'ailleurs une des motivations pour vouloir combiner différents classifieurs ensemble par la suite.

Techniques d'apprentissage automatique

Il existe de nombreux algorithmes d'apprentissage automatique permettant la prédiction de variables, continues ou discrètes. Nous avons retenu 6 d'entre eux : Modèle linéaire généralisé [GLM], Gradient Boosted Trees [GBT], Arbre de décision [DT], K voisins les plus proches [KNN], Règle d'induction [RI] et Naive Bayes [NB]. Nous avons retenu ces algorithmes car :

— Ils peuvent s'adapter aux différents problèmes (classification, régression) et types de caractéristiques (continues, discrètes). Cette flexibilité constitue un avantage par rapport à certaines méthodes de prédiction, prévues pour un seul et même type.

— Ils ne fonctionnent pas sur le même principe, issues de différentes familles ; il est donc possible que certains soient plus adaptés à des données particulières.

— Ils sont déjà implémentés et correctement documentés dans RapidMiner[8], logiciel que nous avons choisi d'utiliser pour cette étape.

Un autre algorithme appelé les machines à vecteurs de support [SVM] est largement utilisé dans la classification automatique de messages (comme vu dans l'état de l'art, section 2.2.3). Nous allons le présenter également ici mais il ne sera utilisé que dans la deuxième partie de ce travail (*cf.* Chapitre 8) puisqu'il n'est pas adapté à tous types de données (données non binaires).

Modèle linéaire généralisé [GLM]

Le modèle linéaire généralisé (GLM), introduit par Nelder & Wedderburn [1972], est une extension des modèles linéaires classiques. Il s'agit d'une généralisation en termes de loi de probabilité d'une part, mais aussi en termes de lien à la linéarité. L'hypothèse sur la distribution associée à chaque modélisation est alors remplacée par une propriété de linéarité commune à tous les modèles, et par une relation espérance-variance. La classe des modèles linéaires généralisés permet l'analyse des données non gaussiennes, et notamment des données discrètes. Un modèle linéaire généralisé est caractérisé par trois hypothèses : une hypothèse sur la distribution de la variable à expliquer, une hypothèse sur l'expression de la linéarité (faisant intervenir les variables explicatives), et une hypothèse sur le lien à la linéarité (c'est-à-dire le lien entre la variable réponse et les variables explicatives).

8. Outil de traitement et d'analyse de données

Avantages : Il s'adapte à plusieurs distributions (normale, poisson, *etc.*). Ce modèle est assez flexible, puisqu'il peut détecter automatiquement le type de distribution adapté aux données et également rapide en exécution.

Inconvénients : GLM est assez sensible à la corrélation entre les variables explicatives. Une forte corrélation entre deux variables entraîne une redondance d'informations et peut fausser l'interprétation des résultats. Il est donc nécessaire de détecter les corrélations entre les variables utilisées (en utilisant des techniques de sélection d'attributs).

Nous avons conservé les paramètres proposés par défaut par RapidMiner, configurés pour choisir automatiquement la valeur de la distribution de la variable et l'expression de la linéarité adaptée aux données.

Gradient Boosted Trees [GBT]

Les forêts boostées sont construites itérativement selon le principe du boosting, qui consiste à construire une famille de modèles qui sont ensuite agrégés par une moyenne pondérée des estimations ou un vote afin d'améliorer l'ajustement et la performance du modèle. Gradient Boosted Trees (GBT) est un ensemble de modèles d'arbres de régression ou de classification. L'objectif du boosting est de réduire l'erreur de prédiction à chaque itération (une itération correspond à la création d'un arbre) afin d'améliorer la précision des arbres, selon le même principe que les algorithmes d'optimisation reposant sur une descente de gradient. La phase de la prédiction de la forêt correspond à la combinaison des prédictions de chaque arbre. Bien que l'augmentation du nombre des arbres augmente leur précision, elle diminue également la vitesse et la capacité d'interprétation de l'homme. La méthode du gradient boosting permet de généraliser l'arborescence pour minimiser ces problèmes.

Avantages : Globalement GBT permet d'améliorer la performance des arbres de décision (boosting)

Inconvénients : Nous avons noté comme inconvénient le temps d'exécution (lent). Cependant, ce critère n'est pas une limite pour notre travail puisque nous disposons du temps nécessaire pour l'exécution du modèle (tous les calculs effectués hors ligne).

Cet algorithme a été testé avec différentes valeurs pour le nombre d'arbres (entre 10 et 20) et la profondeur maximale (entre 3 et 10). Nous constatons que nous obtenons souvent les mêmes performances. Nous avons donc conservé les valeurs par défaut, qui sont de 5 pour la profondeur maximale des arbres et 20 pour le nombre d'arbres nécessaires.

Arbre de décision [DT]

Un arbre de décision est un modèle qui est à la fois descriptif et prédictif. La popularité de la méthode repose en grande partie sur sa simplicité et son interprétabilité par l'humain. Il s'agit de trouver un partitionnement des individus que l'on représente sous la forme d'un arbre de décision. L'objectif est de produire des groupes d'individus (ici, les mots-clés) les plus homogènes possibles du point de vue de la variable à prédire (dans notre cas, catégorie de questions). Il est d'usage de représenter la distribution empirique de l'attribut à prédire sur chaque sommet (nœud) de l'arbre. De nombreux algorithmes d'arbres de décision sont proposés : CART, ID3 (dont les versions ultérieures sont C4.5 et C5.0), CHAID et arbres aléatoires (Random Forest). Dans notre cas, nous utilisons l'algorithme C4.5 qui se base sur une mesure de l'entropie dans l'échantillon d'apprentissage pour produire le modèle. L'avantage du recours à l'entropie est que l'algorithme opère sur des données symboliques, que ce soient

des variables catégorielles (comme les catégories de questions dans notre cas) ou numériques discrètes.

Avantages : Contrairement à beaucoup d'outils de classification, les arbres de décision sont extrêmement intuitifs et fournissent une représentation graphique sous forme d'arbre, facile à lire et à interpréter. Elles permettent d'identifier très rapidement les variables les plus discriminantes d'un jeu de données.

Inconvénients : l'instabilité est un inconvénient non-négligeable. En effet, en modifiant même très légèrement l'échantillon d'apprentissage, il est courant d'obtenir un arbre très différent même si l'efficacité prédictive reste elle, relativement stable.

Différents paramètres permettent de contrôler la construction de chaque arbre. En effet, si les arbres sont trop profonds, des problèmes de sur-apprentissage peuvent apparaître. Pour pallier à cet inconvénient, de nombreuses méthodes d'élagage, appelées aussi post-élagage ont été proposées pour les différents algorithmes d'arbre de décision. Nous avons appliqué ces techniques durant nos expérimentations. Nous avons constaté qu'utiliser comme critère de découpage la "précision" plutôt que les autres ("rapport au gain", "gain d'information", "index de Gini" et "moins carré") donne de meilleurs résultats. Nous conservons la valeur 10 proposée par défaut par comme limite de profondeur.

K plus proches voisins [KNN]

C'est une des méthodes de classification / régression des plus simples qui donne généralement de bons résultats. Il a été employé avec succès dans de nombreux domaines, et a engendré toute une famille de classifieurs connus sous le nom de *classifieurs paresseux* (Lazy). Dans ces systèmes, le seul traitement effectué au cours de la phase d'apprentissage est le stockage des exemples sous une forme optimale de façon à pouvoir les extraire ensuite rapidement. Tous les calculs sont reportés à la phase de classification (d'où le terme de paresseux).

Dans le but de prédire la classe y d'une nouvelle instance x, l'idée principale de la méthode est de trouver les k plus proches voisins de cette instance, puis d'avoir recours à la classe majoritaire parmi les k plus proches voisins trouvés. Pour que cet algorithme soit efficace, il faut une bonne métrique de mesure de distance entre les instances, notamment afin que les attributs non discriminants ne soient pas pris en compte. Il existe plusieurs fonctions de mesure de distance, notamment, la distance euclidienne, la distance de Manhattan, la distance de Minkowski, celle de Jaccard, la distance de Hamming, etc. La mesure de distance est choisie en fonction des types de données manipulées.

Avantages : Cette méthode de classification est facile à comprendre (peut être manipulée facilement par des personnes du non-domaine).

Inconvénients : Le fait de déporter tous les calculs qui pourraient être faits pendant la phase d'apprentissage à la phase de classification, peut éventuellement alourdir la tâche de classification. A noter également, qu'il n'est pas facile de choisir une bonne métrique de classification pour mesurer la similarité entre les données.

Il n'existe pas de solution efficace pour choisir une bonne valeur pour le paramètre k. Ce choix relève d'un compromis (si k est trop petit, le nombre d'exemples qui prennent part à la décision est faible et les exemples bruités peuvent alors jouer un rôle néfaste important. Si k est trop grand, l'hypothèse de localité n'est plus respectée car des exemples très éloignés sont

sélectionnés pour participer au vote). Après plusieurs expérimentations, la valeur de k a été fixée entre 3 et 6 selon la classe à prédire (la valeur par défaut est 5). Bien que nous avons testé plusieurs métriques ("similarité de Jaccard", "distance euclidéenne", *etc.*), la distance euclidienne est celle qui donne les meilleurs résultats et considérée la plus adaptée à nos données ici. Elle peut également être utilisée pour des données numériques et catégorielles.

Règle d'induction [RI]
L'induction de règles de classification fait partie des approches qui produisent de manière incrémentale un ensemble de règles de la forme : *Si* Condition *Alors* Conclusion ; où condition représente une suite de conjonctions de couples «attribut-valeur», et conclusion la classe d'affectation. La méthode produit une base de règles ordonnées. Lors de la classification d'un individu, la première règle est évaluée. Si elle n'est pas déclenchée, on passe à la suivante, etc. Si aucune règle n'est activée, une règle par défaut est utilisée. L'apprentissage de l'ensemble des règles est souvent comparé à l'apprentissage de l'arbre de décision, puisqu'une des méthodes de construction de l'ensemble de règles est de construire des arbres de décision successifs et de conserver la branche qui classifie correctement le plus d'exemples

Avantages : Parmi les avantages de cette méthode, sa représentation facile à lire et à comprendre (même par les non spécialistes du domaine). Notons également, qu'il n'y a pas de collision entre les règles (une seule règle est activée à chaque classification d'individu)

Inconvénients : Son inconvénient principal est de générer un grand nombre de règles, et ce même pour des ensembles de taille moyenne et d'être sensible aux données bruitées.

Nous avons conservé le paramètre par défaut "gain d'information", comme critère de classification.

Naive Bayes [NB]
Les méthodes naïve Bayes sont considérées parmi les méthodes probabilistes les plus connues. Le classifieur bayésien naïf est, comme son nom l'indique, associé au théorème de Bayes, théorème permettant de calculer les probabilités conditionnelles d'un évènement connaissant certains a priori. Ce classifieur repose principalement sur l'hypothèse que les variables sont indépendantes entre elles, ce qui permet alors de simplifier considérablement la détermination des densités de probabilités. Souvent, l'estimation des paramètres de chaque densité de probabilité repose sur la technique du maximum de vraisemblance. Une fois de telles densités estimées, il suffit de calculer les probabilités a posteriori à l'aide de la règle de Bayes pour obtenir le label d'une nouvelle observation.

Avantages : le classifieur bayésien naïf requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification (moyennes et variances des variables). En effet, l'hypothèse d'indépendance des variables permet de se contenter de la variance de chacune d'entre elles pour chaque classe, sans avoir à calculer de matrice de covariance. Il s'agit donc d'un classifieur rapide et robuste.

Inconvénients : l'hypothèse d'indépendance des variables (mots-clés) dans ce modèle explique qu'il soit qualifié de naïf ou de simpliste, puisque c'est une hypothèse très forte et pas toujours vraisemblable. Ses performances sont limitées quand il s'agit d'une grande quantité de données à traiter. En effet, si le nombre de mots augmente, alors le nombre des dépendances entre l'ensemble des mots augmentent également, et donc, la vérification de l'hypothèse de Naïve Bayes est de moins en moins plausible.

Les machines à vecteurs de support [SVM]

Les machines à vecteurs de support ou séparateurs à vaste marge (SVM) est une technique introduite par Vapnik [1998], qui repose sur la recherche d'un hyperplan, qui sépare au mieux les classes entre elles. Elle a tout d'abord été développée dans le cas d'une classification binaire, où l'hyperplan représente une droite, puis étendue au cas multiclasses. Le problème revient alors à trouver une frontière de décision séparant au mieux les différentes catégories. Cette frontière porte le nom d'hyperplan optimal. On parle de maximisation de la marge, où la marge se définit comme étant la distance entre le point le plus proche de l'hyperplan et l'hyperplan lui-même. Les SVM reposent essentiellement sur deux idées clés : la notion de marge maximale et la notion de fonction noyau. Les fonctions noyau permettent de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension, sans la nécessité de savoir la transformation à appliquer pour le changement d'espace. Enfin, les SVM représentent une méthode de classification pour des données binaires seulement ; dans le cas où l'on serait dans un contexte multi-classes, il est nécessaire d'avoir recours à l'algorithme proposé par Friedman [1996].

Avantages : L'avantage est que la solution produite par SVM correspond à l'optimum d'une fonction convexe. Elle ne possède donc pas plusieurs optima locaux, mais un optimum global. Cet optimum est à la recherche d'une hypothèse possédant de bonnes capacités de généralisation à partir d'un espace d'hypothèses donné.

Inconvénients : Cette technique connaît certaines limitations, telles que le choix du noyau, qui peut s'avérer être un problème pour de nombreuses applications réelles (la fonction noyau a souvent des paramètres libres, comme la largeur de marge avec un noyau gaussien, et la recherche de la valeur optimale de ces paramètres libres ne correspond plus à la recherche du minimum d'une fonction convexe).

Nous avons conservé les paramètres par défaut proposés pour le choix du noyau, le noyau linéaire. Cette fonction est considérée comme la plus simple des fonctions noyau et correspond à un produit scalaire dans un espace de grande dimension.

Comme on a pu le voir dans ce résumé des méthodes, chaque algorithme possède des avantages et inconvénients, qui dépendent notamment du type de données à exploiter. Des méthodes sont largement utilisées plus que d'autres, ce qui peut être expliqué par leur facilité d'usage, la facilité du choix des hyperparamètres, la performance du modèle, la stabilité des résultats, le temps de calcul nécessaire, etc.

Mesures d'évaluation

Plusieurs mesures d'évaluation sont disponibles pour quantifier la performance d'un classifieur. Nous allons présenter ici les 4 mesures suivantes :

1. Kappa
2. Précision
3. Rappel
4. F-mesure

Il en existe cependant de nombreuses autres telles que "courbe ROC", "aire sous la courbe ROC (AUC)", *etc.*

Kappa

Le coefficient κ suppose dans sa modélisation du hasard que la répartition des éléments entre catégories peut être différente pour chaque annotateur. Dans ce cas, la probabilité pour qu'une question soit assignée dans une catégorie (k) pour la dimension K est le produit de la probabilité que chaque annotateur l'assigne dans cette catégorie. L'accord aléatoire (A_e^K, liée au hasard) est donc calculé de la façon suivante, n_c1k étant le nombre d'affectations à k pour l'annotateur 1 et i la taille de l'échantillon :

$$A_e^\kappa = \sum_{k \epsilon \kappa} \frac{n_{c1k}}{i} \cdot \frac{n_{c2k}}{i} \tag{4.1}$$

κ se calcule ensuite comme suit, avec A_0 l'accord observé :

$$\kappa = \frac{A_0 - A_e^\kappa}{1 - A_e^\kappa} \tag{4.2}$$

Précision

La précision d'un classifieur par rapport à une certaine classe (autrement dit, par rapport à une certaine modalité de la variable à prédire), se mesure comme la proportion d'individus, parmi tous ceux pour lesquels le classifieur a prédit cette classe, qui appartiennent réellement à celle-ci. Pour la classe c, on calcule la précision comme suit (le nombre d'individus pour lesquels le classifieur a prédit la classe correspond à l'effectif marginal pour $y_1 = c$) :

$$P_c = \frac{\text{Nombre de segments correctement attribués à la classe c}}{\text{Nombre de segments attribués à la classe c}}$$

Rappel

Le rappel d'un classifieur par rapport à une certaine classe se mesure, quant à lui, comme la proportion d'individus, parmi tous ceux qui appartiennent réellement à cette classe, pour lesquels le classifieur a prédit cette classe. Pour la classe c, on calcule le rappel ainsi (le nombre d'individus qui appartiennent réellement à la classe c correspond à l'effectif marginal pour $y_2 = c$) :

$$R_c = \frac{\text{Nombre de segments correctement attribués à la classe c}}{\text{Nombre de segments appartenant à la classe c}}$$

F-mesure

On peut résumer les mesures de précision de rappel par rapport à une classe en un seul indicateur, en calculant la moyenne harmonique :

$$F_c = 2 \cdot \frac{P_c \cdot R_c}{P_c + R_c}$$

4.4.4.3 Annotation automatique à base de TF-IDF

Nous avons utilisé TF-IDF [Salton, 1989] pour calculer les poids des termes. L'objectif de TF-IDF est d'estimer comment les mots d'un document donné sont représentatifs de ce document par rapport à un ensemble plus large de documents. Il combine deux métriques complémentaires : la *Fréquence du Terme* (TF) qui prend en compte le nombre d'occurrences du terme dans le document et l'*Inverse de la Fréquence en Document* (IDF) qui prend en compte le nombre d'occurrences du terme dans le corpus. TF donne donc un poids plus

élevé aux termes courants et un poids plus faible aux termes rares. L'inconvénient est que certains mots qui sont communs dans un document donné mais aussi communs dans tous les documents pourraient finir par avoir un poids qui sur-représente leur importance réelle. IDF règle ce problème en ajustant le poids en fonction de l'importance générale du terme. L'équation 4.3 décrit la méthode de calcul des valeurs de poids TF-IDF individuelles pour chaque terme (mot). Nous avons effectué deux mesures de calcul différentes de TF-IDF sur le corpus de 723 questions.

$$W_{ik} = TF_{ik} \cdot \log\left(\frac{N}{n_k}\right) \tag{4.3}$$

où :

W_{ik} = pondération TF-IDF pour un terme k dans le document Q_i
TF_{ik} = fréquence du terme k dans le document Q_i
IDF_{ik} = inverse de la fréquence du terme k en document $Q_i = \log\left(\frac{N}{n_k}\right)$
N = nombre total de documents dans le corpus de segments
n_k = nombre de questions dans le corpus qui contiennent le terme k

La *première version* consiste à calculer quatre TF-IDF séparément pour chacune des 4 dimensions, pour extraire les mots-clés sur autant de catégories qu'il y a par dimension. Pour une dimension donnée, toutes les questions annotées manuellement dans chaque catégorie (ex : "Ré-expliquer") étaient considérées comme des documents (ex : sur la dimension 1, document1 est l'union des questions annotées comme "Ree"). Les questions non annotées (étiquetées '0') pour cette dimension sont aussi considérées comme un document. Chaque document (ensemble de questions) est converti en un vecteur de poids-mot correspondant, où chaque poids-mot représente la mesure TF-IDF pour le mot dans le document. Le poids TF-IDF (W_{ik}) a été attribué pour chaque terme k dans le document i (i est le nombre de documents dans cette dimension, par ex : i variant de 1 à 3 pour la dimension 1). Afin de classer les nouvelles questions, nous avons utilisé les pondérations TF-IDF calculées sur chaque valeur de dimension séparément de l'échantillon de 723 questions. Nous avons attribué des poids TF-IDF calculés sur l'échantillon d'entraînement pour les mots correspondants sur l'échantillon de test de 200 questions. Ensuite, nous avons choisi la fonction de classement la plus simple qui consiste à additionner les poids TF-IDF pour chaque question sur chaque valeur de dimension. Par conséquent, pour chaque question, pour chaque dimension, nous marquons la question dans cette dimension en fonction de la valeur qui a le poids maximum [MAX]. Enfin, nous avons calculé les valeurs Kappa entre les valeurs trouvées par ce modèle (appelé [**TF-IDF+MAX**] qui correspond à la première version de TF-IDF et qui consiste à annoter une question selon le poids maximum calculé) pour cette dimension, et les valeurs correspondantes trouvées par l'annotation manuelle (*cf.* première colonne du Tableau 4.8). Une question est donc représentée par un vecteur de mots pondérés par TF-IDF+MAX (le processus d'annotation du modèle [**TF-IDF+MAX**] est illustré dans la Figure 4.8).

Trois stratégies de filtres de mots (stopwords) ont été testés avec l'annotation TF-IDF pour avoir les meilleurs résultats.

- Dans la première version, nous n'avons pas appliqué de filtre dans le pré-traitement du texte afin de savoir l'impact des données bruitées sur l'annotation TF-IDF. Les valeurs de Kappa ($\kappa_1 = 0.65$, $\kappa_2 = 0.22$, $\kappa_3 = 0.7$ et $\kappa_4 = 0.58$) ont été calculées sur l'échantillon de test avec l'annotation manuelle comme référence pour les dimensions 1, 2, 3 et 4 respectivement.

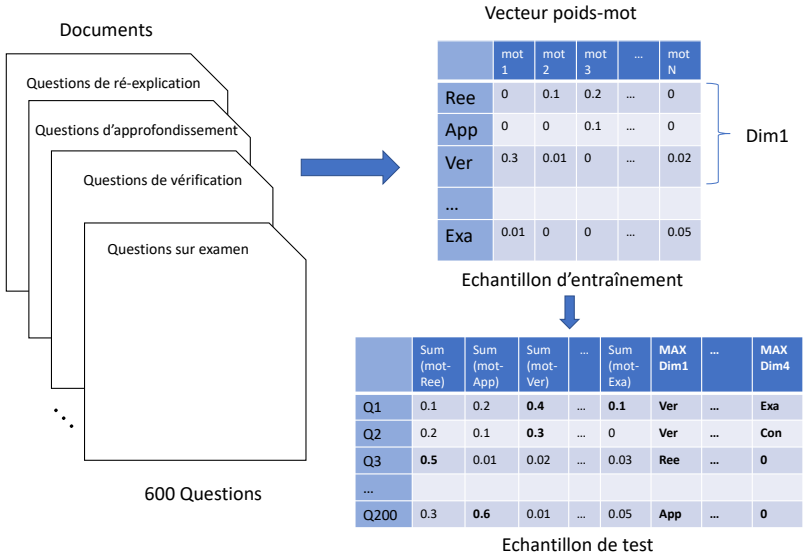


Figure 4.8 – Processus d'annotation TF-IDF avec la valeur maximale des sommes de pondérations sur chaque dimension

- Dans la deuxième version, nous avons filtré les mots vides du texte avant l'extraction de mots-clés. Les valeurs de Kappa obtenues sur chacune des 4 dimensions sont les suivantes : $\kappa_1 = 0.62$, $\kappa_2 = 0.41$, $\kappa_3 = 0.67$ et $\kappa_4 = 0.61$.
- Dans la troisième version, nous avons appliqué deux filtres avant et après l'extraction de mots-clés. Le premier filtre consiste à filtrer les mots vides avant l'extraction de mots-clés et le deuxième est appliqué après le calcul de TF-IDF sur les questions d'entraînement (suppression des mots vides avec une pondération TF-IDF importante [9] pour diminuer l'impact des mots bruités sur l'annotation). Les valeurs de Kappa obtenues sont reportées dans la première colonne du Tableau 4.8.

Les valeurs de Kappa obtenues sur les 4 dimensions sont assez proches sur les trois versions susmentionnées (sauf pour la dimension 2 dans la première version, qui était la plus impactée par les données bruitées). Nous avons choisi la troisième version qui permet d'éliminer le plus de données bruitées sans diminuer la performance. Nous avons également effectué des pré-traitements au corpus de questions avec WordNet (*cf.* section précédente) et sans WordNet. Les résultats obtenus étant similaires en termes de performances, nous avons décidé de conserver la version incluant le pré-traitement avec WordNet, comme il devrait intuitivement mieux généraliser aux variations des questions existantes.

Dans la *deuxième version du calcul*, TF-IDF a été calculé sur le corpus de 723 questions sans distinguer les différentes dimensions. Les questions ne sont pas regroupées par valeur de dimension, mais chaque question du corpus est considérée comme un document (*c-à-d.* 723 documents au total). Le document est ensuite converti en un vecteur de poids-mot correspondant, chaque poids associé à un mot représentant la mesure TF-IDF du mot de la question. Enfin, nous avons utilisé les vecteurs de mots comme entrée pour les techniques d'apprentissage automatique afin de prédire la valeur associée à la question dans cette dimension (comme

9. Les mots avec une pondération TF-IDF supérieur à 0.1 ont été filtrés à la main pour ne pas perdre des mots-clés pertinents

Dim.	GLM	GBT	NB	K-NN	DT	RI	Baseline
Extraction manuelle de mots-clés							
Dim1	**0.72**	0.71	0.23	0.71	0.41	0.70	0.28
Dim2	0.51	0.70	0.06	0.51	**0.72**	0.13	0.58
Dim3	**0.68**	0.60	0.33	0.47	0.55	0.57	0.30
Dim4	0.51	0.36	0.25	**0.61**	0.32	0.11	0.13
Extraction automatique de mots-clés avec WordNet							
Dim1	0.69	0.70	0.28	0.60	**0.73**	0.69	0.28
Dim2	0.10	0.74	0.10	0.50	**0.79**	0.37	0.58
Dim3	**0.68**	0.64	0.37	0.61	0.59	0.60	0.30
Dim4	0.63	**0.66**	0.34	0.60	0.48	0.63	0.13
Extraction automatique de mots-clés sans WordNet							
Dim1	0.73	0.69	0.33	0.56	**0.74**	0.66	0.28
Dim2	0.58	0.81	0.12	0.48	**0.85**	0.29	0.58
Dim3	**0.70**	0.65	0.35	0.60	0.57	0.62	0.30
Dim4	0.63	**0.67**	0.46	0.59	0.10	0.47	0.13

Tableau 4.7 – Kappa entre l'annotation automatique obtenue par différentes méthodes d'apprentissage, une classification baseline et l'annotation manuelle référence

	TFIDF +					
Dim.	Max	GLM	GBT	NB	KNN	DT	RI
Dim1	0.66	0.69	**0.71**	0.47	0.62	0.46	0.61
Dim2	0.39	**0.73**	0.69	0.12	0.56	0.49	0.36
Dim3	**0.66**	0.59	0.60	0.43	0.58	0.37	0.52
Dim4	0.58	**0.71**	0.63	0.37	0.60	0.19	0

Tableau 4.8 – Kappa entre l'annotation automatique obtenue par TF-IDF + différentes méthodes d'apprentissage automatique et l'annotation manuelle référence

décrit dans la section 4.4.4.2).

4.4.4.4 Résultats

Les valeurs kappa trouvées avec les deux annotateurs automatiques individuellement (AA et TF-IDF) sont fournies dans les tableaux 4.7 et 4.8 pour chaque dimension. Les résultats d'une classification basique à base d'AA avec des unigrammes (données brutes dans pré-traitements) et l'algorithme standard arbre de décision (DT) sont également donnés dans le tableau 4.7 (Baseline).

L'annotateur à base d'apprentissage automatique (AA) surpasse nettement l'annotateur à base de TF-IDF sur les trois dimensions 1, 2 et 3 (avec le classifieur DT sur les dimensions 1 et 2 et GLM sur la dimension 3). TF-IDF avec le classifieur GLM donne les meilleures performances sur la dimension 4. De plus, l'annotation basée sur AA sans WordNet fonctionne mieux que celle utilisant WordNet pour toutes les dimensions et en particulier la dimension 2. En revanche, les résultats de la classification basique (DT) sur des données brutes sans largement inférieurs à l'ensemble des résultats obtenus par l'annotateur à base d'AA sur des

données préalablement traitées (*cf.* 4.6), ce qui permet de montrer l'intérêt et l'utilité des pré-traitements effectués.

Nous notons également que l'annotateur expert à base de règles surpasse les deux annotateurs à base d'AA et TF-IDF uniquement sur la dimension 1, alors que leurs performances sont presque similaires sur la dimension 3.

4.4.5 Annotation automatique à base d'approches ensemblistes

L'idée des méthodes ensemblistes est de combiner plusieurs modèles pour obtenir un classifieur global dont la précision de prédiction est supérieure à celle de chacun d'entre eux [Koren, 2009; Munoz *et al.* , 2010]. Parmi les méthodes ensemblistes existantes, l'empilement (ou stacking en anglais) Wolpert [1992] consiste à agréger les prévisions issues de différents modèles. Il est particulièrement utilisé lorsque les types de modèles sont très différents.

Notre prochaine étape consiste à construire un modèle de classification plus performant pour améliorer l'identification automatisée des questions selon le schéma de codage fourni dans le Tableau 4.4. En utilisant l'approche du stacking, nous avons essayé différentes combinaisons de modèles, quel que soit le meilleur classifieur. Le stacking permet normalement systématiquement d'obtenir des meilleurs résultats en combinant des classifieurs, cependant dans notre cas, de part la taille limitée de notre ensemble d'entraînement, il existe une incertitude quant à la possibilité d'obtenir un résultat vraiment meilleur.

4.4.5.1 Méthode de stacking

Dans la première phase, un ensemble de 20 modèles de base a été créé (1 annotation à base de règles d'expert, 7 annotations TF-IDF et 12 annotations à base d'AA). Dans la deuxième phase, nous voulons entraîner un classifieur du méta-niveau qui combine les résultats des modèles de base (des approches ensemblistes sont cachées dans les algorithmes de niveau de base, tel que GBT). En d'autres termes, nous avons 20 prédictions pour chaque dimension pour chacun des 200 segments de questions de l'ensemble de test, ainsi que 20 annotations manuelles pour ces 200 segments qui fournissent une vérité de terrain, et nous voulons entraîner un modèle de classification utilisant certains sous-ensembles de ces 20 caractéristiques. Il est important de travailler sur les 200 segments et pas les 723, car sinon on risque de prendre en compte des questions qui ont déjà été vues par les modèles durant la phase d'apprentissage lorsqu'on teste le méta-classifieur. Nous avons entraîné le classifieur du méta-niveau en utilisant les six mêmes techniques de classification (GBT, GLM, DT, K-NN, NB et RI) mentionnées dans la section 4.4.4.2 pour chaque dimension séparément, en utilisant une validation croisée en 10 fois pour assurer une bonne estimation de la performance (*c-à-d.* entraînement des modèles sur 180 segments et test sur les 20 restantes). Enfin, nous avons calculé les valeurs de Kappa pour chaque modèle entre les valeurs trouvées par ce méta-modèle pour cette dimension, et les valeurs correspondantes trouvées par l'annotation manuelle. En ce qui concerne l'ensemble des caractéristiques que nous avons prises en compte, nous voulions envisager des combinaisons de différents ensembles d'approches. Nous avons donc considéré six combinaisons de méta-apprentissage décrites ci-dessous. Pour chacune des six combinaisons, l'entraînement a été réalisé quatre fois (une fois pour chacune des quatre dimensions - *cf.* Figure 4.9).

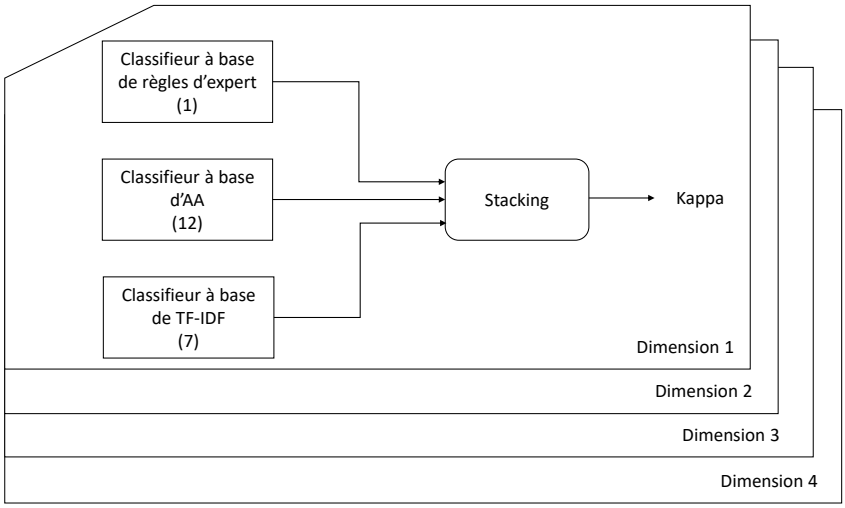


Figure 4.9 – Le processus global de stacking

(1) Empilement des modèles TF-IDF : Nous avons combiné les sorties des méthodes en utilisant chaque classifieur TF-IDF individuellement pour calculer les poids des mots-clés (*c-à-d.* 7 attributs en entrée pour chaque classifieur, *cf.* Tableau 4.8).

(2) Empilement des modèles TF-IDF avec l'annotation à base de règles : Nous avons combiné les sorties des modèles TF-IDF avec la sortie de l'annotateur à base de RE (*c-à-d.* 8 attributs en entrée pour chaque classifieur, *cf.* Tableaux 4.8 et 4.6).

(3) Empilement des méthodes d'AA : Nous avons combiné les sorties des méthodes d'annotation à base d'apprentissage automatique des deux combinaisons : traitement avec et sans WordNet (*c-à-d.* 12 attributs en entrée pour chaque classifieur, *cf.* Tableau 4.7).

(4) Empilement des méthodes d'AA avec l'annotation à base de RE : Nous avons combiné les sorties des méthodes d'annotation à base d'apprentissage automatique (avec et sans WordNet) avec la sortie de l'annotateur à base de règles (*c-à-d.* 13 attributs pour chaque classifieur, *cf.* Tableaux 4.7 et 4.6).

(5) Empilement des méthodes d'AA avec TF-IDF Nous avons combiné les sorties d'annotation à base d'apprentissage automatique (avec et sans WordNet) avec les sorties d'annotation à base de TF-IDF (*c-à-d.* 19 attributs pour chaque classifieur, *cf.* Tableaux 4.7 et 4.8).

(6) Empilement AA, TF-IDF et l'annotation à base de RE : Nous avons combiné les sorties d'annotation à base d'apprentissage automatique (avec et sans WordNet) avec les sorties TF-IDF et la sortie d'annotation à base de règles d'expert (*c-à-d.* 20 attributs pour chaque classifieur, *cf.* Tableaux 4.7, 4.8 et 4.6).

4.4.5.2 Résultats

Les valeurs de kappa trouvées avec les 6 techniques de classification pour chaque dimension sont fournies dans le tableau 4.9. Chaque modèle d'empilage a été entraîné individuellement sur chaque dimension et la valeur obtenue la plus élevée pour chaque dimension parmi les six classifieurs est indiquée en gras, pour chaque ensemble de caractéristiques considérées. Par exemple, sur la première ligne, nous voyons que lorsque l'on combine les 7 classifieurs TF-IDF qui prédisent la dimension 1, le meilleur résultat d'empilement est obtenu avec un arbre de décision (0.75), qui surpasse le meilleur classifieur TF-IDF individuel (0.71 avec GBT, *cf.* Tableau 4.8). Nous constatons que Naïve Bayes est souvent le meilleur classifieur d'ensemble parmi les 6 testés, donnant de meilleures performances sur un petit ensemble de données. Les meilleures performances globales entre les six ensembles de comparaisons sont indiquées par une étoile (*) : pour la dimension 1 et 4, il s'agit de Naïve Bayes combinant les classifieurs à base d'AA et règles d'expert, pour la dimension 2, il s'agit de Naïve Bayes combinant les classifieurs à base de TF-IDF et règles d'expert, et pour la dimension 3 il s'agit de GBT combinant également les classifieurs à base de TF-IDF et règles d'expert.

Lorsque l'on considère les combinaisons TF-IDF, on constate que la combinaison de plusieurs TF-IDF est plus performante que le TF-IDF de base sur les dimensions 1 et 3. Les valeurs de kappa sont globalement inférieures sur les dimensions 2 et 4, ce qui est probablement dû au déséquilibre de données d'entraînement sur ces dimensions (cela explique aussi pourquoi parfois un classifieur obtiendrait un kappa de 0 sur ces dimensions dans les différents tableaux). Par ailleurs, les divers classifieurs de TF-IDF combinés avec l'annotateur à base de RE surpasse à la fois TF-IDF de base et l'annotateur à base de RE, de même pour la combinaison de plusieurs TF-IDF. Des résultats similaires ont été obtenus pour le modèle TF-IDF combiné avec le classifieur à base d'apprentissage automatique, avec une performance légèrement supérieure à celle des classifieurs individuels. Globalement, si nous devons choisir qu'un seul ensemble de caractéristiques, la meilleure option est un ensemble hybride (TF-IDF avec annotateur à base de RE), qui surpasse en moyenne les combinaisons de modèles avec un kappa moyen de 0.77 (à partir des classifieurs donnant les meilleures performances sur chaque dimension, *c-à-d.* NB sur les dimensions 1 et 2, GBT sur la dimension 3 et K-NN sur la dimension 4).

Lorsqu'on considère les combinaisons impliquant des classifieurs basés sur l'AA, l'annotateur à base d'AA combiné avec l'annotateur à base de RE surpasse légèrement le modèle d'apprentissage automatique de base sur les dimensions 1, 3 et 4 par rapport aux autres combinaisons d'AA. Comme pour TF-IDF, l'ensemble hybride (AA avec annotateur à base de RE) donne un kappa moyen de 0.77 au lieu de 0.74 pour l'AA de base.

La combinaison des trois types d'approches (à base de RE, AA et TF-IDF) permet d'obtenir une performance similaire ou inférieure à celle des deux autres ensembles hybrides mentionnés précédemment.

4.4.6 Bilan

La figure 4.10 résume les performances (valeurs de Kappa obtenues avec les intervalles de confiance [Cohen, 1960] calculés) des quatre types de classifieurs (à base de RE, AA, TF-IDF, hybride) sur chacune des quatre dimensions (sur les six techniques de classification utilisées, nous avons reporté à chaque fois la valeur la plus élevée).

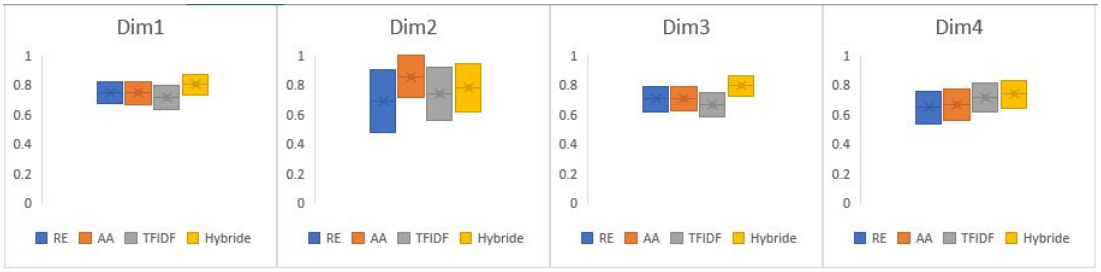


Figure 4.10 – Comparaison des performances des 4 classifieurs en termes de Kappa (centre de la barre) et d'intervalle de confiance (haut et bas de la barre) sur chaque dimension

Globalement, l'annotateur hybride semble surpasser légèrement les autres annotateurs en termes de performance (une amélioration a été observée surtout sur les dimensions 1 et 3). L'annotateur à base d'AA reste le plus performant sur la dimension 2, qui contient le plus grand nombre de données non équilibrées. Cette dernière observation nous a conduit à tester une technique permettant de rééquilibrer les données (SMOTE [Chawla *et al.* , 2002]) sur l'échantillon des 200 segments. Cependant l'utilisation de SMOTE n'a pas vraiment donné de meilleurs les résultats (même sur la dimension 2, $kappa = 0.80$).

4.5 Synthèse

Nous avons développé un schéma de codage de questions posées par les étudiants dans le cadre d'un environnement hybride. Nous avons reporté a posteriori les similarités entre les typologies de questions existantes et notre schéma de codage.

Nous avons également développé un système d'annotation semi-automatique à base de règles d'expert pour annoter l'ensemble de corpus de questions, que nous avons utilisé par la suite pour segmenter et annoter les questions des étudiants de PACES automatiquement (*cf.* chapitres 5 et 6). Bien qu'efficace sur les questions qu'il annote, l'annotation automatique dépend essentiellement des mots-clés pondérés manuellement, ce qui nous a amené à développer des systèmes d'annotation entièrement automatique basés sur des approches statistiques (des techniques d'apprentissage automatique et TF-IDF). La comparaison de performances des différents systèmes d'annotation (presque similaire en termes de performance et certaines d'eux peuvent surpasser les autres sur une dimension donnée, *cf.* Figure 4.10), nous a conduit à combiner ces modèles pour obtenir un annotateur hybride dont la performance est supérieure à celle de chacun d'entre eux.

Nous avons montré que même avec un petit ensemble d'entraînement (moins de 1000 questions), il peut être utile d'ajouter des approches basées sur l'AA pour compléter un annotateur conçu manuellement en utilisant une approche de stacking pour combiner les classifieurs entre eux. L'utilisation d'un ensemble hybride d'annotateurs basés sur l'apprentissage automatique (ou TF-IDF) avec un annotateur existant semble ici être la meilleure approche, en tirant parti des avantages de chaque approche.

Dans notre cas, le modèle ensembliste hybride a permis d'augmenter la performance sur presque toutes les dimensions. Il convient toutefois de noter qu'ici l'utilisation de WordNet pour réduire le vocabulaire n'a pas permis d'améliorer les performances des classifieurs.

L'une des limites à souligner est que nous n'avons considéré qu'un seul ensemble de données. L'augmentation de la valeur des kappas peut aussi parfois être considérée comme modeste, mais il faut mettre cela en perspective avec le fait que les codeurs humains utilisant ce schéma de codage peuvent rarement atteindre un kappa supérieur à 0.75 dans cette tâche. De plus, il faut souligner que les dimensions qui ont été améliorées étaient celles qui étaient les plus éloignées de la performance du codeur humain. Cela renforce donc l'interprétation consistant à dire que les annotateurs plafonnent aux alentours de 0.75 car c'est une limite liée à la subjectivité de la tâche, avec une vérité terrain elle-même potentiellement discutable sur certaines questions.

Modèles TF-IDF						
Dim.	**GLM**	**GBT**	**NB**	**K-NN**	**DT**	**RI**
Dim1	0.73	0.74	0.72	0.73	**0.75**	0.70
Dim2	0	0.35	**0.67**	0.49	0.51	0
Dim3	0.62	**0.70**	0.66	0.67	0.68	0.66
Dim4	0.55	0.67	0.68	**0.69**	**0.69**	0.67
Modèles TF-IDF + RE						
Dim.	**GLM**	**GBT**	**NB**	**K-NN**	**DT**	**RI**
Dim1	0.73	0.72	**0.76**	0.72	0.68	0.71
Dim2	0	0.30	**0.80***	0.66	0.48	0
Dim3	0.70	**0.79***	0.76	0.77	0.75	0.67
Dim4	0.60	0.66	0.72	**0.73**	0.67	0.65
Modèles AA						
Dim.	**GLM**	**GBT**	**NB**	**K-NN**	**DT**	**RI**
Dim1	0.76	0.73	**0.80**	0.76	0.71	0.68
Dim2	0.30	0.48	**0.77**	0.59	0.62	0
Dim3	0.62	0.71	0.71	**0.72**	0.70	0.65
Dim4	0.58	0.65	**0.72**	0.67	0.68	0.57
Modèles AA + RE						
Dim.	**GLM**	**GBT**	**NB**	**K-NN**	**DT**	**RI**
Dim1	0.77	0.77	**0.80***	0.76	0.70	0.69
Dim2	0.16	0.48	**0.77**	0.60	0.62	0
Dim3	0.64	**0.76**	0.71	0.73	0.66	0.64
Dim4	0.60	0.66	**0.74***	0.69	0.63	0.59
Modèles AA + TF-IDF						
Dim.	**GLM**	**GBT**	**NB**	**K-NN**	**DT**	**RI**
Dim1	**0.77**	0.73	**0.77**	0.76	0.71	0.68
Dim2	0.30	0.52	**0.78**	0.61	0.62	0
Dim3	0.66	0.75	0.71	**0.72**	0.70	0.62
Dim4	0.60	0.64	**0.71**	0.71	0.64	0.61
Modèles AA + TF-IDF + RE						
Dim.	**GLM**	**GBT**	**NB**	**K-NN**	**DT**	**RI**
Dim1	0.77	0.75	**0.78**	0.76	0.72	0.68
Dim2	0	0.56	**0.78**	0.58	0.62	0
Dim3	0.65	**0.77**	0.72	0.73	0.67	0.61
Dim4	0.61	0.63	**0.70**	0.69	0.63	0.61

Tableau 4.9 – Les valeurs de Kappa entre les modèles ensemblistes et l'annotation manuelle référence

Chapitre 5

Lien entre questions posées et comportement des étudiants

Pour étudier le lien entre les types de questions posées par les étudiants et leur comportement, nous allons utiliser le jeu de données introduit dans la section 3.1. Nous allons également utiliser l'annotateur à base de règles d'expert pour annoter automatiquement l'ensemble de questions posées par les étudiants de PACES à partir du schéma de codage présenté dans le tableau 4.4. Une fois les questions annotées automatiquement, nous allons tout d'abord comparer les étudiants qui ont posé des questions (Q) à ceux qui n'en pas posées (NQ) et analyser ensuite les types de questions à travers le clustering pour identifier le lien entre ces questions et le comportement des apprenants. Ces résultats ont donné lieu à une publication dans le journal JLA (Journal of Learning Analytics) (sections 5.1 et 5.2.3) [Harrak *et al.* , 2019c] et la conférence EIAH (Environnement Informatique pour Apprentissage Humain) (*cf.* section 5.4) [Harrak *et al.* , 2019b].

Pour aborder la troisième question de recherche posée en section 1.1 (QR3), qui est de savoir si les questions posées par un étudiant peuvent être informatives de ses caractéristiques (*cf.* Tableau 3.4), nous devons identifier des caractéristiques suffisamment génériques, c'est-à-dire qui ne sont pas dépendantes d'un seul cours en particulier (par exemple : nombre de questions posées sur un cours et nombre de séquences enseignées, par opposition à des variables particulières à un cours lié à sa sémantique comme "compréhension des mécanismes biocellulaires" qui ne s'applique qu'à certains cours). Nous avons donc décidé de considérer des données provenant de plusieurs cours (*cf.* Tableaux des unités d'enseignement du premier semestre 3.1 et deuxième semestre 3.2), en l'occurrence les quatre cours qui ont généré le plus de questions (*cf.* Tableau 3.3) : BCH, HBD, BCE et ANT (ce dernier étant le seul cours du 2ème semestre à avoir reçu un grand nombre de questions). Avant de comparer les différentes catégories d'élèves qui posent des questions, nous avons pensé qu'il pourrait être pertinent d'examiner d'abord comment les élèves qui posent des questions diffèrent de ceux qui ne le font pas, afin de contraster les éléments caractéristiques du comportement de questionnement avec ceux associés au simple fait de poser ou non des questions. Nous avons donc effectué une analyse exploratoire (sans hypothèse initiale à vérifier) des deux groupes d'étudiants (Q et NQ) sur l'ensemble de caractéristiques à notre disposition issues de nature différentes : performance, redoublement, assiduité, etc.

Cela nous a amenés à affiner notre QR3 en deux questions :

QR3.1 : Y a-t-il un lien entre le fait que les élèves posent des questions et leurs caractéristiques ?

QR3.2 : Y a-t-il des natures de questions caractéristiques de la performance des élèves ?

5.1 Comparaison des étudiants Q et NQ

5.1.1 Méthode

Pour chacun des 4 cours considérés, nous avons réparti les étudiants entre ceux qui ont posé des questions (groupe Q) sur ce cours et ceux qui n'en ont pas posé (groupe NQ). Ensuite, nous avons analysé ces deux groupes en fonction de 7 des 9 variables présentées dans le tableau 3.4 (NbQst et NbVotRec sont, par définition, différents et donc non pertinents à prendre en compte ici). Pour les deux variables relatives à la note (NotMoy et NotFin) et les deux variables relatives à l'assiduité (AssGlb et AssCou), qui sont des ratios, toutes les distributions ne suivaient pas une loi normale ($p < .05$ dans certains cas lors des tests avec Shapiro-Wilk), ce qui nous a amené à effectuer des tests U de Mann-Whitney au lieu de t-tests. Nous avons également utilisé les tests U de Mann-Whitney pour la variable relative aux votes effectués (NbVotFait), qui est une variable ordinale. Nous rapportons une taille d'effet estimée, calculée comme suit : $r^2 = \eta^2 = \frac{Z^2}{n}$ où Z représente le score z associé à la valeur p du test et n le nombre d'élèves de ce groupe [Fritz *et al.* , 2012]. Pour la variable indiquant si l'étudiant avait redoublé l'année (EtuRed) et celle indiquant si l'étudiant avait finalement été accepté pour passer à la deuxième année après l'examen final (EtuReu), qui étaient des variables catégorielles, les échantillons étaient suffisamment grands pour envisager d'utiliser le test Chi-2 au lieu du test de Fisher. Nous avons également utilisé la correction de Yates sur ces tests pour tenir compte de la continuité lorsqu'une cellule du tableau de contingence avait un nombre inférieur à 5 et reporté l'ampleur d'effet en utilisant le V de Cramér corrigé, noté $\widetilde{V}$ [Bergsma, 2013].

5.1.2 Résultats

En termes de notes, il n'y avait pas de différence statistiquement significative pour NotMoy dans aucun des 4 cours ($p > .05$). Pour NotFin, seul ANT présentait une différence statistiquement significative ($U = 66393$, $p = .020$, $\eta^2 = .003$), les deux autres cours pour lesquels des données étaient disponibles n'en avaient pas.

En termes d'assiduité, pour AssGlb, il n'y avait qu'une différence statistiquement significative pour BCE ($U = 54755$, $p < .001$, $\eta^2 = .024$). Cependant, pour AssCou, il y avait des résultats statistiquement significatifs pour tous les cours : BCH ($U = 213974$, $p < .001$, $\eta^2 = .035$), HBD ($U = 71554$, $p < .001$, $\eta^2 = .097$), BCE ($U = 110005$, $p < .001$, $\eta^2 = .021$) et ANT ($U = 88238$, $p < .001$, $\eta^2 = .042$).

En termes de nombre de votes effectués sur les questions des autres étudiants (NbVotFait), les résultats sont statistiquement significatifs pour tous les cours : BCH ($U = 270020$, $p < .001$, $\eta^2 = .246$), HBD ($U = 225962$, $p < .001$, $\eta^2 = .197$), BCE ($U = 134731$, $p < .001$, $\eta^2 = .215$) et ANT ($U = 84866$, $p < .001$, $\eta^2 = .285$).

En termes d'étudiants redoublants, le test du Chi-2 a révélé une différence significative entre les groupes de BCH ($\chi^2(1) = 35.67$, $p < .001$, $\widetilde{V} = .177$), HBD ($\chi^2(1) = 23.25$, $p < .001$, $\widetilde{V} = .144$), BCE ($\chi^2(1) = 9.11$, $p < .010$, $\widetilde{V} = .088$) mais pas pour ANT. En termes de réussite des élèves, le test du Chi-2 a révélé une différence significative entre les deux groupes dans tous les cours : BCH ($\chi^2(1) = 25.21$, $p < .001$, $\widetilde{V} = .122$), HBD ($\chi^2(1) = 29.05$, $p < .001$, $\widetilde{V} = .132$), BCE ($\chi^2(1) = 49.72$, $p < .001$, $\widetilde{V} = .175$) et ANT ($\chi^2(1) = 86.96$, $p < .001$, $\widetilde{V} = .232$).

Le tableau 5.1 résume les statistiques descriptives pour les deux groupes de chaque cours (proportion pour EtuRed et EtuReu, médiane, premier et troisième quartiles pour les autres variables en raison de la distribution non normale), avec indication de résultats statistiquement significatifs.

5.1.3 Discussion

Les résultats précédents révèlent plusieurs éléments intéressants qu'il convient de mettre en perspective. Le premier point concerne les étudiants qui ont posé des questions dans un cours avaient tendance à avoir des notes finales plus élevées dans ce cours, mais que cette différence n'était significative que pour le seul cours du deuxième semestre (ANT). Une explication envisageable est que les différences s'accentuent avec le temps entre les étudiants qui posent des questions et ceux qui ne le font pas, étant donné que la différence ne devient significative qu'au second semestre. Le fait de poser des questions n'était cependant pas associé à la performance des élèves lors de la séance de QCM en classe (aucun résultat significatif trouvé sur les quatre cours dans le Tableau 5.1). Une autre observation intéressante est le fait de poser des questions est associé à une participation globale plus faible seulement pour BCE, ce qui indique peut-être que ce cours en particulier était plus attrayant pour l'étudiant moyen (seul résultat significatif trouvé sur le groupe d'étudiants qui appartiennent à la médiane), car il n'y a pas de tendance claire autrement. De même, le fait de poser des questions était logiquement associé à une plus grande participation au cours (puisque les réponses sont fournies pendant le cours), sauf dans le cas de HBD où la tendance est inversée pour une raison que nous n'avons pu identifier. Il est intéressant de noter que les élèves ayant voté le plus sont aussi ceux qui ont posé le plus de questions. L'une des limites de ce dernier résultat est que nous manquons de données de navigation pour distinguer parmi les élèves qui n'ont pas posé de questions ceux qui ont réellement passé du temps sur la plate-forme en ligne (nous savons cependant que tous les élèves se sont connectés et ont au moins chargé la page de questions car c'était un pré-requis pour voir leurs notes). Nous ne pouvons donc pas vérifier s'il existe un sous-groupe de "rôdeurs" qui ne posent jamais de questions et ne lisent pas les questions des autres. Nous constatons également qu'il y a une plus forte proportion d'étudiants redoublants parmi ceux qui ne posent pas de questions, du moins au premier semestre, peut-être parce qu'ils peuvent considérer que cela est moins pertinent pour eux. Enfin, les élèves qui posent des questions réussissent toujours mieux à la fin de l'année - bien que nous ne puissions pas dire si les excellents élèves posent naturellement plus de questions ou si c'est le fait de poser des questions qui les a menés à de meilleurs résultats. Cependant, grâce au schéma de codage établi dans le chapitre 4 (*cf.* Tableau 4.4) et à partir d'un outil d'annotation automatique développé à base de règles d'expert (*cf.* section 4.4), nous avons eu la possibilité d'analyser de façon plus fine, dans la section suivante, les élèves du groupe Q, pour voir si nous pouvons les distinguer par le type de questions qu'ils posent.

			NotMoy			NotFin			AssGlb			AssCou			NbVotFait			EtuRed	EtuReu
Cours	**Grp**	*N*	**1er**	**Md**	**3e**	**1er**	**Md**	**3e**	**1er**	**Md**	**3e**	**1er**	**Md**	**3e**	**1er**	**Md**	**3e**	**Prop**	**Prop**
BCH	Q	244	6.67	8.83	11.83	5.50	8.50	12	0.90	0.98	1	1	1**	1	1	2**	9	21%**	22%**
	NQ	1372	5.50	8	10.83	3.75	6.75	10.25	0.76	0.95	1	0.50	1**	1	0	0**	0	42%**	11%**
HBD	Q	201	7.75	10.50	14	6.75	9.75	12.50	0.93	0.98	1	1	1**	1	0	3**	7	23%**	24%**
	NQ	1410	6	9.25	12.25	4	7.75	10.75	0.76	0.95	1	0.50	1**	1	0	0**	0	42%**	11%**
BCE	Q	114	7.75	10	12.50	N/A	N/A	N/A	0.86	0.98**	1	1	1**	1	0	1**	4	26%**	33%**
	NQ	1486	5	7.75	10.75	N/A	N/A	N/A	0.76	0.95**	1	0.40	1**	1	0	0**	0	41%**	11%**
ANT	Q	75	8.60	12	14.60	9.50	13.5**	15.50	0.98	1	1	1	1**	1	0	0**	3	40%**	47%**
	NQ	1528	5	8	11.80	3.50	7.25**	12.19	0.76	0.95**	1	0	0.60**	1	0	0**	0	40%**	11%**

N/A : Pas de données disponible pour ce cours

Médiane (Md), 1^{er} et $3^{ème}$ quartiles des variables descriptives (proportion de EtuRed et EtuReu) pour chaque groupe et cours (* $p < .01$, ** $p < .001$ – en gras, la valeur la plus élevée entre Q et NQ)

Tableau 5.1 – Caractérisation des étudiants qui posent des questions (Q) et de ceux qui n'en posent pas (NQ)

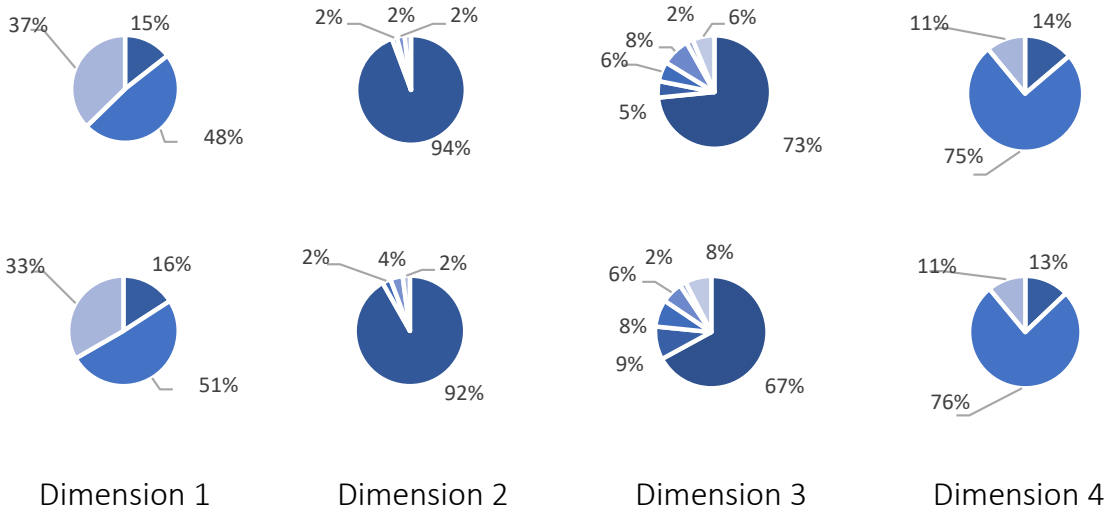


Figure 5.1 – Proportions de questions sur les 4 dimensions à travers les 13 cours (Rangée du haut = les étudiants bons, rangée du bas = les étudiants moyens)(Couleurs : du bleu foncé [0 ou 1] à bleu clair [la valeur maximale pour cette dimension])

5.2 Caractéristiques des étudiants

5.2.1 Proportion de questions posées

L'une des caractéristiques les plus évidentes des étudiants est leur niveau, comme l'indique leur note à l'examen final à la fin de l'année (NotFin). Afin d'étudier si la nature de questions posées, représentée ici par les proportions de types de questions, est une caractéristique corrélée au niveau des étudiants, nous avons décidé de distinguer deux types d'étudiants : les bons étudiants (dont le rang à l'examen final est inférieur à $200^{ème}$ - ce qui correspond aux étudiants autorisés à passer en $2^{ème}$ année du PACES) et les étudiants moyens (dont le rang est entre $200^{ème}$ et $600^{ème}$). Les élèves ayant un rang supérieur au $600^{ème}$ (les élèves de bas niveau ayant de mauvaises notes) n'ont généralement pas posé suffisamment de questions pour être considérés ici. De plus, les enseignants considèrent généralement que les élèves au-delà du $600^{ème}$ rang n'ont pas vraiment de chance de parvenir à réussir l'examen final, même avec leur aide, et souhaitent donc se concentrer davantage sur les élèves moyens. Notre hypothèse relative à la QR3.2 était donc que les bons élèves et les élèves moyens poseraient des questions très différentes.

Malheureusement, l'exploration des données de chaque dimension sur l'ensemble des 13 cours n'a pas permis de révéler une tendance claire (*cf.* Figure 5.1 pour les résultats sur 13 cours - les résultats étaient extrêmement similaires pour les 4 cours principaux [BCE, BCH, HBD, ANT]), ce qui nous a rapidement conduit à répondre négativement à QR3.2 sans avoir besoin de comparaisons statistiques avancées entre les deux groupes.

Cours	Dimension	Bon	Moyen	Tendances opposées
ANT	App	Augmente	Diminue	1 semaine
5 semaines	Sch	Diminue/augmente	stable/diminue	2 semaines
	Lie	Augmente	Diminue/stable	2 semaines
	Exa	Stable	Augmente	1 semaine
BCE	Sch	Diminue/augmente	Augmente/stable	2 semaines
5 semaines	Lie	Augmente/diminue	Diminue/augmente	2 semaines
	Exa	Diminue/augmente	Augmente/diminue	2 semaines
BCH	App	Augmente/diminue	Diminue/augmente	4 semaines
6 semaines	Ver	Augmente/diminue	Diminue/augmente	6 semaines
	Sch	Diminue/augmente	Augmente/diminue	4 semaines
	Lie	Augmente/diminue	Diminue/augmente	2 semaines
	Err	Diminue/augmente	Augmente/diminue	6 semaines
	Exa	Augmente/diminue	Diminue/augmente	5 semaines
	Con	Augmente/diminue	Diminue/augmente	6 semaines
HBD	Lie	Augmente/diminue	Diminue/augmente	3 semaines
4 semaines	Exa	Augmente/diminue	Diminue	1 semaine

Tableau 5.2 – Évolution temporelle des questions posées par les étudiants pour chaque cours et chaque dimension de questions

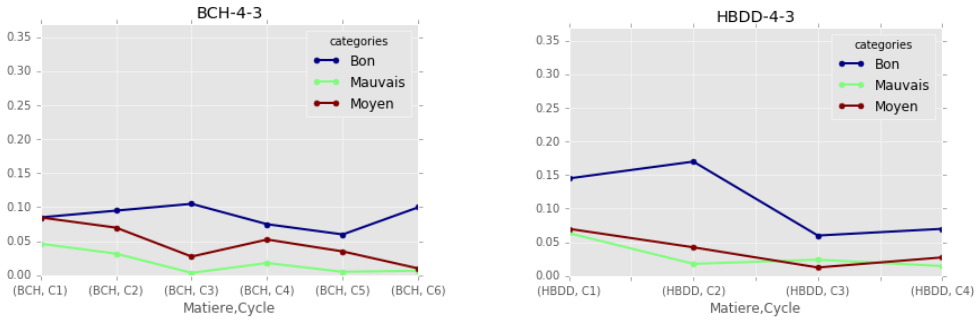
5.2.2 Dynamique des questions posées

Devant l'absence de résultats à la QR3.2, nous avons été amenés à proposer une question de recherche alternative prenant en compte la temporalité :

> **QR3.3** : Est-ce que la temporalité des questions posées est caractéristique de la performance des élèves ?

Notre deuxième hypothèse est que la dynamique des questions posées par les étudiants pourrait être une caractéristique indicative de leur niveau. Nous avons supposé que les bons étudiants pourraient avoir tendance à poser des questions plus complexes vers la fin du cours, parce qu'ils ont déjà acquis une solide compréhension de concepts de base des cours durant les premières semaines. Pour étudier cette question de recherche (QR3.3), nous avons effectué une exploration visuelle de données, comparant cette fois-ci la proportion de questions posées par les élèves chaque semaine dans chaque dimension. Un exemple de visualisation est illustré à la Figure 5.2. Plus précisément, nous recherchions des tendances opposées entre les bons et les moyens élèves - contrairement aux bons étudiants dans notre exemple précédent, on pourrait penser que la proportion des questions complexes posées par les étudiants moyens n'augmenterait pas vers les dernières semaines. Le Tableau présente une vue d'ensemble systématique de la différence de tendances entre les catégories d'étudiants (seules les dimensions montrant des différences de tendances entre les étudiants bons et moyens ont été rapportées ici).

Malheureusement, les résultats ont montré que même si certaines dimensions semblaient importantes pour distinguer les élèves moyens des bons élèves (par exemple la dimension Exa, c'est-à-dire les questions relatives à l'examen final) selon le type de questions posées (par exemple, les bons élèves posent davantage de questions d'approfondissement dans la second moitié des cours d'ANT, alors que l'effet est inverse pour les élèves moyens), les tendances



(a) BCH (b) HBD

Figure 5.2 – Exemple de différence entre les patterns des étudiants bons et moyens sur dim. 4-3 "Exa" pour les cours BCH et HBD

étaient différentes d'un cours à l'autre. Par conséquent, pour répondre à notre QR3.3, il est effectivement nécessaire de considérer non seulement la proportion brute de questions de chaque type, mais aussi leur dynamique à travers le cours. Cependant, cette dynamique semble très liée au cours, ce qui nous a amené à effectuer les analyses suivantes séparément sur chaque cours, au lieu d'essayer de regrouper les questions des différents cours.

5.2.3 Clustering des étudiants selon leur questions

Les algorithmes d'apprentissage automatiques peuvent être répartis en deux grandes catégories : les algorithmes d'apprentissage supervisé et les algorithmes d'apprentissage non supervisé. Nous avons précédemment utilisé des algorithmes d'apprentissage supervisé afin de catégoriser / étiqueter les questions des étudiants (*cf.* Section 4.4). Notre objectif ici est d'obtenir des classes d'étudiants homogènes en se basant sur les types de questions qu'ils ont posées. Nous disposons en effet de plusieurs variables caractérisant nos étudiants, mais sans idée a priori des combinaisons de variables pertinentes à considérer. Nous avons donc fait recours aux algorithmes non supervisés et précisément le clustering.

Cela nous a amené à poser la question de recherche suivante, prenant en considération les résultats de QR3.3 :

> **QR3.4** : Y a-t-il un lien entre la nature des questions posées par les élèves et leurs caractéristiques ?

Les algorithmes de clustering les plus répandus sont K moyennes (K-means), les réseaux de Kohonen (SOM), et la classification ascendante hiérarchique (CAH). Nous allons utiliser dans ce travail K-Means, pour sa simplicité et CAH pour sa représentation visuelle, sous forme d'un arbre de classification, qui facilitera l'interprétation des résultats.

5.2.3.1 Techniques

K-Means

L'algorithme K-Means, défini par Macqueen [1967] est l'un des plus simples algorithmes de classification automatique des données. L'idée principale est de choisir aléatoirement un ensemble de centres fixés a priori et de chercher itérativement la partition optimale. Chaque

individu est affecté au centre le plus proche ; après l'affectation de toutes les données la moyenne de chaque groupe est calculée. Ces moyennes sont considérées par la suite comme les nouveaux centres de groupes. Lorsqu'on aboutit à un état stationnaire (aucune donnée ne change de groupe entre deux itérations), l'algorithme est arrêté.

La principale limite de cette méthode est la dépendance des résultats aux valeurs de départ (centres initiaux). À chaque initialisation correspond une solution différente (optimum local) qui peut dans certains cas être très loin de la solution optimale (optimum global). Une solution naïve à ce problème consiste à lancer l'algorithme plusieurs fois avec différentes initialisations et retenir le meilleur regroupement trouvé. Par ailleurs, nous avons utilisé X-Means (implémenté dans RapidMiner), qui permet de trouver une meilleure partition en une seule exécution. Donc le choix du nombre de clusters optimal (k) est fait automatiquement à partir de X-Means. Nous avons également lancé X-Means à plusieurs itérations avec la même valeur de k pour assurer la stabilité des résultats finaux.

La qualité du clustering dépend également de la mesure de similarité utilisée. Différentes mesures de distance entre objets existent, et le choix de ces distances dépend du type des données considérées et du type de similarité recherchée. Nous avons choisi la distance euclidienne qui est la plus connue et qui s'adapte parfaitement à nos données (numériques).

CAH

La Classification Ascendante Hiérarchique (CAH) est un algorithme classique de clustering hiérarchique. Il se base sur le regroupement pas à pas des deux classes les plus proches (les classes initiales sont les observations) au regard d'une mesure de dissimilarité inter-classe à choisir (saut minimum, saut maximum, lien moyen, ou encore distance de Ward). Nous avons utilisé ici le saut minimum appelé aussi "lien simple" comme mesure de dissimilarité (paramètre par défaut sur RapidMiner) pour fusionner à chaque étape les deux classes dont les deux individus les plus proches ont la plus petite distance. Les résultats de ce type de classification sont habituellement représentés sous la forme d'un dendrogramme (arbre de classification hiérarchique).

5.2.3.2 Méthode

Tout d'abord, nous avons effectué quatre analyses de clustering distinctes à l'aide de X-Means (avec un k compris entre 2 et 10 et lancé à 5 itérations avec la même valeur de k) et CAH sur quatre ensembles de données : les étudiants qui ont posé des questions dans le cours BCH (1227 questions par $N_1 = 244$ étudiants), HBD (979 questions par $N_2 = 201$ étudiants), BCE (685 questions par $N_3 = 114$ étudiants), et ANT (649 questions par $N_4 = 75$ étudiants).

Nous avons effectué le clustering en utilisant comme caractéristiques pour chaque étudiant la proportion de questions posées dans chaque dimension (par exemple, la proportion de questions ayant la valeur 1 ou "Ree" dans la dimension 1) posées (a) globalement, (b) pendant la première moitié du cours et (c) pendant la seconde moitié du cours. Nous avons également considéré la proportion globale des questions posées au cours de la première moitié et de la deuxième moitié du cours (44 caractéristiques générales, comme l'indique la Figure 5.3). Le fait de distinguer (b) et (c), en plus de (a), nous a permis de prendre en compte la dynamique des questions (comme le suggère notre réponse à la question QR3.3), en plus du nombre global.

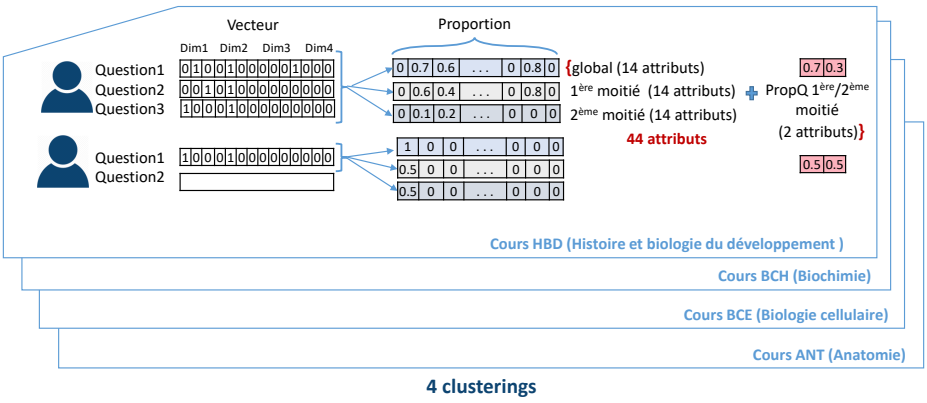


Figure 5.3 – Attributs utilisés pour le clustering des étudiants uniquement en fonction des questions posées

Pour K-Means, nous avons obtenu 4 clusters pour BCH et HBD et 3 clusters pour BCE et ANT, dont les centroïdes sont fournis dans le tableau en annexe A. Le nombre des clusters trouvés par K-Means était identique à l'approche du regroupement hiérarchique (CAH), en découpant le dendrogramme obtenu à une certaine hauteur des branches bien distinctes. Cependant, la qualité des clusters obtenus avec le regroupement hiérarchique était inférieure à celle obtenue avec K-Means (d'après le coefficient de silhouette - *cf.* section 6.1.1 les indices de silhouette variaient entre 0.12 et 0.19 pour le regroupement hiérarchique par rapport à 0.19 et 0.23 pour K-Means). Nous avons également pensé à utiliser une technique de sélection d'attributs pour réduire les 44 attributs étant très important au regard du nombre d'instances. Bien que le clustering effectué après la sélection d'attributs a donné de nouveaux clusters avec des caractéristiques différentes à travers les cours, les clusters obtenus sont d'une qualité intrinsèque inférieure ou égale (d'après l'indice de silhouette). En raison de l'absence d'une meilleure séparation de clusters et de la difficulté supplémentaire de présenter différentes dimensions et caractéristiques des étudiants pour chaque cours, nous nous concentrerons donc dans la suite de ce travail sur les clusters obtenus avec K-Means, utilisant 44 attributs, sans réduction d'attributs.

5.2.4 Caractérisation de clusters

La deuxième étape consistait à caractériser les clusters en considérant les variables des élèves (non utilisés pour le clustering) mentionnées dans le tableau 3.4, en suivant une méthodologie similaire à celle décrite dans la section 5.1. Pour les deux variables relatives à la note (NotMoy et NotFin) et les deux variables relatives à l'assiduité (GlbAtt et CouAtt), qui sont des ratios, toutes les distributions ne suivaient pas une loi normale ($p < .05$ dans certains cas lors des tests avec Shapiro-Wilk), ce qui nous a amené à effectuer des tests Kruskal-Wallis

plutôt que des ANOVAs à sens unique pour les clusters associés aux quatre cours considérés. Nous avons également utilisé les tests Kruskal-Wallis pour les variables relatives aux questions (NbQst) et aux votes (NbVotRec et NbVotFait), qui sont des variables ordinales. Lorsqu'un résultat s'est révélé statistiquement significatif, nous avons effectué des comparaisons post hoc à l'aide du test de Dunn avec correction de Holm-Bonferroni. Nous avons également évalué la taille d'effet pour les tests Kruskal-Wallis H, en utilisant la formule de Cohen pour la mesure d'eta-carré [Cohen, 1988], c'est à dire $\eta^2_H = \frac{H-k+1}{n-k}$, où k représente le nombre de clusters et n le nombre d'élèves dans ce cluster. Pour les deux variables restantes (EtuRed et EtuReu), qui sont catégorielles, nous avons effectué un test du Chi-2 dans des conditions similaires à celles décrites dans la section 5.1. La dénomination des clusters a été choisie a posteriori pour correspondre aux caractéristiques similaires identifiées plus loin.

Résultats

En ce qui concerne les notes, il y avait une différence statistiquement significative entre les clusters pour la note moyenne pour BCH ($\chi(3) = 17.20$, $p < .001$, $\eta^2 = .059$) et pour la note finale ($\chi(3) = 20.71$, $p < .001$, $\eta^2 = .074$). C'était également le cas pour HBD (NotMoy : $\chi(3) = 25.11$, $p < .001$, $\eta^2 = .112$, et NotFin : $\chi(3) = 28.95$, $p < .001$, $\eta^2 = .132$). Pour BCE, il n'y avait pas de différence statistiquement significative entre les clusters pour NotMoy et aucune donnée de NotFin n'était disponible. Pour ANT, il y avait une différence statistiquement significative entre les clusters pour la note moyenne ($\chi(2) = 7.84$, $p = .020$, $\eta^2 = .081$) et pour la note finale ($\chi(2) = 13.50$, $p = .001$, $\eta^2 = .160$).

En ce qui concerne l'assiduité, pour BCH, il y avait une différence statistiquement significative pour AssCou ($\chi(3) = 10.51$, $p = .015$, $\eta^2 = .031$), mais pas pour AssGlb. Pour HBD, le test a montré une différence statistiquement significative pour AssGlb ($\chi(3) = 9.33$, $p = .025$, $\eta^2 = .041$) mais pas pour AssCou. Pour BCE, il y avait une différence significative pour AssCou ($\chi(2) = 13.69$, $p = .001$, $\eta^2 = .105$), mais pas pour AssGlb. Pour ANT, il y avait une différence statistiquement significative pour AssGlb ($\chi(2) = 6.19$, $p = .045$, $\eta^2 = .066$) et AssCou ($\chi(2) = 12.04$, $p = .002$, $\eta^2 = .139$).

En ce qui concerne le nombre de questions/votes, pour BCH, il y avait une différence statistiquement significative pour NbQst ($\chi(3) = 42.12$, $p < .001$, $\eta^2 = .163$) et NbVotRec ($\chi(3) = 12.06$, $p = .038$, $\eta^2 = .037$), mais pas pour NbVotFait. Pour HBD, les différences étaient également statistiquement significatives pour NbQst ($\chi(3) = 33.20$, $p < .001$, $\eta^2 = .153$) and NbVotRec ($\chi(3) = 16.76$, $p < .001$, $\eta^2 = .070$), mais pas pour NbVotDone. Pour BCE, une différence pour NbQst ($\chi(2) = 9.85$, $p = .007$, $\eta^2 = .071$) mais aucune pour NbVotRec et NbVoteDone. Pour ANT, aucune différence statistiquement significative n'a été observée pour NbQst, NbVotRec et NbVotFait.

En ce qui concerne les étudiants redoublants, le test Chi-2 a révélé une différence significative entre les clusters pour BCH ($\chi(3) = 14.43$, $p = .002$, $\widetilde{V} = .217$), HBD ($\chi(3) = 23.72$, $p < .001$, $\widetilde{V} = .322$) et aucune différence pour BCE et ANT. En ce qui concerne les étudiants qui réussissent, le test Chi-2 a révélé une différence significative entre les clusters pour BCH ($\chi(3) = 21.47$, $p < .001$, $\widetilde{V} = .276$), HBD ($\chi(3) = 19.14$, $p < .001$, $\widetilde{V} = .284$) et aucune différence pour BCE et ANT.

Le tableau 5.3 présente un résumé des statistiques descriptives (proportion pour EtuRed et EtuReu, médiane, 1^{er} et $3^{ème}$ quartiles pour les autres variables) pour chaque cluster dans

		N	EtuRed	EtuReu	NotMoy			AssCou			NotFin			AssGlb			NbQst			NbVotFait			NbVotRec		
	Clu.		Prop.	Prop.	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3
BCH	A	44	14%	14%	5.67	7.8	10.25	0.83	1	1	4.12	6.5	10.5	0.74	0.95	1	1	1	2	1	3	7.5	0.25	2	4.38
	B	89	22%	27%	7	9.42	12.33	1	1	1	6.5	8.88	12.38	0.93	1	1	2	4	9	0	3	10.25	0.5	1.83	3.76
	C	77	14%	10%	6.5	8.45	10.67	0.83	1	1	5.5	8	10.5	0.91	0.98	1	1	2	4	1	2	7	0.67	2	3.65
	D	34	44%	47%	8.04	10.75	13.13	1	1	1	8	12	14.25	0.9	0.98	1	2	5.5	13.25	0	1	3.5	0	0.86	1.42
HBD	A	59	7%	7%	7	8.25	10.88	1	1	1	5	7.38	10.31	0.89	0.95	0.98	1	1	3	1	3	6	1	3	8
	B	74	27%	32%	8.75	12.25	14.69	1	1	1	8	11	12.75	0.93	1	1	2	3	9.75	0.25	3	7.5	1	2.89	5
	C	31	16%	16%	7.88	10.25	12.54	1	1	1	7.75	10	12.75	0.95	1	1	1	2	3	0	1	7	0.42	1.6	3.5
	D	37	49%	41%	9.65	13.38	15.56	1	1	1	8.25	12	14.12	0.95	1	1	2	5	7	0	3	8	0	1	2.17
BCE	A	26	15%	19%	6.25	8.6	10.75	0.8	1	1	N/A	N/A	N/A	0.91	0.98	1	1	4	9.25	0	1.5	7	0.49	1.02	3.19
	B	52	31%	31%	7.19	10.3	12.8	1	1	1	N/A	N/A	N/A	0.44	0.94	1	1	1.5	2	0	1	2	0	1.5	3.2
	D	36	28%	47%	8.55	11.2	12.6	1	1	1	N/A	N/A	N/A	0.93	1	1	1	2.5	7.5	0	1.5	3	0.38	1	2
ANT	A	15	20%	27%	4.29	6	13.45	0.8	1	1	6.25	9	12.5	0.91	0.95	1	1	2	3	0	0	3.5	0	0.5	1
	B	30	43%	50%	8.2	11.4	14	1	1	1	11.5	14.5	15.5	0.98	1	1	3	4.5	9	0	0	2	0.14	0.53	0.95
	D	30	47%	53%	11.2	12.7	14.9	1	1	1	12.5	14	16.38	0.98	1	1	2.25	5.5	11.5	0	1	3	0	0.29	0.59

N/A : Aucune donnée disponible pour ce cours

Tableau 5.3 – Statistiques descriptives médiane, 1^{er} et $3^{ème}$ quartiles des variables dépendantes (proportion de EtuRed et EtuReu) pour chaque cluster et chaque cours

Clust.	Clust.	NotMoy			NotFin				AssGlb		AssCou			NbQst			NbVotRec	
		BCH	HBD	ANT	BCH	HBD	BCE	ANT	HBD	ANT	BCH	BCE	ANT	BCH	HBD	BCE	BCH	HBD
A	B	-	***	-	*	***	ND[1]	*	*	-	-	*	-	***	***	**	-	-
A	C	-	-	ND[2]	-	***	ND[1]	ND[2]	-	-	-	A^2	A^2	*	-	A^2	-	-
A	D	**	***	*	***	*	ND[1]	***	*	-	*	***	**	***	***	-	*	***
B	C	-	-	ND[2]	-	-	ND[1]	ND[2]	-	-	-	ND[2]	ND[2]	***	**	ND[2]	-	-
B	D	-	-	-	-	-	ND[1]	-	-	-	-	-	*	-	-	-	**	*
C	D	**	-	ND[2]	**	-	ND[1]	ND[2]	-	-	-	ND[2]	ND[2]	**	**	ND[2]	*	-

ND[1] : Aucune donnée disponible pour ce cours
ND[2] : Pas de cluster C pour ce cours

Tableau 5.4 – Différences par paires pour NotMoy, NotFin, AssGlb, AssCou, NbQst, NbVotRec (* $p < .05$, ** $p < .01$, *** $p < .001$)

chaque cours, et le tableau 5.4 un résumé des résultats statistiques inférentiels susmentionnés, selon les comparaisons post hoc de suivi.

La figure 5.4 présente une synthèse des résultats, décrits plus en détail dans cette section. Le cluster A représente 18 à 29% des étudiants et se caractérise par des notes inférieures à la moyenne (à la fois dans le cours considéré et dans l'ensemble), et une assiduité en cours et globale toujours inférieure - parfois significativement - à celle des étudiants des autres clusters. Les étudiants de ce cluster sont majoritaires (86% en moyenne) à suivre le cours pour la première fois et réussissent moins bien (17% en moyenne, ce qui correspond presque à la proportion d'étudiants redoublants), ils posent moins de questions que la moyenne, mais leurs questions obtiennent plus de votes que la moyenne et sont donc assez populaires. Leurs questions portent principalement sur la ré-explication ou l'approfondissement d'un concept, notamment une demande de définition (Ree, App et Def) et sont souvent posées pendant la première moitié du cours. Ce cluster correspond donc aux *étudiants passifs en difficulté* qui ont besoin d'explications de base, dont de nombreux étudiants pourraient bénéficier.

De l'autre extrémité du spectre, le cluster D représente 14 à 25% des étudiants du premier semestre (et 40% du second semestre, une proportion accrue qui peut être liée à une diminution de l'activité des autres élèves) et se caractérise par des notes de cours et des notes

	Passifs en difficulté	Actifs pointilleux	Actifs en compréhension
	Cluster A	**Cluster D**	**Cluster B**
#étudiants	18-29%	14-25%	36-40%
Notes	inférieur	supérieur	supérieur
Assiduité	faible	élevé	élevé
#questions posées	faible	élevé	élevé
#votes	populaire	non-populaire	populaire
% Redoublants	faible	élevé (42%)	moyen (31%)
% Etudiants réussis	faible	élevé (47%)	moyen (35%)
Type de questions	Ré-explication & définition	Vérification: erreur ou contradiction	Vérification de connaissances & lien entre concepts
Majorité de questions posées	1st moitié	2nd moitié	1st moitié

Figure 5.4 – Tableau récapitulatif des variables similaires/différentes pour les 3 clusters similaires à travers les cours

finales nettement supérieures à celles des autres élèves (et presque toujours statistiquement significativement supérieures aux élèves du cluster A), qui ont tendance à suivre la plupart des classes. Une autre caractéristique distinctive de ce cluster est le fait qu'il contient une proportion importante d'étudiants qui suivent le cours pour la deuxième fois (42% en moyenne) et réussissent davantage (47% en moyenne), qui posent plus de questions que la moyenne mais dont les questions sont moins populaires, avec moins de votes en général. Nous pouvons supposer qu'il s'agit de questions très précises qui exigent déjà une bonne compréhension du contenu du cours et ne sont donc pas considérées comme importantes par les autres étudiants. En effet, les questions posées sont principalement des questions de vérification pour vérifier sa compréhension ou pour pointer une éventuelle contradiction dans le cours (Ver et Err), et moins de questions sur le lien entre les concepts (Lie). Il est intéressant de noter qu'en comparant la proportion de questions posées dans la première moitié de la classe par rapport à la deuxième moitié, ils sont les seuls étudiants à avoir posé plus de questions dans la deuxième moitié des séquences que dans la première moitié, probablement parce que les concepts présentés au début étaient plus simples et faciles à comprendre pour eux. Ce cluster correspond donc à des étudiants actifs pointilleux, qui comprennent bien les bases et signalent des erreurs potentielles dans des concepts avancés. D'après l'équipe pédagogique, ce type de comportement peut même dans certains cas correspondre à une volonté de quelques étudiants de perturber les autres - un comportement très particulier lié à la nature compétitive de la première année de PACES.

Le cluster B représente 36 à 40% des étudiants, dont les notes, l'assiduité et le nombre de questions posées sont similaires à ceux du cluster D. Cependant, la proportion d'étudiants qui suivent le cours pour la deuxième fois et la proportion d'étudiants ayant réussi qui peuvent accéder à la $2^{ème}$ année est inférieure à D (31% et 35% en moyenne respectivement), ils ont voté davantage et leurs questions sont plus populaires. Dans l'ensemble, leurs questions portent principalement sur la vérification des connaissances et sur le lien entre les concepts (Ver, Con et Lie), mais seulement durant la première moitié de la classe. Ils correspondent à des étudiants qui sont en train de développer activement leur compréhension du cours.

Le cluster C (défini uniquement pour les cours BCH et HBD) tend à être un cluster intermédiaire qui se situe toujours entre les clusters A et D.

En comparant ces clusters aux étudiants qui n'ont pas posé de questions (groupe NQ en section 5.1), nous constatons que le groupe NQ est le plus proche du cluster A en termes de notes. Ce groupe d'étudiants ne posant pas de questions se caractérise également par la plus faible proportion d'étudiants ayant réussi (11% en moyenne), ce qui montre que peu importe le type de questions posées, ils préfèrent toujours de ne pas poser de questions. Ils ont également la plus faible proportion d'assiduité en cours, ce qui peut être interprété comme **(a)** un signe qu'ils n'ont pas de question parce que c'est facile pour eux et qu'ils considèrent que la séance de questions et réponses n'est pas nécessaire, ou **(b)** que, n'ayant pas posé de question par eux-mêmes, ils ne sont pas motivés à participer et manquent une occasion pour améliorer leur compréhension. Une analyse complémentaire a révélé que 90% des étudiants de ce groupe n'ont posé aucune question en aucun cours - combiné à la faible performance globale du groupe et au fait qu'ils votent également moins, cela semble donner plus de poids à l'interprétation **(b)**.

Par conséquent, dans l'ensemble, nous pouvons dire que notre clustering (basé uniquement sur les dimensions des questions) et les analyses statistiques inférentielles de suivi (utilisant les caractéristiques des étudiants) nous permettent de répondre positivement à la QR3.4 : la proportion et la dynamique des questions posées par les étudiants sont fortement liées à

certaines de leurs caractéristiques (leur performance, leur présence, le nombre de questions posées, le nombre de votes qu'ils ont obtenus, le nombre de fois où ils ont voté sur les autres questions, s'ils avaient ou non suivi ce cours pour la première fois et s'ils avaient réussi). De plus, bien que les 4 clusterings aient été effectués séparément, et contrairement à nos analyses précédentes, le fait que trois clusters très similaires soient systématiquement apparus dans les 4 cours considérés indique que les relations identifiées ne sont pas dépendantes des cours eux-mêmes. Cela permet donc d'être raisonnablement optimiste quant à la généralisabilité des résultats trouvés (dans le contexte de PACES), aspect qui sera revisité dans le chapitre 6.

5.3 Évolution des questions des étudiants

Nous avons précédemment montré qu'il existait un lien entre les questions posées par les étudiants et leurs profils en termes de performance et divers aspects du comportement du cours (assiduité, nombre de questions et votes). On peut donc s'interroger sur les aspects temporels de ces liens :

> **QR3.5** : Est-ce que la nature des questions posées par les étudiants est plutôt constante ou au contraire très variable au cours du temps ? Et est-ce que ces variations de nature de questions sont associés à des variations de performance des étudiants ?

Pour répondre à cette question, nous avons essayé d'examiner l'évolution des étudiants durant le premier semestre sur les trois mois de cours : Septembre, Octobre et Novembre (puisque les étudiants ne posent pas de questions en Décembre et se préparent pour le concours). Afin d'étudier le lien entre la nature des questions posées par les étudiants et leur évolution dans le temps, nous avons caractérisé les étudiants en fonction de leur notes obtenues sur les QCMs de chaque cours. Nous avons donc calculé pour chaque étudiant, la note moyenne des QCMs (Not1) sur la $1^{\text{ère}}$ moitié du mois (du 1^{er} jusqu'au 15) et la note moyenne (Not2) sur la $2^{\text{ème}}$ moitié du mois (du 15 à la fin du mois) sur toutes matières confondues du premier semestre. Nous avons ainsi pu dinstinguer trois états d'évolution des apprenants :

- En progression : si Not2 de l'étudiant est supérieur à Not1
- En baisse : si Not2 de l'étudiant est inférieur à Not1
- Stable : si Not1 est égale à Not2

Ensuite, nous avons regroupé les étudiants en 4 catégories (range0, range1, range2 et range3) en fonction de leurs proportions de questions posées sur chaque dimension pour chaque mois (toutes matières confondues). Nous avons créé 3 groupes de tailles similaires (range1, range2 et range3) pour les étudiants qui posent des questions, ce qui explique une proportion de questions différente selon chaque dimension (les seuils de proportion ont été fixés automatiquement pour chaque groupe par rapport à chaque valeur de dimension, de manière à donner des groupes d'étudiants de taille identique). Le groupe (range0) pour les étudiants qui n'ont pas posé de questions, contient souvent un nombre important d'étudiants (Comme vu en section 5.1) sur les 4 groupes :

- Range0 : contient les étudiants qui n'ont pas posé de questions sur cette valeur de dimension (ex : "Ree").
- Range1 : contient les étudiants qui ont posé des questions de proportion supérieure à 0 et inférieure à 0.2 (ou 0.5 pour certaines valeurs de dimensions)

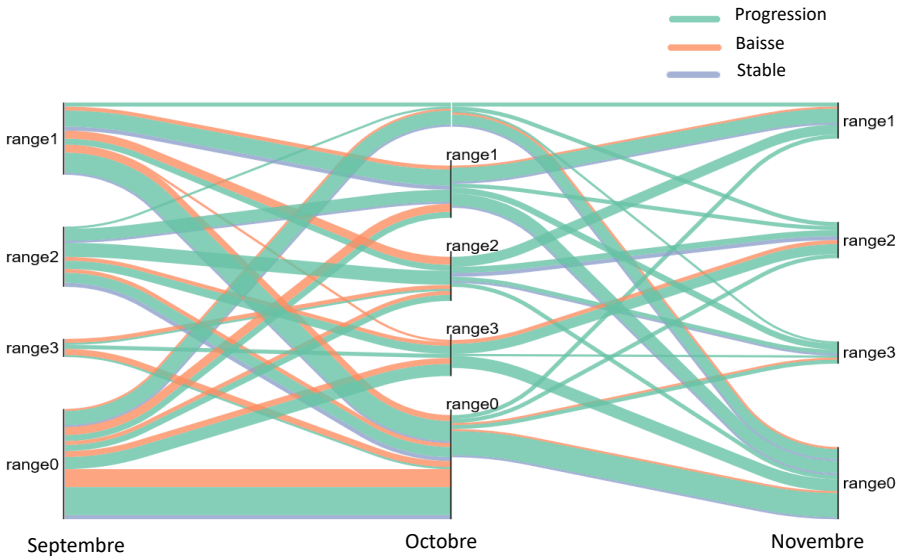


Figure 5.5 – Les flux d'évolution de proportion de questions des étudiants en progression/baisse sur dim. 1-"Ree" (la largeur de flux représente le nombre d'étudiants)

- Range2 : les étudiants de ce groupe ont posé des questions de proportion supérieure à 0.2 (ou 0.5) et inférieure à 0.5 (ou 0.7 pour certaines valeurs de dimensions)
- Range3 : contient les étudiants qui ont posé des questions de proportion supérieure à 0.5 (ou 0.7) et inférieure à 1 selon la valeur de dimension

Notre hypothèse était que les étudiants en progression posaient des questions de nature différente des étudiants en baisse. Pour vérifier cette hypothèse, nous avons représenté l'évolution des étudiants à partir d'un diagramme Sankey [1] (en utilisant la bibliothèque Python floWeaver [2]) sur le premier semestre (*cf.* Figures 5.5 et 5.6). Nous nous sommes focalisés sur les étudiants qui ont posé au moins une question en Septembre (N_{sept}=307 étudiants), en Octobre (N_{oct}=175 étudiants) et en Novembre (N_{nov}=101 étudiants).

Sur la figure 5.5, par exemple les étudiants en baisse (selon les notes obtenues en QCMs) dans range1 (en orange) se sont répartis sur 4 groupes en mois d'Octobre : (1) les étudiants qui continuaient de poser des questions ré-explication (Ree) de même quantité (range1), (2) ceux qui posaient plus de questions de Ree en Octobre (range2) par rapport au mois de Septembre, (3) une minorité d'entre eux qui posaient beaucoup de questions (range3) de Ree sur Octobre et (4) d'autres qui n'en posent plus de questions de ce type (range0). Malheureusement, nous ne pouvons pas distinguer sur ce diagramme si les mêmes groupes d'étudiants de Septembre en Octobre continuaient à poser le même nombre de questions ou le même état d'évolution depuis Octobre à Novembre.

En comparant les flux d'étudiants en baisse/progression dans chacune des trois catégories (range1, 2 et 3) sur dim1 - "Ree" (nous avons choisi de reporter la dimension 1 considérée comme la principale dimension de notre schéma de codage et qui nous paraissent la plus

1. Un diagramme Sankey est un type de diagramme de flux dans lequel la largeur des flèches est proportionnelle au flux représenté.
2. https://sankeyview.readthedocs.io/

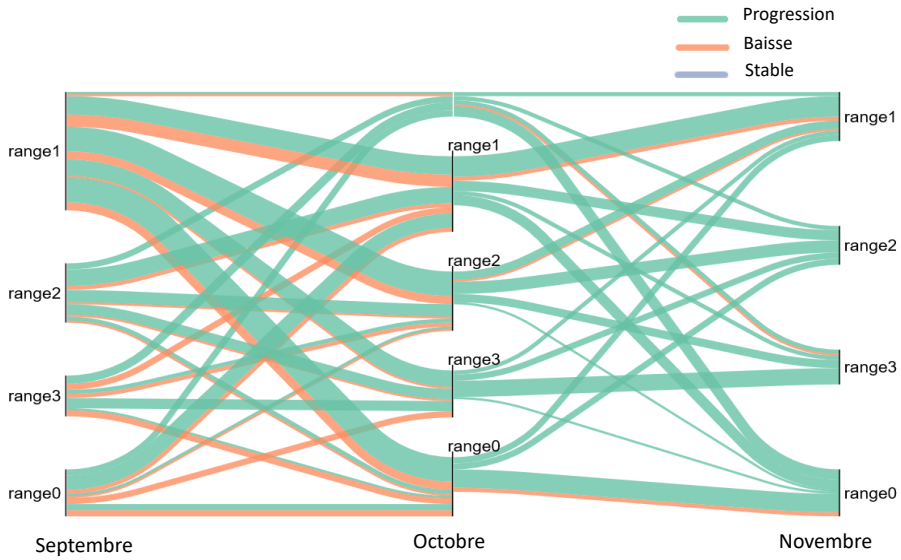


Figure 5.6 – Les flux d'évolution de proportion de questions des étudiants en progression/baisse sur dim. 1-"App" (la largeur de flux représente le nombre d'étudiants)

intéressante en termes de visualisation - les autres dimensions sont également été explorées visuellement), nous constatons que la majorité des étudiants qui posaient des questions de Ree en Septembre, sont des étudiants en baisse. Cependant, la même tendance n'est pas facile à distinguer sur les étudiants qui posaient beaucoup de questions (range3) d'approfondissement (App). Une autre observation relative aux étudiants qui posaient le plus de questions d'approfondissement en Octobre, sont tous des étudiants en progression. Malheureusement, ces visualisations montrent qu'il n'y a pas de différences claires pour distinguer les étudiants en baisse de ceux en progression selon les proportions de questions posées dans chaque dimension et qu'il n'y a pas de lien direct entre les types de questions des étudiants et leur évolution en premier semestre.

5.4 Lien entre vote et comportement d'étudiant

Dans le contexte de notre étude, les enseignants donnent un rôle important aux votes puisque c'est leur seul moyen pour sélectionner les questions des étudiants vu le volume de questions qu'ils reçoivent par mail (une question ayant reçu beaucoup de votes nécessite probablement d'être traitée en session de question-réponse). Pour limiter le nombre de questions et les aider dans ce choix, ils encouragent les étudiants à voter sur les questions déjà posées par d'autres étudiants de leur groupe avant d'en poser de nouvelles. Nous avons conduit des analyses afin d'explorer comment les votes sont associés aux caractéristiques des étudiants (performance, réussite, assiduité, *etc.*) et aux questions qu'ils posent. Plus précisément, notre objectif était de répondre aux deux questions de recherche suivantes :

QR3.6 Est-ce que les votes des étudiants dépendent de la nature des questions posées ? (i.e. est-ce que certains types de questions reçoivent plus/moins de votes que d'autres)

QR3.7 Est-ce que les votes des étudiants dépendent de leurs caractéristiques ?

QR3.8 Le vote est-il identique pour un étudiant qui pose ou ne pose pas de question ?

QR3.9 Y a-t-il des différences dans la nature des questions posées et votées par les étudiants ?

Pour répondre aux questions de recherche, nous avons distingué 4 sous-populations sur chacun des 4 cours considérés précédemment en fonction de l'activité des étudiants en distinguant :

- **QV** : les étudiants qui ont posé au moins une question et voté sur au moins sur une question d'un autre étudiant dans ce cours
- **QNV** : les étudiants qui ont posé au moins une question mais qui n'ont voté sur aucune question dans ce cours
- **NQV** : les étudiants qui n'ont posé aucune question mais ont voté sur au moins une question d'un autre étudiant dans ce cours
- **NQNV** : les étudiants qui n'ont posé aucune question et n'ont voté sur aucune question dans ce cours

Le tableau 5.5 résume les statistiques descriptives des 4 sous-populations considérées.

5.4.1 Caractéristiques des questions votées

Pour aborder la question de recherche QR3.6, qui est de savoir si les questions votées par un étudiant peuvent être liées à un certain types de questions avant d'étudier leurs liens avec les caractéristiques d'apprentissage d'un apprenant. Nous avons comparé le nombre de votes reçus par les questions posées sur les différentes dimensions (*cf.* Tableau 4.4). Nous avons défini 5 catégories de questions : **(1)** non votées [0], **(2)** peu votées [de 1 à 3 votes], **(3)** moyennement votées [de 4 à 6 votes], **(4)** bien votées [de 7 à 10 votes], **(5)** très bien votées [supérieur à 10 votes]. Nous souhaitons voir si des questions d'une certaine nature reçoivent plus de votes que d'autres.

L'exploration visuelle des votes sur chaque dimension de l'ensemble de questions posées sur les 13 cours n'a pas permis d'extraire une caractéristique spécifique à un type de question donné (*cf.* Figure 5.7). C'est le même type de résultat obtenu pendant l'exploration de questions sur la figure 5.1. Par conséquent, ces résultats nous conduisent à répondre négativement à la QR3.6.

5.4.2 Comparaison des votants vs. non votants

Pour examiner la question de recherche QR3.7, qui consiste à vérifier le lien entre le vote et les caractéristiques d'un apprenant, nous n'avons pas fusionné les questions des différents cours mais nous nous sommes concentrés sur les 4 cours ayant généré le plus de questions séparément : BCH, HBD, BCE et ANT (*cf.* Tableau 3.3). De plus, considérer les cours séparément nous permet de vérifier si des tendances similaires apparaissent d'un cours à l'autre.

Méthode. Pour chacun des 4 cours, pour chacune des 8 variables d'étudiants (toutes sauf NbVoteFait, non pertinente pour cette analyse), comme dans la section 5.1, nous avons utilisé des tests statistiques pour comparer la population votante (V, constituée de QV et NQV) et la population non votante (NV, constituée de QNV et NQNV). Pour les deux variables relatives aux notes (NotMoy et NotFin), les deux variables relatives à l'assiduité (AssGlb et AssCou) ainsi que pour les variables relatives aux questions (NbQst) et aux votes (NbVot), toutes les

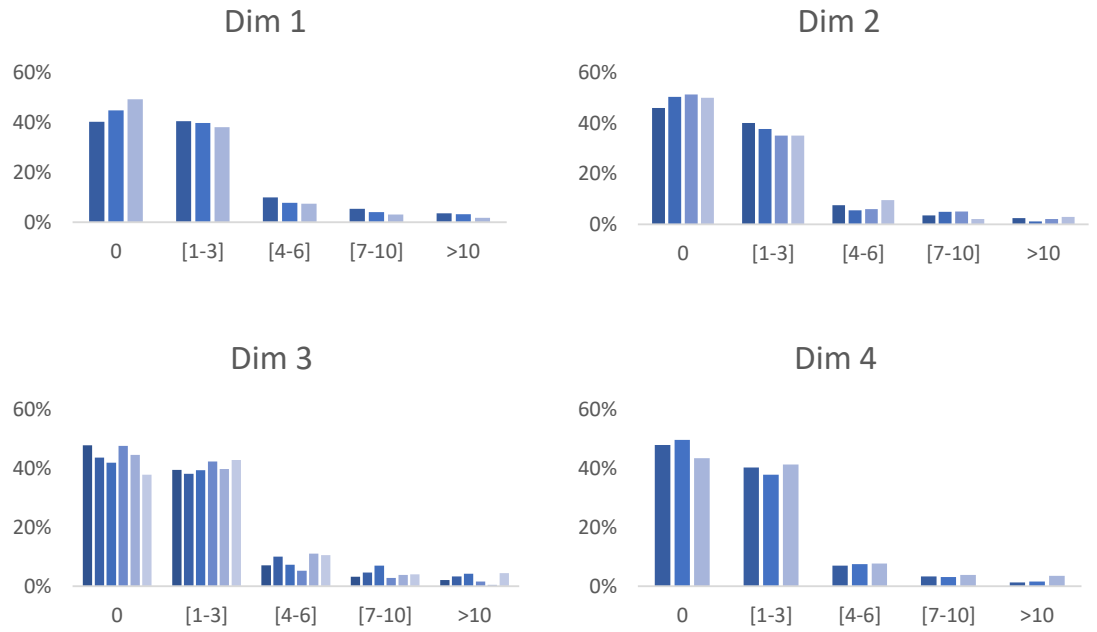


Figure 5.7 – Proportions de votes sur chaque intervalle sur les 4 dimensions à travers les 13 cours (la couleur du bleu foncé [0 ou 1] à bleu clair [la valeur maximale pour cette dimension])

distributions ne suivent pas une loi normale ($p < .05$ dans certains cas lors des tests avec Shapiro-Wilk) qui nous a amené à effectuer les tests Mann-Whitney U plutôt que des t-tests. Pour les deux variables binaires (EtuRed et EtuReu), nous avons effectué des tests Chi-2. Dans chaque cas, l'hypothèse nulle était que V et NV provenaient de la même population, sans l'hypothèse sur la directionnalité pour l'hypothèse alternative (alternative bilatérale). Les seuils de significativité des 31 résultats (et non les 32 car il manque des données pour la NotFin pour BCE) ont été corrigés pour prendre en compte les erreurs de type I en utilisant la méthode de Holm-Šidák.

Résultats. 20 des 31 tests ont révélé des résultats statistiquement significatifs (en utilisant une valeur p critique de 0.05 avec la correction susmentionnée), comme le résume le Tableau 5.6, ce qui nous permet de répondre positivement à la question QR3.7. Pour NotFin, AssGlb, AssCou, NbQst et EtuReu, lorsqu'un résultat est significatif, les élèves qui ont voté (V) avaient toujours une valeur plus élevée associée à ces variables que ceux qui n'ont pas voté (NV). Pour EtuRed et NotMoy, la tendance est inversée (NV $>$ V). Pour NbVotRec, le vote est associé à des votes plus élevés reçus pour BCH, HBD et ANT, mais associés à moins de votes reçus pour BCE.

5.4.3 L'impact du vote sur les caractéristiques d'apprentissage

Méthode. Pour examiner la question QR3.8, pour chacun des 4 cours, nous avons utilisé les mêmes tests qu'en section 5.4.2, mais cette fois en faisant deux comparaisons distinctes en répartissant la population des votants/non-votants entre ceux qui ont aussi posé une question

		N	NotMoy			AssCou			NotFin			AssGlb			NbVotFait			NbQst			NbVotRec			EtuRed	EtuReu
	Grp.		Q1	Md	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Prop.	Prop.
BCH	**QV**	60	6.6	8.67	11.5	5.25	8.25	11.5	0.91	0.98	1	1	1	1	2	4	10.5	1	3	8	0.55	1.88	3.96	0.18	0.20
	QNV	184	7.62	9.75	12.5	6.12	10.5	13.38	0.90	0.98	1	1	1	1	0	0	0	1	2	4	0	1	2.26	0.33	0.28
	NQV	227	5.6	8	10.17	4.25	6.5	9.75	0.79	0.95	1	0.83	1	1	1	3	6	0	0	0	0	0	0	0.24	0.12
	NQNV	1145	5.5	8	10.88	3.75	6.75	10.5	0.76	0.95	1	0.33	0.83	1	0	0	0	0	0	0	0	0	0	0.48	0.10
HBD	**QV**	52	7.75	10.25	13.75	6.62	9.25	12	0.93	0.98	1	1	1	1	2	4	10	1	3	6	1	2	5	0.19	0.22
	QNV	149	8.29	11.67	14	8.44	11.5	13	0.93	0.98	1	1	1	1	0	0	0	1	2	4	0	2	5	0.35	0.29
	NQV	262	6	8.67	11.27	4.5	7.25	10.5	0.88	0.96	1	1	1	1	1	2	6	0	0	0	0	0	0	0.19	0.11
	NQNV	1148	6	9.25	12.5	4	7.75	10.94	0.71	0.95	1	0.5	1	1	0	0	0	0	0	0	0	0	0	0.48	0.10
BCE	**QV**	39	7.29	10.6	12.7	N/A	N/A	N/A	0.83	0.98	1	1	1	1	1	3	8	1	3	10.5	0.38	1	2.41	0.28	0.39
	QNV	75	7.15	9.3	11.95	N/A	N/A	N/A	0.92	0.98	1	1	1	1	0	0	0	1	2	3	0.2	1	2	0.23	0.23
	NQV	118	6	8	10.8	N/A	N/A	N/A	0.9	0.98	1	1	1	1	1	2	3.5	0	0	0	0	0	0	0.23	0.14
	NQNV	1368	5	7.67	10.75	N/A	N/A	N/A	0.74	0.95	1	0.4	1	1	0	0	0	0	0	0	0	0	0	0.44	0.11
ANT	**QV**	38	11.2	13.9	15.4	13.5	15.25	16.5	1	1	1	1	1	1	2	3	7	3	6	16	0.11	0.47	0.7	0.43	0.7
	QNV	37	6	9.2	13.2	7.5	11.5	13.5	0.94	1	1	0.8	1	1	0	0	0	1	3	6	0	0.38	1	0.37	0.24
	NQV	23	8.57	10	12.7	8.12	11.88	14.69	0.98	0.98	1	0.9	1	1	1	1	2.75	0	0	0	0	0	0	0.53	0.22
	NQNV	1505	5	8	11.8	3.5	7	12	0.76	0.95	1	0	0.6	1	0	0	0	0	0	0	0	0	0	0.40	0.1

N/A : Aucune donnée disponible pour ce cours

Tableau 5.5 – Statistiques descriptives des des étudiants à travers des sous-populations pour les 9 variables considérées, pour chacun des 4 cours, pour chacune des 4 sous-populations considérées (QV, QNV, NQV, NQNV)

Cours	NotMoy	NotFin	AssGlb	AssCou	NbQst	NbVotRec	EtuReu	EtuRed
BCH	.005	.027	.023	.000*	.000*	.000*	.022	.000*
HBD	.000*	.015	.425	.000*	.000*	.000*	.026	.000*
BCE	.551	N/A	.000*	.000*	.000*	.000*	.000*	.000*
ANT	.740	.004*	.383	.000*	.000*	.000*	.000*	.280

*significatif avec $p < .05$ après correction Holm-Šidák (.000 signifie $p < .001$)

Tableau 5.6 – Comparaison des caractéristiques des votants et non-votants (V vs. NV)

	QV vs. QNV					
	NotMoy	NotFin	AssGlb	AssCou	EtuReu	EtuRed
BCH	.054	.064	.106	.633	.182	.009
HBD	.199	.002	.490	.588	.329	.026
BCE	.313	N/A	.632	.867	.090	.571
ANT	.000*	.000*	.570	.011	.000*	.572

*significatif avec $p < .05$ après correction Holm-Šidák (.000 signifie $p < .001$)

Tableau 5.7 – Comparaison des caractéristiques des votants et non-votants pour les étudiants qui ont posé des questions (QV vs. QNV)

ou non. Par conséquent, nous avons fait des comparaisons deux à deux entre QV et QNV (pour mesurer l'impact du vote chez les étudiants posant des questions) et entre NQV et NQNV (pour mesurer l'impact du vote chez les étudiants ne posant pas de questions), sur chacun des 4 cours. Deux variables (NbQst et NbVotRec) dépendantes du fait de poser une question n'étaient pas pertinentes et ont donc été exclues de cet ensemble d'analyses. Nous avons effectué un total de 46 tests ici (2 comparaisons de population avec 4 cours et 6 variables à chaque fois, sauf NotFin, manquante pour BCE) et les seuils de significativité ont été corrigés comme précédemment par la méthode de Holm–Šidák pour éviter les erreurs de type I.

Résultats. 13 des 46 tests ont révélé des résultats statistiquement significatifs (en utilisant une valeur p critique de 0.05 avec la correction susmentionnée), comme l'indiquent les Tableaux 5.7 et 5.8. En ce qui concerne les QV par rapport à QNV, les seuls résultats significatifs ont été obtenus sur ANT, le seul cours du deuxième semestre, en dépit d'effectifs plus réduits : les étudiants qui ont posé des questions et voté ont mieux réussi que ceux n'ayant fait que poser des questions, aussi bien aux QCM du cours (NotMoy) qu'au concours final (NotFin) et ont mieux réussi le concours final (EtuReu) (sont acceptés en deuxième année). Aucun résultat significatif n'a été trouvé pour BCH, HBD ou BCE.

En ce qui concerne les NQV par rapport à NQNV, les étudiants qui ont voté ont suivi ce cours plus souvent que ceux qui n'ont pas voté, à travers les quatre cours. Comme nous l'avons déjà observé en 5.4.2, la population des étudiants qui n'ont pas voté contenait plus d'étudiants redoublants que la population des étudiants qui ont voté pour les trois cours du premier semestre ; aucune différence n'a été observée au cours du deuxième semestre (ANT). En termes de note, lorsqu'une différence a été observée, les étudiants n'ayant pas posé de questions et n'ayant pas non plus voté ont eu une meilleure note sur les QCMs ainsi le concours final que ceux ayant uniquement voté. Aucun résultat significatif n'a été trouvé pour AssGlb et EtuRed. Nous pouvons donc répondre positivement à QR3.8, car les patterns entre votants et non-votants varient selon le fait que l'étudiant ait également posé une question ou non.

5.4.4 Comparaison de nature de questions posées et votées

Méthode. Pour comparer la nature des questions que posent les étudiants à la nature des questions sur lesquelles ils votent, nous avons dû nous concentrer sur la population des étudiants qui font les deux (QV). Pour ces étudiants, sur chacun des 4 cours, nous avons considéré toutes les questions sur lesquelles ils ont voté pour calculer la proportion de chaque type de question votée dans chaque dimension. Par exemple, si dans BCH, un élève a voté sur une question de ré-explication et une autre de vérification (étiquetées [Ree,0,Sch,0] et

	NQV vs. NQNV					
	NotMoy	NotFin	AssGlb	AssCou	EtuReu	EtuRed
BCH	.001*	.015	.006	.000*	.503	.000*
HBD	.000*	.000*	.608	.000*	.606	.000*
BCE	.040	N/A	.415	.000*	.308	.000*
ANT	.157	.710	.691	.000*	.086	.240

*significatif avec $p < .05$ après correction Holm-Šidák (.000 signifie $p < .001$)

Tableau 5.8 – Comparaison des caractéristiques des votants et non-votants pour les étudiants qui n'ont pas posé des questions (NQV vs. NQNV)

	Ree	App	Ver	Exe	Sch	Cor	Def	Man	Rai	Rol	Lie	Err	Con	Att
BCH	.277	.966	.100	.000*	.000*	.000*	.151	.945	.002	.008	.182	.004	.000*	.996
HBD	.437	.676	.186	.092	.000*	.345	.779	.282	.000*	.002	.099	.812	.000*	.805
BCE	.444	.062	.111	.344	.116	.133	.156	.014	.052	.523	.045	.544	.000*	.389
ANT	.710	.272	.205	.112	.013	.179	.099	.499	.003	.181	.059	.485	.000*	.433

*significatif avec $p < .05$ après correction Holm-Šidák (.000 signifie $p < .001$)

Tableau 5.9 – Différences entre questions votées et questions posées selon la nature des questions, pour les étudiants faisant les deux

[Ver,0,0,Con]), sur la dimension 1, il aurait voté à 50% sur des questions de valeur "Ree" (ré-explication) et à 50% de ses questions de valeur "Ver" (validation). Ces proportions sont codées entre 0 et 1, de sorte que pour chaque étudiant, sur chaque cours, on obtient un vecteur de vote composé de 14 (3+3+5+3) valeurs comprises entre 0 et 1, et en suivant la même approche pour les questions posées, on peut également obtenir un vecteur de question posée de 14 valeurs (*cf.* section 5.2.3).

Une fois le pré-traitement effectué, la comparaison des questions votées aux questions posées consistait à comparer pour chaque cours, pour chaque valeur d'une dimension (ex : la valeur "Ree" de la dimension 1), la distribution de la proportion des questions posées par les étudiants et celle de questions votées dans cette dimension. En d'autres termes, comparer deux distributions (non distribuées normalement) entre 0 et 1 pour la même population d'élèves, ce qui a été fait en effectuant 14 tests de Wilcoxon, en utilisant une nouvelle fois la méthode Holm-Šidák pour ajuster la valeur p critique.

Résultats. 9 des 56 tests ont révélé des résultats statistiquement significatifs (*cf.* Tableau 5.9), répondant ainsi positivement à la QR3.9. Dans chacun des neuf cas, les élèves posaient plus de questions du type considéré qu'ils ne votaient sur des questions de ce type. Pour la catégorie "vérification des connaissances" (dim4-2, "Con"), dans les 4 cours considérés, les étudiants posaient plus de questions de ce type qu'ils ne votaient sur ces questions. Des tendances similaires ont été observées pour les "exemples" (dim2-1, "Exe") dans BCH, les "schémas" (dim2-2, "Sch") dans BCH et HBD, la "correction" (dim2-3, "Cor") dans BCH et la "raison" (dim3-3, "Rai") dans BCH.

5.4.5 Bilan

En comparant les votants et non-votants sur la plateforme en ligne PACES, et en particulier en distinguant les étudiants qui ont également posé des questions de ceux qui ne l'ont pas fait, des différences claires sont apparues en termes de notes et de présence. Pour les élèves qui ont à la fois voté et posé des questions, des différences ont pu être observées entre la nature des questions sur lesquelles ils ont voté et la nature de celles qu'ils ont choisi de poser eux-mêmes.

En ce qui concerne les notes, la première analyse a révélé que le comportement de vote n'avait pas de lien clair avec celles-ci : voter était négativement associé à la note moyenne pour HBD, mais positivement associé à la note finale pour ANT. Cependant, en distinguant si les étudiants posent également des questions ou non, une image plus claire apparaît : le vote est bénéfique pour les étudiants qui posent également des questions au second semestre, et négatif pour les étudiants qui ne votent qu'au premier semestre. Cela pourrait indiquer qu'au début de l'année, les élèves qui ne font que voter sans poser de questions éprouvent des difficultés et qu'il pourrait être avantageux pour eux d'essayer de formuler leurs propres questions au lieu de voter sur celles des autres élèves. Au contraire, plus tard dans l'année, le vote est une bonne activité complémentaire aux questions posées.

En ce qui concerne l'assiduité, le vote semble être associé aux étudiants qui sont plus susceptibles d'être présents en cours, en particulier pour les étudiants qui ne posent pas de questions. Il est difficile de déterminer si les étudiants votent parce qu'ils ont l'intention d'aller au cours ou s'ils sont plus susceptibles d'y assister parce qu'ils ont voté. Cependant, une autre analyse comparant les étudiants qui posent des questions et ceux qui ne le font pas (Q et NQ) a révélé une relation positive similaire entre le fait de poser des questions et de suivre le cours. Proposer aux votants de poser une question avant de voter n'affecterait donc probablement pas négativement leurs chances d'assister au cours.

En termes de questions posées, les analyses révèlent une forte relation entre le vote et le comportement des personnes qui posent des questions, conformément à l'analyse comparant Q à NQ dans la section 5.1, où Q était positivement associé avec la variable NbVoteFait.

Enfin, nous constatons que les étudiants qui suivent les cours pour la deuxième année sont globalement moins actifs en termes de votes, tout comme ils étaient moins actifs en termes de questions posées. Ce résultat concorde avec les observations des enseignants selon lesquelles ces étudiants ont tendance à être là surtout pour repasser l'examen final et sont donc moins impliqués dans la dynamique des cours qu'ils ont déjà suivis. En analysant les questions sur lesquelles les étudiants votent plus qu'ils ne posent de questions, on voit que les étudiants semblent plus susceptibles de poser une question lorsqu'il s'agit de vérifier les connaissances "Con" ou lorsqu'il s'agit d'un élément très spécifique du cours (un schéma "Sch", un exemple "Exe", une correction "Cor"). Ce résultat est plutôt logique, puisque ces questions sont plus spécifiques et propres à un besoin individuel, qu'une demande de réexplication d'un concept général qui a plus de chance d'être une demande soutenue par plusieurs étudiants.

5.5 Synthèse

Nous avons montré dans ce chapitre, comment à partir de l'annotation automatique à base de règles d'experts et du schéma de codage développé dans le chapitre 4 nous avons pu identifier des profils des étudiants, en utilisant exclusivement les proportions de type de

questions qu'ils ont posées et leur évolution sur le temps. Deux clusters extrêmes (les élèves passifs en difficulté : notes inférieurs à la moyenne ayant des questions populaires et les élèves actifs pointilleux : supérieurs à la moyenne ayant des questions impopulaires) sont toujours apparus, avec parfois un cluster intermédiaire (les élèves supérieurs à la moyenne ayant des questions populaires). Par conséquent, nous avons répondu positivement à la QR3 à l'aide du clustering. Nous avons également caractérisé les étudiants qui n'ont pas posé de questions (moins d'assiduité en cours, moins de réussite et moins de votes) et les avons comparés aux étudiants qui ont posé des questions (forte assiduité, plus de réussite et plus de votes). Nous avons également étudié l'évolution des étudiants durant le premier semestre en utilisant les proportions de questions qu'ils ont posées sur chaque mois. Malheureusement, les résultats n'ont pas révélé de tendances claires entre la nature des questions et l'état d'évolution des étudiants.

D'un autre côté, la comparaison des questions posées et votées par les étudiants révèle que, dans notre cas, le vote ne semble pas être un bon substitut au fait de poser des questions. En résumé, voter est une bonne stratégie pour les étudiants sachant déjà formuler leurs propres questions. En revanche, pour ceux en difficulté, cela peut retarder la prise de conscience de leurs lacunes et leur capacité à les combler activement. Les résultats suggèrent qu'encourager les étudiants à formuler leurs questions, plutôt que de se contenter de voter sur les questions des autres, pourrait être une stratégie positive pour l'apprentissage. En effet, il est possible que pour certains étudiants, voter donne le sentiment de faire ce qui est attendu d'eux, sans pour autant développer les stratégies métacognitives mises en jeu lorsqu'on se pose ses propres questions (identifier les concepts clés, tester sa compréhension, résumer ce qui a été appris, *etc.*). Cela pourrait être fait en encourageant les étudiants à poser une question avant de pouvoir consulter celles des autres. Bien que positive du point de vue des étudiants, cette solution aurait pour effet de densifier encore davantage le « mur de questions » actuellement reçu par les enseignants chaque semaine. Son implémentation ne pourrait donc se faire qu'en lien avec une méthode de visualisation plus efficace des questions posées, ce qui est une des perspectives de ce travail (que nous allons explorer dans le chapitre 7).

Le travail présenté dans ce chapitre se heurte à plusieurs limites : la première est liée aux données collectées par la plateforme. En effet, même si tous les étudiants se connectent à la plateforme de questions, nous n'avons pas accès à des logs permettant de savoir s'ils ont vraiment lu les autres questions posées. La seconde est liée au fonctionnement même de la plateforme qui biaise les votes possibles : en effet, les étudiants qui se connectent en premier n'ont pas la possibilités de voir des questions sur lesquelles ils peuvent voter (sauf s'ils se reconnectent par la suite pour voir les nouvelles questions). On pourrait imaginer une solution consistant à demander aux étudiants à poser des questions pendant 5 jours (sans possibilité de voir les questions des autres), puis se connecter les 2 jours suivants pour pouvoir uniquement voter. Séparer temporellement les phases de questionnement et de vote rejoint aussi l'idée précédemment mentionnée qu'un étudiant devrait pouvoir poser des questions avant de voter.

Chapitre 6

Vers une prédiction de profils des étudiants en ligne

Nous avons montré dans le chapitre 5 que nous avons pu caractériser les élèves à partir de leurs questions à l'aide du clustering. Notre objectif est maintenant d'essayer de répondre à la quatrième question de recherche, qui est de savoir si nous pouvons utiliser les analyses précédentes pour passer à la prédiction de profil de l'apprenant. Cela nous a amené à examiner deux options : premièrement, effectuer un autre clustering sur de nouvelles données issues d'une autre formation en ligne, mais il était peu probable que nous obtenions des résultats similaires en utilisant seulement des données partielles dans un contexte en ligne, puisque le profil d'un apprenant n'est établi qu'à partir des données collectées sur une année entière. En effet, au-delà de l'analyse a posteriori, on souhaite pouvoir outiller l'enseignant pour qu'il puisse avoir plus d'informations sur les élèves qui lui posent des questions en cours d'année. Si les caractéristiques d'élèves sont systématiquement liées à leur profil de questions (c-à-d. pas de variations fortes d'une année sur l'autre), alors il serait possible d'extrapoler en disant non pas *"cet élève a telle ou telle caractéristique"* mais *"les années passées, les étudiants qui posaient ce type de questions avaient plutôt un profil de ce style"*. La deuxième option consistait à appliquer la partition de l'espace entraîné sur l'année N (2012-2013 dans le chapitre 5) à l'année N+1. La validité de cette approche repose sur l'hypothèse que la nature des questions posées dans un cours donné ne varie pas beaucoup d'une année à l'autre. Pour valider cette hypothèse, nous comparerons d'abord les clusters de l'année N avec les clusters de l'année N+1 (*cf.* Figure 6.1 - en bleu et vert). Ensuite, nous comparerons l'application du clustering de l'année N aux données de l'année N+1 (*cf.* Figure 6.1 - en orange), au clustering réalisé sur N+1 (qui ne peut être obtenu avant la fin du cours, *cf.* Figure 6.1 - en vert). Le processus de création d'un modèle prédictif par la réplication des analyses précédentes sur les prochaines itérations des cours est résumé dans la Figure 6.1. Concrètement, lorsqu'on utilise l'algorithme K-Means, l'application d'un clustering comme modèle prédictif consiste simplement à classer tout nouveau point de données (ici, un étudiant) dans un cluster basé sur le centroïde le plus proche (ici en utilisant la même distance euclidienne utilisée pour créer les clusters à l'origine). En d'autres termes, si un profil de question d'étudiant en année N+1 est plus proche du centroïde du cluster A en année N que des centroïdes des clusters B, C et D, alors cet étudiant va être classé dans le cluster A. Ces résultats ont donné lieu à une publication dans le journal JLA (Journal of Learning Analytics) [Harrak *et al.* , 2019c].

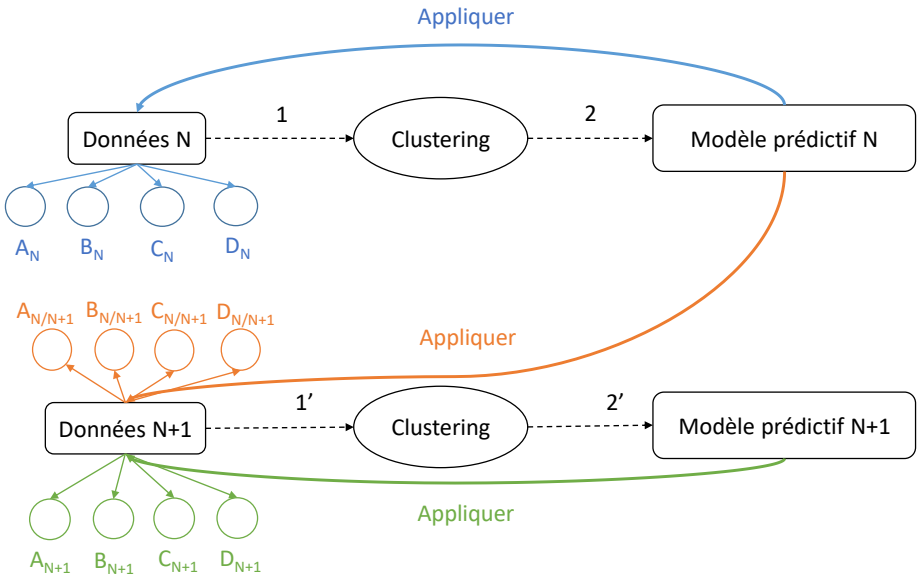


Figure 6.1 – Réplication du clustering sur les prochaines itérations

Plus précisément, notre objectif est de répondre à deux questions :

QR4.1 Lorsque l'on effectue deux clusterings distincts sur deux ensembles de données provenant de deux années différentes, la qualité intrinsèque du clustering est-elle similaire ou varie-t-elle ? En d'autres termes, il est important de regarder si l'application du modèle de clustering de l'année N aux données de l'année N+1 donne des clusters d'une qualité comparable au clustering de l'année N, même comparé à un clustering effectué directement sur l'année N+1 (*cf.* Figure 6.1).

QR4.2 Les caractéristiques de clusters des étudiants sont-elles similaires d'une année à l'autre ?

6.1 Evaluation de clusters : indices de qualité

L'objectif des techniques de classification non supervisée est de produire des classes avec une cohésion maximale (similarité intra-classes) tout en réalisant un maximum de séparation (dissimilarité inter-classes) entre les classes de la partition obtenue (on en a d'ailleurs vu quelques exemples dans la section 5.2.3). Afin de mieux évaluer la qualité des clusters obtenus sur différentes années, nous présentons dans un premier temps un éventail d'indices permettant de mesurer la qualité des clusters et essayons de choisir celui le plus adapté à nos données. Nous allons présenter ici les 4 indices suivants :

1. Inertiels
2. Dunn
3. Davies-Bouldin
4. Silhouette

Il existe d'autres indices de qualité comme les indices opérant sur réseaux de neurones (ex : "performance"), qui ne sont pas pertinents pour notre travail ici.

6.1.1 Indices inertiels

Les indices inertiels [Lebart *et al.* , 1995] sont les plus connus et les plus utilisés pour évaluer la qualité d'un clustering. L'inertie **intra-classes** permet de mesurer le degré d'homogénéité entre les objets appartenant à la même classe. Elle calcule leurs distances par rapport au point représentant le profil de la classe.

$$Intra = \frac{1}{n} \sum_{C \epsilon P} \frac{1}{2n_C} \sum_{i \epsilon C} \sum_{j \epsilon C} d(i,j)^2 \tag{6.1}$$

Où :

$n =$ le nombre total d'objets
$n_c =$ le nombre d'objets dans la classe C
$d =$ mesure de distance

L'inertie **inter-classes** mesure le degré d'hétérogénéité (ou séparation) entre les classes. Elle calcule les distances entre les points représentant les profils des différentes classes de la partition.

$$Inter = \frac{1}{n} \sum_{C \epsilon P} n_c d^2(C, C_G) \tag{6.2}$$

Avec c le centre de la classe C et C_G est le centre du nuage de points. Plus les données à l'intérieur des classes sont homogènes, plus leurs distances par rapport au point représentant la classe sont faibles. Par conséquent, une valeur faible de l'inertie intra-classes décrit une homogénéité des données à l'intérieur des classes. Plus les classes sont hétérogènes entre elles, plus les distances entre les points représentant les profils des classes sont élevées. Donc, une valeur élevée de l'inertie inter-classes traduit une hétérogénéité entre les classes. Cet indice a cependant le défaut de diminuer quand on augmente le nombre de classes.

6.1.2 Indice de Dunn

Cet indice a été créé par Dunn [1974] et consiste à chercher la distance minimale qui sépare deux classes dans la partition tout en tenant compte de la distribution des éléments à l'intérieur des classes. La partition produisant la plus grande valeur de $Dunn$ correspondra à la meilleure partition. L'indice $Dunn$ fournit un bon compromis entre la maximisation de la dissimilarité inter-classe et la minimisation de la dissimilarité intra-classe de la partition. Cependant, il n'est pas normalisé et il est difficile de comparer l'indice de deux jeux de données différents.

$$Dunn = \underset{1 \leq i \leq n}{min} \left\{ \begin{array}{cc} min & \left\{ \frac{d(c_i,c_j)}{max(d'(C_k))} \right\} \\ {\scriptstyle 1 \leq j \leq n} & {\scriptstyle 1 \leq k \leq n} \\ {\scriptstyle j \neq i} & \end{array} \right\}$$

6.1.3 Indice de Davies-Bouldin

L'indice de Davies-Bouldin (DB) [Davies & Bouldin, 1979] traite chaque classe individuellement et cherche à mesurer à quel point elle est similaire à la classe qui lui est la plus proche.

$$DB = \frac{1}{n}\sum_{j=1}^{n} R_{ij}$$

Pour chaque classe i de la partition, on cherche la classe j qui maximise l'« indice de similarité » R_{ij} décrit comme suit :

$$R_{ij} = \frac{I(c_i) + I(c_j)}{I(c_i, c_j)}$$

$I(c_i)$ représente la moyenne des distances entre les individus appartenant à la classe c_i et son centre. $I(c_i, c_j)$ représente la distance entre les centres des deux classes c_i et c_j. La meilleure partition est donc celle qui minimise la moyenne de la valeur calculée pour chaque classe. En d'autres termes, la meilleure partition est celle qui minimise la similarité entre les classes.

Les indices Dunn et Davies-Bouldin mélangent donc à la fois les inerties intra-classes et les inerties inter-classes.

6.1.4 Indice de silhouette

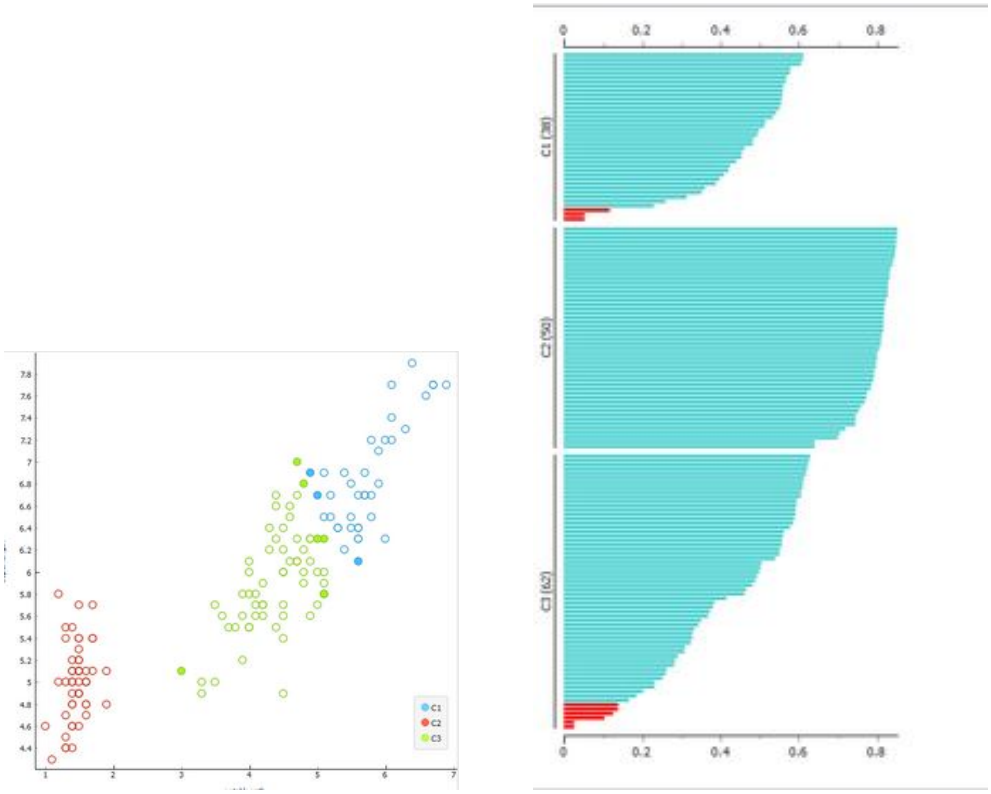
L'indice silhouette [Rousseeuw, 1987] est différent des indices de qualité présentés ci-dessus car il travaille à l'échelle microscopique, c'est-à-dire qu'il s'intéresse aux individus en particulier et non pas seulement aux classes. Le but de l'indice de silhouette est de vérifier si chaque individu a été bien classé. Pour cela, et pour chaque individu i de la partition, on calcule la valeur suivante :

$$-1 \leq S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \leq 1 \tag{6.3}$$

Où : $a(i)$ représente la distance moyenne qui le sépare des autres individus de la classe à laquelle il appartient et $b(i)$ représente la distance qui le sépare des individus appartenant à la classe la plus proche.

Quand $S(i)$ est proche de 1, l'individu est bien lié au cluster dans lequel il est étiqueté et éloigné des autres : la distance qui le sépare de la classe la plus proche est très supérieure à celle qui le sépare de sa classe. Par contre, si $S(i)$ est proche de -1, cela veut dire que l'individu est associé à un autre cluster que celui auquel le partitionnement l'a associé au départ. Mais si S(i) est proche de 0 alors il pourrait également être associé à la classe la plus proche, c'est à dire l'individu est situé entre deux centroïdes. L'indice de silhouette global de la partition est calculé à partir de la moyenne entre les indices de ses éléments. Un exemple [1] de silhouette est illustré par la Figure 6.2 pour évaluer visuellement la qualité du cluster et le degré d'adhésion d'une instance à un cluster.

1. Source : https://blog.biolab.si/tag/visualization/



(a) répartition des instances dans les clusters

(b) indice de silhouette

Figure 6.2 – Exemple de visualisation de silhouette sur un échantillon donné (gauche : indice de silhouette pour chaque instance du cluster ; droite : répartition des instances dans les clusters) – source : biolab.si

La Figure 6.2 affiche la distance moyenne entre les instances au sein du cluster et les instances du cluster le plus proche. Pour une instance donnée, la silhouette proche de 1 indique que l'instance est proche du centre du cluster (Cluster 2). Les instances avec des scores de silhouette proches de 0 (les éléments indiqués en rouge dans les clusters C1 et C3 et représentés par des points remplis dans le scatter plot) sont à la frontière entre deux clusters. Dans l'ensemble, la qualité du clustering pourrait être évaluée par les scores de silhouette moyens des instances dans chaque cluster.

L'un des principaux avantages de l'indice de silhouette, c'est qu'il est normalisé. En effet, une valeur proche de 1 est toujours considérée bien pour cet indice, quel que soit le cluster à évaluer, tandis que les indices Davies-Bouldin et Dunn ne sont pas normalisés (il est donc difficile de comparer deux valeurs issues de partitionnements différents pour ces deux indices). Nous avons donc choisi l'indice silhouette qui nous parait le plus pertinent pour évaluer la qualité des clusters. Une des limites de ce coefficient est qu'il est mieux adapté aux clusters convexes (tel k-means) qu'aux autres (ex : "basés sur densité", DBSCAN), mais cette limite ne pose pas de problème dans notre cas.

6.2 Différence de qualité intrinsèque des clusters d'une année à l'autre

Pour répondre à la question de recherche QR4.1, nous avons comparé avec l'indice de silhouette la qualité de 3 clusterings : le clustering calculé dans la section 5.2.3 sur les données de 2012 (*cf.* Figure 6.1 - flèche 1), un clustering obtenu de manière similaire avec K-Means sur les données de 2013 (*cf.* Figure 6.1 - flèche 1'), et enfin les points de données étiquetées par l'application du modèle prédictif 2012 (*cf.* Figure 6.1 - en orange) sur les données 2013. Nous avons calculé l'indice de silhouette pour chaque cluster et un indice moyen global pour chaque clustering. Les résultats de l'indice de silhouette pour chaque cluster et chaque cours pour les trois partitions de données sont présentés dans le Tableau 6.1. Le nombre de clusters obtenus peut varier d'un clustering à l'autre pour chaque cours.

Les résultats de l'indice de silhouette ont montré que le modèle prédictif de 2012 appliqué à l'ensemble des questions de 2013 (en orange) et le clustering effectué directement sur l'ensemble de données de 2013 (en vert) sont similaires en termes de qualité des clusters (mais qui sont bien évidemment différents en termes de clusters obtenus, même si une comparaison pour chaque cours a révélé quelques similarités). Il semble donc que dans notre contexte, l'approche consistant à utiliser le modèle utilisé à des fins prédictives résultant de l'application du clustering effectué au cours d'une année donnée sur les données de l'année suivante est aussi efficace, en termes de qualité des clusters, tant que le clustering est directement appliqué sur l'ensemble de données de l'année suivante. Cependant, ce résultat est nécessaire mais pas suffisant, car nous ne savons pas encore si les caractéristiques de ces clusters (en termes de notes, d'assiduité, *etc.*) sont identiques, ce qui est obligatoire si nous voulons fournir de façon fiable plus d'informations aux enseignants sur les étudiants qui posent ces questions. Bien que la qualité des clusters obtenus est similaire d'après les résultats, les valeurs de silhouette tournent autour de 0.20, ce qui est assez moyen en termes de qualité.

Cours	Cluster	Modèle clustering 2012 sur données 2013	Clustering 2012	Clustering 2013
BCH	$Cluster_0$	0.19	0.18	0.16
	$Cluster_1$	0.20	0.20	0.29
	$Cluster_2$	0.17	0.31	0.08
	$Cluster_3$	0.13	0.11	0.16
	Indice moyen	0.17	0.19	0.15
HBD	$Cluster_0$	0.06	0.06	0.19
	$Cluster_1$	0.28	0.37	0.20
	$Cluster_2$	0.27	0.15	0.12
	$Cluster_3$	0.25	0.23	0.17
	Indice moyen	0.21	0.23	0.21
BCE	$Cluster_0$	0.21	0.27	0.12
	$Cluster_1$	0.12	0.19	0.15
	$Cluster_2$	0.15	0.17	0.09
	$Cluster_3$	-	-	0.26
	Indice moyen	0.19	0.22	0.20
ANT	$Cluster_0$	0.08	0.18	0.10
	$Cluster_1$	0.15	0.18	0.20
	$Cluster_2$	0.29	0.21	0.10
	$Cluster_3$	-	-	0.26
	Indice moyen	0.16	0.19	0.19

(-) : Pas de cluster disponible pour ce cours

Tableau 6.1 – Résultats de l'indice de silouhette pour chaque cluster et chaque cours de chacun des 3 clusterings

6.3 Caractéristiques des étudiants d'une année à l'autre

Pour répondre à cette question, nous avons caractérisé les clusters du modèle de 2012 sur les données de 2013 selon les 9 variables utilisées pour le clustering de 2012. Ensuite, nous avons essayé de voir s'il y avait une correspondance entre les clusters obtenus sur l'ensemble de données 2013 (*cf.* Figure 6.3 - en orange) et ceux identifiés sur l'ensemble de données 2012 (en bleu). Les clusters de 2012 et 2013 sont comparés en termes de médiane, 1^{er} et $3^{ème}$ quartiles des variables dépendantes (*cf.* Tableau 3.4, moyenne et écart-type de EtuReu et EtuRed) pour BCH uniquement (Figures HBD, BCE et ANT sont en annexe D).

L'analyse de résultats des clusters obtenus sur les données de 2013 correspondaient bien à celles de 2012. Nous constatons que les deux clusterings comparés sont presque similaires sur l'ensemble de variables dépendantes, avec une légère différence observée à chaque fois sur le NbVoteFait. Globalement, les élèves identifiés dans les clusters A, B et D sur les données 2013 avaient des caractéristiques similaires à celles des élèves identifiés sur les données 2012 (en termes de performance, d'assiduité, de nombre de questions posées et de nombre de votes). Par ailleurs, les élèves de 2013 et 2012 partagent les mêmes patterns de questions, c'est-à-dire que la nature des questions posées par les élèves de chaque cluster reste la même (même si les centroïdes peuvent varier). Nous avons donc répondu positivement aux questions de recherche QR4.1 et QR4.2 (*cf.* 6).

6.4 Amélioration de la qualité du partitionnement

Nous avons fait l'hypothèse que la performance du modèle utilisé à des fins prédictives pourrait être améliorée en l'appliquant à des données regroupées à partir des années précédentes. Pour le tester, nous avons annoté automatiquement les données des questions de trois années supplémentaires (2013, 2014 et 2015) et nous avons essayé de former trois modèles de prédiction basés sur trois regroupements différents (comme nous l'avons fait dans la section 5.2.3) : **M1** (à partir des données de 2014 uniquement), **M2** (à partir des données de 2013 combinées avec les données de 2014) et **M3** (à partir des données de 2012, 2013 et 2014) et appliqué chacun des trois modèles aux données de 2015. Ensuite, nous avons comparé les résultats de l'indice de silhouette des trois modèles pour voir s'ils s'amélioraient (c'est à dire si les clusters étaient de meilleure qualité). La comparaison des trois modèles est illustrée dans la Figure 6.4.

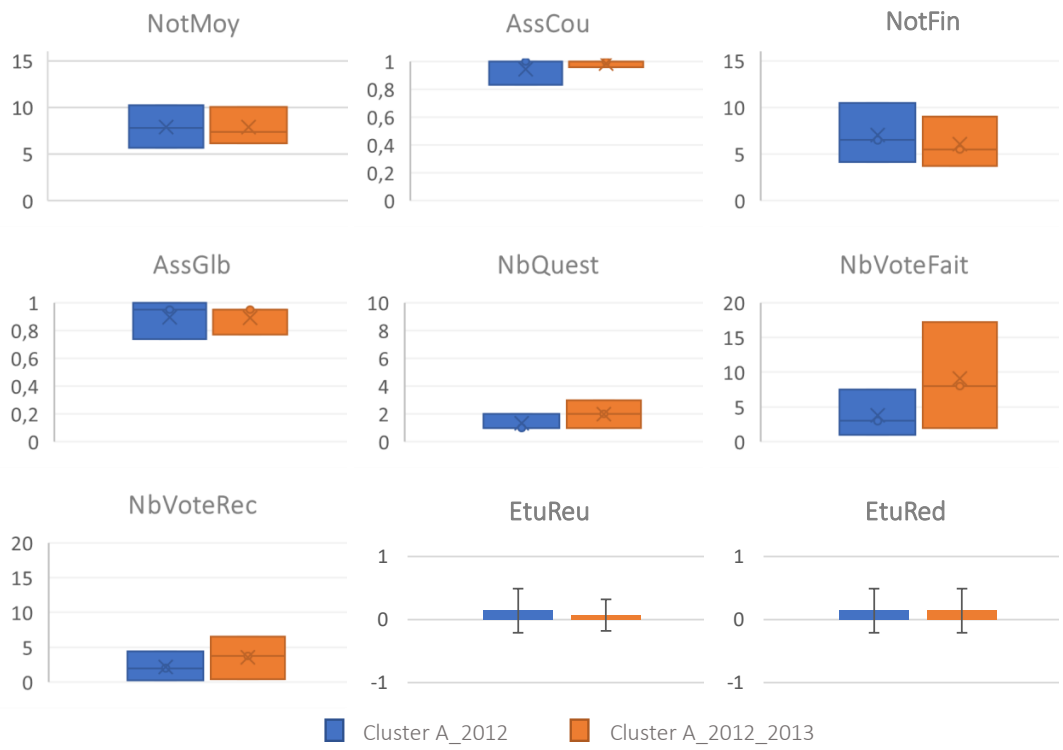
Les résultats obtenus sur la figure 6.4 ont montré que les trois modèles de chaque cours ont presque la même performance, c'est-à-dire que la performance du modèle ne s'améliorait pas avec des données supplémentaires, contrairement à notre hypothèse. D'une part, cela signifie que l'on ne peut pas améliorer la performance du modèle en utilisant les données des années précédentes : une explication envisageable est que les questions posées par les étudiants sont assez similaires d'une année à l'autre, ou que la légère amélioration due à l'augmentation des données est compensée par le fait que le contenu du cours peut évoluer d'une année à l'autre et devenir moins pertinent pour la prédiction. D'autre part, cela signifie qu'il suffit d'utiliser l'ensemble des données d'une seule année (idéalement, la précédente, pour limiter les variations liées aux changements de contenu du cours) pour obtenir des résultats acceptables.

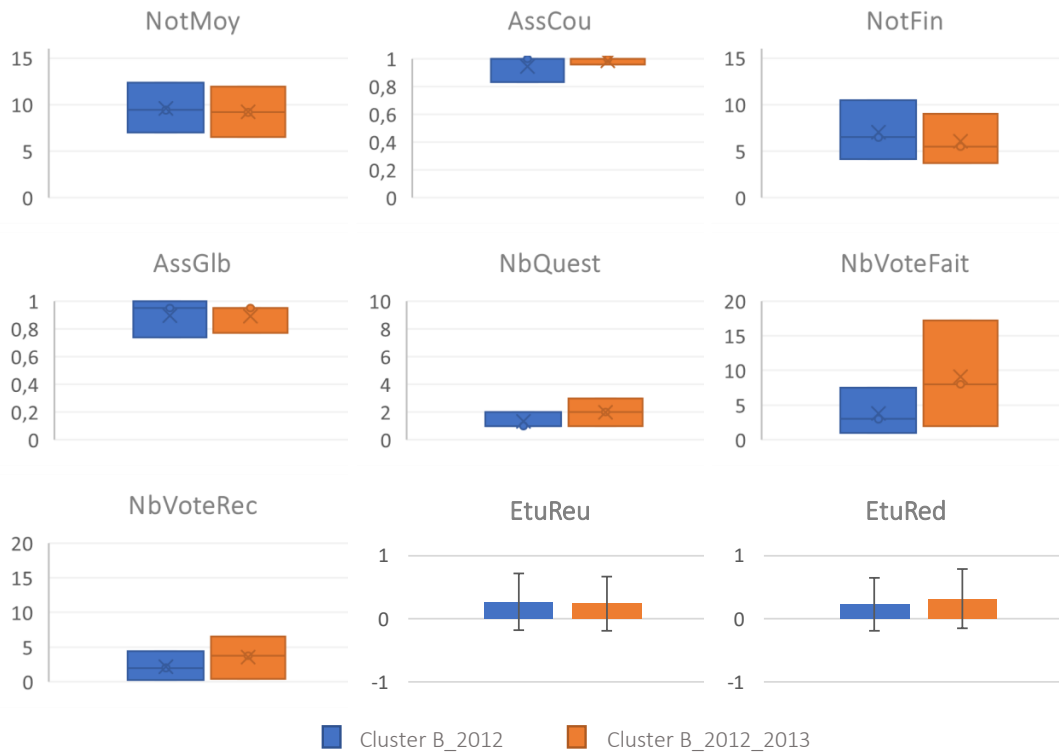
Les résultats ont montré que le modèle est stable d'une année à l'autre, même si les clusters sous-jacents ne sont pas extrêmement distincts (avec un indice de silhouette moyen entre 0.1 et 0.2). Mais ce qui reste particulièrement important pour notre objectif final est que la nature des questions posées par les étudiants et les caractéristiques de leur profil restent les mêmes, ce qui signifie que nous pouvons caractériser les étudiants à partir de leurs questions (prédire les profils des étudiants) en utilisant des analyses des années précédentes. Ceci confirme que nos clusters peuvent être utilisés à des fins prédictives, c'est-à-dire qu'on peut donner avec un niveau de confiance correct des informations sur ce qu'on estime être les caractéristiques de l'étudiant posant une question donnée. Il devrait donc nous permettre de fournir aux enseignants des informations supplémentaires sur les questions qu'ils ont reçues (même au début de l'année - lorsqu'il n'y a pas beaucoup de données disponibles sur les élèves qui les ont posées), sur la base des similitudes des questions avec les questions posées par les élèves dans les années précédentes. Par exemple, en leur faisant savoir qu'il existe une série de questions qui correspondent à celles posées l'année précédente par les élèves qui assistaient régulièrement au cours et qui ont finalement réussi.

6.5 Synthèse

Nous avons présenté dans cette section une approche qui devrait nous permettre de fournir des informations assez fiable sur le profil des élèves en ligne à partir de leurs questions à l'aide d'un modèle utilisé à des fins prédictives dérivé du clustering des données de l'année précédente. Les résultats ont révélé des clusters similaires obtenus sur différentes années, en termes de qualité et de caractéristiques de clusters d'étudiants pour chaque cours. Ceci nous a permis de répondre positivement aux questions de recherche posées au début de ce chapitre. Ce résultat ouvre des perspectives pour aider les enseignants en leur fournissant plus d'informations sur les élèves qui posent des questions en ligne, grâce aux similarités trouvées avec les élèves qui ont posé les mêmes type de questions l'année précédente.

Une des limites à souligner de ce travail est que le profil des étudiants qui posent les premières questions en ligne établi par le clustering n'est probablement pas fiable (car les clusters sont faits sur la base d'étudiants posant plusieurs questions, c'est à dire à 100% de chaque type de questions posées). D'ailleurs, nous n'avons pas mesuré dans cette analyse le nombre de questions nécessaire pour avoir un profil de questionnant fiable, ni le moment estimé pour l'obtenir afin d'utiliser ce modèle.


NotMoy
AssCou
NotFin
AssGlb
NbQuest
NbVoteFait
NbVoteRec
EtuReu
EtuRed
Cluster A_2012
Cluster A_2012_2013


NotMoy
AssCou
NotFin
AssGlb
NbQuest
NbVoteFait
NbVoteRec
EtuReu
EtuRed
Cluster B_2012
Cluster B_2012_2013

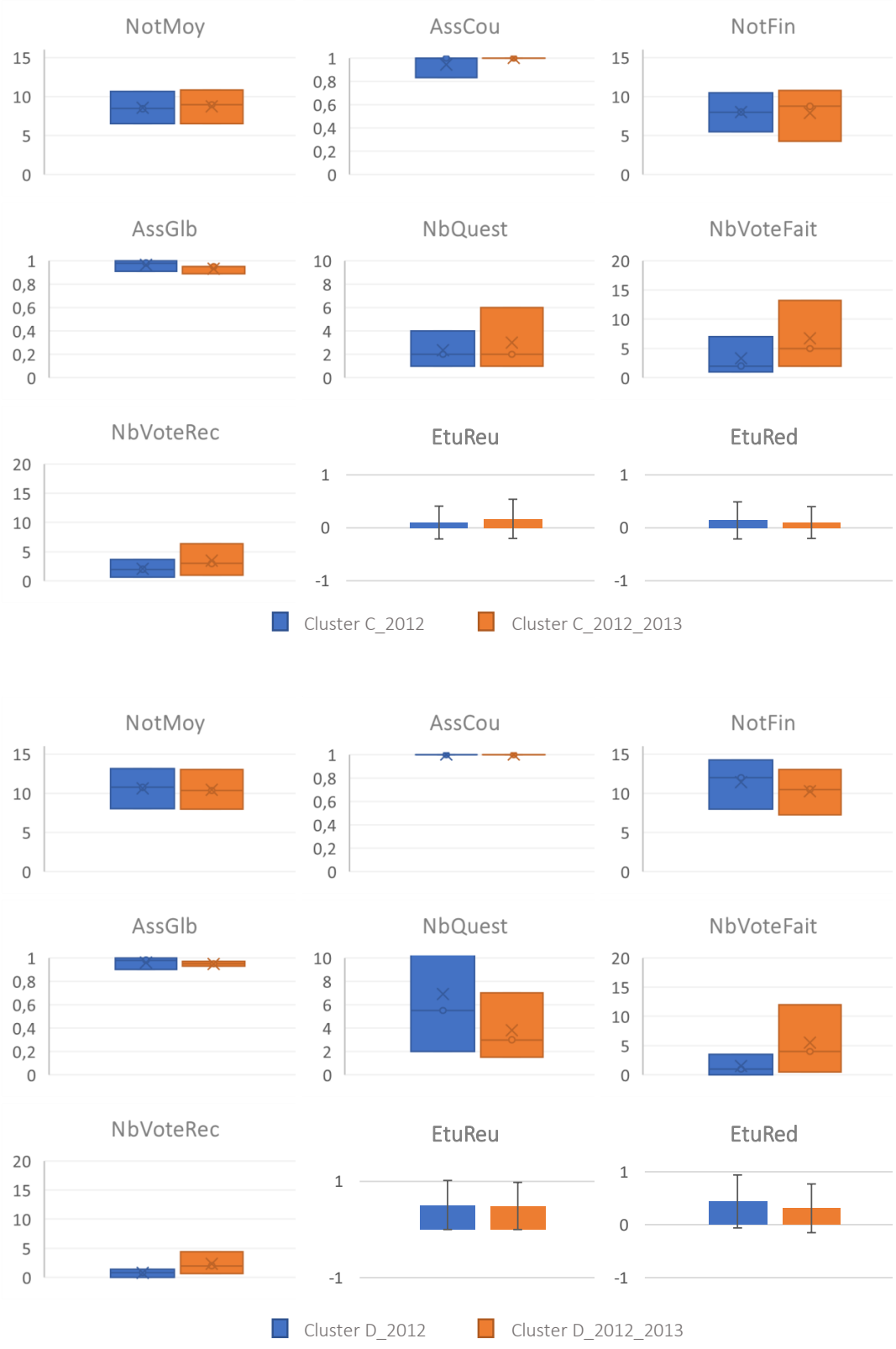


Figure 6.3 – Comparaison des clusters de 2012 et 2013 en termes de médiane (centre de la barre), 1^{er} quartile (bas de la barre) et $3^{ème}$ quartile (haut de la barre) des variables dépendantes et moyenne et écart-type de EtuReu et EtuRed pour chaque cluster de BCH

Figure 6.4 – Indice de silhouette des 3 modèles de clustering utilisés à des fins prédictives réalisés sur les données de 2015 pour chaque cours

Chapitre 7

Evaluation auprès des enseignants

Un des objectifs de ce travail consiste à aider les enseignants à préparer leurs séances de questions-réponses en présentiel (SEPI). Actuellement, les questions envoyées par mail aux enseignants contiennent comme unique information le nombre de votes reçus par question, ce qui permet uniquement une organisation en termes de questions populaires/impopulaires. Cependant, le travail mené dans les chapitres précédents rend possible d'envisager d'autres organisations de questions permettant de traiter les questions des étudiants différemment. Nous avons donc préparé trois organisations de questions alternatives (également appelées dans ce chapitre "visualisation"), en fournissant aux enseignants plus d'informations sur la nature des questions posées et les profils des étudiants. Cela nous a amené à faire un questionnaire à destination des enseignants de PACES leur présentant ces organisations de questions afin de recueillir leurs impressions sur celles-ci. Cette étude nous permettra également d'évaluer l'intérêt de notre travail et notre schéma de codage en particulier auprès des enseignants.

Dans la suite de ce chapitre, nous allons présenter les trois organisations de questions alternatives (textuelle, catégorielle et mixte). Ensuite, l'ensemble des questions posées aux enseignants dans le questionnaire et les données collectées. Nous allons analyser par la suite les réponses des enseignants et leurs choix d'organisations de questions pour répondre à la question de recherche QR5 (*cf.* section 1.1, qui consiste à savoir si les organisations proposées peuvent aider l'enseignant à préparer ses séances questions-réponses. Ces résultats ont donné lieu à une publication dans la conférence LAK20 [Article accepté : Harrak et al. Evaluating Teachers' Perceptions of Students' Questions Organization]

7.1 Organisation de questions

L'université de Grenoble dispose d'un système de formation hybride dans laquelle les étudiants doivent poser chaque semaine des questions à partir de supports de cours étudiés à distance avant le cours (selon une approche de classe inversée), comme présenté dans la section 3.1. Les enseignants de PACES reçoivent par mail la liste des questions par groupe d'étudiants avec pour chaque question le nombre de votes. Cependant, compte tenu du volume de questions posées, les enseignants n'ont souvent pas assez de temps pour répondre à chaque question et doivent donc sélectionner celles auxquelles ils vont répondre. Ils utilisent

donc les votes pour les aider dans ce choix. Nous allons présenter dans cette section trois organisations de questions alternatives que nous avons conçues, basées sur notre schéma de codage (*cf.* Tableau 4.4) et les résultats d'analyses de profils des étudiants (*cf.* section 5.2.3) : une organisation textuelle, une organisation catégorielle et une organisation mixte (mélangeant les deux précédentes).

7.1.1 Organisation textuelle

Dans l'organisation présentée par la figure 7.1, les questions des étudiants passent par l'analyse textuelle décrite dans la section 4.1. Les questions sont alors regroupées en fonction de la nature des questions posées (Ré-explication, Approfondissement, Vérification ou Autres, dimension 1 et 4), tel qu'indiqué par les mots-clés mis en gras dans la figure 7.1. Nous avons illustré cette organisation avec 8 exemples de questions.

7.1.2 Organisation catégorielle

Dans l'organisation illustrée par la figure 7.2, les questions des étudiants sont regroupées en fonction du profil des étudiants des années passées qui posaient ce type de questions (*cf.* chapitre 6). La mention "notes inférieures à la moyenne" n'est donc pas basée sur les notes des étudiants qui posent les questions, et serait donc disponible dès la première séance. L'organisation est illustrée avec les mêmes 8 exemples de questions.

7.1.3 Organisation mixte

Dans l'organisation présentée par la figure 7.3, nous avons combiné les deux informations précédemment présentées : le type de questions posées par l'étudiant, et une estimation de son niveau par rapport à des données des années passées.

7.2 Méthodologie de l'enquête

Les trois organisations ont été proposées aux enseignants via un questionnaire (en annexe E), accompagnées d'autres questions à travers LimeSurvey. Le questionnaire est composé de deux parties visant deux objectifs :

1. évaluer le degré d'utilisation des questions reçues par mail pour la préparation des SEPI (Séances d'Enseignement en Présentiel Interactive), et l'appréciation qu'en ont les enseignants
2. recueillir l'opinion des enseignants sur un système d'organisation de questions alternatif visant à faciliter la préparation des SEPI.

Pour aborder la cinquième question de recherche (QR5) introduite dans la section 1.1, nous avons été amenés à l'affiner en question suivante :

QR5.1 : Est-ce que les catégories de questions identifiées sont pertinentes pour les enseignants ?

La première partie du questionnaire consiste à évaluer l'expérience des enseignants en

Questions de ré-explication

➡ Demande de revenir sur un concept déjà expliqué en cours

Questions :

1. Pourrait-on **ré-expliquer** comment trouver le moment dipolaire d'une molécule ?
2. Pourriez vous **revenir** sur la notion de solutions tampons notamment sur les moyens de réaliser une solution tampon ?

Questions d'approfondissement

➡ Demande plus de précision (clarification) sur un concept donné, enlever une ambiguïté ou pour mieux comprendre

Questions :

3. **Comment** comparer deux atomes qui ne se trouveraient ni dans la même ligne, ni dans la même colonne ?
4. Pourriez-vous **expliquer** ce qui distingue l'atome de l'élément chimique ?

Questions de vérification

➡ Vérifier ou valider une hypothèse

Erreur/ contradiction

Questions :

5. Il semble qu'il y ait une erreur dans le discours de la diapositive 5 : vous dites que `` les ions Na+ et Nacl (Cl- ?)
6. Bonjours , dans l'exemple sur l'électrophorèse vous dites que les aa sont chargé négativement a ph=1 , alors_que leurs ph < phi il ne devrait pas être positif comme présenté sur l'exemple de séparation de mélange ?

Connaissances en cours

Questions :

7. est-ce que tous les métaux de transition sont réducteurs ?

Examen

Questions :

8. Doit on apprendre les métaux du bloc P par coeur ?

Autres Questions

Figure 7.1 – Exemple d'organisation de questions "Analyse Textuelle"

Étudiants en difficulté : notes < à la moyenne

Questions :

2. Pourriez vous **revenir** sur la notion de solutions tampons notamment sur les moyens de réaliser une solution tampon ?
7. est-ce que tous les métaux de transition sont réducteurs ?

Étudiants moyens

Questions :

1. Pourrait-on **ré-expliquer** comment trouver le moment dipolaire d'une molécule ?
3. **Comment** comparer deux atomes qui ne se trouveraient ni dans la même ligne, ni dans la même colonne ?
5. Il semble qu'il y ait une **erreur** dans le discours de la diapositive 5 : vous dites que `` les ions Na+ et Nacl (Cl- ?)
8. Doit on apprendre les métaux du bloc P par coeur ?

Étudiants bons : notes > à la moyenne

Questions :

4. Pourriez-vous **expliquer** ce qui distingue l'atome de l'élément chimique ?
6. Bonjours , dans l'exemple sur l'électrophorèse vous dites que les aa sont chargé négativement a ph=1 , alors_que leurs ph < phi il ne devrait pas être positif comme présenté sur l'exemple de séparation de mélange ?

Figure 7.2 – Exemple d'organisation de questions "Analyse Catégorielle"

N°	Questions de ré-explication	Faible	Moyen	Bon
1.	Pourrait-on **ré-expliquer** comment trouver le moment dipolaire d'une molécule ?	X		
2.	Pourriez vous **revenir** sur la notion de solutions tampons notamment sur les moyens de réaliser une solution tampon ?	X		
	Questions d'approfondissement			
N°	**Questions d'approfondissement**	**Faible**	**Moyen**	**Bon**
3.	**Comment** comparer deux atomes qui ne se trouveraient ni dans la même ligne, ni dans la même colonne ?		X	
4.	Pourriez-vous **expliquer** ce qui distingue l'atome de l'élément chimique ?		X	
N°	**Questions de vérification**	**Faible**	**Moyen**	**Bon**
	• **Erreur/ contradiction**			
5.	Il semble qu'il y ait une erreur dans le discours de la diapositive 5 : vous dites que `` les ions Na+ et Nacl (Cl- ?)		X	
6.	Bonjours , dans l'exemple sur l'électrophorèse vous dites que les aa sont chargé négativement a ph=1 , alors_que leurs ph < phi il ne devrait pas être positif comme présenté sur l'exemple de séparation de mélange ?			X
	• **Connaissances en cours**			
7.	est-ce que tous les métaux de transition sont réducteurs ?			X
	• **Examen**			
8.	Doit on apprendre les métaux du bloc P par coeur ?		X	

Figure 7.3 – Exemple d'organisation de questions "Analyse Mixte"

SEPI, leur appréciation de la classe inversée et la pertinence de nos catégories de questions (*cf.* Tableau 4.4). Nous allons donc présenter brièvement les différentes questions posées aux enseignants dans la première partie.

En ce qui concerne les variables liées à l'expérience des enseignants en SEPI, des questions portent sur le nombre de questions reçues et la proportion de questions traitées, et sont suivies d'une série de 8 affirmations relatives à l'ensemble de questions envoyées par mail pour lesquelles on mesure un degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »).

— nombre trop important

— impression de recevoir trop de questions

— impression de recevoir pas assez de questions

— désorganisation

— manque d'utilité pour la séance de SEPI

— pertinence

— rapport avec le cours

— nouveauté (par rapport aux années précédentes)

En ce qui concerne les variables liées à la classe inversée, une série de trois affirmations relatives à la pédagogie inversée est mesurée par un degré d'accord sur une échelle de 1 à 5.

— Facilité d'enseigner en classe hybride

— Appréciation de l'organisation

— Gain du temps et d'énergie

En ce qui concerne les variables liées à la pertinence perçue des catégories de questions auprès des enseignants, une série d'affirmations relatives à l'utilité de différents types de questions reçues par mail pour préparer la SEPI est mesurée par un degré d'accord sur une échelle de 1 à 5. On mesure donc l'intérêt des enseignants pour des questions :

— cherchant à aller plus loin (dim1 - App)

— demandant un exemple (dim2 - Exe)

— de vérification d'erreur ou contradiction en cours (dim4 - Err)

— populaires (ayant reçu de nombreux votes des autres étudiants, seul élément actuellement présent dans les mails)

Finalement, un dernier ensemble de questions portait sur les types de questions éventuellement ignorées par les enseignants : les questions mal formulées et celles qui traitent de plusieurs sujets.

Dans la deuxième partie du questionnaire, nous avons demandé à l'enseignant d'évaluer l'intérêt d'un système d'organisation de questions alternatif (*cf.* section 7.1) pour faciliter la préparation des SEPI. Le choix de ces organisations est basé essentiellement sur la nature de questions posées par les étudiants et leurs profils en termes de performance (à partir des données collectées en 2012-2013). Nous avons dû nous concentrer sur les catégories de questions principales dans notre schéma de codage (*cf.* Tableau 4.4) pour garder les organisations bien lisibles pour les enseignants. Chaque organisation est accompagnée d'une série d'affirmations, chacune de ces affirmations est mesurée par un degré d'accord sur une échelle de Likert de 1 à 5 - les deux dimensions mesurées sont la facilité de compréhension et l'intérêt de l'organisation en termes d'informations supplémentaires utiles pour la préparation des SEPI. D'autres questions

Temps réponse (s)		Comm -entaire	Âge		Discipline													Ancienneté		
Moy	σ^2	%	Moy	Ect	ANT	BCE	BCH	BPH	BSTAT	HBD	ICM	MAIEU	MAT	ODON	PHAR	PHS	SSH	$[1-5]$	$]5-10]$	$]10+]$
1044	618.15	>50%	48.25	9.79	4	4	5	4	1	4	2	2	1	0	3	4	4	11	13	12

Tableau 7.1 – Statistiques descriptives de la population des répondants en termes de (%) commentaires, (moyenne et écart-type) temps de réponse, âge et (nombre) discipline et ancienneté

portaient sur la manière d'utilisation des catégories présentées dans chacune des organisations en SEPI, l'appréciation de chaque organisation par rapport à ce qu'ils reçoivent actuellement (mieux, pareil, moins bien) et s'il y a des points à améliorer. À la fin de cette deuxième partie, les enseignants sont amenés à choisir une seule organisation (choix final) parmi quatre : les trois nouvelles organisations proposées (textuelle, catégorielle et mixte) et l'organisation actuelle (liste de questions, avec pour chaque question le nombre de votes reçus).

7.3 Données et codage

Le questionnaire a été envoyé à 58 enseignants au total, 37 d'entre eux ont répondu à l'ensemble des questions. Nous avons également filtré et traité les données aberrantes et manquantes dans le questionnaire. En effet, nous avons supprimé un répondant parmi les 37 qui a déclaré ne pas utiliser le système de questions envoyées par mail pour préparer ses SEPI, mais en revanche suivait la méthode ancienne (ex : "les questions sont proposées en tutorat et pas en SEPI"). Par conséquent, nous avons jugé utile de le supprimer puisqu'il n'est pas impliqué dans le système à évaluer. En ce qui concerne les données manquantes, une personne a manqué de préciser son expérience en enseignement, une valeur a été attribuée (]5-10]) en regardant les réponses données par la majorité des enseignants de son âge. Cette variable "expérience" a été recodée en catégorielle (tranches) pour faciliter le remplacement des valeurs manquantes. Les statistiques descriptives de la population des répondants ($N = 36$ enseignants, *cf.* Tableau 7.1) ont montré que les participants couvrent différentes tranches d'âge (de 34 à 69) et d'expérience, ainsi que presque toutes les matières enseignées. De plus, le temps de réponse global est suffisamment important pour estimer que les réponses sont valides, et de nombreux enseignants ont également pris le temps de laisser des commentaires.

7.4 Analyse de questions

La moyenne des notes attribuées par les enseignants dans le questionnaire sur une échelle de 1 à 5 par rapport à leur expérience en SEPI (histogramme de réponses de chaque affirmation en annexe F) :

- nombre trop important (moyenne : 2.65 et écart-type : 1.49)
- impression de recevoir trop de questions (moyenne : 2.73 et écart-type : 1.47)
- impression de recevoir pas assez de questions (moyenne : 2.24 et écart-type : 1.32)
- désorganisation (moyenne : 3.35 et écart-type : 1.38)
- manque d'utilité pour la séance de SEPI (moyenne : 2.32 et écart-type : 1.20)
- pertinence (moyenne : 2.78 et écart-type : 0.95)
- rapport avec le cours (moyenne : 3.70 et écart-type : 1.05)

— nouveauté (moyenne : 1.84 et écart-type : 0.99)

La moyenne des notes attribuées par les enseignants dans le questionnaire sur une échelle de 1 à 5 par rapport à leur expérience en classe inversée (histogramme en annexe F) :

— Facilité d'enseigner en classe hybride (moyenne : 4.03 et écart-type : 1.01)

— Appréciation de l'organisation (moyenne : 3.89 et écart-type : 1.02)

— Gain de temps et d'énergie (moyenne : 2.81 et écart-type : 1.29)

La moyenne des notes attribuées par les enseignants dans le questionnaire sur une échelle de 1 à 5 sur l'intérêt de questions de cette nature pour préparer leur SEPI (histogramme en annexe F) :

— cherchant à aller plus loin (moyenne : 3.43 et écart-type : 1.34)

— demandant un exemple (moyenne : 3.92 et écart-type : 1.09)

— de vérification d'erreur ou contradiction en cours (moyenne : 3.97 et écart-type : 0.96)

— populaires (moyenne : 4.27 et écart-type : 0.96)

L'analyse quantitative et qualitative des réponses des enseignants dans la première partie du questionnaire a révélé des résultats assez intéressants sur les différentes variables (liées à l'expérience en SEPI, classe inversée et pertinence de catégories de questions). En effet, les enseignants considèrent les questions reçues par mail sont moyennement nombreuses, désorganisées mais utiles pour préparer les SEPI. Ils sont partagés entre ceux qui reçoivent trop et pas assez de questions. Les questions posées par les étudiants sont similaires d'une année à l'autre et généralement moyennement pertinentes et déjà abordées en cours (ex : Enseignant29 : *"Les questions se répètent et ne sont pas pertinentes pédagogiquement du à la diversité des publics. Temps très limité")*.

Les enseignants apprécient bien la facilité d'enseigner en classe hybride et son organisation mais ne leur permet pas vraiment de gagner du temps. Cependant, certains n'ont pas bien compris ce qu'est une classe hybride (ex : Enseignant15 : *"je ne suis pas certain de comprendre ce qu'est une classe hybride ! !")* - il peut cependant s'agir d'un problème de terminologie, le questionnaire ne rappelant pas ce qu'on entend par ce terme (préparation des questions à la maison et questions en classe, à l'opposé d'un cours magistral transmissif en amphi).

L'avis des enseignants était plutôt partagé par rapport aux questions cherchant à aller plus loin. Certains les considéraient comme des détails inutiles et d'autres partaient du principe que toute question était par essence même utile. Voici quelques exemples de commentaires : Enseignant08 : *"... Souvent les questions cherchant à aller "plus loin" sont en réalité des questions qui sont de l'ordre du détail. Les demandes d'exemple peuvent être utiles, si elles ne sont pas déjà dans le cours. Au début de la mise en place d'un cours, les questions peuvent permettre d"identifier une erreur, et c'est un système très efficace, ..."* et Enseignant36 : *"Toute question est utile : soit elle permet de détecter une erreur ou imprécision, soit que la notion abordée est difficile à assimiler, soit que le public de l'année est plus faible vis à vis du thème du cours enseigné qui concerne la question."*. Globalement, les résultats ont montré également que les catégories identifiées de questions sont pertinentes et utiles pour la préparation de SEPI pour les enseignants, avec une préférence pour les questions populaires, ce qui nous a permis de répondre positivement à la question de recherche QR5.1, qui consiste à déterminer si certaines catégories de questions de notre schéma de codage sont potentiellement pertinentes pour les enseignants.

En ce qui concerne les types de questions qui peuvent être ignorées par les enseignants, la majorité des enseignants (76%) n'ignoraient pas les questions mal formulées, ni les questions qui

Organisation	Moins bien	Pareil	Mieux	Nombre de choix final
Analyse textuelle	5	12	19	11
Analyse catégorielle	15	16	5	3
Analyse mixte	10	12	14	11

Tableau 7.2 – Caractéristiques des 3 organisations proposées par rapport à l'organisation actuelle

traitent plusieurs sujets (96%), bien que la consigne est de poser des questions uniques portant sur un seul point (sinon il faut normalement poser 2 questions séparées). Donc ces types de questions ne freinent pas les enseignants à traiter les questions des étudiants pour la préparation des SEPI. Par ailleurs, plus que la moitié des enseignants pensent traiter effectivement plus que 80% des questions reçues par mail lors des SEPI.

7.5 Analyse de choix d'organisations

L'analyse quantitative et qualitative de la deuxième partie du questionnaire concernant les évaluations des organisations a révélé une différence assez marquée de choix d'organisations. En effet, le choix final des enseignants était partagé équitablement entre l'organisation "Analyse Textuelle" (N=11), "Analyse Mixte" (N=11) et "actuelle" (N=11). L'organisation "Analyse catégorielle" n'a été choisie que par trois enseignants uniquement.

L'organisation "Analyse Textuelle" était la mieux appréciée de la part des enseignants (19 enseignants l'ont évaluée mieux que l'organisation actuelle). L'analyse qualitative des commentaires associés à cette organisation révèle cependant que certains enseignants n'avaient pas bien compris la manière dont cette organisation serait réalisée, pensant que c'était les étudiants qui devaient réaliser cette catégorisation au moment de poser la question (ex : Enseignant08 *"... Cette organisation aurait peut-être le mérite de faire réfléchir un peu plus l'étudiant sur le type de question qu'il pose, mais ils ne mettent déjà souvent pas le numéro de diapositive, donc ce système ne fonctionnera que s'il est utilisé correctement par les étudiants.*). Sans cette limite perçue dans l'approche, on peut imaginer qu'un nombre plus important d'enseignants aurait pu l'évaluer positivement. Certains enseignants ont d'ores et déjà déclaré avoir des idées sur l'utilisation des questions posées dans chaque catégorie, par exemple : Enseignant25 : *"je crois que ce mode d'organisation correspond plus ou moins au traitement que je fais quand je les reçois ...", "Cas 1 : je réexplique, mais avec une autre approche, et en stigmatisant l'absence apparente d'effort pour comprendre, ce qui se serait traduit par une question moins vague ! Cas 2 : j'approfondis, je précise, je clarifie, j'explique ! en veillant à consolider ainsi l'objectif de départ. C'est le principe "qui peut le plus, peut le moins" Cas n°3 : cas rêvé ! aide à comprendre comment l'étudiant n'a pas compris, et permet de le ramener au bon concept en corrigeant son erreur, ou en lui expliquant comment ce qu'il pensait faux est juste (exception pour les cas où le cours comporte VRAIMENT une erreur, ce qui arrive - en ce cas on conforte les étudiants dans le fait qu'ils sont bien raisonné...)"*. D'autres, étaient plus sceptiques sur son utilisation (Enseignant10 : *"Je tacherai de répondre à toutes, indépendamment de cette catégorisation et de manière chronologique par rapport au cours..."*. On peut cependant noter que cette remarque est transverse à toute organisation alternative de questions, et que les enseignants ayant ce point de vue n'auront de toutes façons aucun intérêt pour l'une d'elles.

L'organisation "Analyse Catégorielle" est assez nettement la moins appréciée des trois.

Certains enseignants pensaient à l'utiliser dans le bon sens en répondant aux questions des étudiants en difficulté en premier jusqu'aux questions posées par les bons. Parmi les enseignants les plus critiques, plusieurs la jugeaient moins bien car ils supposaient que le but de cette organisation était de répondre en priorité aux bons étudiants et étaient inquiets sur l'équité de la proposition (soulignons au passage qu'il ne s'agit que d'une interprétation de la part des enseignants, rien dans la représentation ne suggérait de favoriser cette approche - les questions des étudiants potentiellement en difficulté étant d'ailleurs présentées en premier). Certains étaient opposés ou étaient inquiets sur le fait de se laisser tenter de ne répondre qu'à la catégorie des bons (ex : Enseignant27 : "je ne suis pas très favorable a privilégier les bons par rapport aux moins bons (cela est un risque), notre but est de tous les aider"). Enfin, une piste d'amélioration proposée par les enseignants consisterait à prendre aussi en compte dans cette organisation le niveau des étudiants qui votent sur des questions, et pas seulement ceux qui posent des questions.

L'avis des enseignants par rapport à l'organisation "Analyse Mixte" était contradictoire avec le fait que 11 l'ont choisi comme préférée. En effet, ils l'ont trouvé peu lisible et recommandaient de se concentrer sur un seul aspect. Certains la décrivaient comme : *"une mauvaise méthode qui s'ajoute à une bonne"* (Enseignant44) et même ceux qui la choisissent ne sont certains de s'en servir (Enseignant24 : *Outre le fait que je ne suis pas sûr de m'en servir, Je ne suis pas sûr que l'analyse textuelle permettent une classification toujours pertinente ... donc autant faire une analyse mixte et chacun l'utilisera comme bon lui semble.*).

7.5.1 Lien entre choix d'organisations et expérience en SEPI

La section précédente a mis en avant une variété dans les choix des enseignants (pas d'unanimité pour un choix d'organisation). Il est donc possible qu'il faille envisager de proposer différentes visualisations pour différents types d'enseignants. Mais quels paramètres sont alors pertinents pour choisir d'associer une organisation de question à un enseignant donné ? Nous avons donc été amené à nous poser la question de recherche suivante :

QR5.2 : Est-ce que les différences de choix d'organisations sont liées à une expérience différente en SEPI ?

Comme nous ne souhaitions pas tester individuellement chaque variable explicative, nous avons choisi d'effectuer un clustering (avec K-Means, k compris entre 2 et 10) permettant de prendre en compte différents aspects de l'expérience des enseignants avec la SEPI. Nous avons donc utilisé comme caractéristiques pour chaque enseignant les variables décrivant l'ensemble de questions envoyées par mail (trop nombreuses [+Nb], désorganisées [NOrg], pas utiles [NUti], pertinentes [Per], déjà abordées en cours [AbCou], nouvelles chaque année [NvQst], nombre de questions reçues [QstRec], proportion de questions traitée [QstTra]), les variables décrivant la classe inversée (facile à enseigner [Fac], bien organisée [Org], gain du temps et d'énergie [Tps]) et l'ancienneté [Anc]. Toutes les variables sont normalisées entre 0 à 1, sauf pour le nombre de questions reçues (3 tranches), la proportion de questions traitées (5 tranches) et l'ancienneté (recodée en 3 tranches), qui sont des variables catégorielles (*cf.* Tableaux 7.3 et 7.4).

Nous avons obtenu deux clusters d'enseignants bien distincts en termes de caractéristiques :

Le cluster1 représente des enseignants moins expérimentés (N=22), appréciant moins la

Cluster	Fac	Org	Tps	Norg	NUti	+Nb	Per	AbCou	NvQst
Cluster1	0.576	0.5	0.466	0.739	0.432	0.557	0.364	0.761	0.242
Cluster2	0.857	0.857	0.429	0.393	0.196	0.214	0.554	0.571	0.333

Tableau 7.3 – Centroïdes des 9 variables numériques associées à chaque cluster

Cluster	Anc			QstRec			QstTra				
	[1-5]	]5-10]	]+10]	[1-20]	]20-50]	]+50]	0-20%	20-40%	40-60%	60-80%	80-100%
Cluster1	5	12	5	5	5	12	1	2	2	9	8
Cluster2	6	1	7	6	8	0	0	2	1	1	10

Tableau 7.4 – Distribution des répondants sur les 3 variables catégorielles pour chaque cluster

classe inversée (facilité et organisation), qui reçoivent plus de questions (plus de 50% d'entre eux reçoivent plus de 50 questions par séance) et traitent globalement moins de questions dans leur SEPI. Ils considèrent que les questions envoyées sont très désorganisées, moyennement utiles pour la SEPI et peu pertinentes.

Le cluster2 contient des enseignants (N=14) plus expérimentés, qui apprécient bien la classe inversée (facilité et l'organisation), reçoivent peu de questions (entre 1 et 50 questions reçues par séance) et traitent globalement la majorité de questions posées. Ils considérent que les questions envoyées sont peu désorganisées, utiles pour les SEPI et moyennement pertinentes.

Pour caractériser les deux clusters, nous avons utilisé les variables dépendantes des choix des organisations pour chaque enseignant. Pour chaque cluster, nous avons calculé la moyenne de facilité [MoyFac] (note attribuée entre 1 et 5) et la moyenne d'utilité [MoyUti] (note comprise entre 1 et 5) de chaque organisation. Nous avons également caractérisé chaque cluster selon l'appréciation donnée par les enseignants sur chacune des 3 organisations proposées (moins bien, pareil, mieux) et le choix final d'organisation proposée (organisation "analyse textuelle", "analyse catégorielle", "analyse mixte" et actuelle).

L'analyse des résultats présentés dans le Tableau 7.5, a montré qu'il n'y a pas de différences claires dans les choix effectués par les enseignants des deux clusters. La seule légère tendance visible concerne une appréciation un peu plus forte de l'analyse textuelle par les enseignants du cluster 1, qui semblent au contraire un peu moins convaincus par les deux organisations prenant en compte le niveau estimé de l'élève posant la question. Malheureusement, l'expérience des enseignants en SEPI et en classe inversée n'a donc pas vraiment permis de distinguer leur différences de choix d'organisations. Certaines difficultés mentionnées auparavant, tel que des problèmes de compréhension d'une organisation ou des biais culturels, peuvent avoir un impact sur le choix des organisations. Nous pensons également qu'il y a probablement d'autres facteurs (au-delà de l'expérience en SEPI) qui peuvent expliquer les différences de choix d'organisation (par exemple, les enseignants ayant des idées de la manière à utiliser les questions dans chaque catégorie par rapport aux autres qui n'en ont pas). Nous avons donc répondu négativement à la QR5.2, il n'y a pas de lien évident entre l'expérience des enseignants en SEPI et les différences de choix d'organisations.

Variables		Cluster1 (N=22)	Cluster2 (N=14)
A. Textuelle	MoyFac	4.273	4.143
	MoyUti	3.545	3.286
	Mbien	3	2
	Pareil	8	4
	Mieux	11	8
A. Catégorielle	MoyFac	3.545	4.071
	MoyUti	2.227	2.643
	Mbien	10	5
	Pareil	9	7
	Mieux	3	2
A. Mixte	MoyFac	3.455	3.786
	MoyUti	2.864	3.429
	Mbien	7	3
	Pareil	8	4
	Mieux	7	7
Choix Final	Textuelle	7	4
	Catégorielle	2	1
	Mixte	6	5
	Actuelle	7	4

Tableau 7.5 – Caractérisation des clusters selon les variables dépendantes de chaque organisation : moyenne (MoyFac et MoyUti) et distribution des répondants

7.6 Synthèse

Le travail présenté dans ce chapitre nous a permis d'évaluer la pertinence de notre schéma de codage auprès des enseignants et d'avoir un premier retour sur leurs expériences en SEPI. En effet, l'analyse des résultats du questionnaire a montré l'intérêt que portent les enseignants pour les catégories de questions proposées. Ils apprécient bien la facilité et l'organisation de la classe inversée et trouvent les questions posées par les étudiants utiles pour préparer leurs sessions SEPI. D'un autre côté, l'analyse des choix d'organisations n'a pas révélé une unanimité pour un choix de la part des enseignants. Nous avons donc eu recours au clustering pour étudier le lien entre les choix d'organisations des enseignants et leurs expériences en SEPI. L'analyse des clusters obtenus n'a pas permis non plus d'établir un lien direct entre les différents choix d'organisation et l'expérience en SEPI. Ce résultat peut s'expliquer par le fait qu'il y a probablement d'autres variables qui contribuent aux choix des enseignants au-delà de leur expérience, combiné à un biais de perception quant à la manière dont ces organisations de questions seraient créées (supposition que ce serait à l'étudiant d'annoter ses questions) ou dans l'utilisation potentielle induite (tentation pour soi ou des collègues de se focaliser sur les questions des bons étudiants uniquement).

Une des limites de ce travail est donc liée au fait que certains enseignants avaient mal interprété certaines questions posées et le principe des organisations proposées. Une autre limite éventuelle est liée au fait que nous avons réalisé ce questionnaire sur la base des données collectées en 2012 (pour des raisons pratiques, ces questions ayant déjà été préalablement prétraitées). Depuis, les besoins des enseignants ne sont plus forcément tout à fait les mêmes

et les étudiants posent de moins en moins de questions sur la plateforme en ligne d'après les enseignants.

Chapitre 8

Réplication des processus d'annotation MOOC

Dans les environnements en ligne, il est plus difficile pour les étudiants d'avoir des interactions avec les autres étudiants et l'instructeur par rapport aux cours traditionnels en classe. C'est encore plus critique dans les MOOC (Massive Open Online Courses) où très souvent les étudiants sont socialement isolés (l'une des principales raisons pour lesquelles ils ont tendance à abandonner [Yang *et al.* , 2014]), sans connaître personnellement aucun autre étudiant, même si de plus en plus d'institutions inscrivent leurs élèves à un MOOC pour remplacer une partie de leurs enseignements en présentiel Bouchet & Bachelet [2019]. La nécessité pour les étudiants de poser des questions ou de signaler des erreurs dans le cours est aussi importante en classe que dans des situations moins traditionnelles, comme l'apprentissage à distance. Par conséquent, les forums de discussion sont un élément clé de l'apprentissage en ligne et dans les MOOCs en particulier [Andresen, 2009].

Dans cette partie du travail, nous avons mené des analyses pour explorer la nature des questions posées par les élèves dans le cadre d'un forum de MOOC et nous avons particulièrement essayé de regarder la relation entre ces questions et la performance et les compétences d'autorégulation des élèves. Notre objectif était de tester la réplicabilité du processus d'annotation dans un contexte différent de celui de la classe hybride, tel un MOOC. Pour répondre à la sixième question de recherche (QR6) introduite dans la section 1.1, nous l'avons affinée en quatre questions de recherche :

QR6.1 Est-il possible d'annoter de façon fiable les questions extraites des messages du forum du MOOC selon un schéma de codage multi-niveaux à granularité fine ?

QR6.2 Y a-t-il un lien entre la nature des questions posées dans un MOOC et la performance des élèves ?

QR6.3 Y a-t-il un lien entre la nature des questions posées dans un MOOC et la maîtrise des compétences d'autorégulation par les élèves ?

QR6.4 En termes de réplicabilité d'analyse, y a-t-il des groupes d'étudiants qui posaient les mêmes types de questions sur plusieurs sessions ? Et si oui, est-ce que les caractéristiques des étudiants sont-elles similaires dans différentes sessions et contextes (MOOC et classe hybride) ?

Ce chapitre est organisé comme suit : dans la section 1, nous présentons le contexte du MOOC francophone sur la gestion de projet (GDP) considéré pour cette analyse et l'ensemble de données utilisé. Dans la section 2, nous présentons une extension du schéma de codage utilisé auparavant pour l'annotation des questions des étudiants posées dans le contexte de la classe hybride (PACES, *cf.* 4.4) et comment l'adapter aux besoins de ce nouveau contexte. Dans la section 3 nous regardons les similarités entre notre schéma de codage et ceux présentés dans l'état de l'art (*cf.* 2.2.3). Dans la section 4, nous détaillons l'architecture en cascade du nouveau système d'annotation automatique des questions, permettant d'annoter les questions comme étant tout d'abord liées au cours ou non, et ensuite d'identifier la nature de ces questions. Dans la section 6, à l'aide d'un corpus de questions annotées avec un ensemble de classifieurs, nous analysons la relation qui existe entre les questions posées, la performance et l'autorégulation. Finalement, nous présentons le résultat de la réplication des analyses dans le contexte du MOOC à partir du clustering (*cf.* section 5.2.3) afin de mieux comprendre la nature de questions des étudiants qui réussissent et ceux qui échouent. Ces résultats ont donné lieu à deux publications sous forme de posters dans les conférences internationales EDM (*cf.* section 8.5) [Harrak *et al.* , 2019a] et L@S (Learning at Scale) [Harrak *et al.* , 2019d] (*cf.* section 8.6).

8.1 Contexte GDP

Nous avons considéré dans ce travail quatre ensembles de messages de forum de discussion issus de quatre sessions différentes du même MOOC français sur la Gestion De Projet appelé GDP. Les sessions correspondent à la $5^{ème}$ jusqu'à la $8^{ème}$ édition de ce MOOC semestriel très populaire (plus de 10 000 utilisateurs inscrits à chaque session), tenues respectivement en 2015 et 2016. L'une des particularités de ce MOOC est qu'il permet à ses participants d'obtenir un certificat de base, correspondant à une charge de travail modérée (15-25 heures), ainsi qu'un certificat avancé, qui correspond à une charge de travail plus importante (35-45 heures). Par conséquent, pour chaque participant, nous pouvons déterminer deux notes finales, en plus de l'information relative à l'obtention d'un ou des deux certificats associés.

Le forum de discussion fonctionne de manière typique, organisé autour de fils de discussion créés par l'équipe pédagogique pour répondre à des questions techniques ou administratives, sur les devoirs ou le contenu des cours, entre autres (les étudiants ne peuvent pas créer de fils pour conserver une organisation de haut niveau compréhensible). Au cours de la semaine précédant l'ouverture du MOOC, les élèves sont encouragés à se présenter et à expliquer leur motivation à participer au MOOC, afin de créer un sentiment de communauté. Ce fil "biographique" est de loin le plus populaire (25 à 48% de tous les messages) et représente un exemple de messages qui ne sont pas pertinents pour notre travail (plus de détails sont donnés à la section 8.2.1). Le Tableau 8.1 fournit quelques statistiques de base sur l'utilisation du forum et le nombre d'étudiants inscrits.

De plus, les participants au MOOC sont invités à remplir un questionnaire de recherche au début du MOOC, au milieu du MOOC (après 2 semaines) et vers la fin de celui-ci (semaine 5). Bien que facultatifs, les questionnaires de ces 4 éditions du MOOC ont été remplis par la plupart des étudiants du MOOC encore actifs à ce moment-là, même si leur remplissage n'apporte qu'un très faible bonus de points à la note finale (10 points sur 4000). Ce questionnaire comprend plusieurs contrôles d'attention (ex : "répondre 5 à cette question"), ce qui nous permet de filtrer tous les participants qui n'ont pas répondu comme demandé à ces questions

Session	GDP5	GDP6	GDP7	GDP8
Étudiants inscrits	17579	23315	19392	24603
Étudiants ayant répondu à quiz 1	4842	7537	5951	7998
Certificat de base obtenu	2282	3900	2393	4526
Certificat avancé obtenu	503	697	559	589
Nombre de messages	7655	10597	12224	14072
Nombre de messages uniques	2087	4717	3504	4760

Tableau 8.1 – Statistiques descriptives des 4 sessions du MOOC considérées (inscription, messages et réussite)

pour accroître la fiabilité des données considérées.

Les données examinées plus loin portant sur les compétences d'autorégulation des élèves proviennent des questions posées dans le deuxième questionnaire (G) (après 2 semaines) et sont issues d'un questionnaire francophone validé psychométriquement qui mesure l'autorégulation dans un environnement en ligne [Cosnefroy *et al.* , 2018]. Il est composé de 21 questions à l'aide d'échelles de Likert à 7 points et évalue 4 dimensions : (1) les stratégies cognitives et métacognitives, (2) la procrastination, (3) l'adaptation du contexte et (4) le soutien des pairs. L'utilisation de stratégies cognitives et métacognitives est l'un des éléments clés de l'apprentissage autorégulé, qui englobe aussi généralement des éléments liés à la motivation, la disposition d'objectif, la planification et le suivi [Pintrich, 2004]. Dans un contexte en ligne, la procrastination (qui consiste à retarder les tâches à faire et qui est associée à une mauvaise gestion du temps) peut être vue comme un manque de planification, qui peut être associé à un échec encore plus facilement que dans un contexte normal de classe [You, 2015] et a déjà fait l'objet d'une étude dans certains travaux [Park *et al.* , 2018]. L'adaptation du contexte, qui consiste à trouver de bons moments et des lieux où l'on ne serait pas dérangé pour apprendre efficacement, est spécifique à l'apprentissage en ligne et peut également s'appliquer à la planification de la session d'apprentissage. Enfin, la recherche d'un soutien par les pairs en cas de besoin est particulièrement important pour l'apprentissage en ligne, non seulement parce que l'apprentissage est une activité sociale [Vygotsky, 1978], mais aussi en raison de l'importance des liens sociaux dans des contextes comme les MOOCs [Yang *et al.* , 2014].

8.2 Schéma de codage de questions de MOOC

8.2.1 Méthode

Afin d'identifier la nature des questions posées par les étudiants dans les forums du MOOC, nous avons considéré un échantillon de 500 messages des 4 sessions du MOOC GDP considérées ici (5, 6, 7 et 8). Cet échantillon a été divisé arbitrairement en 3 sous-échantillons (1^{er} sous-échantillon de 200 messages, $2^{ème}$ sous-échantillon de 100 messages et $3^{ème}$ sous-échantillon de 200 messages) pour appliquer 4 étapes successives de catégorisation et définir un schéma de codage comme présenté dans la section 4.1 dans le contexte de PACES. Pour que le corpus brut soit utile pour l'annotation manuelle, certaines étapes de pré-traitement doivent être effectuées.

Nettoyage et pré-traitement des données : Le corpus brut contient des messages envoyés par des étudiants et des instructeurs dans le forum de discussion du MOOC GDP. Les messages envoyés par les étudiants ne sont pas structurés et contiennent du bruit (un message peut contenir plusieurs questions, opinions, réponses à des questions non liées au cours, *etc.*). Afin de réduire le bruit et améliorer la structure des messages, certaines étapes manuelles ont été effectuées. Nous avons filtré les messages provenant des instructeurs, les messages qui sont des réponses à d'autres questions posées (en ne conservant que les messages racine - avec l'hypothèse qu'un message racine doit toujours contenir une forme de question) et les sujets de discussion qui ne sont pas liés au cours (tels que le fil "biographie" mentionné ci-dessus, mais aussi les fils dédiés aux questions techniques).

L'**étape de découverte**, consistait à regrouper des phrases avec des similitudes pour en extraire des concepts significatifs. Bien que les messages des élèves étaient souvent longs et non structurés, nous avons décidé de les segmenter en questions simples (*c-à-d.* centrées sur un seul sujet). Une fois les phrases segmentées, nous avons regroupé les questions en fonction de leur contenu (*c-à-d.* les questions ont été identifiées comme liées au cours ou non). Les groupes de questions basées sur les cours sont ensuite annotés selon le schéma de codage introduit dans le Tableau 8.2, consistant à annoter chaque question selon les quatre dimensions indépendantes.

Nous avons appliqué la même approche sur des questions identifiées non liées au cours et dont la structure et la sémantique semblent similaires (ex : "Je ne trouve pas le lien ?" et "Où est le PDF ?"). Des groupes de questions ont ensuite reçu des "étiquettes" selon ce que nous avons appelé "dimension 0", en référence aux 4 autres dimensions des questions du cours (ex : "ressources non trouvées"). Puis, nous avons identifié les exclusions mutuelles entre étiquettes (par exemple, une question simple ne peut pas être à la fois une demande de "ressources non trouvées" et une demande d'"administration"). En résumé, le nouveau schéma de codage est composé de deux schémas de codage indépendants : le schéma de codage pour les questions liées au cours (annoté comme un vecteur de 4 valeurs) et le schéma de codage des questions non liées au cours identifiées par les catégories de dimension 0 (*cf.* Tableau 8.3).

L'**étape de consolidation** consistait à annoter le $2^{ème}$ sous-échantillon de 100 messages (349 segments) pour valider les catégories et les dimensions identifiées précédemment. Cela a conduit à divers ajustements aux catégories et dimensions comme l'ajout de la catégorie "phatique" au schéma de codage des questions non liées au cours (une expression phatique dans notre contexte est typiquement une question qui implique simplement qu'une réponse est attendue à la fin du message mais sans valeur sémantique intrinsèque, ex : "que pensez-vous de cela ?", "quelqu'un peut-il me répondre ?").

Dans l'**étape de validation**, deux annotateurs humains ont effectué une annotation séparée pour valider la généralité de nos catégories sur le troisième sous-échantillon de 200 messages (626 segments). Les deux annotateurs humains ont utilisé comme référence unique le schéma de codage créé à la fin de l'étape précédente pour annoter chacun de ces segments. Cela fournit également une base de référence sur le niveau d'accord auquel on peut raisonnablement s'attendre avec le classifieur automatique. Une des 4 dimensions a été ajustée (dimension 4 - Att) et 7 catégories sont issues de l'étape de consolidation : Dim1, Dim2, Dim3, Dim4 (relatives aux questions du cours) et les catégories de Dim0 (socialisation, examen, problème technique, ressources non trouvées, questions administratives, outils et phatiques, relatives aux questions non liées au cours). Les annotateurs humains ont fait deux annotations distinctes et indépendantes sur chaque dimension, et leur accord a été évalué à l'aide du Kappa de Cohen. Tout d'abord, le kappa a été calculé pour déterminer si un segment était une question ou non

($\kappa = 0.85$) et ensuite sur des questions explicites [1] seulement ($\kappa = 0.96$). Pour les 13 segments pour lesquels les deux annotateurs n'étaient pas d'accord sur le fait qu'il s'agissait de questions explicites, une discussion a rapidement abouti à un accord. Les étapes suivantes ont été réalisées sur les 301 segments identifiés comme questions. La deuxième étape consistait à vérifier l'accord sur le sujet de la question (cours *vs.* non-cours), pour lequel le kappa obtenu était $\kappa = 0.85$. Pour les 22 segments pour lesquels les deux annotateurs n'étaient pas d'accord, une nouvelle phase de discussion a conduit à choisir une annotation plutôt que l'autre pour chacun d'entre eux. Il est important de noter que la difficulté principale pour ces segments provient du sujet du MOOC lui-même, lequel enseigne entre autres quelques concepts (réunions, rapports...) et l'utilisation d'outils (Google Drive, anti-virus...). Les mots-clés associés à ces éléments de contenu, dans de nombreux autres MOOCs, ne seraient pas associés au contenu du cours mais aux questions administratives qui l'entourent. De même, le concept du "projet" dans ce MOOC peut faire référence soit à un "projet" (devoir d'évaluation), soit à une gestion de projet (concept de cours). Ces ambiguïtés propres au contenu du MOOC ont donc certainement un effet négatif sur la valeur des kappa obtenus, qui seraient certainement plus élevées sur un MOOC de science "dure" par exemple. Enfin, les kappas ont été calculés pour chacune des 4 dimensions ($\kappa_1 = 0,70$, $\kappa_2 = 0,61$, $\kappa_3 = 0,69$ et $\kappa_4 = 0,57$) et pour la dimension 0 des questions ne portaient pas sur le cours ($\kappa_0 = 0,58$). Le premier et le deuxième sous-échantillon, qui ont été utilisés respectivement à l'étape de la découverte et de la consolidation (684 segments) ont également été annotés à nouveau sur des questions explicites, le sujet de la question (cours / non-cours) et les 5 dimensions (Dim0 à 4) pour considérer les changements et fournir une vérité terrain à laquelle comparer la notation automatique dans les sections suivantes.

8.2.2 Résultats

Le principal résultat de cette étape a été la création d'un schéma de codage pour les questions pas liées au cours et quelques ajustements pour les questions liées au cours (mis en gras), fournis dans les Tableaux 8.2 et 8.3 respectivement. En effet, suite au changement de corpus (GDP), nous avons été amenés à adapter le sujet de la dimension 2 de "modalités d'explication" au "sujet de questions". Idéalement il aurait fallu faire une autre dimension mais la dimension 2 avait changé de nature, on avait trouvé plus des questions sur "le sujet de question" mais pas sur "la modalité d'explication". Les 4 principes de l'étape d'annotation manuelle établies avant dans la section 4.2 tiennent toujours, et nous avons en rajouté deux autres pour annoter automatiquement le nouveau corpus :

- Seules les questions explicites peuvent être annotées en tant que questions de cours ou non

- Seules les questions de cours peuvent être annotées sur Dim1, Dim2, Dim3 et Dim4. Les questions non liées au cours ne doivent être annotées que sur Dim0 (catégories)

1. Une *question explicite* correspond à une demande avec une formulation claire de question. Un exemple de question explicite est "Pourriez-vous expliquer X ?", à l'inverse, "Je ne comprenais pas X." serait annotées comme *question implicite*

8.3 Similarités entre taxonomies

Comme dans le cas de la version originale du schéma de codage (*cf.* chapitre 4), nous avons constaté a posteriori un certain chevauchement entre certaines typologies destinées aux forums de discussions présentées dans l'état de l'art (*cf.* section 2.2.3) et notre schéma de codage. En particulier, les catégories de la dimension 0 sont globalement incluses dans la catégorie "non cours" [Cui & Wise, 2015]. De même, plusieurs catégories de la dimension 0 correspondent à des sous-catégories dans "sujet de message" [Stump *et al.* , 2013], tels que "socialisation", "examen" et "problème technique". En revanche, les catégories concernant les questions de cours ne sont pas bien spécifiques dans la plupart des cas et elles sont largement désignées par "contenu" [Stump *et al.* , 2013] ou "cours" [Cui & Wise, 2015] dans les autres typologies de messages de MOOC. D'autres schémas de codage existants permettaient de catégoriser les messages dans les forums de discussion [Agrawal *et al.* , 2015; Almatrafi *et al.* , 2018]. Cependant, ils ont utilisé des indicateurs pour détecter la "confusion" ou "urgence" dans les messages, ce qui ne correspond pas vraiment au principe de notre schéma de codage. Le tableau 8.4 résume les chevauchements identifiés avec certains schémas de codage dans l'état de l'art.

Il est important de noter que toutes les typologies présentées concernent des "messages", alors que notre travail se situe à un niveau de granularité plus fin, en s'intéressant aux questions dans ces messages. Nous proposons donc à la fois une granularité fine sur ce qu'on annote, et dans la manière de les annoter.

8.4 Annotation automatique

Pour annoter l'ensemble des questions posées par les élèves (et à terme pour l'utiliser en ligne pour analyser les questions collectées), nous avons utilisé dans la première partie de ce travail (PACES) un outil d'annotation semi-automatique à base de règles d'expert. Bien qu'efficace sur les questions qu'il annote (Kappa moyen de 0.70), une proportion importante de questions ne sont pas annotées par cet outil. Par ailleurs, il y a une difficulté particulière à l'utilisation d'une approche basée sur des mots-clés identifiés manuellement dans le contexte du MOOC GDP en raison du sujet du MOOC lui-même (comme mentionné dans l'étape de validation, *cf.* section 8.2.1). Ceci nous a conduit à utiliser ici une annotation entièrement automatique basée sur des techniques d'apprentissage automatique (*cf.* section 4.4.4.2) sur le corpus de questions (*cf.* La Figure 8.1 résume les différentes étapes de l'annotation automatique).

La première étape a consisté à transformer les 1307 segments annotés manuellement en vecteurs de mots. Tout d'abord, nous avons segmenté automatiquement l'ensemble des messages (*cf.* comme décrit dans la section 4.4. Nous avons effectué par la suite un ensemble de pré-traitements classiques sur le corpus de 1307 segments (500 messages) annotés manuellement : tokenisation, stemmatisation, suppression de ponctuation (sauf pour ' ?') et de stopwords (mots creux non porteurs de sens), *etc.* . Nous avons ensuite extrait tous les unigrammes et bigrammes, avec une approche de type sac de mots, et compté leurs occurrences dans cet échantillon. Chacun des 1307 segments était représenté par un vecteur de mot binaire ('1' si le mot est dans le segment,'0' sinon). Nous avons finalement réduit le nombre de mots-clés extraits (nombre élevé de mots-clés par rapport au nombre de segments) pour conserver les plus importants et les plus significatifs en utilisant une technique de sélection d'attributs (suppression des unigrammes/bigrammes moins fréquents et corrélés).

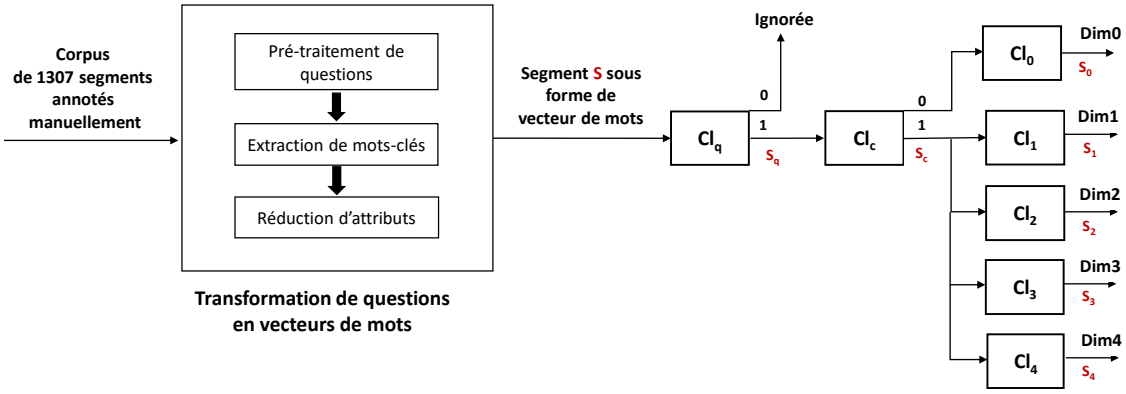


Figure 8.1 – Architecture globale en cascade de l'annotation automatique

Ensuite, nous avons conçu un annotateur en 3 phases pour identifier les segments avec des questions, questions liées au cours et non-cours et la nature de ces questions. Au total, 7 classifieurs ont été entraînés pour annoter un segment en 4 étapes successives :

Cl$_q$ **:** le classifieur de segments en question/pas-question

Cl$_c$ **:** le classifieur de questions cours/non-cours

Cl$_0$ **:** le classifieur de questions non liées au cours, selon la dimension 0

Cl$_1$ **:** Le classifieur de questions liées au cours selon la dimension 1

Cl$_2$ **:** Le classifieur de questions liées au cours selon la dimension 2

Cl$_3$ **:** Le classifieur de questions liées au cours selon la dimension 3

Cl$_4$ **:** Le classifieur de questions liées au cours selon la dimension 4

Les 7 classifieurs sont structurés selon une architecture en cascade, comme illustré dans la Figure 8.1.

Nous avons essayé 7 techniques de classification différentes pour chaque classifieur (introduites dans la *cf.* section 4.4.4.2) : machines à vecteurs de support (SVM), modèle linéaire généralisé (GLM), Gradient Boosted Trees (GBT), arbre de décision (DT), K-NN, Naive Bayes (NB) et Règle d'induction (RI), chacun avec différentes valeurs d'hyperparamètres. Pour chaque classifieur, l'entrée était les vecteurs de mots représentant les segments en termes de mots-clés, et l'étiquette à prédire était la valeur associée au segment dans cette dimension. Nous avons considéré les 1307 segments annotés manuellement comme des données étiquetées. Nous avons utilisé une validation croisée pour chaque classifieur, en gardant 10% des segments pour les tests et en utilisant le reste des segments pour entraîner le classifieur. La première étape consistait à entraîner le classifieur (Cl_Q) sur 1307 segments avec la validation croisée pour identifier si un segment était une question ou non ('1' si question,'0' sinon). Dans la deuxième étape, nous avons entraîné le classifieur (Cl_C) sur seulement 704 segments identifiés comme des questions avec la validation croisée pour identifier le sujet de la question (étiqueté

"1" pour cours et "0" pour non-cours). La troisième étape consistait à entraîner le classifieur (Cl_0) sur 293 questions non-cours avec la validation croisée pour annoter chaque catégorie de la dimension 0. Enfin, dans la dernière étape, nous avons entraîné un classifieur sur chaque dimension séparément (les dimensions étant conçues comme indépendantes, Cl_1, Cl_2, Cl_3 et Cl_4 pour Dimension 1, 2, 3 et 4 respectivement) sur 412 questions de cours avec validation croisée pour noter chaque valeur (ex : "ré-expliquer" en "Ree").

Nous avons ensuite calculé les valeurs Kappa entre les prédictions des modèles de classification et les valeurs correspondantes trouvées par l'annotation manuelle (*cf.* Tableau 8.5). L'ensemble du corpus de segments de messages a été annoté par les techniques les plus performantes pour chaque classification (pour le classifieur Cl_Q, comme il y avait égalité entre plusieurs, nous avons pris GBT).

Les valeurs Kappa obtenus par l'annotation automatique peuvent sembler basses sur certaines dimensions, mais pour un corpus de messages particulièrement difficile, comme celui du MOOC GDP, ces valeurs peuvent être acceptables. Nous avons donc répondu positivement à la première question de recherche, qui consiste à annoter les questions extraites des messages du forum du MOOC selon un schéma de codage multi-niveaux à granularité fine.

8.5 Lien entre les questions, l'autorégulation et la réussite

Dans la suite de ce travail, nous nous sommes intéressés à étudier le lien entre le type de questions posées par les étudiants et l'autorégulation et la réussite. De nombreux travaux montrent que l'apprentissage autorégulé (AAR) est un indicateur clé de la réussite des élèves. Les apprenants autorégulés ont la capacité de planifier, de s'organiser, s'auto-instruire et s'auto-évaluer, ce qui est essentiel pour réussir à long terme [Zimmerman & Martinez-Pons, 1988]. En pratique, l'AAR a été associé aux meilleurs résultats scolaires, l'engagement et la motivation [Zimmerman, 2002]. Plusieurs recherche ont permis d'obtenir un large éventail de définitions [Pintrich, 1999; Zimmerman & Martinez-Pons, 1988] et modèles [Boekaerts, 1997; Puustinen & Pulkkinen, 2001], qui servent à souligner l'importance d'aider les gens à apprendre comment initier une grande variété de facteurs méta-cognitifs, cognitifs, affectifs et motivationnels pour atteindre leurs objectifs d'apprentissage [Kizilcec *et al.* , 2017].

Nous avons étudié ici plus particulièrement comment l'AAR était utilisé pour analyser les interactions des élèves dans les environnements en ligne. van den Boom *et al.* [2004] a étudié les effets des incitations à la réflexion et de la rétroaction du tuteur sur le développement de la compétence d'AAR dans un environnement en ligne. Les résultats indiquent que l'intervention avec des incitations à la réflexion ainsi que la rétroaction du tuteur affectent de façon différenciée le développement des aspects AAR. L'auteur décrivait également comment les contributions affichées dans les forums facilitent la référence précise pendant les discussions ainsi que la révision et la réflexion ultérieures. Dettori & Persico [2008] ont examiné si l'analyse des interactions peut aider à comprendre la pratique et le développement de l'AAR dans les communautés d'apprentissage virtuelles. Les chercheurs ont utilisé une taxonomie d'indicateurs de l'AAR pour analyser directement quels types d'élèves sont autorégulés à partir de leurs messages. Ils ont constaté que les messages des élèves ne reflétaient pas entièrement les indicateurs de l'AAR, mais leur approche reste valide pour enquêter sur la présence de l'AAR dans de plus grands ensembles de messages et dans des contextes différents. Bouchet *et al.*

[2013] ont caractérisé les élèves à partir du clustering en fonction de leurs interactions avec un système de tuteur intelligent favorisant l'AAR. Ils ont montré qu'il existe des variations entre les clusters extraits d'élèves en ce qui concerne les instructions reçues par le système pour exécuter les processus d'AAR.

L'AAR est particulièrement important dans les environnements d'apprentissage en ligne, principalement pour les cours à grande échelle comme les MOOCs [Kizilcec *et al.* , 2017]. Cross *et al.* [2017] ont étudié la recherche d'aide comme forme du processus d'AAR dans les communautés en ligne. Les chercheurs ont développé un schéma de codage pour étiqueter les demandes d'aide en "question directe" et "demande d'aide implicite" et les réponses fournies en "partage de connaissances", "échange de ressources" et "retour/validation". Ils ont automatisé par la suite la détection du comportement du chercheur/fournisseur d'aide à partir de techniques d'apprentissage automatique supervisées (SVM et forêts aléatoires) dans le discours en ligne. Les résultats ont montré que des caractéristiques de base (n-grammes et nombre de questions) se sont avérées intéressantes pour détection automatiquement les comportements étudiés.

Dans l'ensemble, l'AAR a donc été étudié à des fins très diverses dans les environnements en ligne. Toutefois, le lien entre l'AAR et les questions posées dans les forums de discussion n'a apparemment pas encore été directement exploré. Nous nous intéressons donc à l'analyse de la nature des questions posées par les étudiants à partir du schéma de codage développé et l'AAR.

8.5.1 Codage de données

Afin d'étudier la relation entre chaque type de question et la réussite d'une part et l'autorégulation d'autre part, pour chaque étudiant de la 8ème session du MOOC qui a envoyé un message sur le forum, nous avons codé le nombre de segments catégorisés comme :

- une question explicite (NbQ)
- pas une question ou une question implicite (NbNQ)
- une question ne portant pas sur le contenu du cours (NbQ-NC)
- une question ne portant pas sur le contenu du cours correspondant à un acte de socialisation (NbQ-NC-Soc), une question administrative (NbQNC-Adm), un examen (NbQ-NC-Exa), un problème technique (NbQ-NC-PbT), une ressource non trouvée (NbQ-NC-Res), un outil (NbQ-NC-Out) ou une phatique (NbQ-NC-Pha),
- une question portant sur le cours (NbQ-C),
- une question de cours sur une ré-explication (NbQ-C1-Ree), une demande d'approfondissement d'un concept (NbQ-C1-App), ou une vérification (NbQ-C1-Ver)
- une question de cours pour demander un exemple (NbQ-C2-Exe), un schéma (NbQ-C2-Sch), ou une correction (NbQ-C2-Cor),
- une question de cours pour demander une définition (NbQ-C3-Def), comment procéder (NbQ-C3-Man), la raison de quelque chose (NbQ-C3-Rai), le rôle de quelque chose (NbQ-C3-Rol), ou le lien entre concepts (NbQ-C3-Lie),
- une question de cours pour demander une vérification concernant une erreur apparente (NbQ-C4-Err), une connaissance en cours (NbQ-C4-Con) ou relative au besoin de connaître ou non celle-ci en vue de l'évaluation (NbQ-C4-Att).

Nous avons également codé les variables binaires correspondantes, notées Pos-cat-val, qui indiquent si l'étudiant a posé au moins une fois une question classée comme telle (ex : Pos-C3-

Def vaut 1 si l'étudiant a posé au moins une question sur une définition, c-à-d. si NbQ-C3-Def > 0).

En plus des variables relatives aux questions posées, nous avons également pour chaque étudiant :

- quatre scores d'AAR, mesurés par un questionnaire Cosnefroy *et al.* [2018]. Ce questionnaire est inclus dans le 2e questionnaire recherche (en annexe G) du MOOC (2 semaines après le début), comme mentionné en introduction de ce chapitre. Développé spécifiquement pour la mesure de différentes facettes liées à l'auto-régulation dans les environnements en ligne, il présente en plus l'avantage d'être étalonné en français, les traductions pouvant parfois avoir un effet sur la validité psychométrique des questionnaires. Bien que les auteurs ont calculé la moyenne des 4 scores individuels pour obtenir un score AAR global (avec un code de procrastination inversé), nous avons pensé qu'ils ont capturé différentes facettes de l'AAR qui pouvaient être associées individuellement à différents comportements de questionnement. Nous avons donc défini pour chaque étudiant son score d'absence de procrastination (Sco_{NPr}), son score de contexte (Sco_{Ctx}), son score de stratégie (Sco_{Str}) et son score de soutien des pairs (Sco_{Pai}). Chaque score est la moyenne de 5 à 6 questions et prend une valeur comprise entre 1 et 7,
- le score obtenu pour le parcours de base du MOOC (Sco_{Bas}), codé comme une valeur entre 0 et 100,
- le score obtenu pour le parcours avancé du MOOC (Sco_{Adv}), codé comme une valeur entre 0 et 100.

Comme pour les variables liées à la question, nous avons aussi codé les variables binaires correspondantes pour la performance, avec Cer_{Bas} (*resp.* Cer_{Ava}) qui indique si l'étudiant a obtenu la certification de base (*resp.* avancée) (c-à-d. si $Sco_{Bas} \geq 50$ (*resp.* si $Sco_{Adv} \geq 50$))

8.5.2 Corrélation et contingence

8.5.2.1 Méthode

Nous avons calculé pour chaque variable de question (NbQ-*) son coefficient de corrélation de Pearson (r) avec les 2 variables de performance et les 4 variables d'autorégulation (Sco_*). De plus, nous avons séparé la population entre ceux qui ont réussi et ceux qui ont échoué sur un parcours (basique ou avancé) en utilisant Cer_* et nous avons calculé le coefficient de contingence ϕ (équivalent au V de Cramér pour une matrice 2x2) avec chacune des variables binaires correspondant au fait de poser ou non une question de ce type (Pos-*).

8.5.2.2 Résultat 1 : performance

286 étudiants ont envoyé au moins un message avec un segment contenant une question explicite. Les résultats (résumés dans le Tableau 8.6) révèlent que poser des questions explicites (plus que des messages non catégorisés comme des questions), et des questions sur un sujet pertinent en cours, est un comportement en corrélation positive avec la performance (sur le parcours de base du MOOC et encore plus sur le parcours avancé). Plus particulièrement, les questions les plus fortement positivement corrélées à la performance sont celles qui permettent de vérifier la compréhension (Ver) d'un thème du cours ou d'une compétence que l'on doit maîtriser pour l'examen final.

Enfin, nous observons que les valeurs les plus élevées sont globalement similaires entre les valeurs r et ϕ, ce qui indique que c'est plus le fait de poser une question de l'un ou l'autre des types mentionnés ci-dessus que le fait d'en poser plusieurs qui pourrait être un indicateur intéressant de la performance finale. Les n/a pour ré-expliquer 'Ree' et schéma 'Sch' indiquent qu'aucun segment n'a été annoté avec ces étiquettes.

8.5.2.3 Résultat 2 : autorégulation

123 élèves parmi ceux qui ont envoyé au moins un message avec une question explicite ont également répondu au questionnaire d'AAR. Les résultats (résumés dans le Tableau 8.7) révèlent que la procrastination (ou l'absence de procrastination) n'est liée à aucun type particulier de question. Il est cependant, comme on pouvait s'y attendre, corrélé négativement avec le score dans les parcours de base ($r = -.349$) et avancé ($r = -.372$) du MOOC, c'est-à-dire que les étudiants qui procrastinent ont des scores plus faibles, ce qui est conforme aux autres travaux dans ce domaine [You, 2015]. La facette contexte de l'AAR n'est corrélée positivement qu'avec le nombre de messages ne contenant pas de question (NQ) ou une question implicite. En regardant d'un peu plus près la distribution des questions de non-cours, on voit que le seul taux de corrélation élevé (bien que non significatif) est lié aux questions d'administration. On peut donc supposer que s'intéresser au fonctionnement du MOOC est effectivement quelque chose qui se rattache, marginalement, à la prise en compte du contexte par l'apprenant.

Les deux autres facettes sont plus intéressantes, car les étudiants qui déclarent savoir utiliser des stratégies cognitives et méta-cognitives (comme la prise de notes) lorsqu'ils apprennent en ligne posent moins de questions sur l'organisation de l'examen final, moins de questions phatiques et moins de questions de vérification. En d'autres termes, il semble qu'en étant plus organisés, peut-être lorsqu'ils regardent la vidéo ou parcourent des pages de contenu, ils ont moins besoin de vérifier des informations qui sont probablement déjà mentionnées quelque part.

Quant aux étudiants qui déclarent être capables d'interagir avec les autres afin d'apprendre de manière plus efficace, ils ont logiquement tendance à envoyer plus de messages (questions et non-questions), qui peuvent concerner ou non le cours. En examinant plus en détail la nature des questions qu'ils posent, ils socialisent davantage avec les autres et posent davantage de questions administratives. Ils ont aussi tendance à poser des questions très pratiques sur comment effectuer une certaine tâche ou sur la raison pour laquelle un concept fonctionne d'une certaine manière.

L'analyse de ces résultats nous permet donc de répondre positivement aux questions de recherche 2 et 3 relatives au lien entre la nature des questions posées et la performance et les capacités d'auto-régulation. Bien que la valeur de certaines corrélations soit assez faible sur certaines catégories, nous pouvons dire globalement qu'il y a un lien entre la nature de questions posées et la performance et le comportement d'autorégulation chez les étudiants.

8.6 Profils des apprenants en termes de performance

Dans cette section, nous essayons de reproduire la méthodologie suivie dans la section 5.2.3 sur les données PACES, et construire des clusters sur la base des profils de questions posées par les étudiants. Nous regardons par la suite si les clusters obtenus peuvent être caractérisés

en termes de différence de performance. D'un autre côté, nous avons trouvé dans la section 8.5 qu'il existe un lien entre les questions des étudiants sur les forums et leur performance. Nous supposons donc qu'une analyse plus fine du contenu des messages du MOOC peut aider en particulier à prédire la réussite.

Pour répondre à la quatrième question de recherche, nous avons d'abord effectué quatre clusterings à l'aide de l'algorithme K-Means (k variant entre 2 et 10) sur quatre ensembles de données : les élèves qui ont posé des questions dans GDP5 (N_5= 278 élèves), GDP6 (N_6= 275), GDP7 (N_7=314) et GDP8 (N_8=287). Nous avons effectué le clustering en utilisant comme caractéristiques pour chaque élève la proportion de chaque question posée dans chaque dimension (par exemple : la proportion de questions avec la valeur "App" dans la dimension 1) posée globalement. Les résultats révèlent que deux clusters similaires se retrouvent dans chaque session du MOOC, appelés C1 et C2.

La deuxième étape consistait à caractériser les clusters en analysant les 19 dimensions utilisées (aucune question annotée sur Sch et Ree) pour extraire les différences significatives. Nous avons effectué 76 (19 fois 4) tests Mann-Whitney U pour chaque dimension pour chacune des 4 sessions, et ajusté la valeur du seuil p avec la correction de Bonferroni (valeur p ajustée $=$ .0007). Le Tableau 8.8 résume les résultats pour les dimensions avec une différence statistiquement significative dans au moins une des quatre sessions. Par rapport au cluster C2, les étudiants en C1 posent toujours plus de questions sur les examens, moins de questions de vérification en particulier sur les concepts du cours. Ils posent aussi parfois plus de questions administratives (GDP5 et 6), moins de questions reliant deux concepts (GDP7) et sur la manière de procéder (GDP8) ainsi que moins de questions pour approfondir leur compréhension (GDP8).

La troisième étape consistait à caractériser les clusters en termes d'attributs non utilisés pour le clustering.

- **NotFin :** note des élèves à l'examen final (note sur 100)
- **Reu :** proportion d'étudiants ayant réussi
- **QstCou :** proportion de questions sur le cours

Nous avons fait 4 tests de Chi-square pour la réussite (variable binaire, 1 si réussi et 0 sinon) et 4 Mann-Whitney U pour la dernière année. Ces tests ont révélé une différence statistiquement significative pour la note finale des sessions 6 et 8 seulement ($p = .014$ et $.040$ et $\eta^2 = 0.018$ et 0.010 respectivement), avec une note finale plus élevée pour C2, et une proportion plus élevée d'étudiants de C2 qui obtiennent leur certificat à la fin pour la session 8 seulement ($\chi^2 = 6.77$, $p = .009$, 79.9% contre 65.5%). Nous constatons également que la proportion de questions de cours est significativement plus élevée pour C2 pour chaque session.

Nous avons donc constaté que des clusters consistants (avec les mêmes caractéristiques) de questions apparaissent. En plus de leurs caractéristiques similaires, ces clusters sont dans certains cas en corrélation avec la performance. Bien que le fait de savoir si les questions liées au cours ou non peut être suffisant pour aider à prédire la réussite, nous soutenons que notre approche offre une meilleure compréhension de la nature des questions des élèves qui réussissent ou échouent, ouvrant la voie à une interprétation plus fine de ce que certains élèves font mal. Nous envisageons de reproduire cette analyse sur différents MOOCs pour voir si des patterns similaires peuvent être trouvés.

En ce qui concerne la réplication des analyses dans des contextes différents, nous avons

retrouvé certains patterns caractérisant les étudiants qui réussissent bien de ceux en difficulté dans les deux contextes étudiés (PACES et MOOC GDP) à partir du clustering. Les bons étudiants posaient souvent plus de questions de vérification de connaissances en cours et lien entre concepts. Ce qui peut s'expliquer ici, qu'ils ont une bonne compréhension de concepts de cours et posaient des questions bien ciblées sur les notions de cours. Nous avons également constaté que le nombre d'étudiants qui posaient des questions de vérification est en augmentation de la session GDP 5 (avec 35%) à la session GDP 8 (70%). Cela peut également expliquer la corrélation trouvée entre les étudiants qui réussissent bien en GDP8 et le fait de poser des questions de vérification et d'approfondissement. Malheureusement, nous ne disposons pas d'autres caractéristiques d'apprentissage (celles présentées dans le Tableau 3.4) pour comparer les clusters consistants du MOOC GDP avec ceux trouvés sur PACES.

8.7 Synthèse

Nous avons montré dans cette partie qu'il était possible d'annoter non seulement les messages des forums du MOOC, mais aussi des questions individuelles posées dans ces messages parfois longs. La segmentation des messages permet de distinguer l'intention de l'étudiant avec une granularité fine, en utilisant un schéma de codage adapté pour les questions liées ou non au cours. Par ailleurs, ce résultat ouvre la voie à l'annotation automatique des messages du MOOC, par exemple pour aider l'équipe pédagogique à déterminer rapidement l'intention des messages qui n'ont pas encore reçu de réponse. Un autre aspect intéressant est le fait que la nature des questions posées dans les messages fournit des informations sur la performance finale de l'élève, ainsi que sur certains aspects de son autorégulation (en particulier sa tendance à interagir avec les autres pour apprendre et son utilisation de stratégies cognitives et métacognitives). Il convient également de noter que certains patterns trouvés ici, comme le fait que les étudiants qui posent des questions de vérification ont tendance à réussir globalement mieux que les autres, sont cohérents avec les résultats que nous avons obtenus dans la première partie de ce travail (*cf.* section 5.2.3), malgré la différence forte de contexte (un MOOC vs. une plateforme de questions en ligne dans un contexte hybride).

Nous avons également montré une consistance au niveau de caractéristiques des clusters d'étudiants sur les différentes sessions de MOOC, qui sont également en corrélation avec la performance dans certains cas. Ces résultats sont cohérents avec les clusters des étudiants trouvés dans le contexte de la classe hybride (PACES). Il est important de noter que nous avons trouvé des patterns similaires en termes de nature de questions posées. En effet, les étudiants qui réussissent bien posent souvent plus de questions de vérification de connaissances que les autres.

Globalement, la réplication du processus suivi sur les données PACES appliqué aux données du MOOC GDP, a permis l'extension du schéma de codage de questions pour identifier les questions explicites dans les messages des forums de discussions et de les annoter séparément selon les catégories de questions de cours et non cours. Nous avons suivi le même processus d'annotation pour définir les catégories de questions non liées au cours. Bien que les messages des forums de discussions sont organisés autour de fils de discussion (questions techniques ou administratives, sur les devoirs ou le contenu des cours), les étudiants ne respectent souvent pas cette organisation. Nous avons donc effectué certains pré-traitements pour que le corpus soit utile à l'annotation manuelle et rajouter une nouvelle dimension pour annoter les questions non liées au cours (*cf.* section 8.2.1), ce qui n'a pas été nécessaire dans PACES puisque toutes

les questions posées en ligne portaient sur le cours. Par ailleurs, nous avons réussi à répliquer le processus d'analyse de PACES sur le MOOC GDP et obtenir des profils des étudiants similaires sur plusieurs sessions et contextes à partir de la nature des questions posées.

Certains aspects limitent cependant la généralisation de ces résultats : tout d'abord, comme nous l'avons déjà expliqué, le sujet du MOOC a probablement limité la performance de certains classifieurs en raison de son faible vocabulaire technique, avec des mots qui recouvrent le contenu et le contexte du cours. Deuxièmement, les valeurs kappa obtenues pour ces classifieurs étant moyennes, cela peut diminuer l'impact de certaines corrélations observées dans la section 8.5, valeurs de corrélation qui ne sont elles-mêmes jamais extrêmement élevées même lorsque $p < .05$. Enfin, comme c'est toujours le cas pour tout résultat relatif au forum dans les MOOC, ils ne sont généralement utilisés que par une petite minorité d'apprenants actifs dans le MOOC.

Les futures directions consistent à considérer certains des messages exclus ici pour simplifier cette première approche (messages qui ne sont pas à la racine dans le fil de discussion, fils techniques ou de socialisation puisque nous avons finalement dû concevoir un schéma de codage non liées aux questions de cours). Une autre direction que nous envisageons consisterait à envisager également des forums d'autres MOOC sur différents thèmes pour constituer un plus grand corpus de messages qui pourraient potentiellement améliorer la performance globale du système d'annotation.

Dim1	Type de question	Description
Ree	Ré-expliquer / redéfinir	Demander de revenir sur un concept déjà expliqué en cours
App	Approfondir un concept	Approfondir une connaissance, clarifier une ambiguïté ou demander plus de détails pour mieux comprendre
Ver	Validation / vérification	Vérifier ou valider une hypothèse
Dim2	**Sujet de question**	Description
Exe	Exemple	Exemple d'application (cours/exercice)
Sch	Schéma	Schéma d'application ou explication sur ce dernier
Cor	Correction	Correction d'un exercice en cours/examen
Dim3	Type d'explication	Description
Def	Définir	Définir un concept ou un terme
Man	Manière (comment ?)	La manière comment procéder
Rai	Raison (pourquoi ?)	Demander la raison
Rol	Rôles (utilité ?)	Demander l'utilité / fonction
Lie	Lien entre concepts	Vérifier le lien entre deux concepts, le définir
Dim4	Type de vérification (facultatif)	Description
Err	Erreur / contradiction	Détecter une erreur/ contradiction en cours ou dans l'explication de l'enseignant
Con	Connaissances liées au cours	Vérifier une connaissance
Att	**Connaissances attendues**	Vérifier une information dans un examen **ou quiz (évaluation)**

Tableau 8.2 – Schéma de codage présenté dans le chapitre 4 (adaptation propre au MOOC GDP en gras), utilisé pour annoter les questions des étudiants liées au cours

Dim0	Catégories	Description
Soc	Socialisation	Questions d'ordre social
Adm	Administratif	Administratif MOOC : inscription, certificat, modules, *etc.*
Exa	Examen/ quiz	Modalités d'évaluation : notes, format, *etc.*
PbT	Problème technique	Détecter un problème technique ou demander une solution
Res	Ressources non trouvées	Demander des ressources non trouvées
Out	Outils	Demander des outils
Pha	Phatique	Question qui n'a pas de valeur réelle ou information

Tableau 8.3 – Schéma de codage crée à partir de l'annotation manuelle pour annoter les questions des étudiants non liées au cours

Dim0	Questions non-cours	[Stump *et al.* , 2013]	[Cui & Wise, 2015]
Soc	Socialisation	Inclus dans social	Inclus dans non cours
Adm	Administratif	-	Inclus dans non cours
Exa	Examen/ quiz	Structure (partiellement)	Inclus dans non cours
PbT	Problème technique	Plateforme (partiellement)	Inclus dans non cours
Res	Ressources non trouvées	-	Inclus dans non cours
Out	Outils	Inclus dans technologie	Inclus dans non cours
Pha	Phatique	-	-
Dim1-4	**Questions cours**		
Dim1	Type de questions	Contenu (partiellement)	Cours (partiellement)
Dim2	Sujet de question	Contenu (partiellement)	-
Dim3	Type d'explication	Contenu (partiellement)	-
Dim4	Type de vérification	Contenu (partiellement)	-

Tableau 8.4 – Similarités entre les schéma de codage existants pour les forums de discussion et notre schéma de codage étendu

Classifieur	SVM	DT	GLM	GBT	K-NN	NB	RI
Cl_Q	0.60	**0.91**	**0.91**	**0.91**	0.80	0.54	**0.91**
Cl_C	**0.66**	0.40	0.62	0.51	0.33	0.47	0.43
Cl_0	-	0.28	0.35	**0.37**	0.22	0.21	0.11
Cl_1	-	0.61	0.63	**0.68**	0.27	0.07	0
Cl_2	-	0.30	0.09	**0.39**	0.05	0.03	0.37
Cl_3	-	0.50	**0.56**	0.54	0.14	0	0.26
Cl_4	-	**0.48**	0.45	0.47	0.19	0.07	0.35

(-) : Pas adapté aux données non binaires

Tableau 8.5 – Kappas obtenus entre l'annotation automatique et la référence annotation manuelle

	Sco_{Bas}		Cer_{Bas}	Sco_{Ava}		Cer_{Ava}
Cat.	r	p	ϕ	r	p	ϕ
NQ	.122	**.039***	.085	.160	**.007***	.043
Q	.148	**.012***	.037	.237	**.000***	.046
NC	.083	.164	.071	.121	**.041***	.018
NC-Soc	.082	.168	.091	.076	.198	.036
NC-Adm	.052	.384	.038	.016	.782	.013
NC-Exa	.061	.303	.057	.129	**.030***	.050
NC-PbT	.079	.180	.091	.090	.128	.086
NC-Res	-.001	.982	.004	.009	.885	.083
NC-Out	-.013	.821	.036	.025	.676	.060
NC-Pha	.013	.826	.023	.051	.391	.046
C	.153	**.010***	.134	.253	**.000***	.192
C1-Ree	n/a	n/a	n/a	n/a	n/a	n/a
C1-App	.031	.605	.063	.066	.264	.008
C1-Ver	.181	**.002***	.182	.292	**.000***	.276
C2-Exe	.047	.425	.052	.100	.092	.107
C2-Sch	n/a	n/a	n/a	n/a	n/a	n/a
C2-Cor	.053	.370	.024	.080	.179	.060
C3-Def	.049	.407	.037	.025	.674	.046
C3-Man	.006	.925	.078	.037	.538	.028
C3-Rai	.045	.451	.033	.031	.600	.034
C3-Rol	.051	.391	.052	.028	.641	.021
C3-Lie	.065	.273	.042	.097	.101	.074
C4-Err	.045	.444	.037	.077	.197	.075
C4-Con	.174	**.003***	.156	.282	**.000***	.252
C4-Att	.133	**.024***	.141	.209	**.000***	.222

n/a : aucun segment annoté avec ce code

Tableau 8.6 – Corrélation entre les types de questions et le score et la certification de base et avancé du MOOC

Cat.	Sco_{NPr}		Sco_{Ctx}		Sco_{Str}		Sco_{Pai}	
	r	p	r	p	r	p	r	p
NQ	-.064	.480	.178	**.049***	-.137	.131	.259	**.004***
Q	-.064	.485	.056	.541	.169	.061	.247	**.006***
NC	-.104	.254	.123	.174	-.161	.076	.212	**.019***
NC-Soc	-.095	.295	.081	.376	-.107	.239	.198	**.028***
NC-Adm	-.058	.520	.162	.074	.018	.840	.258	**.004***
NC-Exa	-.090	.322	-.006	.951	-.216	**.016***	.078	.392
NC-PbT	-.060	.513	.023	.802	.001	.995	.155	.087
NC-Res	.067	.459	.138	.128	.039	.665	.166	.067
NC-Out	.007	.939	.103	.257	.016	.858	-.050	.584
NC-Pha	-.018	.844	.079	.385	-.226	**.012***	.093	.307
C	-.028	.761	.005	.958	-.143	.115	.222	**.014***
C1-Ree	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
C1-App	-.040	.660	.075	.407	-.001	.987	.219	**.015***
C1-Ver	-.020	.660	-.022	.805	-.179	**.048***	.196	**.030***
C2-Exe	.005	.957	.123	.174	-.123	.174	.068	.487
C2-Sch	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
C2-Cor	-.105	.247	.038	.675	-.045	.619	.118	.193
C3-Def	.005	.957	-.130	.151	.135	.136	.055	.547
C3-Man	-.089	.329	.111	.221	-.002	.984	.187	**.038***
C3-Rai	-.068	.457	.140	.123	.024	.796	.222	**.014***
C3-Rol	n/a	n/a	n/a	n/a	.n/a	n/a	n/a	n/a
C3-Lin	-.009	.918	.117	.197	-.067	.459	.076	.406
C4-Err	-.112	.216	-.063	.491	-.067	.460	-.048	.599
C4-Con	-.009	.919	-.048	.597	-.176	.052	.191	**.034***
C4-Att	-.024	.795	.097	.288	-.136	.135	.169	.062

n/a : aucun segment annoté avec ce code

Tableau 8.7 – Corrélation entre les types de questions et les quatre scores d'AAR

		dim0$_{exa}$			**dim0**$_{adm}$			**dim1**$_{ver}$			**dim1**$_{app}$			**dim3**$_{man}$			**dim3**$_{lie}$			**dim4**$_{con}$			**NotFin**			**Reu**	**QstCou**
Cluster	*N*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Prop*	*Prop*
$C1_{GDP5}$	189	0	0.25*	1	0	0*	0.33	0	0*	0	0	0	0.18	0	0	0	0	0	0	0	0*	0	47.2	66.3	90.8	0.74	0.20*
$C2_{GDP5}$	96	0	0*	0	0	0*	0	0	0.93*	1	0	0	0.12	0	0	0	0	0	0	0.5	0.67*	1	52.82	82.4	92.22	0.78	0.86*
$C1_{GDP6}$	177	0	0.33*	1	0	0*	0.17	0	0*	0	0	0	0.14	0	0	0	0	0	0	0	0*	0	46.77	56.47*	89.67	0.70	0.22*
$C2_{GDP6}$	98	0	0*	0	0	0*	0	0.50	1*	1	0	0	0	0	0	0	0	0	0	0.50	0.75*	1	51.88	84.61*	93.64	0.78	0.86*
$C1_{GDP7}$	189	0	0.33*	1	0	0	0	0	0*	0	0	0	0.33	0	0	0	0	0*	0	0	0*	0	79.80	91.15	94.70	0.80	0.28*
$C2_{GDP7}$	125	0	0*	0.15	0	0	0	0.50	0.70*	1	0	0	0.17	0	0	0	0	0*	0	0.50	0.67*	1	81.72	91.92	96.20	0.81	0.84*
$C1_{GDP8}$	88	0.67	1*	1	0	0	0	0	0*	0	0	0*	0	0	0*	0	0	0	0	0	0*	0	33.30	84.68*	91.93	0.66*	0.08*
$C2_{GDP8}$	199	0	0*	0	0	0	0	0	0.44*	1	0	0*	0.33	0	0*	0	0	0	0	0	0.33*	1	77.90	88.33*	92.61	0.80*	0.66*

* signifie résultat significatif avec la correction de Bonferroni

Tableau 8.8 – Résumé de la médiane, du 1^{er} et $3^{ème}$ quartiles des variables utilisées pour le clustering et pour les variables dépendantes (NotFin, proportion de réussite et proportion de QstCou) pour chaque cluster et pour chaque session

Chapitre 9

Conclusion et perspectives

Nous avons montré dans cette thèse que l'utilisation d'un schéma de codage de questions et un système d'annotation automatique pourraient aider les enseignants à structurer l'ensemble des questions des étudiants et à mieux préparer leurs sessions de questions-réponses. Notre approche permet en effet de combiner les bénéfices de l'usage de notre typologie de questions, basée essentiellement sur des mots-clés indépendants du domaine, pour catégoriser les questions des étudiants. D'autre part, l'utilisation d'un système d'annotation automatique permet d'établir des liens entre les questions posées et le profil des apprenants.

Dans un premier temps, nous avons développé un schéma de codage des questions posées par les étudiants dans un environnement d'apprentissage hybride et nous avons reporté a posteriori les similitudes entre les typologies de questions existantes et notre schéma de codage. Bien que plusieurs travaux avaient définis des typologies de questions d'étudiants, aucun de ceux-ci ne correspondaient entièrement à notre contexte ici puisqu'ils dépendent essentiellement du contexte étudié et sont difficiles à réutiliser et automatiser hors du contexte de classe. Nous avons donc défini notre propre schéma de codage de questions, utilisant une approche ascendante fondée sur les données avec certaines caractéristiques : granularité assez fine, générique (en termes de contextes et domaine) et permettant une automatisation

Dans un deuxième temps, nous avons conçu et mis à l'épreuve quatre systèmes différents d'annotation automatique : un annotateur semi-automatique à base de règles d'expert et mots-clés pondérés manuellement, deux autres systèmes d'annotation entièrement automatiques basés sur des approches statistiques (des techniques d'apprentissage automatique et TF-IDF) et un annotateur hybride qui permet de combiner ces différents modèles pour en créer un nouveau plus performant en utilisant une approche de stacking. Afin d'assurer une bonne estimation de la performance sur des données non vues, les quatre systèmes ont été entraînés individuellement sur un ensemble de questions annotées manuellement sur chaque dimension séparément et ensuite évalués sur un échantillon indépendant de questions. L'utilisation d'un ensemble hybride d'annotateurs basés sur l'apprentissage automatique (ou TF-IDF) avec un annotateur semi-automatique à base de règles était la meilleure approche dans notre cas d'étude.

Nous avons montré également comment l'annotation automatique de questions nous permettait d'identifier des caractéristiques du profil des étudiants en termes de performance et d'autres aspects de leur comportement. Cela en utilisant exclusivement les proportions du type

de questions posées et leur évolution dans le temps. Ainsi, deux groupes extrêmes trouvés à partir du clustering (les élèves inférieurs à la moyenne ayant des questions populaires et les élèves supérieurs à la moyenne ayant des questions non populaires) sont toujours apparus sur différents cours et années, avec parfois un groupe intermédiaire (les élèves supérieurs à la moyenne ayant des questions populaires). Nous avons également caractérisé les étudiants qui n'ont pas posé de questions (assister à moins de cours, réussir moins bien et voter moins bien) et nous les avons comparés aux étudiants qui ont posé des questions (participation élevée, plus de succès et aussi voter davantage). Par ailleurs, nous avons pu constater dans les contextes étudiés que voter sans poser de questions n'est pas suffisant pour l'apprentissage, le vote ne semblant pas être un bon substitut au fait de poser des questions. En effet, il semble dans notre contexte que des activités «actives» (au sens de Chi [2009] –ici, voter) complémentaires à des activités «constructives» (ici, poser des questions) soient plus efficaces que des activités «constructives» seules, mais qu'une activité «active» seule soit plus négative en termes d'apprentissage que de la passivité. Pour résumer, voter est une bonne stratégie pour les étudiants sachant déjà formuler leurs propres questions. En revanche, pour ceux en difficulté, cela peut retarder la prise de conscience de leurs lacunes et leur capacité à les combler activement.

Par ailleurs, nous avons présenté un modèle dérivé du clustering des données et entraîné sur des sessions antérieures du même cours afin de prédire le profil des élèves en ligne à partir de leurs questions. Nous avons montré que les clusters obtenus sont similaires en termes de qualité et de caractéristiques d'étudiants d'une année à l'autre, ce qui signifie que la nature des questions posées par les étudiants et les caractéristiques de leur profil restent les mêmes. Cette analyse devrait nous permettre de fournir aux enseignants des informations supplémentaires sur les questions qu'ils ont reçues même en début d'année.

De plus, pour évaluer l'utilisabilité de nos propositions et particulièrement le schéma de codage développé auprès des enseignants de PACES, nous avons proposé via un questionnaire trois organisations pour regrouper les questions des étudiants par catégories de questions ou par profil d'étudiant (niveau d'apprenant). L'objectif de cette enquête était de proposer aux enseignants des alternatives d'organisations de questions pour remplacer le "mur de questions" qu'ils reçoivent actuellement afin de les aider à mieux préparer leurs sessions en présentiel (SEPI). L'analyse des résultats révèle la satisfaction des enseignants de la classe inversée et la pertinence des catégories définies dans le schéma de codage. Par ailleurs, bien que la segmentation de questions permet de distinguer les intentions de l'étudiant à granularité fine pendant l'annotation de questions, il s'est avéré que ce n'était pas important pour les enseignants. En revanche, nous n'avons pas pu déterminer un choix préféré d'organisation de questions auprès des enseignants : il ne faudrait donc probablement pas une seule mais plusieurs organisations à proposer aux enseignants, pour que chacun puisse choisir celle qui correspond le mieux à son intention pédagogique.

Pour vérifier la réplicabilité du processus d'annotation, nous avons étendu notre schéma de codage destiné aux questions des étudiants posées en classe hybride, pour l'adapter aux questions posées dans les forums de discussions dans le contexte d'un MOOC. Le nouveau schéma de codage consiste à catégoriser les questions des étudiants en distinguant celles liées au cours et non-cours. Ensuite, nous avons conçu un annotateur en trois phases pour identifier les messages avec des questions, questions liées au cours et non-cours et la nature de ces questions. Ce nouveau système permet non seulement d'annoter les messages des forums du MOOC, mais aussi des questions individuelles posées dans ces messages parfois longs. Il est donc important de noter que notre processus d'annotation et le schéma de codage utilisé pour les questions posées par les étudiants de PACES dans le cadre d'une classe hybride, peuvent

être facilement répliqués et réutilisés dans d'autres contextes (dans notre cas ici, MOOC).

En termes de réplication d'analyses, nous avons des caractéristiques similaires d'étudiants sur plusieurs sessions de MOOC, qui sont également en corrélation avec la performance dans certains cas à partir du clustering. Ces résultats sont cohérents avec les patterns extraits pour les étudiants bons et moyens dans le contexte de PACES. Bien que les messages postés dans les forums du MOOC nécessitent des pré-traitements différents des questions posées en classe hybride, nous avons réussi à répliquer le processus d'annotation et d'analyse de PACES sur le MOOC GDP. Cette approche semble donc être robuste et pouvoir s'appliquer à plusieurs contextes.

Certains aspects peuvent cependant limiter la généralisation de ces résultats.
Dans le cadre de PACES, tout d'abord, le contexte compétitif du concours (dans le cas de notre population, sur les 1600 inscrits, seulement les 200 premiers étudiants peuvent accéder en deuxième année) peut éventuellement biaiser certains résultats (par exemple : les étudiants qui posent des questions pour perturber les autres, le nombre important des redoublants, le système en lui-même qui ne favorise pas la collaboration et le partage de connaissances entre les étudiants). Un autre point important à signaler est que le nombre de questions posées par les étudiants est en augmentation chaque année dans la plupart des plateformes en ligne. En revanche, les enseignants de PACES constataient une baisse générale d'activité des étudiants sur la plateforme depuis l'année 2012 jusqu'à 2019.
Deuxièmement, nous n'avons pas eu accès aux logs permettant de savoir si les étudiants de PACES ont vraiment lu les autres questions posées. Les étudiants qui se connectent en premier n'ont également pas de questions sur lesquelles ils peuvent voter, sauf s'ils se reconnectent par la suite pour voir les nouvelles questions.

Du côté du travail de réplication, l'une des principales limites du MOOC GDP est le sujet du MOOC en lui-même. En raison de son faible vocabulaire technique, avec des mots qui recouvrent le contenu et le contexte du cours (par exemple : le concept du "projet" dans ce MOOC peut faire référence soit à un "projet" (devoir d'évaluation), soit à une gestion de projet (concept de cours)), a probablement limité la performance de certains classifieurs. Les valeurs kappa obtenues pour ces classifieurs étant moyennes, cela se répercute certainement sur la valeur de certaines corrélations observées dans la section 8.5. Enfin, comme c'est toujours le cas pour tout résultat relatif au forum dans les MOOC, ils ne sont généralement utilisés que par une petite minorité d'apprenants actifs dans le MOOC.

Perspectives

Nous clôturons ce manuscrit par deux principaux ensembles de perspectives issues de nos travaux que permettront de poursuivre le travail mené dans le cadre de cette thèse, à savoir des perspectives plutôt liées aux aspects informatique de la thèse, et des perspectives plus généralement associées au domaine de l'enseignement et l'apprentissage.

Perspectives informatiques

Ces perspectives visent à améliorer ou étendre les systèmes présentés dans cette thèse à des fins informatiques :

- L'intégration ou l'association de plusieurs taxonomies afin d'étendre notre schéma de codage de questions présenté dans le chapitre 8. En effet, nous avons pu identifier des similarités entre certaines taxonomies existantes (*cf.* section 2.1 et 8.4) et notre schéma de codage de questions. Le fait d'intégrer les catégories des autres taxonomies ou également le rajout de nouveaux éléments nécessitant une analyse sémantique (par exemple déterminer la nature de fautes dans les questions pour traiter la redondance ou les contradictions au sein d'une question) permettra d'élargir les dimensions et donner une granularité plus fine à notre schéma de codage pour le rendre plus générique.

- Améliorer l'annotation automatique du système à base de techniques d'apprentissage automatique en utilisant différentes sources de données (corpus variés de questions). Par exemple, entraîner le modèle de classification sur des questions issues de différents domaines : médecine, physique, gestion, etc. L'analyse des types des erreurs faites par les classifieurs permettrait d'étudier les limites de chacun et par conséquent choisir des méthodes plus adaptées pour améliorer leur performances.

- Réutiliser le schéma de codage pour catégoriser les questions des étudiants posées en d'autres langues et répliquer ainsi l'ensemble du processus d'annotation de questions pour rendre notre approche plus générique. Ce genre de tâche est facile à réaliser, puisqu'il est possible de ré-appliquer l'ensemble de méthodes et outils utilisés ici pour annoter des questions posées en langues différentes du Français.

- Améliorer la performance du modèle de clustering utilisé à des fins prédictives, en rajoutant par exemple de nouvelles caractéristiques pour prédire les profils des étudiants. Nous avons vu dans le chapitre 6 que le fait de rajouter plus de données n'a pas aidé à améliorer la performance du modèle. Cependant, il serait intéressant de mesurer le nombre minimum de questions que nécessite notre modèle pour rester fonctionnel et à partir de combien de semaines de cours nous pourrions l'obtenir.

- Utiliser les questions des étudiants comme des caractéristiques supplémentaires pour améliorer la prédiction de la performance ou le comportement d'auto-régulation chez les étudiants, puisque la nature des questions posées peut être corrélée à la performance et certains aspects de l'autorégulation.

Perspectives enseignement et apprentissage

La deuxième partie des perspectives concerne d'un côté l'enseignant, afin de lui fournir des outils de suivi ou d'aide dans le choix des questions à traiter en priorité, et d'un autre côté fournir une rétroaction à l'apprenant pour améliorer son apprentissage :

- Aider l'enseignant à cibler les questions pertinentes (par exemple : les questions qui nécessitent un minimum de connaissances de domaines pour les formuler, en utilisant des techniques tel que Word Embeddings pour une analyse sémantique des questions). Cela peut nécessiter de faire évoluer la classification actuelle présentée dans ce travail (comme déjà fait lors du passage au contexte de MOOC), mais aussi de croiser les transcriptions de cours avec les questions des étudiants pour identifier les catégories de questions qui nécessitent le plus de notions de cours dans la formulation de questions des étudiants.

- Fournir des informations supplémentaires à l'équipe pédagogique sur les apprenants

(paires questions-réponses caractérisant chaque étudiant) qui participent aux forums de discussions par le fait de poser des questions, répondre aux questions des autres ou faire les deux. Cela demanderait de repérer les patterns linguistiques des réponses caractéristiques des étudiants (en utilisant par exemple des techniques tel que Word Embeddings) selon leurs performances et les croiser également avec les types de questions posées afin de prédire le profil de questions-réponses pour chaque étudiant.

- Aider les étudiants qui ne savent pas formuler leurs propres questions à poser des questions en ligne et ne pas se contenter seulement de voter sur les questions des autres. En effet, comme vu dans le chapitre 5, des activités "actives" au sens de Chi [2009] (ici, voter) ne sont pas suffisantes en termes d'apprentissage et il faudra les accompagner à des activités « constructives » (ici, poser des questions). D'ailleurs, plusieurs travaux avaient étudié la valeur de questionnement pour l'apprentissage afin d'encourager les étudiants à poser des questions de haut niveau [Graesser & Person, 1994] et à reconnaître les bonnes questions [Marbach-Ad & Sokolove, 2000].

- Aider l'équipe pédagogique à répondre aux questions urgentes des étudiants afin de libérer le temps et l'attention de l'enseignant pour s'engager dans d'autres activités. Cela en combinant notre travail et celui de Almatrafi *et al.* [2018] par exemple qui ont déjà détecté "l'urgence" dans les messages de forums. Notre approche permettrait donc d'annoter les questions des étudiants dans ces messages et les annoter d'une manière plus fine selon notre schéma de codage et système d'annotation automatique.

- Évaluer comment forcer à poser des questions avant de voter peut aider les étudiants. Une expérience davantage contrôlée à mettre en place après avoir vu une vidéo de cours, qui nécessiterait deux groupes d'étudiants, dans la première moitié du cours, le premier devrait poser des questions, le second devrait voter sur les questions précédemment posées pour passer à la suite. Le rôle des deux groupes devrait s'inverser dans la deuxième moitié du cours pour assurer la même chance à tous les étudiants. L'objectif serait de vérifier si une différence apparaît en termes de performance entre les groupes (*cf.* section 5.4 du chapitre 5). Par conséquent, ce genre d'intervention permettrait également aux enseignants de choisir la bonne stratégie d'enseignement (encourager les étudiants à poser leurs propres questions plutôt que de voter directement sur les questions des autres sans faire un effort préalable de formulation de question).

- Envisager des tableaux de bords personnalisés pour les enseignants à partir des organisations proposées. En effet, l'analyse des résultats du questionnaire (*cf.* Chapitre 7) a révélé qu'il n'y a pas d'unanimité de choix d'organisations de la part des enseignants. Il serait donc plus pertinent de présenter différentes visualisations alternatives de questions aux enseignants et laisser à chacun d'entre eux choisir celle qui convient le plus à sa stratégie en classe au lieu de présenter une seule visualisation pour tous.

Bibliographie

Agrawal, Akshay, Venkatraman, Jagadish, Leonard, Shane, & Paepcke, Andreas. 2015. YouEDU : Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips. Stanford InfoLab.

Almatrafi, O., & Johri, A. 2018. Systematic Review of Discussion Forums in Massive Open Online Courses (MOOCs). *IEEE Transactions on Learning Technologies*, 1–1.

Almatrafi, Omaima, Johri, Aditya, & Rangwala, Huzefa. 2018. Needle in a haystack : Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education*, **118**(Mar.), 1–9.

Anderson, Lorin W., Krathwohl, David R., Airasian, Peter W., Cruikshank, Kathleen A., Mayer, Richard E., Pintrich, Paul R., Raths, James, & Wittrock, Merlin C. 2001. A taxonomy for learning, teaching, and assessing : A revision of Bloom's taxonomy of educational objectives, abridged edition. *White Plains, NY : Longman*.

Andresen, Martin A. 2009. Asynchronous discussion forums : success factors, outcomes, assessments, and limitations. *Journal of Educational Technology & Society*, **12**(1), 249–257.

Antonenko, Pavlo D., Toy, Serkan, & Niederhauser, Dale S. 2012. Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, **60**(3), 383–398.

Arguello, Jaime, & Shaffer, Kyle. 2015. Predicting speech acts in MOOC forum posts. *In : Ninth International AAAI Conference on Web and Social Media*.

Artstein, Ron, & Poesio, Massimo. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, **34**(4), 555–596.

Atapattu, Thushari, Falkner, Katrina, & Tarmazdi, Hamid. 2016. Topic-wise Classification of MOOC Discussions : A Visual Analytics Approach. *Pages 276–281 of : International Conference on Educational Data Mining*.

Azzag, Hanene, Guinot, Christiane, & Venturini, Gilles. 2006 (Jan.). Classification hiérarchique et visualisation de pages Web. *Pages 5–16 of : 6ème journées francophones Extraction et Gestion des Connaissances*.

Bayat, Berken, Krauss, Christopher, Merceron, Agathe, & Arbanowski, Stefan. 2016. Supervised Speech Act Classification of Messages in German Online Discussions. *Pages 204–209 of : FLAIRS Conference*.

Bergmann, Jonathan, & Sams, Aaron. 2012. *Flip your classroom : reach every student in every class every day*. 1. ed edn. Eugene, Or. : ISTE. OCLC : 802351882.

Bergsma, Wicher. 2013. A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*, **42**(3), 323–328.

Bihani, Ankita, Ullman, Jeffrey D., & Paepcke, Andreas. 2018 (Mar.). *FAQtor : Automatic FAQ generation using online forums*.

Bloom, B. S., & Engelhart, M. B. 1956. Furst, EJ. *Hill, WH, & Krathwohl, DR Taxonomy of educational objectives. The classification of educational goals. Handbook I : Cognitive domain. New York : Longmans Green*.

Boekaerts, Monique. 1997. Self-regulated learning : A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, **7**(2), 161–186.

Bouchet, François. 2009. Characterization of conversational activities in a corpus of assistance requests. *Proceedings of the 14th Student Session of the European Summer School for Logic, Language, and Information (ESSLLI)*, 40–50.

Bouchet, François, & Bachelet, Rémi. 2019. Socializing on MOOCs : Comparing University and Self-enrolled Students. *Pages 31–36 of :* Calise, Mauro, Delgado Kloos, Carlos, Reich, Justin, Ruiperez-Valiente, Jose A., & Wirsing, Martin (eds), *Digital Education : At the MOOC Crossroads Where the Interests of Academia and Business Converge*. Lecture Notes in Computer Science. Springer International Publishing.

Bouchet, François, Harley, Jason M., Trevors, Gregory J., & Azevedo, Roger. 2013. Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *JEDM-Journal of Educational Data Mining*, **5**(1), 104–146.

Bouchet, François, Labarthe, Hugues, Yacef, Kalina, & Bachelet, Rémi. 2017. Comparing Peer Recommendation Strategies in a MOOC. *Pages 129–134 of : Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*. UMAP '17. New York, NY, USA : ACM. event-place : Bratislava, Slovakia.

Breslow, Lori, Pritchard, David E., DeBoer, Jennifer, Stump, Glenda S., Ho, Andrew D., & Seaton, Daniel T. 2013. Studying Learning in the Worldwide Classroom Research into edX's First MOOC. *Research & Practice in Assessment*, **8**, 13–25.

Cao, Meng, Tang, Yun, & Hu, Xiangen. 2017. An Analysis of Students' Questions in MOOCs Forums. *Pages 56–63 of :* Hu, Xiangen, Barnes, Tiffany, Hershkovitz, Arnon, & Paquette, Luc (eds), *Proceedings of the 10th International Conference on Educational Data Mining*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, **16**(June), 321–357.

Chi, Michelene T. H. 2009. Active-Constructive-Interactive : A Conceptual Framework for Differentiating Learning Activities. *Topics in Cognitive Science*, **1**(1), 73–105.

Chin, Christine, & Brown, David E. 2000. Learning deeply in science : An analysis and reintegration of deep approaches in two case studies of grade 8 students. *Research in Science Education*, **30**(2), 173–197.

Chin, Christine, & Brown, David E. 2002. Student-generated questions : A meaningful aspect of learning in science. *International Journal of Science Education*, **24**(5), 521–549.

Chin, Christine, & Kayalvizhi, G. 2002. Posing Problems for Open Investigations : What questions do pupils ask ? *Research in Science & Technological Education*, **20**(2), 269–287.

Chin, Christine, & Osborne, Jonathan. 2008. Students' questions : a potential resource for teaching and learning science. *Studies in science education*, **44**(1), 1–39.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. 2nd edn. Hillsdale, NJ : Erlbaum.

Colbert, James T., Olson, Joanne K., Clough, Michael P., & Tomanek, Debra. 2007. Using the Web to Encourage Student-generated Questions in Large-Format Introductory Biology Classes. *CBE—Life Sciences Education*, **6**(1), 42–48.

Cook, Connor, Kelly, Sean, Olney, Andrew M., & D'Mello, Sidney K. 2018 (July). An Open Vocabulary Approach for Estimating Teacher Use of Authentic Questions in Classroom Discourse. *Pages 116–126 of :* Boyer, Kristy Elizabeth, & Yudelson, Michael (eds), *Proceedings of the 11th International Conference on Educational Data Mining*.

Cormack, Gordon V. 2008. Email Spam Filtering : A Systematic Review. *Foundations and Trends® in Information Retrieval*, **1**(4), 335–455.

Cosnefroy, Laurent, Fenouillet, Fabien, & Heutte, Jean. 2018. Développement et validation d'une échelle d'apprentissage autorégulé en ligne. *In : 2e Colloque international e-Formation des Adultes et Jeunes Adultes.*

Cross, Sebastian, Waters, Zak, Kitto, Kirsty, & Zuccon, Guido. 2017. Classifying help seeking behaviour in online communities. *Pages 419–423 of : Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM.

Cui, Yi, & Wise, Alyssa Friend. 2015. Identifying Content-Related Threads in MOOC Discussion Forums. *Pages 299–303 of : Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*. Vancouver, BC, Canada : ACM Press.

Dascalu, Mihai, Dessus, Philippe, Trausan-Matu, Ștefan, Bianco, Maryse, & Nardy, Aurélie. 2013. ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. *Pages 379–388 of : Artificial Intelligence in Education*. Springer, Berlin, Heidelberg.

Davies, D. L., & Bouldin, D. W. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-1**(2), 224–227.

Dettori, G., & Persico, D. 2008. Detecting Self-Regulated Learning in Online Communities by Means of Interaction Analysis. *IEEE Transactions on Learning Technologies*, **1**(1), 11–19.

Dringus, Laurie P., & Ellis, Timothy. 2005. Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education*, **45**(1), 141–160.

Dunn, J. C. 1974. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, **4**(1), 95–104.

Eicher, Bobbie, Polepeddi, Lalith, & Goel, Ashok. 2017. Jill Watson doesn't care if you're pregnant : grounding AI ethics in empirical studies. *In : AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, New Orleans, LA*, vol. 7.

Elgort, Irina, Lundqvist, Karsten, McDonald, Jenny, & Moskal, Adon Christian Michael. 2018. Analysis of student discussion posts in a MOOC : Proof of concept. *In : Proceedings of 8th International Conference on Learning Analytics & Knowledge (LAK18), Sydney, Australia*.

Etkina, Eugenia, & Harper, Kathleen Andre. 2002. Weekly Reports : Student Reflections on Learning. An Assessment Tool Based on Student and Teacher Feedback. *Journal of College Science Teaching*, **31**(7), 476–80.

Fort, Karën. 2012 (Dec.). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. phdthesis, Université Paris-Nord - Paris XIII.

Friedman, J. 1996. *Another approach to polychotomous classification. Department of Statistics*. Stanford University Stanford, CA, USA :.

Fritz, Catherine O., Morris, Peter E., & Richler, Jennifer J. 2012. Effect size estimates : Current use, calculations, and interpretation. *Journal of Experimental Psychology : General*, **141**(1), 2–18.

Garcia-Fernandez, Anne. 2010 (Jan.). *Génération de réponses en langue naturelle orales et écrites pour les systèmes de question-réponse en domaine ouvert*. thesis, Paris 11.

Gautam, Dipesh, Maharjan, Nabin, Graesser, Arthur C., & Rus, Vasile. 2018. *Automated Speech Act Categorization of Chat Utterances in Virtual Internships*. International Educational Data Mining Society.

Gillois, Pierre, Pagonis, Daniel, Vuillez, Jean-Philippe, Bosson, Jean-Luc, & Romanet, Jean-Paul. 2013. Réforme pédagogique et «e-learning» pour le concours de première année à la faculté de médecine de Grenoble : satisfaction des étudiants et des enseignants. *La presse médicale*, **42**(2), e44–e52.

Goel, Ashok K., & Polepeddi, Lalith. 2016. *Jill Watson : A Virtual Teaching Assistant for Online Education*. Technical Report. Georgia Institute of Technology.

Graesser, Art, Ozuru, Yasuhiro, & Sullins, Jeremiah. 2010. What is a good question ? *Bringing reading research to life.*

Graesser, Arthur C., & Person, Natalie K. 1994. Question asking during tutoring. *American educational research journal*, **31**(1), 104–137.

Hakkarainen, Kai. 2003. Progressive inquiry in a computer-supported biology class. *Journal of Research in Science Teaching*, **40**(10), 1072–1088.

Harper, Kathleen A., Etkina, Eugenia, & Lin, Yuhfen. 2003. Encouraging and analyzing student questions in a large physics course : Meaningful patterns for instructors. *Journal of Research in Science Teaching*, **40**(8), 776–791.

Harrak, Fatima, Bouchet, François, & Luengo, Vanda. 2017. Identifying relationships between students' questions type and their behavior. *Pages 402–403 of : 10th International Conference on Educational Data Mining*.

Harrak, Fatima, Bouchet, François, Luengo, Vanda, & Gillois, Pierre. 2018. Profiling Students from Their Questions in a Blended Learning Environment. *Pages 102–110 of : Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. LAK '18. Sydney, Australia : ACM.

Harrak, Fatima, Bouchet, François, Luengo, Vanda, & Bachelet, Rémi. 2019a (July). Automatic identification of questions in MOOC forums and association with self-regulated learning. *In : Educational Data Mining*.

Harrak, Fatima, Bouchet, François, & Luengo, Vanda. 2019b (June). Comparaison de questions posées et votées en ligne dans le cadre d'une classe inversée. *In : Environnements Informatiques pour l'Apprentissage Humain*.

Harrak, Fatima, Bouchet, François, & Luengo, Vanda. 2019c. From Students' Questions to Students' Profiles in a Blended Learning Environment. *Journal of Learning Analytics*, **6**(1), 54–84–54–84.

Harrak, Fatima, Bouchet, François, Luengo, Vanda, & Bachelet, Rémi. 2019d (June). Towards Improving Students' Forum Posts Categorization in MOOCs and Impact on Performance Prediction. *In : Learning @ Scale*.

Hecking, T., Chounta, I., & Hoppe, H. U. 2015 (Sept.). Analysis of User Roles and the Emergence of Themes in Discussion Forums. *Pages 114–121 of : 2015 Second European Network Intelligence Conference*.

Huang, Jonathan, Dasgupta, Anirban, Ghosh, Arpita, Manning, Jane, & Sanders, Marc. 2014. Superposter Behavior in MOOC Forums. *Pages 117–126 of : Proceedings of the First ACM Conference on Learning @ Scale Conference*. L@S '14. New York, NY, USA : ACM. event-place : Atlanta, Georgia, USA.

Ishola, Oluwabukola Mayowa, & McCalla, Gordon. 2017a. Personalized Tag-Based Knowledge Diagnosis to Predict the Quality of Answers in a Community of Learners. *Pages 113–124 of :* André, Elisabeth, Baker, Ryan, Hu, Xiangen, Rodrigo, Ma. Mercedes T., & du Boulay, Benedict (eds), *Artificial Intelligence in Education*. Lecture Notes in Computer Science. Springer International Publishing.

Ishola, Oluwabukola Mayowa, & McCalla, Gordon. 2017b (June). Predicting Prospective Peer Helpers to Provide Just-In-Time Help to Users in Question and Answer Forums. *Pages 283–243 of :* Hu, Xiangen, Barnes, Tiffany, Hershkovitz, Arnon, & Paquette, Luc (eds), *Proceedings of the 10th International Conference on Educational Data Mining*.

Jenders, Maximilian, Krestel, Ralf, & Naumann, Felix. 2016. Which Answer is Best ? : Predicting Accepted Answers in MOOC Forums. *Pages 679–684 of : Proceedings of the 25th International Conference Companion on World Wide Web*. WWW '16 Companion. Republic and Canton of Geneva, Switzerland : International World Wide Web Conferences Steering Committee. event-place : Montréal, Québec, Canada.

Jiang, Zhuoxuan, Zhang, Yan, Liu, Chi, & Li, Xiaoming. 2015. Influence Analysis by Heterogeneous Network in MOOC Forums : What Can We Discover ?. *In : International Conference on Educational Data Mining*.

Kessler, Rémy, Torres-Moreno, Juan-Manuel, & El-Bèze, Marc. 2006. Classification automatique de courriers électroniques par des méthodes mixtes d'apprentissage. *Ingénierie des systèmes d'information*, **11**(2), 93–112.

Khosravi, Hassan. 2017. Recommendation in Personalised Peer-Learning Environments. *arXiv :1712.03077 [cs]*, Dec. arXiv : 1712.03077.

Kim, Jihie, Shaw, Erin, & Ravi, Sujith. 2010a. Mining Student Discussions for Profiling Participation and Scaffolding Learning. *Pages 299–310 of : Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Kim, Jihie, Li, Jia, & Kim, Taehwan. 2010b. Towards identifying unresolved discussions in student online forums. *Pages 84–91 of : Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Kim, Jungjoo. 2013. Influence of group size on students' participation in online discussion forums. *Computers & Education*, **62**(Mar.), 123–129.

Kiss, Tibor, & Strunk, Jan. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, **32**(4), 485–525.

Kizilcec, René F., Pérez-Sanagustín, Mar, & Maldonado, Jorge J. 2017. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, **104**(Jan.), 18–33.

Klüsener, Marcus, & Fortenbacher, Albrecht. 2015. Predicting students' success based on forum activities in MOOCs. *Pages 925–928 of : Intelligent Data Acquisition and Advanced Computing Systems : Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on*, vol. 2. IEEE.

Kolb, David A. 1984. *Experiential learning : Experience as the source of learning and development*. Englewood Cliffs.

Kolb, David A. 1985. *Learning-style inventory : Self-scoring inventory and interpretation booklet : Revised scoring*. Boston : McBer and Co.

Koren, Yehuda. 2009. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, **81**(2009), 1–10.

Labarthe, Hugues, Bouchet, François, Bachelet, Rémi, & Yacef, Kalina. 2016. Does a Peer Recommender Foster Students' Engagement in MOOCs ? 418-423.

Lai, Ming, & Law, Nancy. 2013. Questioning and the quality of knowledge constructed in a CSCL context : a study on two grade-levels of students. *Instructional Science*, **41**(3), 597–620.

Lebart, Ludovic. 2001. Traitement statistique des questions ouvertes : quelques pistes de recherche. *Journal de la société française de statistique*, **142**(4), 7–20.

Lebart, Ludovic, Morineau, Alain, & Piron, Marie. 1995. *Statistique exploratoire multidimensionnelle*. Paris : Dunod. OCLC : 729665369.

Lebrun, Marcel. 2015. L'école de demain : entre MOOC et classe inversée. *Économie et management. La révolution numérique*, 41–47.

Li, Haiying, Duan, Ying, Clewley, Danielle N., Morgan, Brent, Graesser, Arthur C., Shaffer, David Williamson, & Saucerman, Jenny. 2014. Question Asking During Collaborative Problem Solving in an Online Game Environment. *Pages 617–618 of : Intelligent Tutoring Systems*. Lecture Notes in Computer Science. Springer, Cham.

Lin, Fu-Ren, Hsieh, Lu-Shih, & Chuang, Fu-Tai. 2009. Discovering genres of online discussion threads via text mining. *Computers & Education*, **52**(2), 481–495.

Liu, Qian, Peng, Weijun, Zhang, Fan, Hu, Rong, Li, Yingxue, & Yan, Weirong. 2016a. The effectiveness of blended learning in health professions : systematic review and meta-analysis. *Journal of medical Internet research*, **18**(1).

Liu, Weizhe, Kidziński, Łukasz, & Dillenbourg, Pierre. 2016b. Semiautomatic Annotation of MOOC Forum Posts. *Pages 399–408 of :* Li, Yanyan, Chang, Maiga, Kravcik, Milos, Popescu, Elvira, Huang, Ronghuai, Kinshuk, & Chen, Nian-Shing (eds), *State-of-the-Art and Future Directions of Smart Learning*. Singapore : Springer Singapore.

Longstaff, Emily. 2017. Ritual in Online Communities : A Study of Post-Voting in MOOC Discussion Forums. *International Journal of Human–Computer Interaction*, **33**(8), 655–663.

Macina, Jakub, Srba, Ivan, Williams, Joseph Jay, & Bielikova, Maria. 2017. Educational Question Routing in Online Student Communities. *Pages 47–55 of : Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys '17. New York, NY, USA : ACM. event-place : Como, Italy.

Macqueen, J. 1967. SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS. *MULTIVARIATE OBSERVATIONS*, **1**(14), 281–297.

Manning, Christopher, Raghavan, Prabhakar, & Schütze, Hinrich. 2010. Introduction to information retrieval. *Natural Language Engineering*, **16**(1), 100–103.

Marbach-Ad, Gili, & Sokolove, Phillip G. 2000. Good science begins with good questions. *Journal of College Science Teaching*, **30**(3), 192.

Marbach-Ad, Gili, & Sokolove, Phillip G. 2000. Can undergraduate biology students learn to ask higher level questions ? *Journal of Research in Science Teaching*, **37**(8), 854–870.

Mazur, Eric. 1997. Peer Instruction : A User's Manual. *American Journal of Physics*, **67**(4), 359–360.

Merceron, Agathe. 2014. Connecting Analysis of Speech Acts and Performance Analysis-An Initial Study. *In : LAK Workshops*.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, & Dean, Jeffrey. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv :1301.3781 [cs]*, Jan. arXiv : 1301.3781.

Miller, George A. 1995. WordNet : A Lexical Database for English. *Commun. ACM*, **38**(11), 39–41.

Mitchell, Tom M. 1997. Does Machine Learning Really Work ? *AI Magazine*, **18**(3), 11–11.

Munoz, Daniel, Bagnell, J. Andrew, & Hebert, Martial. 2010. Stacked hierarchical labeling. *Pages 57–70 of : European Conference on Computer Vision*. Springer.

Mustafaraj, Eni, & Bu, Jessica. 2015. The Visible and Invisible in a MOOC Discussion Forum. *Pages 351–354 of : Proceedings of the Second (2015) ACM Conference on Learning @ Scale*. L@S '15. New York, NY, USA : ACM. event-place : Vancouver, BC, Canada.

Nelder, J. A., & Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society : Series A (General)*, **135**(3), 370–384.

Nizet, Isabelle, & Meyer, Florian. 2016. Inverser la classe : effets sur la formation de futurs enseignants. *Revue internationale de pédagogie de l'enseignement supérieur*, **32**(32(1)).

Nylén, Aletta, Thota, Neena, Eckerdal, Anna, Kinnunen, Päivi, Butler, Matthew, & Morgan, Michael. 2015. Multidimensional Analysis of Creative Coding MOOC Forums : A Methodological Discussion. *Pages 137–141 of : Proceedings of the 15th Koli Calling Conference on Computing Education Research*. Koli Calling '15. New York, NY, USA : ACM. event-place : Koli, Finland.

Oleksandra, Poquet, & Shane, Dawson. 2016. Untangling MOOC learner networks. *Pages 208–212 of : Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. Edinburgh, United Kingdom : ACM Press.

Otero, Jose, & Graesser, Arthur C. 2001. PREG : Elements of a model of question asking. *Cognition and instruction*, **19**(2), 143–175.

Park, Jihyun, Yu, Renzhe, Rodriguez, Fernando, Baker, Rachel, Smyth, Padhraic, & Warschauer, Mark. 2018. *Understanding Student Procrastination via Mixture Models*. Proceedings of the 11th International Conference on Educational Data Mining.

Pashler, Harold, McDaniel, Mark, Rohrer, Doug, & Bjork, Robert. 2008. Learning Styles : Concepts and Evidence. *Psychological Science in the Public Interest*, **9**(3), 105–119.

Paternostre, Marjorie, Francq, Pascal, Lamoral, Julien, Wartel, David, & Saerens, Marco. 2002. Carry, un algorithme de désuffixation pour le français. *Rapport technique du projet Galilei*.

Pedrosa de Jesus, Helena, Teixeira-Dias, José J. C., & Watts, Mike. 2003. Questions of chemistry. *International Journal of Science Education*, **25**(8), 1015–1034.

Pedrosa de Jesus, Helena, Almeida, Patricia, & Watts, Mike. 2004. Questioning Styles and Students' Learning : Four case studies. *Educational Psychology*, **24**(4), 531–548.

Pennington, Jeffrey, Socher, Richard, & Manning, Christopher. 2014. Glove : Global Vectors for Word Representation. *Pages 1532–1543 of : Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics.

Peters, Vanessa L., & Hewitt, Jim. 2010. An investigation of student practices in asynchronous computer conferencing courses. *Computers & Education*, **54**(4), 951–961.

Pintrich, Paul R. 1999. The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, **31**(6), 459–470.

Pintrich, Paul R. 2004. A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review*, **16**(4), 385–407.

Pizzini, Edward L., & Shepardson, Daniel P. 1991. Student Questioning in the Presence of the Teacher During Problem Solving in Science. *School Science and Mathematics*, **91**(8), 348–352.

Porter, M F. 1980. An algorithm for suffix stripping. 157.

Porter, M. F. 2006. *Stemming algorithms for various European languages*.

Puustinen, Minna, & Pulkkinen, Lea. 2001. Models of Self-regulated Learning : A review. *Scandinavian Journal of Educational Research*, **45**(3), 269–286.

Qadir, Ashequl, & Riloff, Ellen. 2011. Classifying sentences as speech acts in message board posts. *Pages 748–758 of : Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reich, Justin, Tingley, Dustin H., Leder-Luis, Jetson, Roberts, Margaret, & Stewart, Brandon. 2014 (Sept.). *Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses*. SSRN Scholarly Paper ID 2499725. Social Science Research Network, Rochester, NY.

Rousseeuw, Peter J. 1987. Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**(Nov.), 53–65.

Rus, Vasile, Moldovan, Cristian, Niraula, Nobal, & Graesser, Arthur C. 2012. *Automated Discovery of Speech Act Categories in Educational Games*. International Educational Data Mining Society.

Sagot, Benoît, & Fišer, Darja. 2008. Building a free French wordnet from multilingual resources. *In : OntoLex*.

Salton, Gerard. 1989. Automatic text processing : The transformation, analysis, and retrieval of. *Reading : Addison-Wesley*.

Scardamalia, Marlene, & Bereiter, Carl. 1992. Text-Based and Knowledge Based Questioning by Children. *Cognition and Instruction*, **9**(3), 177–199.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *In : Proceedings of International Conference on New Methods in Language Processing*.

Searle, John R. 1969. Speech act theory. *Cambridge : Cambridge UP*.

Sindhgatta, Renuka, Marvaniya, Smit, Dhamecha, Tejas I., & Sengupta, Bikram. 2017 (June). Inferring Frequently Asked Questions from Student Question Answering Forums. *In :* Hu, Xiangen, Barnes, Tiffany, Hershkovitz, Arnon, & Paquette, Luc (eds), *Proceedings of the 10th International Conference on Educational Data Mining*.

Stump, Glenda S, DeBoer, Jennifer, Whittinghill, Jonathan, & Breslow, Lori. 2013. Development of a Framework to Classify MOOC Discussion Forum Posts : Methodology and Challenges. 20.

Supraja, S., Hartman, Kevin, Tatinati, Sivanagaraja, & Khong, Andy WH. 2017. Toward the Automatic Labeling of Course Questions for Ensuring their Alignment with Learning Outcomes. *Pages 56–63 of : Proceedings of The 10th International Conference on Educational Data Mining*. Wuhan, China : Xiangen Hu, Tiffany Barnes, Arnon Hershkovitz, Luc Paquette.

Séguéla, Julie. 2017. Fouille de données textuelles et systèmes de recommandation appliqués aux offres d'emploi diffusées sur le web. 203.

Teixeira-Dias, José J.C., Pedrosa de Jesus, Helena, Neri de Souza, Francislê, & Watts, Mike. 2005. Teaching for quality learning in chemistry. *International Journal of Science Education*, **27**(9), 1123–1137.

Tian, Yuan, Kochhar, Pavneet Singh, Lim, Ee-Peng, Zhu, Feida, & Lo, David. 2013. Predicting Best Answerers for New Questions : An Approach Leveraging Topic Modeling and Collaborative Voting. *Pages 55–68 of :* Nadamoto, Akiyo, Jatowt, Adam, Wierzbicki, Adam, & Leidner, Jochen L. (eds), *International Conference on Social Informatics*. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

van Aalst, Jan. 2009. Distinguishing knowledge-sharing, knowledge-construction, and knowledge-creation discourses. *International Journal of Computer-Supported Collaborative Learning*, **4**(3), 259–287.

van den Boom, Gerard, Paas, Fred, van Merriënboer, Jeroen J. G., & van Gog, Tamara. 2004. Reflection prompts and tutor feedback in a web-based learning environment : effects on students' self-regulated learning competence. *Computers in Human Behavior*, **20**(4), 551–567.

Vapnik, Vladimir. 1998. The Support Vector Method of Function Estimation. *Pages 55–85 of :* Suykens, Johan A. K., & Vandewalle, Joos (eds), *Nonlinear Modeling : Advanced Black-Box Techniques*. Boston, MA : Springer US.

Villiot-Leclercq, Emmanuelle, Dhorne, Lucie, Charroud, Christophe, & Dessus, Philippe. 2012. Les formations préparatoires au C2i2e entre présence et distance : quels scénarios d'hybridation ? *In : Actes du 4e Colloque International de l'Université à l'ère du numérique (CIUEN'12)*.

Vygotsky, Lev Semenovich. 1978. *Mind in society : The development of higher psychological processes*. Harvard university press, new edition.

Watts, Mike, Gould, Gillian, & Alsop, Steve. 1997. Questions of Understanding : Categorising Pupils' Questions in Science. *School Science Review*, **79**(286), 57–63.

Wen, Miaomiao, Yang, Diyi, & Rosé, Carolyn Penstein. 2014. Sentiment Analysis in MOOC Discussion Forums : What does it tell us ? 8.

White, Richard, & Gunstone, Richard. 1992. *Probing Understanding*. Routledge.

Wise, Alyssa Friend, Cui, Yi, Jin, WanQi, & Vytasek, Jovita. 2017. Mining for gold : Identifying content-related MOOC discussion threads across domains through linguistic modeling. *The Internet and Higher Education*, **32**(Jan.), 11–28.

Wolpert, David H. 1992. Stacked generalization. *Neural Networks*, **5**(2), 241–259.

Wong, Jian-Syuan, Pursel, Bart, Divinsky, Anna, & Jansen, Bernard J. 2015. An analysis of MOOC discussion forum interactions from the most active users. *Pages 452–457 of : International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer.

Yang, Diyi, Wen, Miaomiao, & Rose, Carolyn. 2014. Peer influence on attrition in massive open online courses. *Pages 405–406 of : Proceedings of the 7th International Conference on Educational Data Mining*, vol. 5.

You, Ji Won. 2015. Examining the effect of academic procrastination on achievement using LMS data in e-learning. *Journal of educational technology & society*, **18**(3), 64.

Zeng, Ziheng, Chaturvedi, Snigdha, & Bhat, Suma. 2017. Learner Affect Through the Looking Glass : Characterization and Detection of Confusion in Online Courses. *Pages 272–277 of : International Conference on Educational Data Mining*.

Zhang, Jianwei, Scardamalia, Marlene, Lamon, Mary, Messina, Richard, & Reeve, Richard. 2007. Socio-cognitive dynamics of knowledge building in the work of 9- and 10-year-olds. *Educational Technology Research and Development*, **55**(2), 117–145.

Zimmerman, Barry J. 2002. Becoming a Self-Regulated Learner : An Overview. *Theory Into Practice*, **41**(2), 64–70.

Zimmerman, Barry J., & Martinez-Pons, Manuel. 1988. Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, **80**(3), 284–290.

Annexe A

Centroïdes

Le Tableau des centroïdes des 44 caractéristiques associées à chaque cluster de chaque cours.

Caractéristiques	BCH				HBD				BCE			ANT		
	A	B	C	D	A	B	C	D	A	B	D	A	B	D
$prop\_Ree$	.80	.12	.10	.16	.03	.09	.50	.09	.13	.08	.19	.21	.08	.16
$prop\_App$	.15	.34	.83	.23	.94	.29	.42	.24	.75	.26	.46	.71	.27	.38
$prop\_Ver$	.02	.54	.05	.60	.03	.62	.04	.66	.09	.66	.32	.08	.65	.46
$prop\_Exe$	.06	.14	.09	.07	.02	.04	.00	.06	.04	.03	.00	.10	.10	.03
$prop\_Sch$	.04	.22	.16	.35	.07	.28	.26	.37	.04	.10	.10	.17	.36	.27
$prop\_Cor$	.02	.23	.23	.12	.02	.02	.00	.04	.08	.09	.04	.00	.01	.03
$prop\_Def$	.23	.09	.14	.13	.36	.17	.11	.15	.21	.11	.10	.12	.10	.11
$prop\_Man$	.16	.21	.28	.10	.06	.11	.10	.06	.24	.09	.04	.17	.20	.13
$prop\_Rai$	.04	.20	.22	.10	.07	.05	.12	.09	.11	.14	.23	.19	.06	.27
$prop\_Rol$	.02	.04	.04	.08	.10	.06	.09	.07	.10	.11	.18	.08	.10	.05
$prop\_Lie$	.15	.20	.16	.23	.16	.35	.14	.23	.12	.16	.12	.25	.25	.14
$prop\_Err$	.07	.06	.12	.19	.05	.06	.06	.14	.08	.03	.15	.12	.14	.08
$prop\_Con$	.01	.84	.03	.73	.05	.91	.02	.78	.04	.90	.36	.07	.80	.64
$prop\_Att$	.03	.10	.09	.08	.03	.03	.08	.08	.12	.06	.07	.08	.07	.05

Annexe B

Expressions régulières

Liste des expressions régulières à partir des mots-clés associés à chaque dimension.

Liste des expressions régulières	Dim1	Dim2	Dim3	Dim4
[Rr].[-_]*expliqu.*| refai.*| reven.*| redi.*| re-?pr.cis.*| [rR]appele.*| repren.*| re-?d.fini.*	1	0	0	0
expliqu.*| explic.*| pr.cis.*| d.velopp.*| d.taill.*| clarifi.*| [Qq]uoi| .u'entend.*| [Qq]u(') ?el(le) ?s ?| [Ll]es ?quels ?| r.sum.| [qQ]uand| [Cc]ombien| (D') ?[Oo]ù| ne_comprends ?_pas| donn.*| [Qq]ue_représent.*	2	0	0	0
d.fini.*| signifi*| .u'appell ?e.*| [Qq]ue_veut| vaut| .u'est*	2	0	1	0
..urquoi| [rR]aison	2	0	3	0
(l') ?erreur| comment_est-ce_possible| Quelle_possibilité| cependant| [Pp]ourtant| contradiction| alors_qu'ils ?| alors_que	3	0	0	1
[Cc]omment| cmt| d.marche| m.canisme	2	0	2	0
[Oo][uU]$| [pP]eu.*[-_].*| [Dd]oit[-_].*| dit_on| S'agit*| .st[-_]| N'est.*| c'est_bien	3	0	0	0
correspond(ent) ?-.*| concerne(nt) ?-.*| s ?ont-.*	3	0	5	0
concours| coeur| apprendre| retenir| toujours_donn*| ._savoir| [fF]aut[-_]il| conna.t-on| exam.*	3	0	0	3
.xemple	0	1	0	0
sch.ma| repr.sent.*| configur.*| [Ff]igure.*| image*| | l.gend.*	0	2	0	0
correct.*| corrig.*| r.ponse| raisonnement| (l') ?.xercice| .uestion	0	3	0	0
ce_qu'est| ce_que_sont	0	0	1	0
ser.*| [Rr].l.*| (l') ?utilit.*| fonction.*	0	0	4	0
diff.ren.*| lien.*| (l') ?.quivalen.*| m.me_chose| m.me_.*| relat.*| simil.*| limite_.*| .ntre_.*| synonym.*	0	0	5	0

Annexe C

WordNet

Ce Tableau représente les mots-clés représentatifs de chaque dimension utilisés en entrée de WordNet (Mots-entrés), la liste des mots trouvés par WordNet (Mots-sortis) et les mots que nous avons conservés (mis en gras). Chaque liste des mots conservés dans chaque dimension est remplacée par un seul mot-clé. Par exemple : les mots conservés de la dimension réexpliquer "Ree" : "revenir", "rappeler", "repréciser", "refaire", *etc.* sont remplacés par le mot "réexpliquer".

Catégorie	**Mots-entrés**	**Mots-sortis**
Ree	Réexpliquer	-
	Repréciser	-
	Refaire (14)	aboutir, agir, créer, escroquer, estamper, filouter, gérer, intriguer, magouiller, piquer, **refaire**, rouler, se faire pistonner, truander
	Revenir (28)	aller, amener, changer, commencer, communiquer, confronter, créer, devenir, donner, dénommer, engendrer, générer, livrer, offrir, passer, présenter, ramener, rapatrier, *rappeler*, rattraper, remettre, rendre, rentrer, **reprendre**, retourner, **revenir**, réintégrer, tomber
	Redire (6)	communiquer, **redire**, **reformuler**, réitérer, **répéter**, 'informer'
	Rappeler (26)	commémorer, définir, effectuer, envoyer, exclure, exposer, faire, mener, penser, ramener, **rappeler**, remarquer, remémorer, rendre, respecter, retirer, **revenir**, réaliser, réfléchir, **répéter**, se rappeler, se souvenir, signaler, sortir, souligner, souvenir

	Reprendre (80)	abattre, aborder, accuser, acquérir, admonester, adopter, agir, ajouter, apprivoiser, appuyer, assumer, blâmer, cacher, cambrioler, changer, choisir, compenser, comporter, comprendre, confirmer, confisquer, corriger, couvrir, critiquer, dater, demeurer, devenir, dissimuler, domestiquer, dompter, donner, détourner, encourager, engager, englober, engueuler, entourer, envelopper, exhorter, gronder, généraliser, inciter, inclure, inscrire, pinailler, prendre, ramener, rattraper, **recommencer**, recouvrer, **redire**, redémarrer, regagner, rembourser, remettre, **remonter**, rendre, renfermer, renforcer, renouer, rentrer, **reprendre**, reprocher, restituer, retourner, retrouver, **revenir**, récompenser, récupérer, régénérer, réprimander, réprouver, résumer, saisir, salarier, sermonner, supposer, trouver, usurper, étendre
	Redonner (9)	donner, recommencer, reconduire, **redonner**, rendre, restaurer, restituer, régénérer, **rétablir**
	Revoir (22)	accroître, analyser, assainir, bouger, changer, considérer, examiner, penser, rafraîchir, reconsidérer, remplacer, remédier, repenser, **revoir**, rédiger, réexaminer, réfléchir, réviser, **réécrire**, **réévaluer**, transformer, vérifier
App	Expliquer (26)	affirmer, **argumenter**, changer, **clarifier**, conjecturer, débattre, débrouiller, défendre, députer, **développer**, excuser, **expliquer**, **formuler**, illuminer, informer, **justifier**, plaider, **préciser**, présenter, prétendre, redéfinir, **résoudre**, **éclaircir**, **élaborer**, élucider, être
	Approfondir (4)	**approfondir**, **clarifier**, disserter, **détailler**
	Décrire (16)	apporter, causer, compter, correspondre, dessiner, distinguer, **décrire**, **identifier**, illustrer, inclure, provoquer, présenter, rapporter, reposer, représenter, sembler, **étudier**, éviter
	Préciser (62)	accommoder, advenir, affermir, affirmer, **analyser**, appeler, **appliquer**, arguer, attacher, attirer, attribuer, avoir en tête, baptiser, changer, charger, choisir, **citer**, **clarifier**, conseiller, **constituer**, contracter, disserter, décider, **déclarer**, définir, délimiter, déléguer, dénommer, **désigner**, **déterminer**, **expliquer**, figurer, fixer, **identifier**, **indiquer**, limiter, lister, **mentionner**, mettre, montrer, nommer, orienter, proclamer, **préciser**, **présenter**, qualifier, regrouper, remarquer, référer, "s identifier à", se trouver, servir, **situer**, **spécifier**, stipuler, succéder, viser, **éclaircir**, **éclairer**, élucider, énumérer, être

	Développer (53)	aboutir, accroître, acquérir, affecter, aggraver, agrandir, apparaître, **appuyer**, augmenter, avoir, baser, bâtir, casser, confirmer, construire, creuser, créer, cultiver, devenir, disserter, déceler, découvrir, dépouiller, **développer**, dévoiler, **expliquer**, fabriquer, fonder, **formuler**, **fournir**, gonfler, grandir, jeter, mettre, moderniser, mûrir, naître, obtenir, pousser, produire, publier, recevoir, rendre, reposer, rédiger, **révéler**, surgir, édifier, **élaborer**, **élargir**, **évaluer**, évoluer, être
	Détailler (6)	**appliquer**, **approfondir**, choisir, **clarifier**, disserter, **détailler**
	Clarifier (11)	changer, **clarifier**, débrouiller, **expliquer**, illuminer, **préciser**, redéfinir, **résoudre**, **éclaircir**, **éclairer**, élucider
	Résumer (10)	abstraire, affirmer, créer, jouer, manquer, recommencer, redire, reprendre, répéter, **résumer**
	Argumenter (13)	affirmer, arguer, **argumenter**, débattre, défendre, députer, **expliquer**, indiquer, **justifier**, **montrer**, plaider, **présenter**, prétendre
	Justifier (28)	absoudre, accomplir, accoter, affirmer, aider, appeler, approuver, **argumenter**, changer, confirmer, débattre, défendre, députer, excuser, **expliquer**, "faire l apologie de", faire valoir, fixer, formaliser, **justifier**, pardonner, plaider, protéger, présenter, prétendre, relever, tolérer, **valider**
	Résoudre (26)	**aborder**, adopter, adresser, appréhender, attaquer, **calculer**, cerner, **clarifier**, **conclure**, débrouiller, décider, **déterminer**, **expliquer**, finir, illuminer, juger, redéfinir, remédier, **régler**, répondre, **résoudre**, servir, solutionner, **traiter**, élucider, "être daccord"
Exe	Exemple (11)	cas, **exemple**, exercice, illustration, leçon, **modèle**, prise, prélèvement, spécimen, échantillon, échantillonnage
Sch	Schéma (11)	contenu, **dessin**, **diagramme**, esquisse, idée, **illustration**, plan, projet, **représentation**, **schéma**, ébauche
	Figure (17)	alentour, apparence, art, attribut, déplacement, effet, figue, figuier, **figure**, forme, hypothèse, **illustration**, manoeuvre, **représentation**, scène, scénario, simulation
Cor	Correction (10)	amendement, avenant, châtiment, correcteur, **correction**, peine, punition, rectification, sanction, supplément

Def	Définir (32)	caractériser, demeurer, distinguer, **définir**, délimiter, déterminer, effectuer, envoyer, exclure, exposer, fixer, limiter, mener, mentionner, modifier, paraître, privilégier, préciser, qualifier, rappeler, remarquer, reposer, ressembler, restreindre, réaliser, réduire, signaler, sortir, souligner, spécifier, stipuler, être
Man	Comment	-
	Manière (16)	air, chemin, **façon**, **manière**, mine, **modalité**, **mode**, mode de vie, **moyen**, ordre, piste, présence, sens, style, voie, vue
	Démarche (9)	allure, avancée, coup, **démarche**, pas, passage, progrès, à deux pas, étape
Rai	Raison (4)	avertissement, **cause**, **pourquoi**, **raison**
Rol	Rôle (23)	acte, action, activité, caractérisation, **contribution**, cotisation, devoir, droite, essai, **fonction**, intérimaire, lieu, mission, part, participation, partie, personnage, propriété, **rôle**, travail, **usage**, **utilisation**, **utilité**
Lie	Lien (16)	alliance, connexion, contact, hyperlien, hyperlier, hypertexte, **liaison**, **lien**, lien hypertexte, linkup, manille, nexus, obligataire, obligation, **rapport**, **relation**
Err	Erreur (4)	**erreur**, **faute**, **malentendu**, méfait
	Contradiction (11)	ampleur, conflit, **contradiction**, contraire, contraste, contrevérité, déclaration, démenti, négation, **opposition**, rapport

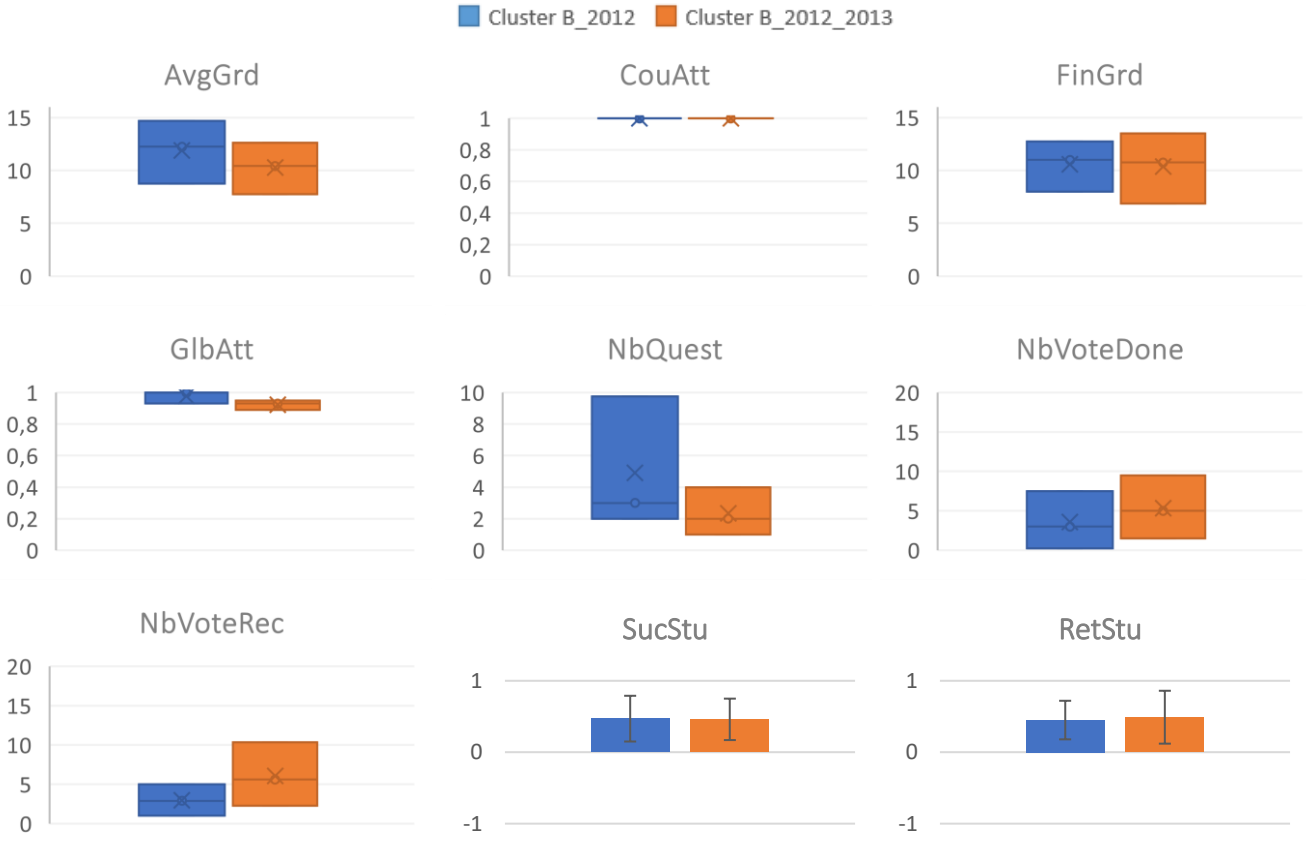
Annexe D

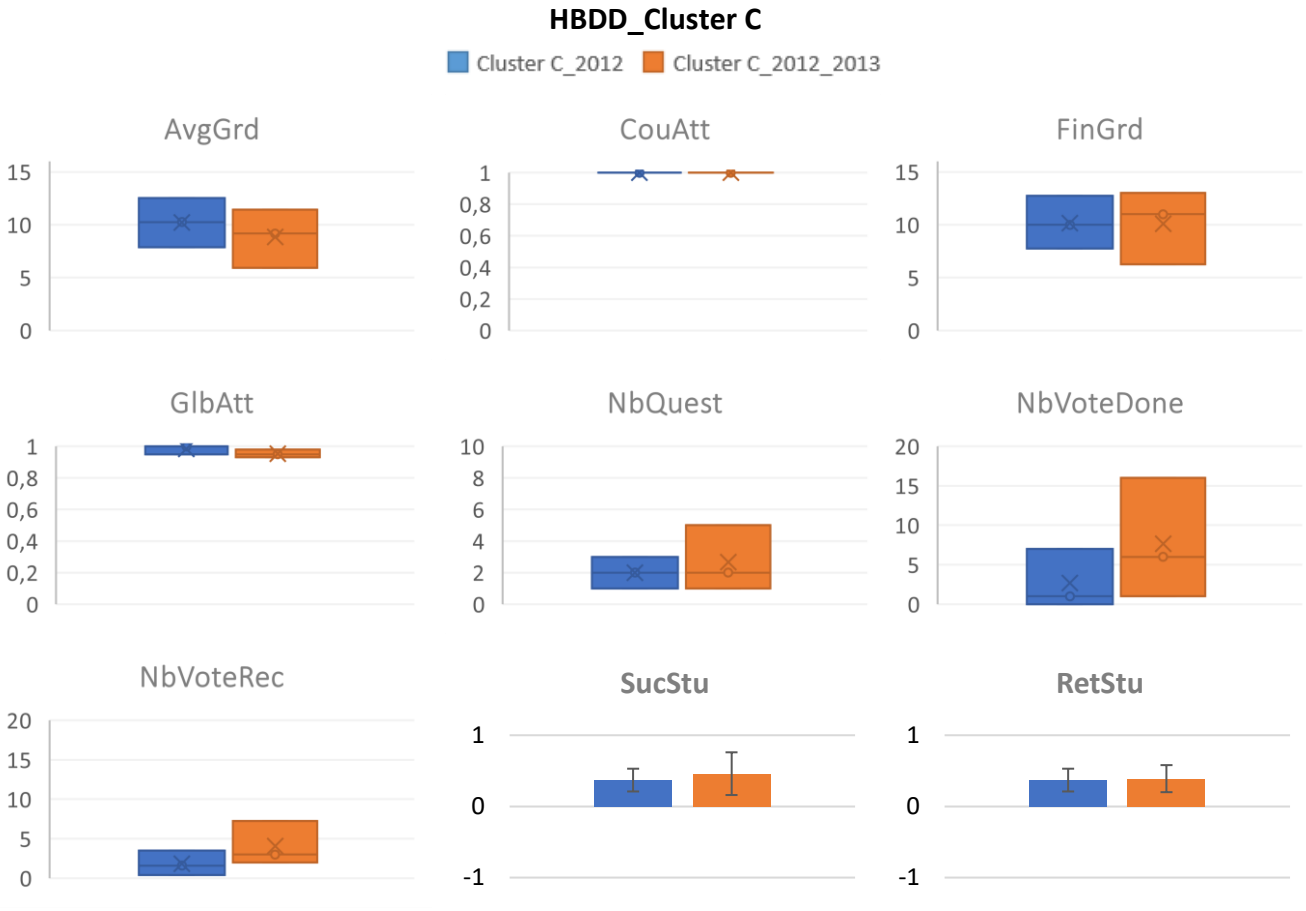
Comparaison des clusters 2012 et 2012 sur 2013

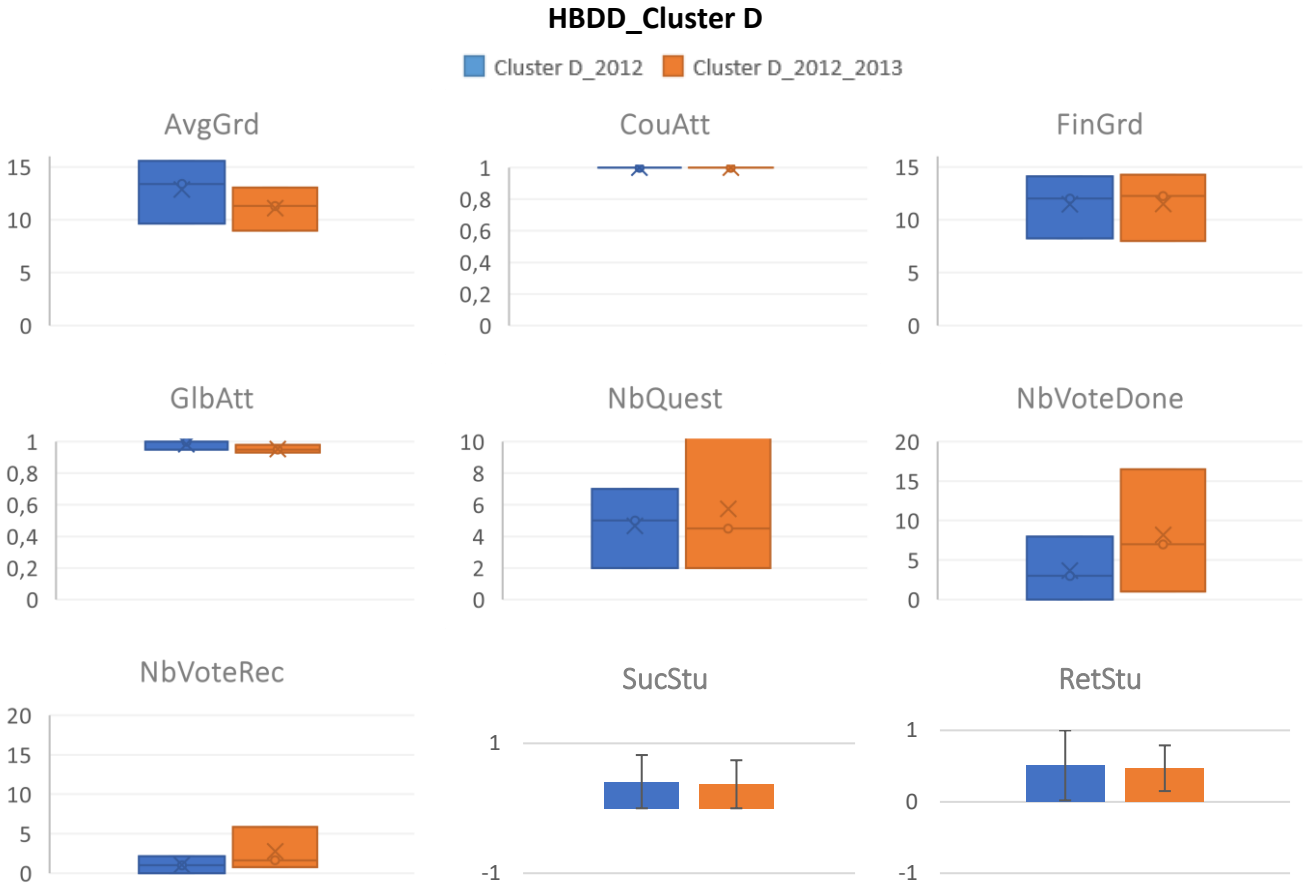
Comparaison des clusters de 2012 et 2013 en termes de médiane (centre de la barre), 1^{er} quartile (bas de la barre) et $3^{ème}$ quartile (haut de la barre) des variables dépendantes et moyenne et écart-type de EtuReu et EtuRed pour chaque cluster des 3 cours

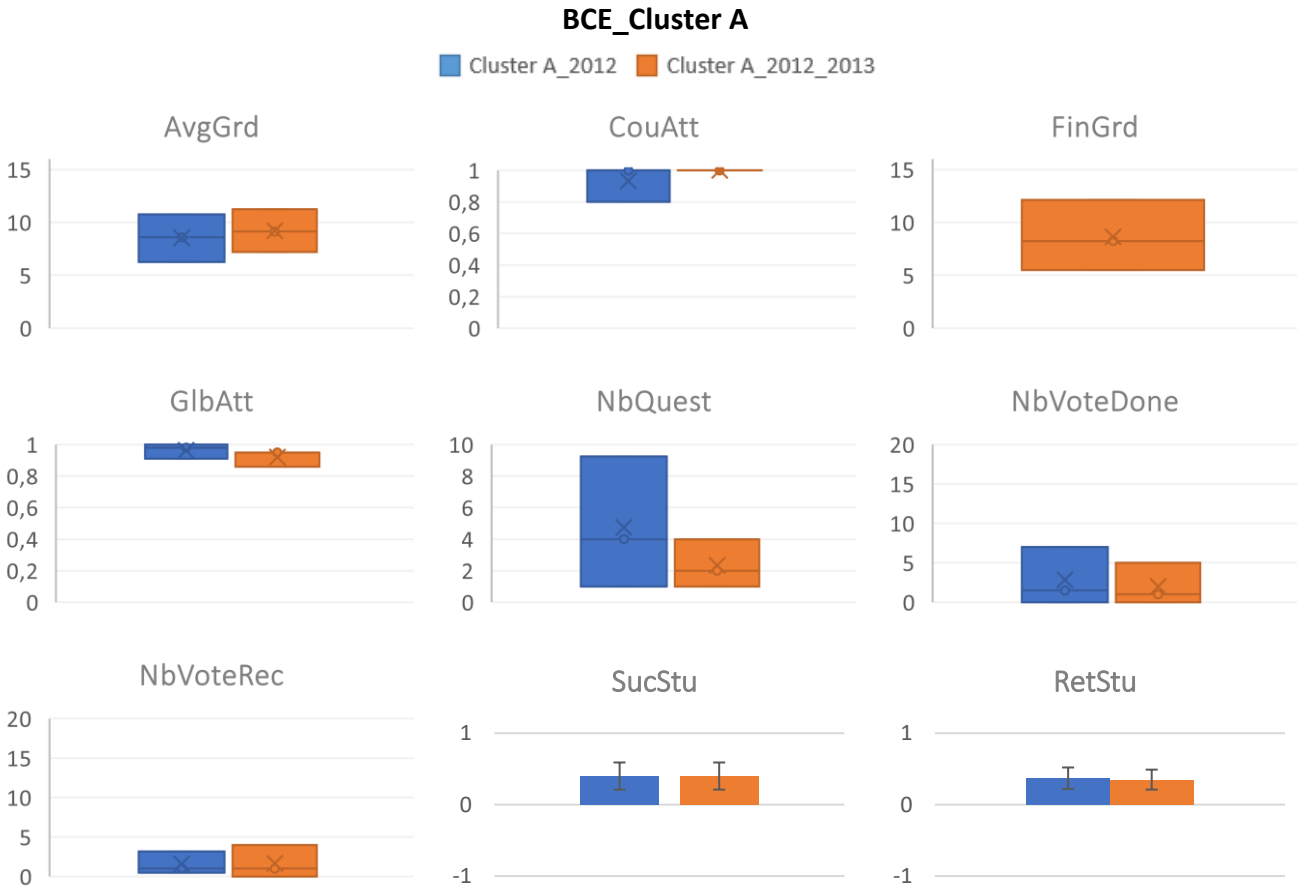
HBDD_Cluster A

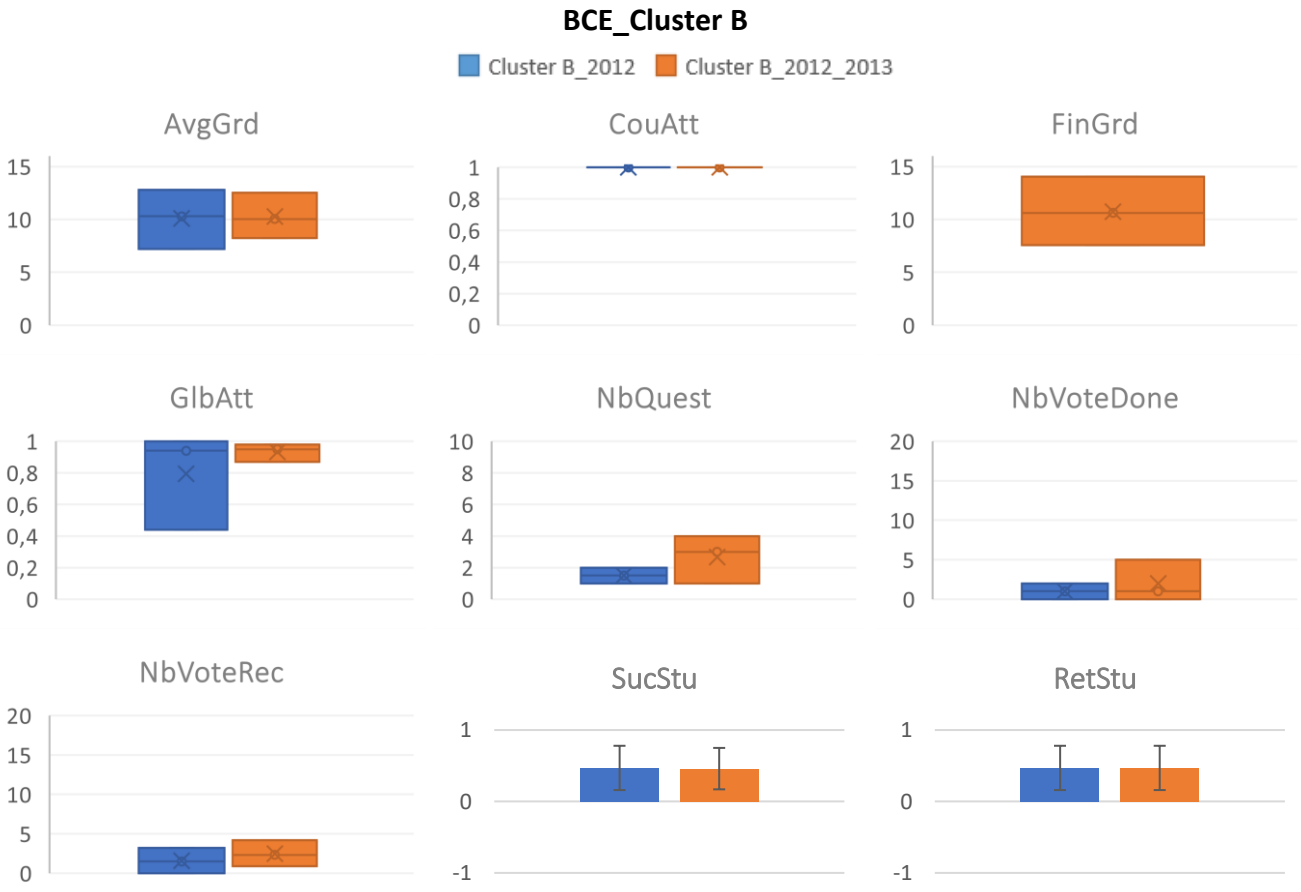
HBDD_Cluster B

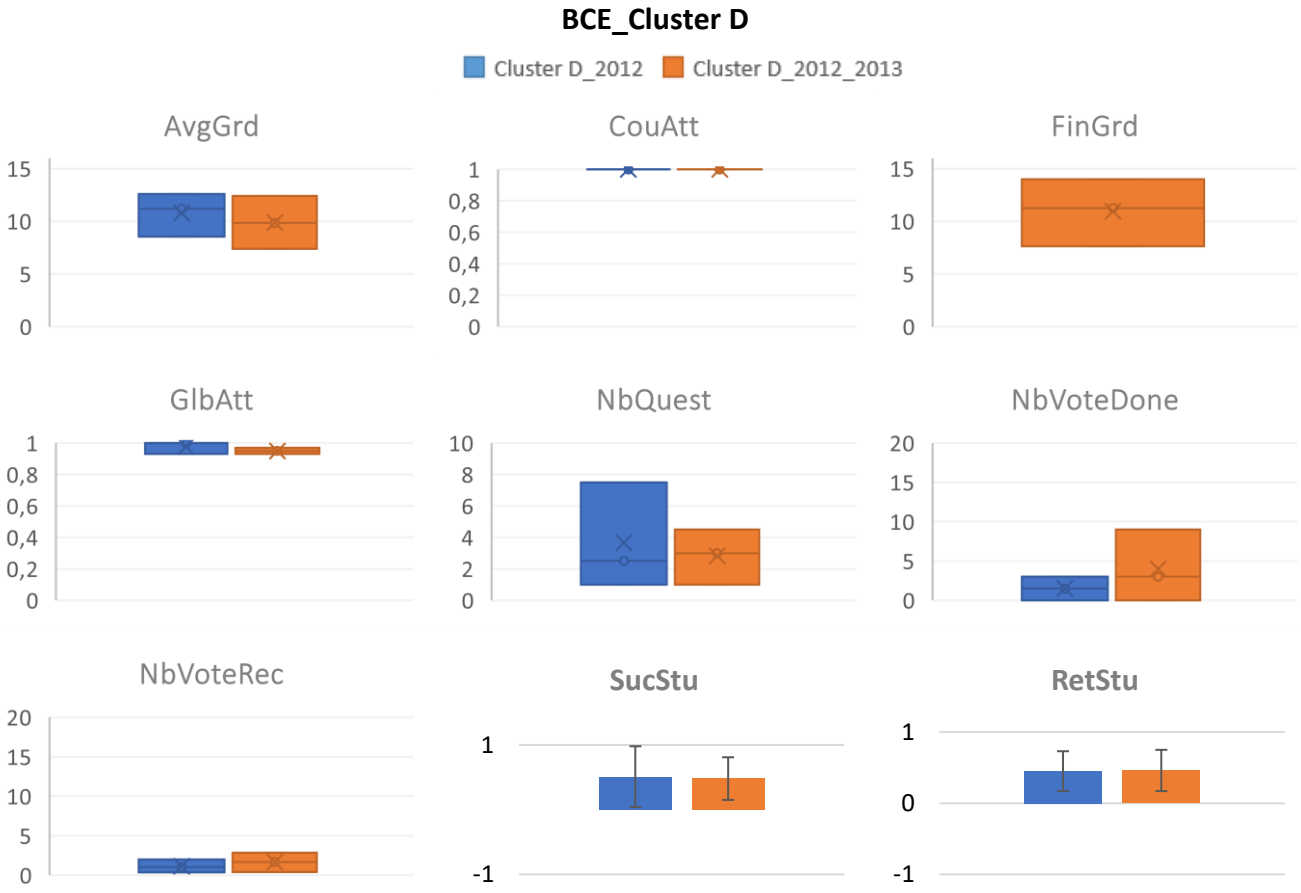


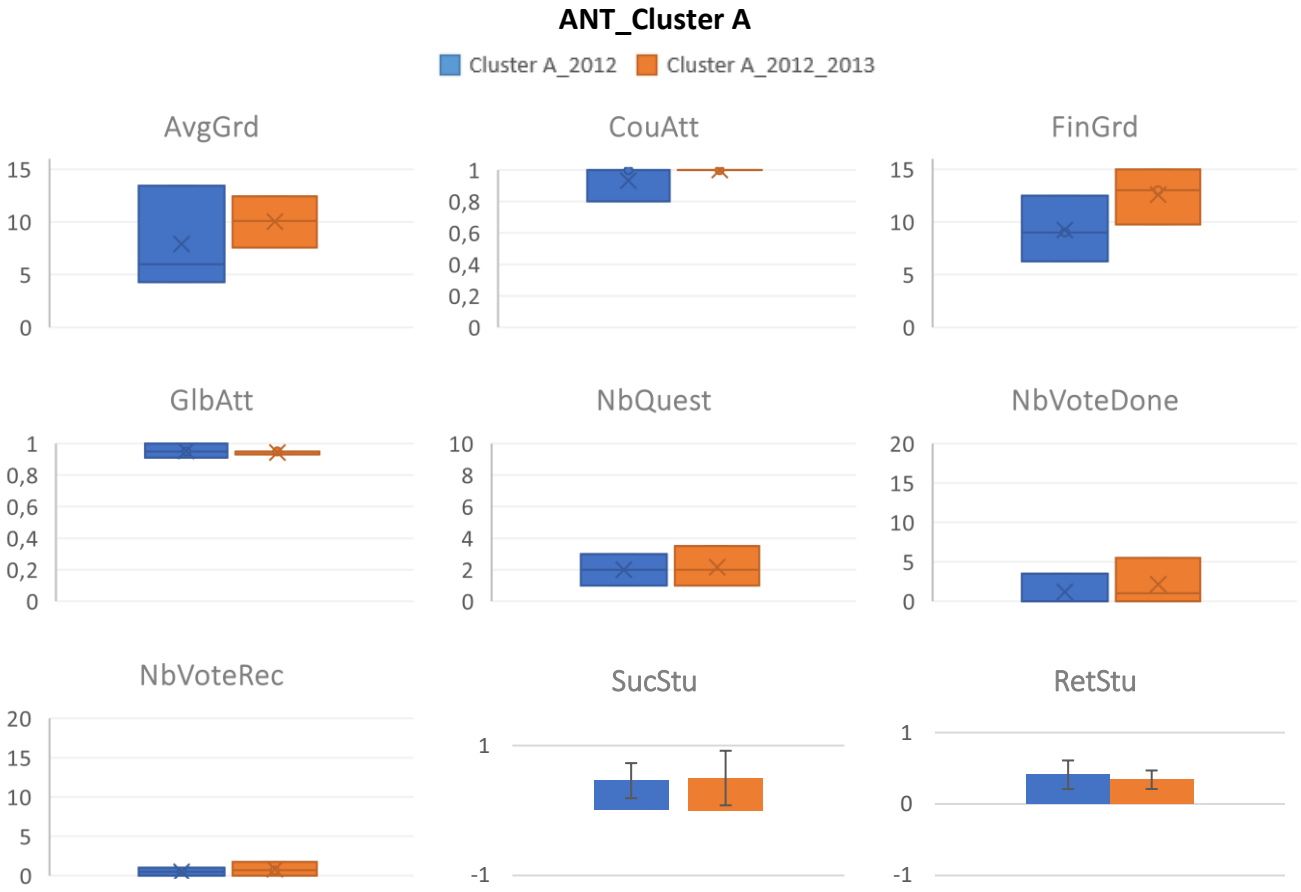

HBDD_Cluster C
Cluster C_2012
Cluster C_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu

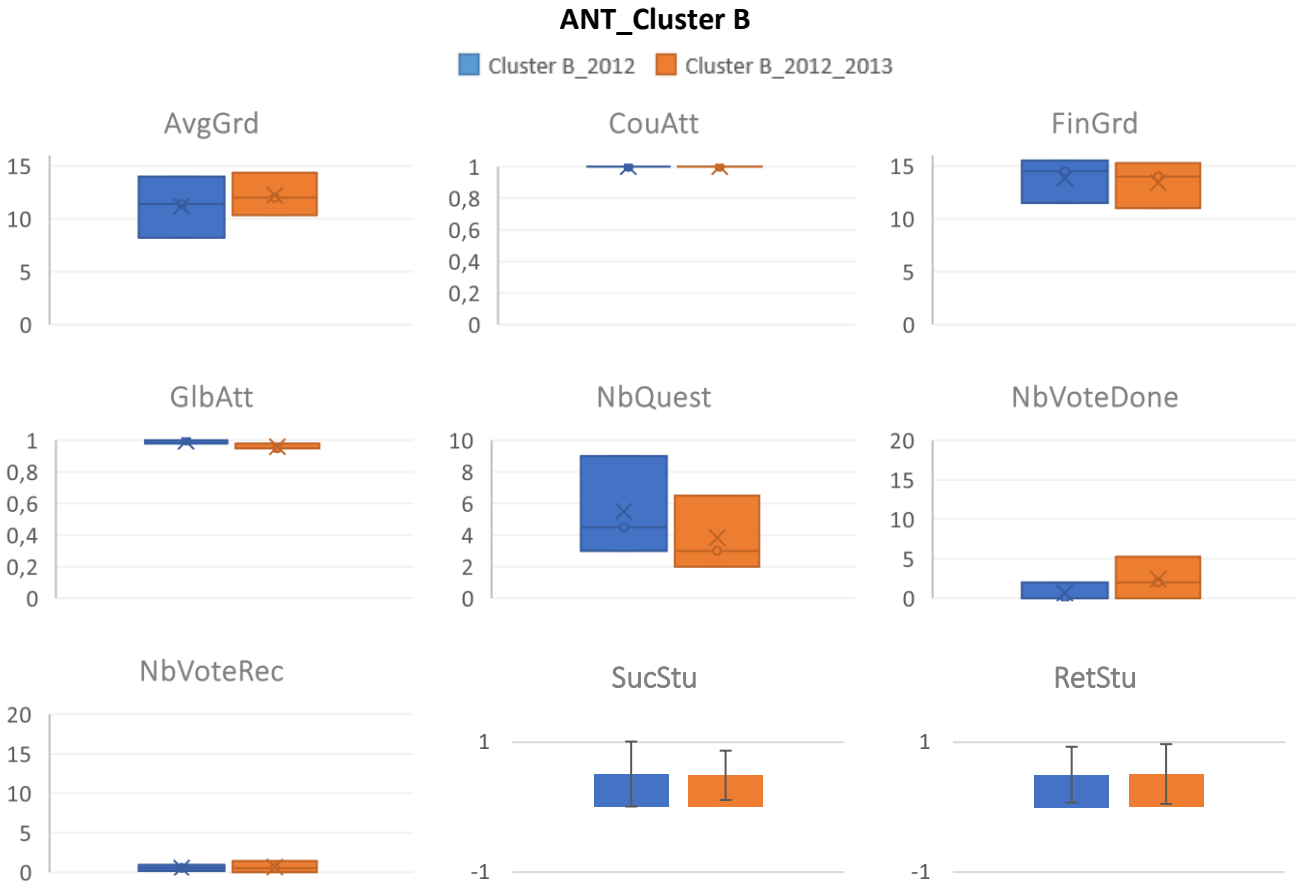

HBDD_Cluster D
Cluster D_2012
Cluster D_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu

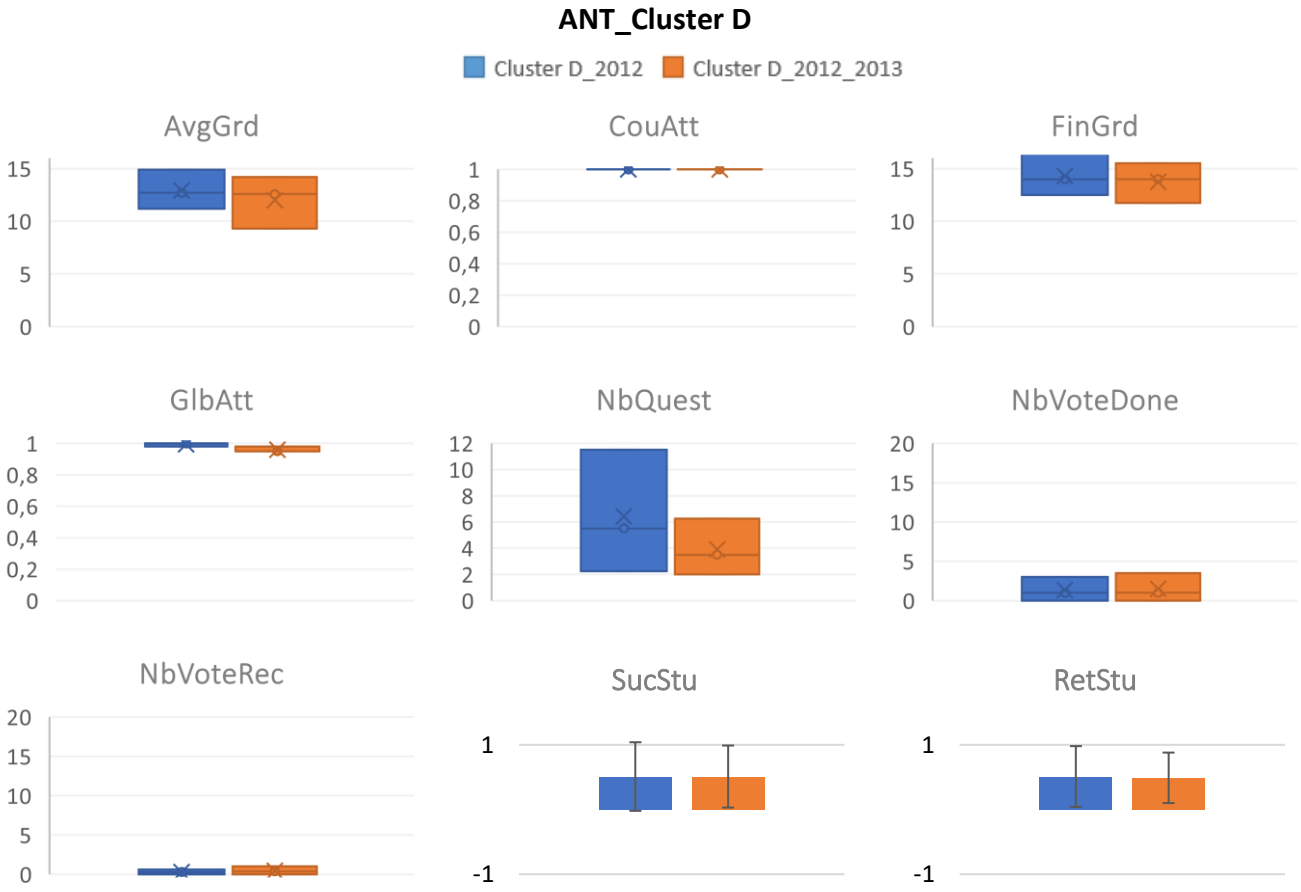

BCE_Cluster A
Cluster A_2012
Cluster A_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu


BCE_Cluster B
Cluster B_2012
Cluster B_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu


BCE_Cluster D
Cluster D_2012
Cluster D_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu


ANT_Cluster A
Cluster A_2012
Cluster A_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu


ANT_Cluster B
Cluster B_2012
Cluster B_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu


ANT_Cluster D
Cluster D_2012
Cluster D_2012_2013
AvgGrd
CouAtt
FinGrd
GlbAtt
NbQuest
NbVoteDone
NbVoteRec
SucStu
RetStu

Annexe E

Questionnaire SEPI

Le questionnaire présenté aux enseignants de PACES sur leurs séances SEPI.

PREMIERE PARTIE

1. Quel est votre âge ?

Votre réponse doit être comprise entre 18 et 90
Seul un nombre entier peut être inscrit dans ce champ.

*2. Quelle(s) matière(s) enseignez-vous ?

Cochez la ou les réponses

- [ ] ANT
- [ ] BCE
- [ ] BCH
- [ ] BPH
- [ ] BSTAT
- [ ] HBDD
- [ ] ICM
- [ ] MAIEU
- [ ] MAT
- [ ] ODON
- [ ] PHAR
- [ ] PHS
- [ ] SSH

*3. Vous intervenez au :

Cochez la ou les réponses

- [ ] 1er semestre
- [ ] 2nd semestre

*4. Depuis combien de temps enseignez-vous cette/ces matière(s) ?

*5. Nous allons vous proposer une série d'affirmations relatives à la pédagogie inversée mise en oeuvre sur Grenoble en PACES. Pour chacune de ces affirmations veuillez indiquer votre degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »)

	1	2	3	4	5
Il me semble facile d'enseigner en classe hybride	○	○	○	○	○
L'organisation de la classe hybride me convient telle qu'elle fonctionne actuellement	○	○	○	○	○
La classe hybride me permet de gagner du temps et de l'énergie	○	○	○	○	○

Ajout PG: Le terme « Classe hybride » correspond au terme national pour ce que nous appelons « la pédagogie inversée » sur Grenoble, c'est-à-dire une formation associant du contenu en e-learning et du présentiel.

6. Autres remarques sur la pédagogie inversée en PACES

*7. Nous allons vous proposer une série d'affirmations relatives aux mails reçus pour préparer les SEPI. Pour chacune de ces affirmations veuillez indiquer votre degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »)

	1	2	3	4	5
J'ai l'impression que les questions envoyées par mail sont trop nombreuses	○	○	○	○	○
J'ai l'impression que les questions envoyées par mail sont désorganisées	○	○	○	○	○
J'ai l'impression que les questions envoyées par mail ne sont pas utiles pour préparer ma SEPI	○	○	○	○	○

8. Autres difficultés rencontrées pour préparer les SEPI

*9. Combien de questions recevez-vous en moyenne par séance ?

Veuillez sélectionner une réponse ci-dessous

- Aucune
- Entre 1 et 20
- Entre 20 et 50
- Plus de 50

*10. Nous allons vous proposer une série d'affirmations relative à l'ensemble des questions reçues chaque semaine par mail pour préparer la SEPI. Pour chacune de ces affirmations veuillez indiquer votre degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »)

	1	2	3	4	5
J'ai l'impression de ne pas recevoir assez de questions pour préparer ma SEPI	○	○	○	○	○
J'ai l'impression de recevoir trop de questions pour préparer ma SEPI	○	○	○	○	○
J'ai l'impression que les questions posées sont généralement pertinentes	○	○	○	○	○
J'ai l'impression que les questions posées sont généralement déjà abordées en cours	○	○	○	○	○
J'ai l'impression qu'il y a beaucoup de nouvelles questions posées chaque année	○	○	○	○	○

*11. Nous allons vous proposer une série d'affirmations relatives à l'utilité de différents types de questions reçues par mail pour préparer la SEPI. Pour chacune de ces affirmations veuillez indiquer votre degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »)

	1	2	3	4	5
J'ai l'impression qu'une question cherchant à aller plus loin, m'est utile pour la préparation de la SEPI	○	○	○	○	○
J'ai l'impression qu'une question demandant un exemple, m'est utile pour la préparation de la SEPI	○	○	○	○	○
J'ai l'impression qu'une question de vérification d'erreur/contradiction en cours, m'est utile pour la préparation de la SEPI	○	○	○	○	○
J'ai l'impression que de savoir qu'une question est populaire (a reçu de nombreux votes des autres étudiants) m'est utile pour la préparation de la SEPI	○	○	○	○	○

12. Autres critères non évoqués qui rendent une question utile d'après vous

*13. Ignorez-vous les questions mal formulées reçues par mail pour préparer la SEPI ?

✔ Oui ⊘ Non

*14. Ignorez-vous les questions qui traitent de plusieurs sujets reçues par mail pour préparer la SEPI ?

Exemple : « Pourrait-on revenir sur la notion d'effet d'écran svp. Dans la formule Z*=Z-sigma connaît-on Z* et Z ? »

✔ Oui ⊘ Non

*15. En moyenne, quelle proportion de questions reçues par mail pensez-vous traiter effectivement lors de votre SEPI ?

Veuillez sélectionner une réponse ci-dessous

- ○ 0-20%
- ○ 20-40%
- ○ 40-60%
- ○ 60-80%
- ○ 80-100%

16. Comment utiliseriez-vous les questions posées dans chaque catégorie ?

*17. Nous allons vous proposer une série d'affirmations sur la proposition d'organisation "**Analyse Textuelle**" présentée ci-dessus. Pour chacune de ces affirmations veuillez indiquer votre degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »)

Cette question est obligatoire
Veuillez compléter toutes les parties.

	1	2	3	4	5
Cette visualisation me semble facile à comprendre	○	○	○	○	○
Cette visualisation m'apporte des informations supplémentaires qui me semblent utiles pour préparer ma SEPI	○	○	○	○	○

*18. Par rapport à ce que vous recevez actuellement, l'organisation de questions "**Analyse Textuelle**" vous semble

Veuillez sélectionner une réponse ci-dessous

Cette question est obligatoire

- Mieux
- Pareil
- Moins bien

19. Y a-t-il des points qui vous semblent à améliorer par rapport à l'organisation de questions "**Analyse Textuelle**" ?

20. Comment utiliseriez-vous les questions posées dans chaque catégorie ?

*21. Nous allons vous proposer une série d'affirmations sur l'organisation de questions "**Analyse Catégorielle**" reçues par mail pour préparer la SEPI. Pour chacune de ces affirmations veuillez indiquer votre degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »)

Cette question est obligatoire
Veuillez compléter toutes les parties.

	1	2	3	4	5
Cette visualisation me semble facile à comprendre	○	○	○	○	○
Cette visualisation m'apporte des informations supplémentaires qui me semblent utiles pour préparer ma SEPI	○	○	○	○	○

*22. Par rapport à ce que vous recevez actuellement, l'organisation de questions "**Analyse Catégorielle**" vous semble

Veuillez sélectionner une réponse ci-dessous

Cette question est obligatoire

- Mieux
- Pareil
- Moins bien

24. Comment utiliseriez-vous les questions posées dans chaque catégorie ?

✱25. Nous allons vous proposer une série d'affirmations sur l'organisation de questions "**Analyse Mixte**" reçues par mail pour préparer la SEPI. Pour chacune de ces affirmations veuillez indiquer votre degré d'accord sur une échelle de 1 à 5 (où 1 = « pas du tout d'accord » et 5 = « tout à fait d'accord »)

Cette question est obligatoire
Veuillez compléter toutes les parties.

	1	2	3	4	5
Cette visualisation me semble facile à comprendre	○	○	○	○	○
Cette visualisation m'apporte des informations supplémentaires qui me semblent utiles pour préparer ma SEPI	○	○	○	○	○

✱26. Par rapport à ce que vous recevez actuellement, l'organisation de questions "**Analyse Mixte**" vous semble

Veuillez sélectionner une réponse ci-dessous
Cette question est obligatoire

○ Mieux
○ Pareil
○ Moins bien

Organisation de questions "Analyse Mixte"

Dans cette organisation, on combine les deux informations déjà précédemment présentées : le type de questions posées par l'étudiant, et une estimation de son niveau par rapport à des données des années passées.

Pour illustrer cette organisation, voici 8 exemples de questions, extraites des années précédentes:

N°	Questions de ré-explication	Faible	Moyen	Bon
1.	Pourrait-on **ré-expliquer** comment trouver le moment dipolaire d'une molécule ?		X	
2.	Pourriez vous **revenir** sur la notion de solutions tampons notamment sur les moyens de réaliser une solution tampon ?	X		
	Questions d'approfondissement			
N°	**Questions d'approfondissement**	**Faible**	**Moyen**	**Bon**
3.	**Comment** comparer deux atomes qui ne se trouveraient ni dans la même ligne, ni dans la même colonne ?		X	
4.	Pourriez-vous **expliquer** ce qui distingue l'atome de l'élément chimique ?			X
N°	**Questions de vérification**	**Faible**	**Moyen**	**Bon**
	• **Erreur/ contradiction**			
5.	Il semble qu'il y ait une erreur dans le discours de la diapositive 5 : vous dites que `` les ions Na+ et Nacl (Cl- ?)		X	
6.	Bonjours , dans l'exemple sur l'électrophorèse vous dites que les aa sont chargé négativement a ph=1 , alors_que leurs ph < phi il ne devrait pas être positif comme présenté sur l'exemple de séparation de mélange ?			X
	• **Connaissances en cours**			
7.	est-ce que tous les métaux de transition sont réducteurs ?	X		
	• **Examen**			
8.	Doit on apprendre les métaux du bloc P par coeur ?		X	

Organisation de questions "Analyse Catégorielle"

Dans cette organisation, les questions des étudiants sont regroupées en fonction du profil des étudiants des années passées qui posaient ce type de questions.

La mention "notes inférieures à la moyenne" n'est donc pas basée sur les notes des étudiants qui posent les questions, et serait donc disponible dès la première séance.

Pour illustrer cette organisation, voici 8 exemples de questions, extraites des années précédentes:

Étudiants en difficulté : notes < à la moyenne

Questions :

2. Pourriez vous **revenir** sur la notion de solutions tampons notamment sur les moyens de réaliser une solution tampon ?
7. est-ce que tous les métaux de transition sont réducteurs ?

Étudiants moyens

Questions :

1. Pourrait-on **ré-expliquer** comment trouver le moment dipolaire d'une molécule ?
3. **Comment** comparer deux atomes qui ne se trouveraient ni dans la même ligne, ni dans la même colonne ?
5. Il semble qu'il y ait une **erreur** dans le discours de la diapositive 5 : vous dites que `` les ions Na+ et Nacl (Cl- ?)
8. Doit on apprendre les métaux du bloc P par coeur ?

Étudiants bons : notes > à la moyenne

Questions :

4. Pourriez-vous **expliquer** ce qui distingue l'atome de l'élément chimique ?
6. Bonjours , dans l'exemple sur l'électrophorèse vous dites que les aa sont chargé négativement a ph=1 , alors_que leurs ph < phi il ne devrait pas être positif comme présenté sur l'exemple de séparation de mélange ?

DEUXIEME PARTIE

Nous allons maintenant vous présenter 3 organisations alternatives à l'e-mail que vous avez l'habitude de recevoir.
Pour chacune de ces organisations, nous vous demanderons de les évaluer selon différents critères (lisibilité, intérêt perçu pour l'organisation de la SEPI, etc.).

Organisation de questions "Analyse Textuelle"

Dans cette organisation, les questions des étudiants subissent une analyse textuelle. Les questions sont alors regroupées en fonction de la **nature des questions posées (re-Explication, Approfondissement, Vérification ou Autres)**, tel qu'indiqué par les mots-clés mis en gras ci-dessous.

Pour illustrer cette organisation, voici 8 exemples de questions, extraites des années précédentes:

27. Y a-t-il des points qui vous semblent à améliorer par rapport à l'organisation de questions "**Analyse Mixte**" ?

*28. Pour finir, si vous ne deviez choisir qu'une seule organisation parmi les 3 proposées et l'organisation actuelle, vous choisiriez :

Veuillez sélectionner une réponse ci-dessous

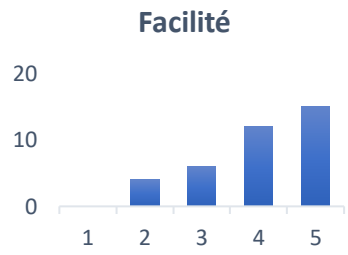
Cette question est obligatoire

- ○ L'organisation actuelle
- ○ L'organisation "Analyse Textuelle"
- ○ L'organisation "Analyse Catégorielle"
- ○ L'organisation "Analyse Mixte"

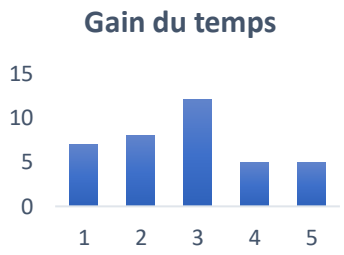
Annexe F

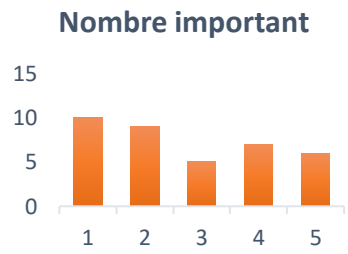
Distribution des variables

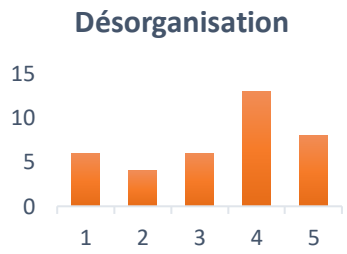
Histogrammes de distribution sur une échelle de 1 à 5 des variables liées à la classe inversée (couleur bleu), l'expérience des enseignants en SEPI et leur impression sur les questions envoyées par courriel (couleur orange) et la pertinence perçue des catégories de questions auprès des enseignants (couleur verte).


Facilité
20
10
0
1 2 3 4 5

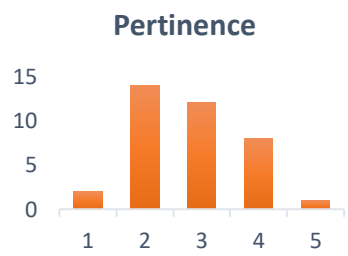

Organisation
15
10
5
0
1 2 3 4 5

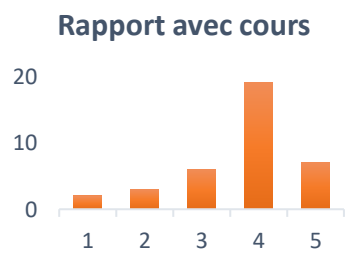

Gain du temps
15
10
5
0
1 2 3 4 5


Nombre important
15
10
5
0
1 2 3 4 5


Désorganisation
15
10
5
0
1 2 3 4 5


Manque d'utilité
15
10
5
0
1 2 3 4 5


Pertinence
15
10
5
0
1 2 3 4 5

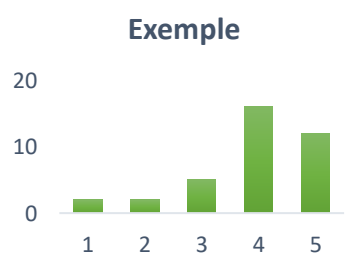

Rapport avec cours
20
10
0
1 2 3 4 5

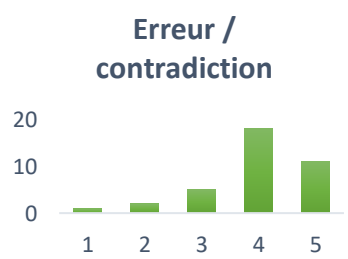

Nouveauté
20
10
0
1 2 3 4 5

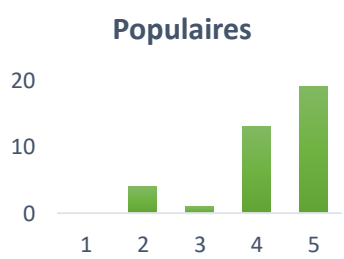

Trop de questions
15
10
5
0
1 2 3 4 5


Peu de questions
20
10
0
1 2 3 4 5


Aller plus loin
15
10
5
0
1 2 3 4 5


Exemple
20
10
0
1 2 3 4 5


Erreur / contradiction
20
10
0
1 2 3 4 5


Populaires
20
10
0
1 2 3 4 5

Annexe G

Questionnaire d'autorégulation

Questionnaire utilisé pour mesurer l'autorégulation des étudiants du MOOC GDP.

Description de l'échelle

Cette échelle mesure 4 différents construits, soit la procrastination, le contrôle du contexte d'apprentissage, les stratégies d'apprentissage et le soutien des pairs. Elle contient 21 énoncés (5 ou 6 énoncés par sous-échelle) et est mesurée sur une échelle de 1 à 7 points.

Référence

Cosnefroy, L., Fenouillet & Heutte, J. (2018). *Développement et validation d'une échelle d'apprentissage autorégulé en ligne*. 2e Colloque international e-Formation des Adultes et Jeunes Adultes, Lille, France.

Vos méthodes de travail en ligne

Ce questionnaire a pour but de mieux connaître vos méthodes de travail en ligne, celles que vous utilisez habituellement pour étudier un cours, réaliser les travaux demandés ou vous préparer à un examen. Il n'y a pas de bonne réponse, chacun a sa façon de faire personnelle et plusieurs méthodes peuvent mener au succès.

Lisez attentivement chaque phrase et répondez, sur l'échelle située en face, en entourant un nombre de 1 = pas du tout à 7 = tout à fait.

pas du tout	très peu	un peu	moyennement	assez	fortement	tout à fait
1	2	3	4	5	6	7

		1= pas du tout						7= tout à fait
1	Quand j'étudie un cours en ligne, je commence par prendre des notes à partir des différents documents fournis.	1	2	3	4	5	6	7
2	Quand je ne vois vraiment pas comment m'y prendre dans mes cours en ligne, je demande conseil à d'autres étudiants.	1	2	3	4	5	6	7
3	Je choisis un moment où je pense ne pas être trop distrait pour étudier sur mes cours en ligne.	1	2	3	4	5	6	7
4	Sur mes activités en ligne me mettre au travail me demande en général beaucoup d'efforts.	1	2	3	4	5	6	7
5	Quand j'ai un cours en ligne à étudier, j'imagine à quel moment je devrais le faire pour être le plus efficace possible	1	2	3	4	5	6	7
6	Je cherche à m'approprier le contenu des cours en ligne en prenant des notes.	1	2	3	4	5	6	7
7	Je discute avec d'autres étudiants de certains points de cours qui ne sont pas clairs.	1	2	3	4	5	6	7
8	Je me mets dans un endroit confortable pour étudier en ligne.	1	2	3	4	5	6	7
9	Je n'arrive pas à me motiver pour travailler sur les cours en ligne.	1	2	3	4	5	6	7
10	J'échange avec les autres étudiants pour savoir si nous avons compris la même chose.	1	2	3	4	5	6	7
11	Je fais des brefs résumés ou des schémas pour organiser les connaissances lorsque j'étudie un cours en ligne.	1	2	3	4	5	6	7
12	J'ai du mal à rester concentré et à aller jusqu'au bout lorsque j'étudie un cours en ligne.	1	2	3	4	5	6	7

		1= pas du tout						7= tout à fait
13	Pour étudier en ligne, je choisis un endroit qui me protège des distractions	1	2	3	4	5	6	7
14	Je vais sur les réseaux sociaux pour partager mes problèmes avec les autres étudiants.	1	2	3	4	5	6	7
15	Je fais un résumé de ce que j'ai appris dans les cours en ligne afin de vérifier ma compréhension des cours.	1	2	3	4	5	6	7
16	Sur mes activités en ligne, souvent j'éprouve un tel sentiment d'ennui en pensant au travail à faire que je n'arrive pas à m'y mettre.	1	2	3	4	5	6	7
17	J'arrange l'endroit où je vais étudier en ligne pour qu'il soit agréable.	1	2	3	4	5	6	7
18	Je copie les parties des documents que je trouve intéressantes pour les intégrer dans mes notes.	1	2	3	4	5	6	7
19	Quand j'étudie un cours en ligne, je relis mes notes encore et encore pour m'aider à me souvenir du contenu.	1	2	3	4	5	6	7
20	Sur mes activités en ligne je recule sans cesse le moment d'étudier et fais tout au dernier moment quelle que soit la discipline.	1	2	3	4	5	6	7
21	J'échange avec les autres étudiants pour savoir comment s'y prendre dans les cours en ligne.	1	2	3	4	5	6	7

Nous vous remercions pour votre participation !

CLÉ DE CODIFICATION

EAREL

# 4, 9, 12, 16, 20	**Procrastination (PROC) ➔ 5 items reverses (PROC(r))**
# 3, 5, 8, 13, 17	**Contrôle du contexte d'apprentissage (CTXTE)**
# 1, 6, 11, 15, 18, 19	**Stratégies cognitives/métacognitives (COGN)**
# 2, 7, 10, 14, 21	**Soutien des pairs (PAIRS)**

FORMULE POUR CALCULER
l'indice d'apprentissage autorégule en ligne
(IAREL)

L'indice d'apprentissage autorégule en ligne (IAREL) peut être calculé après avoir inversé au préalable le score de procrastination (PROC(r)).

IAREL = Moyenne (PROC(r) , CTXTE , COGN , PAIRS)

Merci d'utiliser cette référence pour citer ***l'Échelle d'apprentissage autorégulé en ligne (EAREL)***

Cosnefroy, L., Fenouillet & Heutte, J. (2018). *Développement et validation d'une échelle d'apprentissage autorégulé en ligne.* 2e Colloque international e-Formation des Adultes et Jeunes Adultes, Lille, France