Liens entre performance, assiduité et questions posées et/ou questions votées en ligne dans le cadre d'une classe inversée

▶ **Fatima HARRAK, François BOUCHET, Vanda LUENGO**
(LIP6, Sorbonne Université)

■ **RÉSUMÉ** • Les questions des élèves sont utiles pour leur apprentissage et pour l'adaptation pédagogique des enseignants. Nous étudions ici la nature des questions posées en ligne par les étudiants et comment le vote sur ces questions peut être lié à l'apprentissage. Nous avons donc développé un schéma de codage, puis conçu un annotateur automatique que nous avons appliqué à l'ensemble du corpus. Le résultat révèle que les votants réussissent mieux et assistent plus souvent au cours, mais le fait de poser des questions est associé à un apprentissage plus important.

■ **MOTS-CLÉS** • Question d'élève, vote d'élève, classe inversée.

■ ***ABSTRACT*** • *Students' questions are useful for their learning experience as well as to help teachers to adapt their pedagogy. We study here a corpus of questions asked online by students and how voting on these questions can be associated to learning. We have therefore developed a coding scheme of questions and built an automatic annotator to tag the whole corpus. The result reveals the voters perform better and attend class more often, and asking more questions is associated with better learning.*

■ ***KEYWORDS*** • *Student's question, student's vote, blended learning.*

1. Introduction

Les questions des élèves jouent un rôle important dans le processus d'apprentissage, non seulement pour aider les élèves à mieux apprendre (Sullins *et al.,* 2015), mais aussi pour aider l'enseignant à déterminer ce qui a été compris (ou non) et à adapter sa pédagogie en conséquence. Les environnements en ligne et autres environnements informatiques pour l'apprentissage humain (EIAH) peuvent éliminer de nombreux obstacles qui empêchent les élèves de poser des questions en classe (Otero et Graesser, 2001). Nous nous intéressons ici à une formation hybride dans laquelle les étudiants doivent poser chaque semaine des questions à partir de supports de cours étudiés à distance avant le cours, pour aider les enseignants à préparer leurs séances de questions-réponses en présentiel. 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. Pour les aider dans ce choix et limiter le nombre de questions, ils encouragent les étudiants à voter sur les questions déjà posées avant d'en poser de nouvelles. 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 compréhension des points abordés par leurs camarades. Mais on peut aussi penser qu'un vote n'est pas exactement équivalent à une question. En effet, dans le cadre théorique Interactive-Constructive-Active-Passive (ICAP) proposé par Chi et Wylie (2014), 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), prémisse d'une activité « interactive » lorsque cette question recevra une réponse, potentiellement dans le cadre d'un échange dialogique. Ces types d'activités sont 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), qui elle-même est préférable à un engagement « passif » où l'on se contente de lire les questions des autres. D'après ce cadre théorique, il pourrait donc être contre-productif d'encourager un étudiant à voter plutôt qu'à poser une question. Face à cette contradiction sur la valeur d'un vote, nous avons conduit des analyses afin d'explorer comment les votes sont associés à la performance, à l'engagement des étudiants et aux questions qu'ils posent.

Plus précisément, notre objectif était de répondre aux trois questions de recherche suivantes :

(QR1) Quel est le lien entre le vote et la performance d'un étudiant, notamment en comparant les performances des étudiants votants qui posent des questions et de ceux qui n'en posent pas ?

(QR2) Le vote est-il lié à l'engagement de l'apprenant en classe et globalement ?

(QR3) Les étudiants votent-ils sur des questions dont la nature est différente de celle des questions qu'ils posent ?

Pour traiter ces questions de recherche, nous avons défini un schéma de codage adapté aux questions des étudiants et un système d'annotation automatique pour annoter l'ensemble du corpus de questions dont nous disposions. Dans la suite de cet article, nous proposons dans la section 2 un état de l'art introduisant plus en détail le cadre ICAP sur lequel s'appuie ce travail, comparant les différentes taxonomies de questions d'élèves existantes, et s'intéressant à l'utilité possible des questions ou des votes sur des questions dans un contexte éducatif. Nous présentons dans la section 3 le contexte de l'étude et les données utilisées, avant de décrire dans la section 4 la méthodologie de catégorisation de questions et d'annotation automatique. Enfin, nous présentons dans la section 5 les résultats des analyses effectuées pour répondre aux trois questions de recherche ci-dessus, et nous concluons avec quelques perspectives et limites de ce travail en section 6.

2. État de l'art

2.1. Cadre théorique : ICAP

Un aspect essentiel de notre travail est relatif à la distinction entre la valeur d'apprentissage intrinsèque associée au fait de poser une question, par opposition au fait de simplement voter sur une question déjà posée par un tiers. Dans ce cadre, les travaux de Chi et Wylie (2014) fournissent un cadre théorique particulièrement pertinent. En effet, dans sa théorie ICAP, Chi différencie 4 types d'activités : « interactive » (I), « constructive » (C), « active » (A) et « passive » (P). Dans les *activités passives*, l'apprenant se contente de recevoir le savoir sans comportement visible attestant d'un travail d'intégration des nouvelles connaissances, ce qui est souvent lié à un apprentissage en surface. Les *activités actives* sont celles qui attirent l'attention de l'apprenant (impliquant souvent un mouvement physique),

telles que regarder ou manipuler certains aspects du matériel d'apprentissage, répéter, voter, etc. Les *activités constructives* sont celles qui demandent aux apprenants d'aller au-delà de ce qui était explicitement présenté dans les supports d'apprentissage, qui peuvent contenir de nouvelles idées, telles que s'auto-expliquer, induire de nouvelles hypothèses, poser des questions, réfléchir, etc. Enfin les *activités interactives* se focalisent sur le dialogue en deux types, soit avec des experts (dialogues d'instruction), soit avec des pairs (dialogues conjoints). Les 4 types d'activités ne sont pas mutuellement exclusifs, mais au contraire hautement inclusifs : ainsi être interactif subsume être constructif (e.g. pour poser une question susceptible d'entraîner un dialogue d'instruction, il faut déjà avoir fait un travail de synthèse), et être constructif subsume également être actif (e.g. pour faire un schéma de synthèse d'un cours, il faut déjà avoir retranscrit en partie celui-ci) qui subsume le fait d'être passif (e.g. pour retranscrire il faut écouter). Dans le cadre de cet article, ni les activités passives (pour lesquelles nous n'avons pas de traces) ni les activités interactives ne sont considérées, vu le contexte particulier de cette formation hybride qui ne favorise pas la collaboration et les échanges (les étudiants ne peuvent pas répondre aux questions posées par les autres étudiants, et les enseignants ne peuvent pas répondre à toutes les questions pendant les séances dédiées aux questions-réponses).

2.2. Typologies de questions

Les chercheurs ont étudié le comportement de questionnement des élèves dans divers contextes éducatifs, tels que la classe (Chin et Kayalvizhi, 2002), le tutorat (Graesser et Person, 1994) et les environnements d'apprentissage en ligne (Li *et al.,* 2014). En particulier, plusieurs taxonomies ou schémas de codage présentant différents degrés de granularité ont ainsi été proposés. Scardamalia et Bereiter (1992) se concentrent sur la distinction entre les questions fondées sur le texte et celles fondées sur le savoir (ces dernières ayant un potentiel éducatif plus fort). Bien que cette distinction soit pertinente dans notre contexte, n'ayant pas accès aux transcriptions des vidéos et diapositives avec lesquelles les étudiants ont interagi avant de poser leurs questions, il était difficile d'identifier ceci automatiquement.

D'autres chercheurs ont proposé une typologie de questions distinguant celles pouvant faire l'objet d'une investigation scientifique (par ex. : comparaison, cause à effet, prédiction, exploration) des autres (Chin et Kayalvizhi, 2002). Bien que notre but ne soit pas d'encourager à

poser un certain type de questions, cette distinction pourrait s'appliquer à notre travail, mais est difficile à réaliser sans experts du domaine. Graesser et Person (1994) ont pour leur part élaboré une taxonomie de questions posées pendant les séances de tutorat, utilisée pour la génération automatique de questions. Bien que leur taxonomie puisse être pertinente ici, certaines catégories comprenaient des « questions de raisonnement approfondi » de haute qualité, associées à des modèles de raisonnement difficiles à identifier automatiquement. Enfin, des recherches récentes (Supraja *et al.,* 2017) ont utilisé une version réduite de la taxonomie de Bloom (Bloom *et al.,* 1956) pour établir un lien entre rétroaction pratique et performance de l'apprenant en matière d'évaluation. Cette taxonomie, en raison de son origine, tend toutefois à être plus appropriée aux questions de l'enseignant qu'à celles des élèves.

2.3. Utilité pédagogique des questions d'élèves

L'analyse des questions d'apprentissage a été utilisée à des fins très diverses afin d'améliorer l'efficacité de l'enseignement et l'apprentissage des élèves. Ainsi, Harper e*t 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 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 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. Graesser et 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 autorégulé 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.

Chin et 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 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 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 montrent que les questions posées par les élèves peuvent également refléter leur engagement, cependant poser des questions « superficielles » est généralement peu utile.

Teixeira-Dias *et al.* (2005) ont exploré les questions formulées par les élèves au cours de l'élaboration des projets de groupe pour analyser leurs comportements au lieu de leur compréhension. Les auteurs ont trouvé que les questions avaient contribué à l'engagement des étudiants en chimie, permettant d'accroître l'interaction entre l'enseignant et les étudiants et leur confiance en eux-mêmes pour formuler des questions. Par conséquent, la qualité de l'interaction en classe pendant l'apprentissage et l'enseignement de la chimie a été améliorée.

2.4. Utilité pédagogique des votes d'élèves

Si l'on s'intéresse maintenant aux travaux centrés sur la valeur potentielle du vote sur une question, on peut voir que les votes des élèves ont notamment été étudiés pour analyser le comportement des élèves dans les forums en ligne. Bihani *et al.* (2018) ont utilisé le nombre de votes sur les questions et réponses des étudiants et sur les réponses de l'enseignant pour révéler les paires de questions/réponses pertinentes pour les futurs cours. Zeng *et al.* (2017) ont également utilisé le nombre de votes pour détecter les messages exprimant un sentiment comme la confusion dans les forums de discussions. Ils ont constaté que les messages exprimant la confusion reçoivent un nombre important de votes. Jiang *et al.* (2015) ont analysé les étudiants considérés comme des « 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 et négatifs) sur les messages et les commentaires des utilisateurs actifs. Contrairement à Jiang *et al.* (2015), ils ont constaté que les utilisateurs actifs sont aussi des utilisateurs influents qui apportent généralement une contribution positive aux discussions du forum du MOOC. Les votes des élèves ont donc

surtout été utilisés pour analyser le comportement des élèves, mais la nature des questions votées n'a apparemment pas encore été explorée.

Dans l'ensemble, les typologies de questions proposées jusqu'à présent dépendent essentiellement du contexte, et nous avons décidé de définir un nouveau schéma de codage utilisant une approche fondée sur les données. Dans cet article, nous nous intéressons à l'analyse de la nature des questions posées et des questions votées par les étudiants.

3. Contexte et données de l'étude

Nous avons considéré l'ensemble des questions posées par des étudiants de 1re année de médecine et pharmacie d'une université française en 2012-2013. 1608 étudiants étaient inscrits cette année-là, une partie d'entre eux seulement ayant posé des questions. La Faculté de médecine dispose d'un système de formation hybride pour ses étudiants de 1re année (PACES). Chaque semestre se termine par un concours (en janvier et mai) éliminatoire (seule une partie des étudiants qui ont échoué au concours est autorisée à repasser l'année une seule fois, les autres devant se réorienter). Chaque unité d'enseignement est composée de deux à six séquences de 4 semaines (cf. figure 1).

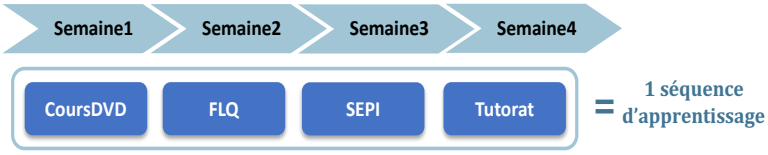

Figure 1 • Les quatre activités d'une séquence d'apprentissage sur quatre semaines

Dans chaque séquence, la première semaine consiste à étudier le cours sur DVD-ROM ou sur le site Medatice (diapositives + vidéo du professeur). La deuxième semaine est consacrée à la Formulation en ligne des questions (FLQ) pour les enseignants : ces questions concernent exclusivement les cours multimédias étudiés la semaine précédente. Les élèves peuvent voir les questions posées uniquement par les élèves de leur groupe (environ 200 élèves par groupe) et voter pour celles auxquelles ils veulent aussi une réponse, mais il ne leur est pas possible de les commenter ou d'y répondre. En fin de semaine, les questions sont envoyées par courriel aux enseignants intervenant la troisième semaine, qui les utilisent pour structurer leurs

sessions d'enseignement interactives en classe (SEPI). Au cours de ces sessions, l'enseignant répond à certaines questions posées en ligne par les étudiants. La quatrième semaine est consacrée à des séances de tutorat afin de tester les connaissances acquises lors de la séquence de formation par le biais d'un autotest utilisant des questions à choix multiples (QCM), qui sont ensuite corrigées par un professeur auxiliaire. Il y a deux séances de tutorat de 2 heures par semaine. Chaque étudiant peut vérifier individuellement ses notes et son classement par rapport à l'ensemble de la promotion, et il est nécessaire de s'être connecté à la plateforme de questions pour pouvoir consulter ses notes.

Pour chacun des 13 cours, nous avons donc 2 à 6 ensembles de questions (un par séquence) posées au total par 429 étudiants (6457 questions au total) et votées par 672 étudiants (10 951 votes) pendant la deuxième semaine de chaque séquence. La répartition des questions est inégale (cf. Tableau 1), avec plus de questions au 1er semestre, notamment car certains étudiants sont obligés d'arrêter à la fin de celui-ci, ce qui explique qu'il y ait moins d'étudiants au 2e semestre. On note que seul un élève sur quatre a posé au moins une question, ce qui peut être lié à l'encouragement à voter au lieu de poser des questions (pour forcer à lire les questions des autres mais aussi pour réduire le nombre de questions reçues par courriel).

Tableau 1 • Distribution des questions posées par cours

BCH	BPH	HBD	BCE	ANT	PHS	SSH	ICM	MAT	Spec.
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 = Pharmacie, Odontologie, Maïeutique

4. Méthode de catégorisation et d'annotation des 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. Notre objectif est de fournir des catégories de questions qui prennent en compte l'intention de l'élève. Par conséquent, nous avons décidé de définir notre propre schéma de codage pour identifier le type des questions posées par les étudiants, en utilisant une approche ascendante fondée sur les données. Nous présentons ici la démarche suivie pour construire l'annotateur associé.

4.1. Méthodologie de catégorisation

Afin d'identifier la nature des questions posées par les étudiants, nous avons travaillé sur un échantillon de 800 questions (12 % du corpus) issues de deux cours (BCH et HBD), considérés par l'équipe pédagogique comme étant parmi les plus difficiles et ayant suscité le plus de questions (cf. Tableau 1). Cet échantillon a été divisé en 4 sous-échantillons de 200 questions pour appliquer 4 étapes successives de catégorisation.

(1) L'**étape de découverte** consistait à regrouper empiriquement des phrases ayant des similitudes pour en extraire des concepts significatifs. Bien que les enseignants demandent aux étudiants de poser des questions simples (c.-à-d. d'éviter des questions comme « Pourriez-vous expliquer à nouveau X ? De plus, Y n'était pas clair »), 40 % des questions pouvaient être divisées en plusieurs questions indépendantes. Une fois les phrases segmentées en questions dites simples, nous avons regroupé celles dont la structure (par ex. « qu'est-ce que X ? » et « qu'est-ce que Y ? ») et la sémantique (par ex. « qu'est-ce que X ? » et « pourriez-vous définir X ? ») semblent similaires. Des groupes de questions ont ensuite reçu des *étiquettes* (par ex. « définition d'un concept »), chaque étiquette est associée à un groupe de questions. Puis nous avons identifié les exclusions mutuelles entre étiquettes (par ex. une question simple ne peut pas être à la fois une vérification et une demande de réexplication). Cela nous a conduits à définir le concept de « dimension », ensemble d'étiquettes de type de questions similaires mais mutuellement exclusives (par ex. une question ne peut pas être à la fois une « vérification » et une « ré-explication » au sein de la première dimension car vérifier suppose de proposer soi-même une réexplication en premier lieu – bien sûr il peut cependant y avoir un enchaînement des deux demandes dans 2 propositions différentes). Chacune de ces étiquettes individuelles (« vérification », « réexplication »...) sont des valeurs pouvant être associées à une dimension. Chaque question simple peut alors être associée à une annotation dans ce schéma de codage en choisissant une valeur par dimension.

(2) L'**étape de consolidation** consistait à annoter le deuxième sous-échantillon 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 la dimension 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 et ainsi pouvoir intervenir ensuite.

(3) Dans l'**étape de validation**, nous avons effectué deux annotations indépendantes pour valider l'ensemble de nos catégories sur le troisième sous-échantillon de 200 phrases. Deux annotateurs humains ont utilisé comme référence unique le schéma de codage créé à la fin de l'étape précédente pour annoter chaque segment (238 au total). À l'issue de l'étape précédente, trois 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). 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 (Arstein et Poesio, 2008). Le Kappa est un score d'accord entre -1 et 1, où 1 correspond à un accord parfait et 0 à un accord uniquement explicable par le hasard (ex : en prédisant systématiquement « pile » après un lancer de pièce, bien que l'on ait raison 1 fois sur 2, cet accord entre la prédiction et la réalité s'explique uniquement par le hasard et correspondrait à un kappa de 0). Les Kappas obtenus ici sont $\kappa 1 = 0.72$, $\kappa 2 = 0{,}62$ (où $\kappa 1$ et $\kappa 2$ correspondent respectivement au Kappa de Dim1 et Dim2) soit bien au-dessus de 0 et témoignent donc d'un accord fort non uniquement explicable par le hasard. Pour Dim4, en raison de son caractère facultatif, les deux annotateurs n'ont pas nécessairement annoté les mêmes questions (annotateur 1 : 82 questions ; annotateur 2 : 68 questions) : sur les 68 en commun, le kappa valait 0,66. Puis ils se sont rencontrés pour discuter et résoudre les désaccords, ce qui a conduit à un affinement final des catégories (par exemple, séparation des catégories Dim1 et Dim4, ajout de la catégorie Dim3). Finalement, tout l'échantillon (600 phrases) a été réannoté sur les 4 dimensions pour tenir compte des changements et fournir une référence à laquelle comparer l'annotation automatique. Cette version finale du schéma de codage est présentée dans le Tableau 2. Une annotation de question peut donc être vue comme un vecteur de 4 valeurs (ex : « Pourriez-vous réexpliquer la différence entre un composé ionisable et un partiellement ionisable ? » marquée comme « Ree » sur Dim1, « Lie » sur Dim3 et aucune valeur « 0 » pour les dimensions 2 et 4, c.-à-d. [Ree,0,Lie,0]).

(4) Finalement, dans l'**étape d'évaluation**, le dernier sous-échantillon de 200 segments a été annoté manuellement par les deux annotateurs experts (avec un kappa accru de 0,83 sur Dim1, 0,76 sur Dim2 et 0,47 sur Dim3). Ce sous-échantillon, non utilisé pour l'entraînement de l'annotateur automatique, a été utilisé pour son test.

Le schéma de codage proposé est donc issu de l'expertise humaine puisque les 14 catégories ont été définies par les chercheurs, puis revues et validées par un enseignant expert du domaine.

Tableau 2 • Schéma de codage créé à partir de l'annotation manuelle

Dim1	Type de question	Description
Ree	Réexpliquer/redéfinir	Demander de revenir sur un concept déjà expliqué
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	**Modalité d'explication**	**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 ?)	Demander 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 dans cours ou dans l'explication de l'enseignant
Con	Connaissances du cours	Vérifier une connaissance
Exa	Examen	Vérifier une connaissance attendue à l'examen

4.2. Annotation automatique

Afin d'annoter l'ensemble de questions posées par les étudiants, un outil semi-automatique à base de règles et de mots clés pondérés manuellement a été utilisé dans un premier temps pour segmenter et annoter les questions automatiquement. Bien qu'efficace sur les questions qu'il annote (kappa élevé), certaines questions ne sont pas annotées par cet outil (Harrak *et al.,* 2018) : en effet, cet outil dépend de mots-clés pondérés manuellement, et certaines dimensions dans notre schéma de codage n'ont pas de mots clés explicites pour les annoter (par ex. connaissances en cours dans la Dim4). Par conséquent, nous avons envisagé d'utiliser une annotation entièrement automatisée basée sur des techniques d'apprentissage automatique sur le corpus des questions, indépendantes de mots-clés pondérés manuellement. Les différentes étapes suivies sont décrites dans ce qui suit et sont résumées dans la figure 2.

Pour ce qui est de l'étape de segmentation préalable à l'annotation, nous utilisons un système de détection de la limite de la phrase intégré dans NLTK (Kiss et Strunk, 2006), qui est l'un des systèmes de traitement automatique de la langue fonctionnant en français. Il repose sur une approche dite non supervisée, et 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 fourni aurait dû être plus segmenté qu'il ne l'était.

La première étape a consisté à transformer les 923 segments annotés manuellement en vecteurs de mots. Tout d'abord, nous avons utilisé la version française de WordNet (Sagot et Fišer, 2008), base de données lexicale reliant des concepts sémantiques entre eux dans une ontologie selon une variété de relations sémantiques (telles que synonymie et hyperonymie) afin de ramener différentes expressions synonymes à une même expression dans les questions. Par exemple pour la valeur « Rai » dans Dim3, les mots synonymes « cause », « raison » et « motif » sont remplacés dans le texte par « pourquoi ». L'objectif étant de diminuer la diversité lexicale et de renforcer certaines expressions pour le traitement. Nous avons effectué par la suite un ensemble de prétraitements classiques

sur le corpus de 923 segments : *tokenisation*, *racinisation*, suppression de ponctuation et de *stopwords* (mots creux non porteurs de sens), etc. Puis, nous avons extrait tous les unigrammes et bigrammes (n-grammes avec n = 1 et n = 2 respectivement), avec une approche de type *sac de mots*, et compté leurs occurrences dans l'échantillon de 600 questions (723 segments) de l'étape de validation. Chaque segment est représenté par un vecteur de mots (nombre d'occurrences de chaque unigramme/bigramme extrait sur chaque segment). Le nombre de n-grammes étant très important par rapport au nombre de segments, nous avons réduit celui-ci pour conserver les mots-clés les plus importants et les plus significatifs en utilisant une technique de sélection d'attributs (suppression des n-grammes les moins fréquents et corrélés).

La deuxième étape a consisté à entraîner un classifieur pour annoter automatiquement chaque valeur (ou étiquette) de dimension (par ex. « réexpliquer »). Nous avons testé 6 techniques de classification différentes telles qu'implémentées dans RapidMiner (Modèle linéaire généralisé, Gradient Boosted Trees, Arbre de décision, K-NN, Règle d'induction et Naïve Bayes, avec différentes valeurs d'hyperparamètres testées pour chacune) sur chaque dimension séparément, les dimensions étant conçues comme indépendantes. Chaque classifieur est entraîné en prenant en entrée un ensemble de vecteurs de mots représentant les 723 segments de l'ensemble d'entraînement, et l'étiquette à prédire est la valeur associée manuellement au segment dans cette dimension.

Le modèle est ensuite évalué sur un échantillon indépendant de 200 segments 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. Les meilleurs résultats ont été obtenus par l'algorithme Gradient Boosted Trees avec un Kappa moyen sur chaque dimension de 0,70 (cf. Tableau 3) – une valeur suffisamment élevée pour appliquer l'annotation automatique au corpus complet.

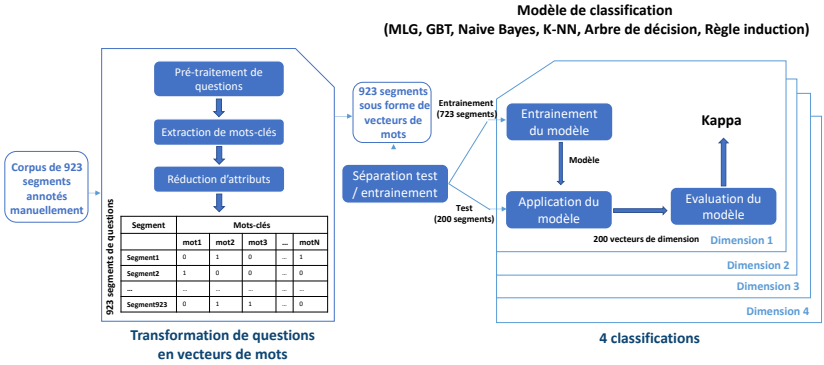

Figure 2 • Processus d'annotation à base d'apprentissage automatique

Tableau 3 • Kappas obtenus entre les différentes techniques de classification utilisées et l'annotation experte de référence

Dimension	Modèle linéaire généralisé (GLM)	Gradient Boosted Trees (GBT)	Naive Bayes	K-NN (K = 2)	Arbre décision (C4.5)	Règle induction
Dim1	0,68	**0,70**	0,29	0,57	0,36	0,70
Dim2	0,17	0,77	0,10	0,43	**0,79**	0,37
Dim3	**0,69**	0,63	0,37	0,61	0,63	0,58
Dim4	0,62	**0,66**	0,38	0,60	0,13	0,66

5. Analyse de différences entre les questions sur lesquelles votent les étudiants et les questions qu'ils posent eux-mêmes

À l'issue des travaux exposés dans la section précédente, nous disposions d'un annotateur automatique que nous avons pu appliquer au corpus complet de questions, ce qui a permis d'obtenir un ensemble de 6457 questions annotées automatiquement. Nous avons concentré notre analyse sur quatre cours ayant généré le plus de questions (cf. Tableau 1) et considérés par les enseignants comme les plus difficiles : BCH, HBD, BCE et ANT. Les trois premiers ont lieu au premier semestre et ANT au second semestre. Nous n'avons pas fusionné les questions des différents cours, car des études antérieures sur ces données avaient montré des différences

significatives entre les cours (Harrak *et al.,* 2019 ; Harrak *et al.,* 2018). Ceci est lié au fait que la dynamique des questions semblait être un indicateur pour distinguer les étudiants, mais celle-ci est très liée au cours. De plus, considérer les cours séparément permet de vérifier si des tendances similaires apparaissent d'un cours à l'autre.

Pour répondre aux questions de recherche, nous avons distingué 4 sous-populations sur chacun des cours considérés en fonction de l'activité des étudiants en distinguant : Q pour les étudiants « ayant posé au moins une question » (NQ sinon), et V pour les étudiants « ayant voté sur au moins une question » (NV sinon). En croisant les deux, cela donne donc les 4 sous-populations suivantes : QV, QNV, NQV et NQNV. Dans la suite de cet article, nous présentons l'analyse du lien entre vote et performance en section 5.1, la relation entre vote et engagement en section 5.2, la comparaison de la nature entre questions posées et questions votées en section 5.3 et un tableau de synthèse des résultats en section 5,4.

5.1. Analyse de lien entre questions, vote et performance

Pour examiner la QR1 (c.-à-d. le lien entre questions/vote et performance d'un étudiant), nous avons étudié dans un premier temps le lien entre questions et performance pour les étudiants ayant posé des questions (Q) et ceux qui n'en ont pas posées (NQ) et le lien entre vote et performance pour les votants et non votants en section 5.1.1. Ensuite, nous avons analysé de manière plus fine le lien entre le vote et la performance en section 5.1.2, notamment pour les étudiants qui posent des questions (QV vs. QNV) et ceux qui n'en posent pas (NQV vs. NQNV).

5.1.1. Lien entre questions et performance et lien entre vote et performance

5.1.1.1. Méthode

Pour évaluer la performance, nous avons considéré pour chaque étudiant et sur chaque cours : (1) la note moyenne obtenue sur les QCM du cours (NotMoy, sur 20), qui peut donner une mesure de l'impact à court terme des questions posées, et (2) la note finale obtenue au concours à cette matière (NotFin, sur 20), qui peut donner une mesure de l'effet à plus long terme.

Ensuite, pour chacun des 4 cours considérés, nous avons effectué des comparaisons 2 à 2 de la performance obtenue entre d'une part les étudiants qui ont posé des questions (Q, constituée de QNV et QV) sur ce

cours et ceux qui n'en ont pas posé (NQ, constituée de NQV et NQNV), puis d'autre part la population votante (V, constituée de QV et NQV) et la population non-votante (NV, constituée de QNV et NQNV). Pour ces 2 variables (NotMoy et NotFin), nous avons utilisé des tests Mann-Whitney U (MacFarland et Yates, 2016) au lieu de t-tests (les distributions ne suivant pas une loi normale). Nous avons effectué 2 fois 7 tests (2 comparaisons de population avec 4 cours et 2 variables à chaque fois, sauf NotFin, manquante pour BCE). Nous rapportons une taille d'effet estimée, calculée comme suit : $r^2 = \eta^2 = Z^2/n$ 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). Les seuils de significativité ont été corrigés par la méthode de Holm-Šidák (Abdi, 2007) pour éviter les erreurs de type I (rejet de l'hypothèse nulle alors qu'elle est vraie). 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 $\tilde{V}$ (Bergsma, 2013).

5.1.1.2. Analyse des résultats

Les résultats des tests pour effectuer les 2 comparaisons de population (Q vs. NQ et V vs. NV) avec 4 cours et deux variables à chaque fois (NotMoy et NotFin) sont décrits dans ce qui suit et sont résumés dans le Tableau 4. Les résultats significatifs sont mis en gras et associés à une valeur $p < .001$ après la correction de Holm-Šidák.

Pour les étudiants ayant posé des questions (Q) et ceux qui n'en ont pas posées (NQ), il n'y avait pas de différence statistiquement significative pour les deux variables NotMoy et NotFin dans chacun des 4 cours.

En ce qui concerne les étudiants ayant voté à des questions (V) par rapport à ceux qui n'ont voté à aucune (NV), 2 résultats significatifs (cf. Tableau 4) ont été obtenus (sur 7 tests) : les étudiants qui ont voté (V) ont des notes finales plus élevées que ceux qui n'ont pas voté (NV) pour le cours ANT ($U = 56357.5$, $p = .004$, $\eta^2 = .004$). La tendance est inversée pour la NotMoy pour le cours HBD ($U = 211938.5$, $p < .001$, $\eta^2 = .010$). La valeur de p indique que les tests sont statistiquement significatifs, la valeur U du test est à comparer à la valeur maximale qui est le produit de la taille des 2 échantillons considérés, tandis que η^2 indique la force de l'effet d'une variable sur l'autre.

Tableau 4 • Comparaison des étudiants (Q vs. NQ et V vs. NV) en termes de performance

Cours	Q vs. NQ		V vs. NV	
	Not-Moy	NotFin	NotMoy	NotFin
BCH	.477	.584	.005	.027
HBD	.408	.066	**.000***	.015
BCE	.080	N/A	.551	N/A
ANT	.540	.020	.740	**.004***

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

5.1.2. Liens croisés entre questions, votes et performance

5.1.2.1. Méthode

Afin d'analyser de manière plus fine la relation entre les votes et les questions en termes de performances, nous avons fait des comparaisons 2 à 2 entre QV et QNV (pour analyser le vote chez les étudiants ayant posé des questions) et entre NQV et NQNV (pour analyser le vote chez les étudiants qui n'ont pas posé de questions), sur chacun des 4 cours, pour les deux variables NotMoy et NotFin. Pour ces 2 variables, nous avons utilisé les mêmes tests qu'en section précédente et effectué 2 fois 7 tests (2 comparaisons de population avec 4 cours et 2 variables à chaque fois, sauf NotFin, manquante pour BCE) et les seuils de significativité ont été corrigés par la méthode de Holm-Šidák pour éviter les erreurs de type I.

5.1.2.2. Analyse des résultats

Les résultats des tests de comparaison des votants (QV vs. QNV et NQV vs. NQNV) en termes de performance (note moyenne et note finale) et statistiques descriptives (quartiles et médiane) des 4 sous-populations sur chacun des 4 cours sont présentés dans les Tableaux 5 et 6.

En ce qui concerne QV par rapport à QNV, 1 seul résultat significatif a été obtenu sur les cours du premier semestre (BCH, HBD, BCE) : pour HBD, parmi les étudiants ayant posé des questions, ceux n'en ayant pas voté ont eu une meilleure note finale que les étudiants qui en ont voté (U = 2977,5, p = .002, η^2 = .040). En revanche, pour ANT, 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 (U = 1452, p < .001, η^2 = .127) qu'au concours final (U = 1494,5, p < .001, η^2 = .155).

En ce qui concerne les NQV par rapport à NQNV, 3 résultats significatifs ont également été obtenus (sur 7 tests) : pour BCH (U = 112 024, p = 0,001, η^2 = 0,006) et HBD (U = 121258.5, p < .001, η^2 = .016), les étudiants n'ayant pas posé de questions et n'ayant pas non plus voté ont eu une meilleure note sur les QCM que ceux ayant uniquement voté. Ce résultat se retrouve également au niveau du concours final pour HBD (U = 129 974, p < .001, η^2 = .007). Aucune différence n'a été observée au cours du deuxième semestre (ANT). En résumé, lorsqu'une différence a été observée, les élèves qui ont voté (sans poser de question) ont obtenu des résultats inférieurs, tant dans le cours que dans l'ensemble.

Tableau 5 • Comparaison des votants en termes de Note Moyenne (NotMoy) pour les étudiants qui posent des questions (QV vs. QNV) et ceux qui n'en posent pas (NQV vs. NQNV

	QV vs. QNV	NQV vs. NQNV	QV				QNV				NQV				NQNV			
	p	P	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3
BCH	.049	**.001***	181	6,5	8,5	11,5	59	7,7	9,7	12,5	217	5,6	8,0	10,1	980	5,5	8,0	10,9
HBD	.176	**.000***	154	7,7	10,1	13,8	53	8,2	11,7	14,0	252	6,0	8,7	11,2	956	6,0	9,2	12,5
BCE	.128	.039	83	7,4	10,4	12,6	47	6,4	9,0	11,7	117	6,0	8,0	10,8	1133	5,0	7,7	10,8
ANT	**.001***	.005	42	11,2	13,8	15,3	46	6,4	10,2	13,3	23	8,6	10,0	12,7	968	5,0	8,0	11,8

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

Tableau 6 • Comparaison des votants en termes de la Note Finale (NotFin) pour les étudiants qui posent des questions (QV vs. QNV) et ceux qui n'en posent pas (NQV vs. NQNV)

	QV vs. QNV	NQV vs. NQNV	QV				QNV				NQV				NQNV			
	p	P	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3
BCH	.080	.015	168	5,2	8,2	11,5	56	6,3	10,3	13,3	197	4,2	6,5	9,7	880	3,7	6,7	10,5
HBD	**.002***	**.000***	144	6,5	9,2	12,1	44	8,2	11,5	13,1	229	4,5	7,2	10,5	886	4,0	7,7	10,9
BCE	-	-	83	-	-	-	49	-	-	-	118	-	-	-	1368	-	-	-
ANT	**.000***	.706	42	13	15,3	16,2	45	7,5	11,5	14,2	22	8,1	11,9	14,7	1116	3,5	7,0	12,0

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

5.2. Analyse de lien entre questions, vote et engagement

Pour examiner la QR2 (c.-à-d. le lien entre questions/vote et engagement), nous avons comparé les populations Q vs. NQ et V vs. NV en termes d'engagement pour étudier respectivement le lien entre questions et engagement et lien entre vote et engagement en section 5.2.1. Ensuite, nous avons analysé de manière plus fine le lien entre le vote et l'engagement en section 5.2.2, notamment pour les étudiants qui posent des questions (QV vs. QNV) et ceux qui n'en posent pas (NQV vs. NQNV).

5.2.1. Lien entre questions et engagement et lien entre vote et engagement

5.2.1.1. Méthode

Pour évaluer l'engagement, nous avons considéré cette fois, pour chaque étudiant et sur chaque cours, des variables liées à l'assiduité (utilisée comme un marqueur de l'engagement, mais l'engagement est un concept plus large). Il s'agit d'une variable déclarative, établie à partir des appels faits en cours en présentiel. On en extrait 2 variables : (1) le ratio de l'assiduité globale (AssGlb) sur les deux semestres, de 0 (jamais là) à 1 (toujours là) et (2) le ratio de l'assiduité (AssCou) sur ce cours, de 0 (jamais là) à 1 (toujours là). Nous avons utilisé les mêmes tests qu'en section 5.1 (Mann-Whitney U puisque les distributions ne suivaient pas une loi normale) pour comparer les populations Q et NQ ainsi que V et NV en termes d'engagement. Nous avons effectué 2 fois 8 tests (2 comparaisons de population avec 4 cours et 2 variables à chaque fois). Nous rapportons également la taille d'effet estimée et corrigeons les seuils de significativité par la méthode de Holm-Šidák.

5.2.1.2. Analyse des résultats

Les résultats des tests de comparaisons de population (Q vs. NQ et V vs. NV) sur les 4 cours et avec deux variables à chaque fois (AssGlb et AssCou) sont décrits dans ce qui suit et sont résumés dans le Tableau 7. Les résultats significatifs sont mis en gras et associés à une valeur $p < .001$ après la correction de Holm-Šidák.

En ce qui concerne les Q par rapport à NQ, pour la variable assiduité globale, il n'y avait qu'une différence statistiquement significative pour BCE ($U = 54\,755$, $p < .001$, $\eta^2 = .024$). Cependant, pour l'assiduité en cours, il y avait des résultats statistiquement significatifs pour les 4 cours : BCH ($U = 213\,974$, $p < .001$, $\eta^2 = .035$), HBD ($U = 71\,554$, $p < .001$, $\eta^2 = .097$), BCE ($U = 110\,005$, $p < .001$, $\eta^2 = .021$) et ANT ($U = 88\,238$, $p < .001$, $\eta^2 = .042$).

En ce qui concerne les V par rapport à NV, nous avons trouvé également que les étudiants qui ont voté assistaient plus souvent au cours que ceux qui n'ont pas voté, sur les 4 cours considérés : BCH ($U = 299\,389$, $p < .001$, $\eta^2 = .029$), HBD ($U = 304\,339$, $p < .001$, $\eta^2 = .044$), BCE ($U = 179\,322$, $p < .001$, $\eta^2 = .040$) et ANT ($U = 72\,080$, $p < .001$, $\eta^2 = .036$) et de manière globale uniquement pour BCE ($U = 113\,413$, $p < .001$, $\eta^2 = .008$).

Tableau 7 • Caractérisation des étudiants (Q vs. NQ et V vs. NV) en termes d'engagement

Cours	Q vs. NQ		V vs. NV	
	AssGlb	AssCou	AssGlb	AssCou
BCH	.821	**.000***	.023	**.000***
HBD	.113	**.000***	.425	**.000***
BCE	**.000***	**.000***	**.000***	**.000***
ANT	.361	**.000***	.383	**.000***

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

5.2.2. Relation entre vote, questions et l'engagement

5.2.2.1. Méthode

Pour analyser de manière plus fine le comportement des élèves en termes d'engagement, nous avons suivi la même démarche qu'en section 5.2.1 et fait des comparaisons 2 à 2 entre QV et QNV et entre NQV et NQNV, sur chacun des 4 cours, pour les deux variables AssCou et AssGlb. Pour ces 2 variables, nous avons utilisé les mêmes tests Mann-Whitney U (distributions ne suivant pas une loi normale) et effectué 2 fois 8 tests (2 comparaisons de population avec 4 cours et 2 variables à chaque fois). Les seuils de significativité ont été corrigés par la méthode de Holm-Šidák pour éviter les erreurs de type I.

5.2.2.2. Analyse des résultats

Les résultats des tests de comparaison des votants (QV vs. QNV et NQV vs. NQNV) en termes d'engagement (assiduité globale et assiduité en cours) et statistiques descriptives (quartiles et médiane) des 4 sous-populations sur chacun des 4 cours sont présentés dans les Tableaux 8 et 9.

En ce qui concerne les QV par rapport à QNV, aucun résultat statistiquement significatif n'a été obtenu.

En ce qui concerne les NQV par rapport à NQNV, les étudiants qui ont voté suivaient le cours plus souvent que ceux qui n'ont pas voté, sur les 4 cours considérés (comme déjà observé dans 5.2.1.2) : BCH (U = 151652.5, p < .001, η^2 = .013), HBD (U = 181 119, p < .001, η^2 = .025), BCE (U = 107180.5, p < .001, η^2 = .028) et ANT (U = 25 908, p < .001, η^2 = .011). En revanche, aucun résultat statistiquement significatif n'a été obtenu pour l'assiduité globale.

Tableau 8 • Caractérisation des votants en termes d'assiduité globale (AssGlb) pour les étudiants qui posent des questions (QV vs. QNV) et ceux qui n'en posent pas (NQV vs. NQNV)

Cours	QV vs. QNV	NQV vs. NQNV	QV				QNV				NQV				NQNV			
	p	p	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3
BCH	.106	.006	185	0,91	0,98	1	61	0,90	0,98	1	227	0,79	0,95	1	1147	0,76	0,95	1
HBD	.490	.608	154	0,93	0,98	1	54	0,93	0,98	1	262	0,88	0,96	1	1150	0,71	0,95	1
BCE	.632	.415	83	0,83	0,98	1	49	0,92	0,98	1	118	0,9	0,98	1	1368	0,74	0,95	1
ANT	.570	.691	43	1	1	1	47	0,94	1	1	23	0,98	0,98	1	1507	0,76	0,95	1

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

Tableau 9 • Caractérisation des votants en termes d'assiduité en cours (AssCou) pour les étudiants qui posent des questions (QV vs. QNV) et ceux qui n'en posent pas (NQV vs. NQNV)

Cours	QV vs. QNV	NQV vs. NQNV	QV				QNV				NQV				NQNV			
	p	p	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3	N	Q1	Md	Q3
BCH	.555	**.000***	185	1	1	1	61	1	1	1	227	0,83	1	1	1147	0,33	0,83	1
HBD	.697	**.000***	154	1	1	1	54	1	1	1	262	1	1	1	1150	0,5	1	1
BCE	.249	**.000***	83	1	1	1	49	1	1	1	118	1	1	1	1368	0,4	1	1
ANT	.026	**.000***	43	1	1	1	47	0,8	1	1	23	0,9	1	1	1507	0	0,6	1

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

5.3. Comparaison de la nature des questions posées et des questions votées

Pour examiner la QR3 (c.-à-d. pas de différences qui apparaissent entre la nature des questions posées et celle des questions votées par les étudiants), nous avons analysé la nature des questions posées et la nature des questions votées par les étudiants qui font les deux (QV).

5.3.1. 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 réexplication et une autre de vérification (étiquetées [Ree,0,Sch,0] et [Ver,0,0,Con]), sur la dimension 1, il aurait voté à 50 % sur des questions de valeur « Ree » (réexplication) et à 50 % sur des 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. En suivant la même approche pour les questions posées, on peut également obtenir un vecteur de questions posées de 14 valeurs.

Une fois le prétraitement effectué, la comparaison des questions votées aux questions posées consistait à comparer pour chaque cours, pour chaque valeur d'une dimension (par 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 comme en 5,1 la méthode Holm-Šidák pour ajuster la valeur *p* critique.

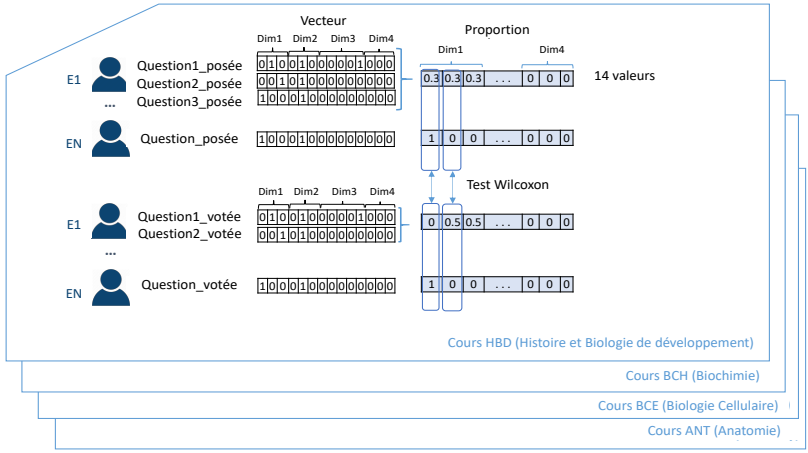

Figure 3 • Codage des questions posées et des questions votées (par QV) en termes de proportion

5.3.2. Analyse des résultats

Un seul test sur 56 a révélé un résultat statistiquement significatif (cf. Tableau 10), répondant ainsi négativement à la QR3.

Tableau 10 • Différences entre voter et poser une question selon la nature des questions (pour QV)

	Ree	App	Ver	Exe	Sch	Cor	Def	Man	Rai	Rol	Lie	Err	Con	Exa
BCH	.176	.427	.11	.31	.411	.017	.92	**.000***	.236	.352	.295	.514	.259	.078
HBD	.382	.851	.717	.755	.809	.225	.093	.728	.007	.285	.003	.043	.941	.706
BCE	.476	.067	.015	.515	.723	.929	.89	.652	.797	.681	.118	.51	.686	.033
ANT	.826	.087	.551	.204	.795	.18	.076	.485	.212	.198	.039	.611	.691	.701

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

5.4. Synthèse des résultats

Une synthèse de l'ensemble des résultats obtenus précédemment (cf. sections 5.1, 5.2 et 5.3) pour les différentes populations, résumant les différences significatives (jamais, parfois et toujours) pour les 4 variables considérées (note moyenne, note finale, assiduité globale et assiduité en cours), est présentée dans le Tableau 11.

Tableau 11 • Synthèse des résultats pour les différentes populations sur les 4 variables considérées

	Performance		Engagement	
	NotMoy	NotFin	AssGlb	AssCou
Q	-	-	+	++
NQ	-	-	-	-
V	-	+	+	++
NV	+	-	-	-
QV	+	+	-	-
QNV	-	+	-	-
NQV	-	-	-	++
NQNV	+	+	-	-

« - » : « jamais » (sur aucun cours), « + » : « parfois » (sur certains cours), « ++ » : « toujours » (sur tous les cours)

6. Discussion et conclusion

Les résultats obtenus (cf. Tableau 11) révèlent plusieurs éléments intéressants qu'il convient de mettre en perspective. En termes de performance, la première analyse a révélé que le fait de poser des questions n'était pas associé à la performance des élèves (aucun résultat significatif trouvé pour la note obtenue lors de la séance de QCM en classe et la note finale). Le comportement de vote n'avait pas non plus de lien clair avec la performance : 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 votants qui posent également des questions ou non, une image plus claire apparaissent : le vote est plutôt négatif, surtout lorsqu'ils n'ont pas posé de questions par ailleurs. Cette tendance semble

cependant s'inverser plus tard dans l'année, où le fait de voter en complément du fait de poser des questions entraîne de meilleurs résultats, tant aux QCM de cours qu'à l'examen final. La différence entre le premier et le second semestre pourrait être liée au fait que les étudiants les plus en difficulté ont été obligés de quitter la formation à la fin du premier semestre, ainsi qu'à la baisse générale d'activité sur la plateforme : les étudiants qui continuent à y participer sont donc probablement les plus motivés de ceux ayant suivi les cours du 1er semestre.

En ce qui concerne l'engagement, le fait de poser des questions est associé à une plus grande participation au cours (puisque les réponses sont fournies pendant le cours). La comparaison des étudiants qui votent sur des questions et ceux qui ne votent pas (V et NV) a révélé une relation positive similaire. Le vote semble être associé également aux étudiants qui sont plus susceptibles d'être présents en cours, en particulier pour les étudiants qui ne posent pas de questions et sont souvent présents au cours. Cependant, 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. Il est intéressant de noter que le fait de voter et poser des questions ne semble pas lié à un engagement supérieur au fait de ne faire que l'une des deux activités. Dans notre contexte, il semble donc que des activités « actives » (au sens de Chi – ici, voter) complémentaires à des activités « constructives » (ici, poser des questions) soient plus efficaces que des activités « constructives » seules en termes d'apprentissage, mais qu'une activité « active » seule soit plus positive que de la passivité en termes d'engagement et négative en termes de performance.

L'analyse de la nature des questions posées et des questions votées par les étudiants qui font les deux (QV), c'est-à-dire les plus impliqués et ceux qui réussissent le mieux (notamment au second semestre), montre qu'il n'y a globalement pas de différence de nature entre les questions sur lesquelles ils votent et celles qu'ils posent eux-mêmes. Ce résultat est intéressant car il suggère une interprétation possible du résultat précédent, à savoir que les votes de ces étudiants correspondent effectivement bien à des questions qu'ils se posent vraiment. Il est possible que des étudiants se contentant de voter ne font pas l'effort de formuler leurs propres questions, et que s'ils le faisaient, elles seraient d'une nature différente.

Ce travail est exploratoire et présente donc plusieurs limites : 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. 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. Une expérience davantage contrôlée dans laquelle les étudiants doivent poser des questions et/ou voter sur des questions précédemment posées après avoir vu une vidéo, pour pouvoir passer à la suite, et complétée d'une approche qualitative (entretiens des étudiants), permettrait de vérifier les interprétations précédentes. Néanmoins le travail préalable réalisé ici rend désormais possible ce type d'expérience. Enfin, il est probable que les votes enregistrés soient en fait la manifestation de deux processus bien différents. Le premier correspond aux étudiants qui font l'effort de se poser des questions, se connectent à la plateforme et trouvant que celle-ci a déjà été posée, ne peuvent plus que la voter. Ce type de vote masque en fait une activité réelle « constructive » (au sens de Chi). Au contraire, les étudiants qui se connectent éventuellement sans question préalable, et découvrent à la lecture d'une question qu'ils se la posent également, sont dans une activité « active » (toujours au sens de Chi). Là aussi, forcer les étudiants à poser leurs questions éventuelles avant de lire celles des autres permettrait d'éviter cette ambiguïté dans le sens à donner au vote. Néanmoins il est important de souligner que le fait que des différences ont été observées entre les simples votants (les NQV qui mélangent donc des activités « actives » et « constructives ») et les poseurs de questions (les QV et QNV, entièrement dans une démarche « constructive ») plaident a priori en faveur d'un écart réel en fait encore plus important entre activités purement « actives » et purement « constructives ».

Dans notre contexte, ces résultats suggèrent qu'encourager les étudiants à formuler leurs questions, plutôt que de se contenter de voter sur les questions des autres, serait une stratégie positive pour l'apprentissage et permettrait également aux enseignants de choisir la bonne stratégie d'enseignement. 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...). Cela pourrait être fait en encourageant les étudiants à poser une question avant de pouvoir consulter celles des autres. Du point de vue des enseignants, cela signifie qu'il est d'autant plus critique de leur proposer une visualisation plus efficace que le « mur de questions » actuel pour les aider à mieux organiser leurs interventions durant la 3e semaine, ce qui a été abordé dans un autre travail (Harrak *et al.,* 2020).

Nous avons proposé des organisations alternatives de questions aux enseignants *via* un questionnaire pour évaluer l'utilisabilité de nos propositions et particulièrement le schéma de codage développé. Envisager des tableaux de bord personnalisés pour les enseignants à partir des organisations proposées est l'une des perspectives principales de ce travail.

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.

Globalement, notre schéma de codage permettrait d'annoter les questions des étudiants de manière plus fine en termes d'intentions et de nourrir la réflexion de l'enseignant pour lui proposer éventuellement une réaction pédagogique différente sur les questions posées. L'annotation automatique de questions permettrait également d'identifier des caractéristiques du profil des étudiants en termes de performance et d'autres aspects de leur comportement. 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 inversée, devraient pouvoir être facilement répliqués et réutilisés dans d'autres contextes et travaux.

REMERCIEMENTS

Nous remercions Pierre Gillois de nous avoir fourni les données.

RÉFÉRENCES

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics, 3,* 103-107.

Artstein, R. et Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555-596. https://doi.org/10.1162/coli.07-034-R2

Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *J Korean Statistical Society, 42*(3), 323-328. https://doi.org/10.1016/j.jkss.2012.10.002

Bihani, A., Ullman, J. D. et Paepcke, A. (2018). FAQtor : Automatic FAQ generation using online forums. Dans K. E. Boyer et M. Yudelson (dir.), *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)* (p. 529-532). ERIC.

Bloom, B. S. et Engelhart, M. B., Furst, E. J., Hill, W. H. et Krathwohl, D. R. (1956). *Taxonomy of educational objectives. The classification of educational goals. Cognitive domain: 1.* Addison-Wesley Longman Ltd.

(Chi et Wylie, 2014)

Chi, M. T. H. et Wylie, R. (2014). The ICAP framework: linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Chin, C. et Brown, D. E. (2002). Student-generated questions: a meaningful aspect of learning in science. *International Journal of Science Education, 24*(5), 521-549. https://doi.org/10.1080/09500690110095249

Chin, C. et Kayalvizhi, G. (2002). Posing problems for open investigations: what questions do pupils ask? *Research in Science & Technological Education, 20*(2), 269-287. https://doi.org/10.1080/0263514022000030499

Fritz, C. O., Morris, P. E. et Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2-18. https://doi.org/10.1037/a0024338

Graesser, A. C. et Person, N. K. (1994). Question asking during tutoring. *American educational research journal, 31*(1), 104-137.

Harper, K. A., Etkina, E. et Lin, Y. (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. https://doi.org/10.1002/tea.10111

Harrak, F., Bouchet, F., Luengo, V. et Gillois, P. (2020). Evaluating teachers' perceptions of students' questions organization. Dans C. Rensing et H. Drachsler (dir.), *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK 2020)* (p. 11-16). ACM. https://doi.org/10.1145/3375462.3375509

Harrak, F., Bouchet, F. et Luengo, V. (2019). From students' questions to students' profiles in a blended learning environment. *Journal of Learning Analytics, 6*(1), 54-84. https://doi.org/10.18608/jla.2019.61.4

Harrak, F., Bouchet, F., Luengo, V. et Gillois, P. (2018). Profiling students from their questions in a blended learning environment. Dans A. Pardo K. Bartimote-Aufflick (dir.), *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK 2018)* (p. 102–110). ACM. https://doi.org/10.1145/3170358.3170389

Jiang, Z., Zhang, Y., Liu, C. et Li, X. (2015). Influence analysis by heterogeneous network in MOOC forums: what can we discover? Dans O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros (dir)., *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)* (p. 242-249). ERIC.

Kiss, T. et Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Comput Linguist, 32*(4), 485–525.

Li, H., Duan, Y., Clewley, D. N., Morgan, B., Graesser, A. C., Shaffer, D. W. et Saucerman, J. (2014). Question asking during collaborative problem solving in an online game environment. Dans S. Trausan-Matu, K. E. Elizabeth Boyer, M. Crosby, K. Panourgia (dir.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)* (p. 617–618). Springer.

MacFarland, T. W. et Yates, J. M. (2016). Mann–Whitney U Test. Dans T. W. MacFarland et J. M. Yates (dir.), *Introduction to Nonparametric statistics for the biological sciences using R* (p. 103–132). Springer International Publishing. https://doi.org/10.1007/978-3-319-30634-6_4

Otero, J. et Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition and instruction, 19*(2), 143–175.

Sagot, B. et Fišer, D. (2008). Building a free French wordnet from multilingual resources. *OntoLex*. https://hal.inria.fr/inria-00614708/document

Scardamalia, M. et Bereiter, C. (1992). Text-based and knowledge based questioning by children. *Cognition and Instruction, 9*(3), 177-199. https://doi.org/10.1207/s1532690xci0903_1

Sullins, J., McNamara, D., Acuff, S., Neely, D., Hildebrand, E., Stewart, G. et Hu, X. (2015). Are you asking the right questions: The use of animated agents to teach learners to become better question askers. Dans I. Russell and W. Eberle (dir.). *Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2015)* (p. 479-481). AAAI Press. https://asu.pure.elsevier.com/en/publications/are-you-asking-the-right-questions-the-use-of-animated-agents-to-

Supraja, S., Hartman, K., Tatinati, S. et Khong, A. W. (2017). Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes. Dans X. Hu, T. Barnes, A. Hershkovitz et L. Paquette (dir.), *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)* (p. 56-63). ERIC.

Teixeira-Dias, J. J. C., Pedrosa de Jesus, H., Neri de Souza, F. et Watts, M. (2005). Teaching for quality learning in chemistry. *International Journal of Science Education, 27*(9), 1123-1137. https://doi.org/10.1080/09500690500102813

Wong, J.-S., Pursel, B., Divinsky, A. et Jansen, B. J. (2015). An analysis of MOOC discussion forum interactions from the most active users. Dans N. Agarwal, K. Xu et N. Osgood (dir.), *Proceedings of the 8th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP 2015)* (p. 452–457). Springer.

Zeng, Z., Chaturvedi, S. et Bhat, S. (2017). Learner affect through the looking glass: Characterization and detection of confusion in online courses. Dans X. Hu, T. Barnes, A. Hershkovitz et L. Paquette (dir.), *Proceedings of the International Conference on Educational Data Mining (EDM 2017)* (p. 272-277). ERIC.