Connaissances pour enseigner l'informatique : analyse textuelle de productions d'enseignants de l'école primaire

9

Gabriel Parriaux[1,2]
Christophe Reffay[3]
Béatrice Drot-Delange[4]
Mehdi Khaneboubi[1]

1. Université Paris Cité, Laboratoire EDA, 75006 Paris, France
2. Haute École Pédagogique (HEP) du canton de Vaud, 1007 Lausanne, Suisse
3. Université de Franche-Comté, ELLIADD, 25000 Besançon, France
4. Université Clermont-Auvergne, INSPÉ Clermont-Auvergne, Laboratoire ACTÉ, 63400 Chamalières, France

Dans ce chapitre, nous portons nos analyses sur ce que des documents pédagogiques (scénarios pédagogiques, fiches de préparation d'activités, retours d'expérience sur une activité mise en œuvre) pour l'enseignement de l'informatique, produits par des enseignants novices, peuvent révéler de leurs connaissances didactiques. Partant d'un ensemble de documents produits par les étudiants – futurs enseignants de trois instituts de formation en France et en Suisse, nous rendons compte d'un parcours de recherche en quatre étapes.

Dans une première étude (Drot-Delange *et al.*, 2021), ces productions ont été inventoriées et classées manuellement selon l'institution d'origine, le niveau des élèves ciblés (entre 6 et 11 ans), le type d'activités proposées (informatique débranchée, programmation, jeux sérieux ou robotique pédagogique), le matériel utilisé (robot ou application spécifique), les connaissances ou compétences ciblées (en informatique ou non) et les éventuels concepts informatiques manipulés. Dès lors, nous avons relevé dans quelques-unes de ces productions ce qui, d'un point de vue didactique, nous semble s'apparenter à des mécompréhensions, voire des manques. Deux de ces erreurs sont apparues plusieurs fois dans notre corpus et sont en lien avec un usage de vocabulaire informatique peu approprié, voire erroné, selon le contexte. La première consiste en une confusion au niveau des termes utilisés pour parler d'une donnée, en tant que représentation statique d'une information, ou d'une instruction, en tant qu'élément dynamique dans un programme, qui peut être exécuté. La seconde consiste en une utilisation erronée de vocabulaire tiré du paradigme de programmation séquentielle dans un contexte de programmation événementielle, en particulier dans des activités de robotique avec Thymio. Ce constat nous amène à penser que les spécificités des robots pédagogiques et leurs liens avec différents paradigmes

de programmation devraient être davantage explicités dans les formations d'enseignants.

Lors d'une seconde étude (Reffay *et al.*, 2023), afin de mieux investiguer cette dernière problématique, nous avons restreint le corpus aux seuls documents portant sur des activités de robotique, puis y avons ajouté des productions d'enseignants plus expérimentés, recueillies sur le web, en vue d'effectuer des comparaisons. Ayant inventorié le lexique de ce nouveau corpus à l'aide de méthodes d'analyse textuelle, nous en avons extrait les mots en lien avec l'informatique et avons pu vérifier la manière dont le lexique, ainsi que les notions informatiques qu'il véhiculait, différaient pour chaque robot.

Une troisième étude (Parriaux *et al.*, 2023) est venue prolonger la seconde, toujours sur le même corpus de robotique éducative. Adoptant une approche plus globale, considérant la totalité du lexique présent dans le corpus, nous avons extrait les thématiques à l'aide d'une méthode de *clustering*. Avec cette technique inductive et partiellement automatisée, il a ensuite été possible de montrer l'association qui existait en termes de lexique entre chacun des robots et l'une de ces thématiques.

Ce chapitre intervient comme la quatrième et dernière étape de ce parcours de recherche. Nous revenons au corpus de départ constitué de productions incluant une diversité d'activités débranchées ou non, des outils, comme les robots, mais aussi différents logiciels d'apprentissage de la programmation. Nous y appliquons les méthodes d'analyse textuelle de l'étude précédente, comme le *clustering*, afin de déterminer de manière inductive les thématiques abordées dans ces documents, ainsi que les liens entre le lexique et le type d'outil mobilisé ou le degré scolaire visé.

Cadre théorique et questions de recherche

Nous nous intéressons aux connaissances didactiques en lien avec l'enseignement de l'informatique. Par connaissances didactiques, nous entendons les connaissances professionnelles des enseignants qui sont spécifiques à l'enseignement d'un contenu donné, dans notre cas l'informatique. Nous considérons qu'un moyen d'accéder à ces connaissances, c'est d'en chercher les traces par exemple dans les productions des enseignants.

Pour approfondir ce que nous entendons par « connaissances didactiques », nous inscrivons notre étude au sein du cadre conceptuel des connaissances pédagogiques du contenu ou *Pedagogical Content Knowledge* (PCK) proposé par Shulman (1986, 2007) et enrichi par Magnusson *et al*. (1999). Ces derniers le décrivent comme constitué de cinq composants : la connaissance

des curricula, la connaissance de la compréhension des élèves, la connaissance des stratégies d'enseignement, la connaissance de l'évaluation ainsi que l'orientation en matière d'enseignement de la discipline. Ce dernier composant se réfère aux connaissances et aux croyances d'un enseignant envers les finalités de l'enseignement d'une discipline à un certain degré scolaire.

Plus récemment, une communauté de chercheurs en didactique des sciences a poursuivi le développement de ce cadre théorique pour le formaliser dans un « modèle de consensus affiné » ou *Refined Consensus Model* (RCM) des PCK (Carlson *et al.*, 2019). Ce nouveau modèle a la particularité de distinguer trois domaines de PCK : le PCK collectif (cPCK), le PCK personnel (pPCK) et le PCK en acte *enacted* PCK (ePCK). Ce dernier domaine comprend l'ensemble des connaissances qui se manifestent dans un contexte précis de l'activité de l'enseignant auprès d'une classe déterminée d'élèves, sur un sujet particulier et dans un contexte bien défini. Le ePCK déborde du lieu et du temps de la classe pour concerner également les activités en lien avec la planification et l'évaluation d'une leçon.

Plusieurs traductions en français cohabitent pour l'expression *Pedagogical Content Knowledge* (Kermen et Izquierdo-Aymerich, 2017). Nous choisissons la traduction en « connaissances didactiques » qui nous semble la plus accessible en français et la plus proche de la signification originale de Shulman.

Nos questions de recherche sont les suivantes :

1. QR1 : que pouvons-nous dire des thématiques abordées par les enseignants dans les ressources pédagogiques qu'ils ont produites pour enseigner l'informatique ?
2. QR2 : qu'est-ce que l'analyse du lexique composant ces ressources nous apprend au sujet des connaissances des enseignants pour enseigner l'informatique ?
3. QR3 : quelles sont les relations que l'on peut établir entre le lexique présent dans les ressources, le type d'activités, les outils et l'âge des élèves concernés ?

Corpus et méthode

Présentation du corpus

Notre corpus est constitué d'un total de 140 ressources pédagogiques pour enseigner l'informatique produites par des étudiants – futurs enseignants du primaire en formation initiale dans les Instituts nationaux supérieurs du professorat et de l'éducation (INSPE) de Clermont Auvergne et de Besançon en France, ainsi qu'à la Haute école pédagogique (HEP) du canton de Vaud en

Suisse, entre 2019 et 2020. Ces ressources, ainsi que les connaissances dont elles sont le recueil, sont influencées par de multiples facteurs, parmi lesquels le contexte de formation dans lequel elles ont été produites. Issues de formations initiales de futurs enseignants, elles rendent compte de la création par les étudiants de scénarios pédagogiques pour enseigner la science informatique à l'école. Pour l'INSPE de Besançon, il s'est agi de devoirs donnés aux étudiants dans le cadre d'un module obligatoire du master métiers de l'enseignement de l'éducation et de la formation (MEEF). Pour l'INSPE de Clermont Auvergne, ce sont des comptes rendus d'activités de programmation visuelle réalisées dans le cadre du module certificat informatique et Internet niveau 2 (C2i2e) du master MEEF. Pour la HEP du canton de Vaud, il s'est agi de présentations effectuées pour un examen oral validant un module facultatif d'introduction à la science informatique. Ces scénarios étaient en général réalisés par deux et étaient parfois expérimentés en classe.

De ce fait, il est tout à fait imaginable que le contenu de ces ressources ait subi l'influence du contexte de formation et des exigences institutionnelles formulées aux étudiants. Dans le cas de l'INSPE de Besançon, les étudiants étaient amenés à produire leur séquence à l'aide d'un outil auteur qui, par son format et ses rubriques, impose l'adjonction de certaines informations. Dans le cas de la HEP Vaud, les étudiants recevaient des critères précis d'évaluation de leur présentation qui les incitait à traiter de certains contenus. Nous pouvons notamment penser que la définition d'objectifs d'apprentissage, en lien avec les programmes scolaires, faisait partie des contraintes formulées. Nous pouvons également nous questionner sur le fait de savoir si nous aurions trouvé les mêmes éléments dans des ressources d'enseignants produites hors d'un tel contexte de formation. L'impact du contexte sur les résultats n'est donc pas totalement négligeable.

Nous précisons ci-après les variables catégorielles associées à ces ressources avec leurs différentes modalités. Pour chaque modalité de variable, nous précisons entre parenthèses le nombre de ressources associées à cette modalité. Certaines ressources peuvent contenir plusieurs activités et peuvent donc être associées à plusieurs modalités d'une même variable.

— Domaine informatique : algorithmique-programmation (131) et représentation des données (14).
— Type d'activité : informatique débranchée (117), robotique (76), programmation visuelle (39) et micromonde (19) (application incluant consignes, aides, évaluation et retours automatiques permettant à l'élève de s'exercer en autonomie).
— Outil : jeu du robot (74), déplacement sur grille (61), Bluebot (29), BeeBot (22), Thymio (17), ScratchJr (27), Scratch (15),

cryptographie (10). Dix-sept autres outils qui apparaissaient de manière très ponctuelle ($n \leq 5$) n'ont pas été retenus dans les analyses.
— Genre des groupes d'étudiants : masculin (6), féminin (108) et mixte (26).
— Formation suivie : huit heures (82) (INSPE de Besançon), 28 heures (46) (HEP Vaud) et 4 heures (12) (INSPE de Clermont Auvergne).
— Niveau scolaire des élèves : élémentaire (41) (3-6 ans), primaire inférieur (58) (6-9 ans) et primaire supérieur (50) (9-11 ans).

En dehors des niveaux scolaires très également répartis, les effectifs (à plat) de ces modalités montrent des disparités importantes pour chaque variable : davantage de programmation (90 %) que de représentation des données (10 %), une forte proportion d'activités débranchées (47 %) avec le jeu du robot[1] (25 %) et les déplacements sur grille (21 %), ainsi que de la robotique (30 %). Parmi les sept robots mentionnés dans notre corpus, les plus utilisés sont : Bluebot (37 %), BeeBot (28 %) et Thymio (22 %).

1. Le jeu du robot est une activité d'apprentissage de la programmation dans laquelle un élève, ou l'enseignant, joue le rôle d'un robot qui doit être guidé par des instructions données par le reste de la classe.

Les documents produits sont issus de formations ayant des durées très variables : 59 % sont issus de la formation de Besançon (8 h), 33 % de Lausanne (28 h) et 9 % de Clermont (4 h). Ce déséquilibre est encore accentué par le fait que les productions de Lausanne ont pris la forme de présentations orales dans le cadre d'un examen. Leur contenu est donc, par le format de diapositives et la transcription de l'audio, plus court en termes de texte que le contenu des documents plus détaillés constitués à Besançon et à Clermont où ils prenaient la forme de scénarios pédagogiques déroulés sur plusieurs pages et incluant les objectifs pédagogiques, les ressources utilisées, le rôle de l'enseignant et des élèves et les différentes activités à mettre en œuvre.

Analyse lexicale et textométrie

Nous choisissons d'analyser notre corpus à l'aide de méthodes issues du domaine de l'analyse des données textuelles (ADT). Leur particularité consiste à appliquer une approche statistique à l'étude du texte (Lebart *et al.*, 2019).

Cette méthode statistique considère le texte comme un ensemble de mots que l'on dénombre pour en analyser le contenu. La fréquence d'apparition de mots du lexique peut être comparée à une fréquence attendue dans le cas d'une distribution aléatoire de vocabulaire, et cet écart à la moyenne, mis en relation avec les caractéristiques des ressources observées, permet de mettre en évidence des proximités ou au contraire des oppositions entre certaines de ces caractéristiques et les termes du lexique.

Classification de Reinert

L'algorithme de classification de Reinert permet de regrouper des segments d'un corpus dans des catégories qui se distinguent par le lexique utilisé (Reinert, 1983). Par le biais de cette classification ou *clustering*, le chercheur voit surgir des thématiques qui constituent l'univers linguistique investi par les individus ayant généré les données du corpus. Cette méthode doit donc nous permettre de répondre à la première question de recherche (QR1).

Cet algorithme est fondé sur une classification descendante hiérarchique[2]. De manière itérative, il divise le corpus en deux clusters puis va identifier un maximum d'hétérogénéité en optimisant l'homogénéité de chaque cluster. Cette opération est répétée jusqu'à ce que le nombre de clusters demandé par le chercheur soit atteint. Il s'agit d'une méthode de classification non supervisée, ce qui signifie que les catégories ne sont pas connues à l'avance et sont déterminées par le corpus lui-même. C'est la raison pour laquelle nous parlons ici d'une méthode inductive.

Dans notre cas, le prétraitement des données et les analyses sont conduits dans le langage R (R Core Team, 2022)[3]. Nous procédons à une segmentation du corpus en segments d'environ 40 mots, puis nous réalisons sa lemmatisation, à savoir la réduction de ses termes à leur forme générique, à l'aide de la bibliothèque *SpacyR* dans R et le modèle de langage *fr_dep_news_trf* pour le français. Le corpus est ainsi découpé en *tokens* avec séparation aux espaces. Les mots vides (ne véhiculant pas de sens) sont retirés ainsi que les termes apparaissant dans moins de trois segments (considérés non significatifs).

Une fois ces prétraitements effectués, le lexique est extrait des segments et une matrice est constituée, appelée matrice termes-documents, avec, pour colonnes, les termes du lexique et en lignes, les segments. Chaque cellule, à l'intersection d'une ligne (segment) et d'une colonne (terme), indique alors la présence d'un terme du lexique par un 1 lorsqu'il apparaît (une ou plusieurs fois) dans le segment ou son absence par un 0.

La méthode de Reinert cherche à maximiser la distance du χ^2 entre deux regroupements de segments. Pour choisir les segments à regrouper, Max Reinert propose d'effectuer une analyse factorielle des correspondances (AFC) de la matrice termes-documents, puis d'ordonner les segments en fonction de leurs coordonnées sur le premier axe issu de cette AFC, c'est-à-dire celui qui maximise les oppositions entre les segments. Pour réaliser la scission du corpus en deux clusters, Reinert réalise alors différents regroupements en se fondant sur cet ordre-là en cherchant toujours à maximiser la distance du χ^2 entre les deux clusters. Cette étape est reproduite, de manière itérative, jusqu'à obtenir le nombre de clusters souhaité.

2. Les méthodes de classification descendante hiérarchique (CDH) constituent une famille de méthodes de classification qui permettent de trier une population en différentes classes sur la base de critères de ressemblance entre individus. La CDH procède par bipartitions successives, c'est-à-dire que l'on part du haut vers le bas : on commence par considérer toute la population dans son ensemble, puis on la divise en deux classes, et ce de manière itérative jusqu'au nombre de classes souhaité. Ces méthodes se différencient des techniques de classification ascendante hiérarchique (CAH) qui, à l'inverse, fonctionnent du bas vers le haut : partant des individus considérés individuellement les uns des autres, on constitue une classe en regroupant les deux individus les plus ressemblants. De manière itérative à nouveau, ces regroupements sont opérés jusqu'à obtenir un nombre de classes qui est pertinent pour l'interprétation du chercheur.

3. R est un langage de programmation particulièrement utilisé dans le domaine des analyses de données quantitatives.

Chaque cluster est caractérisé par le nombre de segments qui le composent et la liste des termes qui s'y trouvent surreprésentés. Ces termes, ainsi que le contexte des segments dans lesquels ils apparaissent, sont ceux qui permettent d'interpréter le sens que prend le cluster et d'en déduire la thématique à laquelle il se rattache.

Analyse des correspondances

Une fois le lexique extrait, nous nous intéressons aux relations entre celui-ci et les différentes caractéristiques de nos ressources.

En observant la présence de termes spécifiques dans certains documents et pas dans d'autres, ou leur fréquence d'apparition plus élevée dans certaines parties de corpus que dans d'autres, et en croisant cette information avec les caractéristiques des différents documents ou parties de documents, il est possible d'établir des proximités entre certains mots du lexique et certaines ressources. Pour mener une telle analyse, il est nécessaire de disposer d'une méthode qui permet de visualiser ces distances entre lexique et ressources, ou caractéristiques des ressources, et c'est ici qu'intervient l'analyse des correspondances.

L'analyse des correspondances (Benzécri, 1973) est une méthode statistique exploratoire utilisée pour analyser les relations entre des variables catégorielles. Son objectif est de mettre en évidence les relations entre les différentes modalités des variables étudiées et de les représenter graphiquement. On peut considérer que l'analyse factorielle des correspondances multiples est homologue de l'analyse en composantes principales appliquée à des variables catégorielles.

À partir du tableau de contingence, dont la composition est précisée plus loin, l'algorithme va procéder à une réduction de la dimensionnalité des données et permettre leur projection sur un espace à deux dimensions. Le graphique obtenu (tel que celui de la figure 2) permet de visualiser les oppositions ou au contraire les rapprochements entre les différentes catégories étudiées (Beaudouin, 2016).

Dans le contexte de l'ADT, l'analyse des correspondances permet une représentation conjointe des termes et des variables dans un plan factoriel qui montre les attractions entre les termes du lexique et les modalités de variables.

Résultats et discussion

QR1 : quels sont les thèmes abordés dans notre corpus ?

Les prétraitements appliqués à notre corpus nous permettent d'obtenir une matrice termes-documents qui croise 2407 termes et 5358 segments. Le *clustering* de Reinert est réalisé dans R à l'aide de la bibliothèque Rainette. Après plusieurs essais, pour éviter des clusters trop volumineux (n > 1000), nous décidons de diviser notre corpus en un total de 34 clusters.

Ainsi que le présente la figure 1, nous constatons que la taille des clusters produits est très inégale.

Pour la suite des analyses, nous décidons de ne retenir que les dix premiers clusters contenant plus de 200 segments et présentés ci-après par ordre décroissant du nombre de segments. Pour chacun d'entre eux, nous avons analysé les segments où apparaissent les termes surreprésentés pour proposer un titre. Les cinq premiers mots surreprésentés de chaque cluster sont indiqués entre parenthèses après la description du cluster.

- Cluster 9 — déplacements avec directions : description d'activités de déplacement, parfois avec appui sur des touches, pour effectuer diverses actions comme tourner, avancer. Des directions sont indiquées, le référentiel est relatif au robot ou personnage qui se déplace (avancer, tourner, action, droite, touche).
- Cluster 11 — déplacements sur des cases : description d'activités de déplacement sur quadrillage, depuis une case de départ à un point d'arrivée en suivant un certain parcours. Le référentiel est absolu et ne fait pas référence à la position du personnage ou du robot (case, arrivée, départ, point, fleur).
- Cluster 4 — séquence pédagogique, contexte et apprentissage : présentation d'une séquence pédagogique composée de plusieurs leçons avec une mise en contexte (culture numérique, monde numérique, science informatique) et des apprentissages qui y sont liés (séquence, numérique, apprentissage, printemps, discipline).
- Cluster 17 — institutionnalisation : moments collectifs de mise en commun, d'institutionnalisation (commun, mise, institutionnalisation, collectif, réponse).
- Cluster 19 — activité enseignants et élèves : description de l'activité des enseignants et des élèves et de leurs interactions ; explicitation des consignes, organisation du travail seul, en binôme (élève, enseignant, recherche, consigne, groupe).

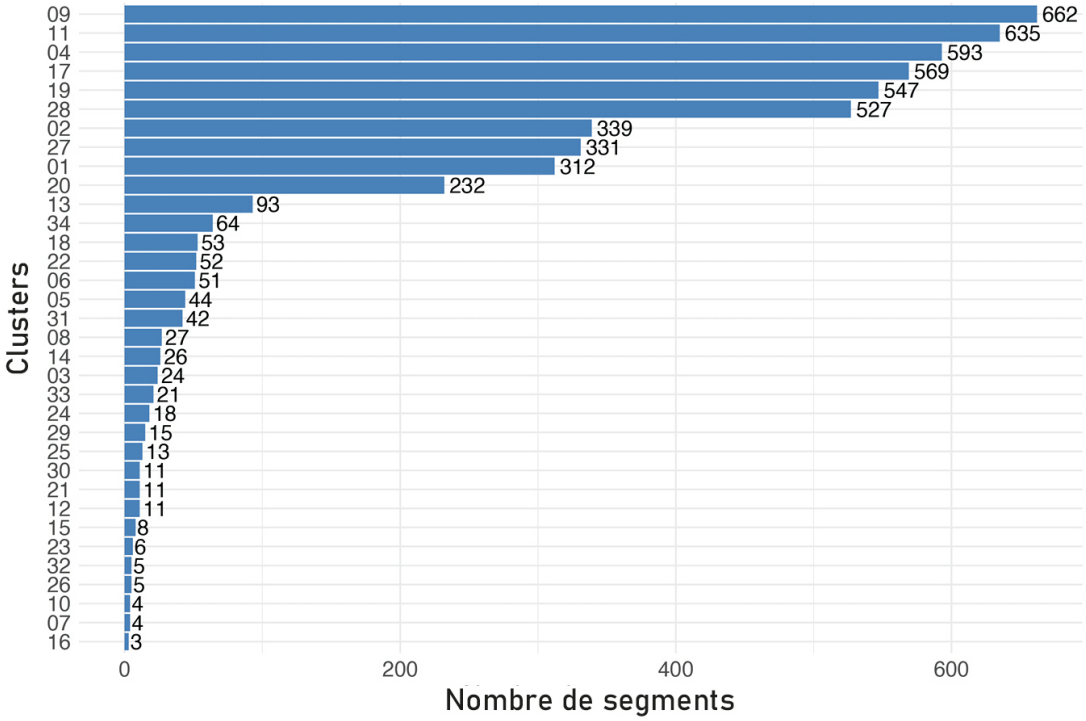

Figure 1 – Les 34 clusters de la classification de Reinert et leurs tailles.

— Cluster 28 — séance, phase, matériel et objectifs : segments en lien avec la présentation d'une ou de plusieurs séances d'enseignement avec le matériel, le découpage en phases d'un certain nombre de minutes, les objectifs et l'évaluation (séance, matériel, minute, phase, objectif).
— Cluster 2 — domaines et compétences du socle : éléments du socle de compétences, notamment dans le domaine scientifique et mathématique, en référence à une leçon. Liens avec le langage, la communication (domaine, compétence, exprimer, socle, scientifique).
— Cluster 27 — rappels et réinvestissement : activités de réinvestissement, rappel de la séance précédente ; mention de la tablette soit dans une liste de matériels, soit dans la description d'une activité ; logiciel ScratchJr ou autres logiciels et applications (précédent, rappel, tablette, réinvestissement, séance).
— Cluster 1 — cycle et repérage dans l'espace : segments qui font mention du ou des cycle(s) d'enseignement pour lesquels ces activités sont prévues. Ils font également allusion au repérage dans l'espace ou le plan ainsi qu'à des déplacements qu'il s'agit de décrire (cycle, espace, repérer, relation, familier).

— Cluster 20 — difficultés, autonomie et remédiation : gestion des difficultés des élèves par l'enseignant, remédiation ; différenciation et autonomie (difficulté, auprès, autonomie, travail, moment).

L'interprétation du sens de ces clusters nous permet de répondre à notre première question de recherche : que pouvons-nous dire des thématiques abordées par les enseignants dans les ressources pédagogiques qu'ils ont produites pour enseigner l'informatique ?

— La première thématique se rapporte à l'échelle de la séquence, en tant qu'enchaînement de séances. Les segments qui s'y rattachent se situent plutôt tout au début des ressources desquelles ils proviennent. Ils donnent le contexte général et mentionnent les apprentissages visés (cluster 4). Nous situons cette thématique dans une granularité plus grossière que les suivantes et la nommons « contexte général d'une séquence ».
— Une deuxième thématique concerne l'échelle de la leçon, donc avec une granularité plus fine que la précédente, et est en lien avec les domaines et compétences de la séance, ses phases, son matériel et ses objectifs, ainsi qu'avec les activités de l'enseignant et des élèves (clusters 2, 19 et 28). Nous l'intitulons « caractéristiques d'une leçon ».
— Une troisième thématique est constituée par les moments clés d'une leçon autour de l'institutionnalisation, de la gestion des difficultés et de la remédiation, ainsi que les rappels et réinvestissements (clusters 17, 20 et 27). Nous lui donnons le nom de « moments clés d'une leçon ».
— Une quatrième thématique est celle des déplacements, typique des activités d'apprentissage de la programmation chez des enfants, principalement sous forme d'activités débranchées (jeu du robot) ou de robotique (BeeBot, Bluebot) (clusters 9 et 11). Nous la nommons « notions de déplacement ».

À noter que le cluster 1 – cycle et repérage dans l'espace – est un peu à la frontière des première, seconde et quatrième thématiques, puisqu'il mentionne des éléments plutôt généraux en lien avec la séquence, d'autres plus spécifiquement en lien avec la séance, et d'autres encore en lien avec le repérage et les déplacements.

Nous avons donc quatre thématiques que l'on pourrait assimiler à différents plans de cadrage sur l'objet d'enseignement : un plan d'ensemble à l'échelle de la séquence composée de plusieurs leçons, un plan moyen à l'échelle d'une leçon ; deux plans rapprochés sur les moments clés d'une leçon et sur les notions de déplacements.

QR2 : qu'est-ce que l'analyse du lexique composant ces ressources nous apprend des connaissances des enseignants ?

Les résultats de nos analyses des ressources pédagogiques montrent que les enseignants utilisent davantage de termes informatiques dans les parties en lien avec des savoirs institutionnalisés que dans celles en lien avec des connaissances contextualisées.

Nous allons expliciter ce résultat en explorant plus avant la signification de chacune des thématiques, ainsi que des clusters qui s'y rattachent, par l'analyse des segments qui les composent et par les termes surreprésentés.

Les termes assimilés à des notions informatiques sont surreprésentés dans la première thématique portant sur l'échelle de la séquence cluster 4 — séquence pédagogique, contexte et apprentissage[4] et en partie cluster 1 — cycle et repérage dans l'espace[5]. En dehors de cette thématique, très peu de termes informatiques apparaissent ailleurs. On trouve le terme de « boucle » dans la quatrième thématique des déplacements (cluster 9 — déplacements avec directions).

Les noms des outils employés figurent parmi les termes surreprésentés des thématiques des déplacements et des moments clés (clusters 11, 27 et 20) : « BeeBot », « ScratchJr » et « Bluebot » respectivement. Il peut sembler étonnant que le terme « BeeBot » soit associé au cluster 11, en lien avec des déplacements dans un référentiel absolu, alors qu'on aurait plutôt tendance à l'associer à un référentiel relatif que l'on retrouve dans le cluster 9, au vu de la manière dont ce robot se programme avec des boutons gauche-droite, notamment. 15 % des segments du cluster 11 contiennent le terme « BeeBot », contre 4 % seulement du cluster 9. De façon similaire, l'association entre le terme « BeeBot » et le cluster 20 — difficultés, autonomie et remédiation surprend. Nous faisons l'hypothèse qu'elle est due plutôt au contexte de formation qu'à une spécificité de ce robot.

Ces analyses apportent des éléments de réponse à notre seconde question de recherche : qu'est-ce que l'analyse du lexique composant ces ressources nous apprend au sujet des connaissances des enseignants pour enseigner l'informatique ?

Les ressources pédagogiques produites par des enseignants constituent des traces de leurs connaissances didactiques-en-acte (ePCK). L'analyse de leur lexique nous montre que les enseignants utilisent davantage de termes informatiques lorsqu'ils sont dans une vision éloignée de l'activité de terrain, de l'acte d'enseigner. C'est dans la première thématique « contexte général d'une séquence » et du cluster 4 — séquence pédagogique, contexte et apprentissage, qui traitent de la séquence d'enseignement sous forme d'une

4. Avec les termes informatique, programmation, cryptage, machine, technologie, décryptage, algorithme, débrancher.

5. Avec les termes algorithmique, programme, coder, programmer, automatique.

introduction et d'une mise en contexte d'une série de leçons, que le plus grand nombre de termes informatiques apparaissent. Souvent, les termes font partie d'expressions reprises à l'identique d'objectifs inscrits dans les textes officiels. Selon le cadre conceptuel des PCK, les éléments tirés des programmes ont à voir avec la connaissance du curriculum, qui recouvre des connaissances institutionnalisées. Celles-ci peuvent également être vues comme des connaissances didactiques collectives ou *collective* PCK (cPCK), une base de connaissance spécialisée qui est partagée par un groupe de professionnels (Carlson *et al.*, 2019).

Notre étude montre que lorsque l'on se rapproche du niveau de la séance elle-même, du détail des activités de classe et de leurs spécificités, comme dans les deuxième et troisième thématiques, « caractéristiques d'une leçon » et « moments clés d'une leçon », les termes informatiques ressortent beaucoup moins. Cela ne veut pas dire qu'ils n'existent pas ou n'y sont pas du tout présents, mais qu'ils ne sont pas surreprésentés dans l'un ou l'autre des clusters constituant ces parties de productions. Selon le cadre des PCK, nous rapprocherions ces thématiques du niveau des ePCK, connaissances didactiques-en-acte, définies comme les connaissances spécifiques d'un enseignant dans un contexte particulier, avec un groupe d'élèves déterminé ayant pour but d'apprendre un concept ou un aspect particulier d'une discipline (Carlson *et al.*, 2019).

Nous voyons une possible convergence dans la distinction entre les connaissances didactiques collectives et en-acte (cPCK *vs.* ePCK) du modèle de consensus affiné des PCK (Carlson *et al.*, 2019) et la distinction savoir/connaissance du cadre de la didactique française des mathématiques (Margolinas, 2014). Pour celle-ci, les savoirs existent en tant que texte écrit et constituent une construction sociale, partagée, validée. On les retrouve, par exemple, dans les curricula des différentes disciplines. Les connaissances, elles, sont davantage personnelles et incarnées. Elles représentent une mise en contexte d'un savoir institutionnel par un acteur dans une situation bien définie.

Dans notre corpus, les portions de textes en lien avec la première thématique (cluster 4 — séquence pédagogique, contexte et apprentissage) portant un regard d'ensemble sur la séquence contiennent un certain nombre de notions informatiques et peuvent être rapprochées de savoirs ou de connaissances didactiques collectives (cPCK), tandis que les portions de texte portant un regard rapproché sur la leçon ne contiennent que peu de notions informatiques et sont à rapprocher de connaissances ou de connaissances didactiques-en-acte (ePCK).

Pour ces raisons, nous pouvons dire que les enseignants utilisent davantage de notions informatiques lorsqu'ils considèrent la séquence dans son ensemble et que l'on a plutôt à faire à des savoirs

institutionnalisés. Ils en utilisent moins lorsqu'ils se rapprochent du détail de la leçon et que ce sont des connaissances contextualisées qui sont en jeu.

QR3 : quelles relations entre le lexique et les modalités de variables du corpus ?

Pour répondre à notre troisième question de recherche (QR3), nous opérons une analyse des correspondances sur les données lexicales tirées de notre corpus. Nous souhaitons observer les relations qu'il peut y avoir entre les mots du lexique et les caractéristiques des productions. Compte tenu du nombre plus important de termes dans les productions issues de la formation qui a duré 8 h (Besançon) et des biais que cela risquerait de générer au niveau des interprétations, nous décidons de ne pas tenir compte de la variable liée au type de formation suivie.

Nous constituons donc un tableau lexical des questions, selon la méthode décrite par Cibois (1990), qui croise les mots du lexique avec l'ensemble des modalités de nos variables. Il est constitué des 21 modalités de variables en lignes et des 2407 mots du lexique en colonnes. Chaque cellule compte le nombre d'occurrences du mot du lexique dans les productions associées avec la modalité de la variable concernée. Nous avons conservé les variables : « domaine informatique », « genre », « type d'activité », « outil » et « niveau scolaire des élèves ».

Nous appliquons l'analyse des correspondances sur un tableau de 18 modalités et 77 termes. Les deux premiers axes produits par l'analyse des correspondances représentent une inertie cumulée de 56 %.

Nous présentons sur la figure 2 les associations et oppositions apparentes du premier plan factoriel.

Lors d'une analyse des correspondances, l'interprétation se base sur les distances entre les objets. Les proximités au niveau des positions sur le plan rendent compte de potentielles associations entre des termes ou des caractéristiques des ressources. Au contraire, des positions opposées par rapport aux axes représentent des oppositions au niveau du lexique ou des variables.

Dans une première lecture correspondant à l'axe 1 (horizontal) de l'analyse des correspondances, les activités de robotique et le jeu du robot sont plutôt en lien avec le niveau scolaire élémentaire et un vocabulaire de découverte, alors que les activités de type micromonde et de programmation visuelle visent plutôt les niveaux scolaires supérieurs et sont en lien avec un vocabulaire de déplacement, de repérage, d'informatique et de mathématique.

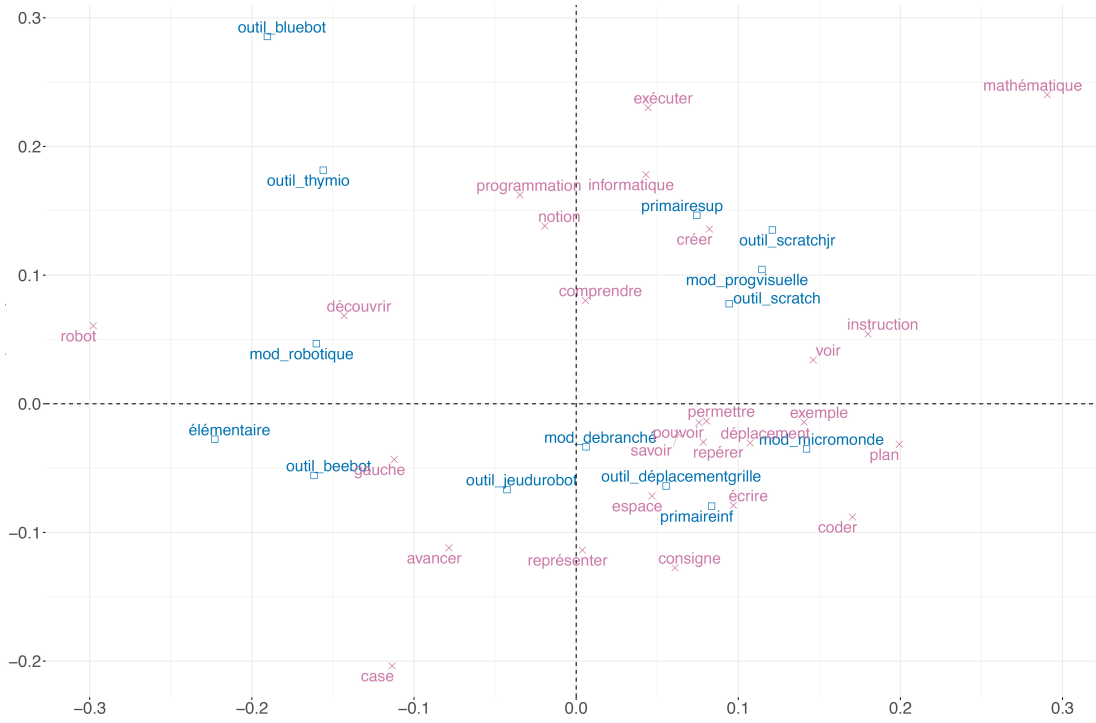

Figure 2 – Analyse des correspondances croisant le lexique et les variables. Le lexique est en violet avec des puces en croix ; les variables sont en bleu avec des puces carrées. Les 14 modalités de variable et les 26 termes du lexique avec les plus hautes valeurs de $\cos^2$ sont représentés.

Selon une seconde lecture correspondant à l'axe 2 (vertical) de l'analyse des correspondances, les activités de robotique et de programmation visuelle sont plutôt mises en lien avec le niveau scolaire du primaire supérieur et un vocabulaire informatique, alors que les activités débranchées sont plutôt en lien avec les niveaux scolaires inférieurs et un vocabulaire de déplacement.

L'axe 1 oppose la robotique dans les classes de niveau élémentaire et les autres activités de type programmation visuelle et micromonde avec des élèves plus âgés.

À gauche de l'axe, on trouve le type d'activité robotique, le niveau scolaire élémentaire (3-6 ans), les outils Thymio et BeeBot, le jeu du robot. À droite de l'axe se situent les types d'activité micromonde et programmation visuelle, le niveau de scolarité primaire inférieur (6-9 ans), les déplacements sur grille, les outils ScratchJr et Scratch.

Au niveau du lexique, à la robotique sont associés les termes « robot », « découvrir » et « gauche ». Aux activités de programmation visuelle et micromonde sont associés les termes « coder », « déplacement », « instruction », « plan », « savoir », « repérer » et « mathématique ».

L'axe 2 oppose quant à lui la programmation visuelle et la robotique avec des élèves de la fin du primaire aux activités débranchées avec des élèves plus jeunes.

Au-dessus de l'axe, on trouve les outils Bluebot, Thymio et ScratchJr, le niveau scolaire primaire supérieur (9-11 ans) et le type d'activité programmation visuelle. Au-dessous de l'axe, on trouve le type d'activité débranché, les déplacements sur grille et le jeu du robot, ainsi que le niveau scolaire primaire inférieur.

En ce qui concerne le lexique, les termes « notion », « programmation », « comprendre », « informatique », « exécuter », « créer », « programme », « machine » sont à proximité de la programmation visuelle, alors que les termes « consigne », « représenter », « case », « espace », « parcours » et « avancer » sont du côté des activités débranchées.

Il est à noter que les variables de genre et de domaine de l'informatique ne ressortent pas de ces analyses et ne constituent pas des facteurs de différenciation en lien avec le lexique utilisé dans nos ressources.

Conclusion

Cette recherche constitue la quatrième et dernière étape d'un processus d'analyse de ressources pédagogiques pour enseigner l'informatique produites par des étudiants-futurs enseignants de l'école primaire. Ces ressources constituent des traces de leurs connaissances didactiques-en-acte et l'analyse de leur lexique nous donne à voir le contenu et l'organisation de ces connaissances.

Nous avons appliqué des méthodes d'analyse des données textuelles telles que la classification de Reinert et l'analyse des correspondances pour investiguer le lexique présent dans ces ressources, extraire les grandes thématiques abordées et établir des liens entre le lexique et certaines caractéristiques de notre corpus.

En ce qui concerne notre première question qui s'intéresse aux thématiques abordées par les enseignants auteurs de ces ressources, nous faisons le constat que nous pouvons catégoriser ces thématiques en fonction de leur granularité : une première thématique concerne la séquence pédagogique en tant que suite de séances (plan d'ensemble), une seconde thématique est située à une échelle plus fine de la séance (plan moyen). Deux autres thématiques sont en lien avec des moments clés d'une leçon et des activités de déplacement (plans rapprochés).

Pour répondre à notre deuxième question sur ce que ce lexique nous apprend au sujet des connaissances didactiques des enseignants pour enseigner l'informatique, nous faisons le constat

que les termes informatiques sont davantage présents dans les clusters qui sont éloignés de la granularité fine de la séance. C'est lorsqu'ils présentent le contexte général d'une séquence pédagogique que les enseignants mentionnent des termes informatiques, notamment en tant que savoirs institutionnalisés. Mais plus l'on se rapproche du niveau de la séance en classe et des connaissances contextualisées, moins ces termes apparaissent. Nous pouvons faire l'hypothèse que la présence des termes informatiques en tant que savoirs institutionnalisés représente une influence du contexte de la formation et que les étudiants – futurs enseignants ont répondu à une nécessité de formation de coller au curriculum prescrit. Ceci pourrait constituer une future piste d'investigation, notamment par l'analyse des supports de formation. En ce qui concerne l'absence de termes informatiques au niveau de la leçon, nous pouvons faire l'hypothèse que les enseignants de l'école primaire opèrent une sorte de transposition didactique-en-acte, simplifiant leur expression pour se rapprocher de ce qu'ils considèrent comme accessible aux élèves.

Dans les résultats des recherches antérieures (Drot-Delange *et al.*, 2021), des difficultés de compréhension des concepts informatiques de la part des enseignants ont été relevées. En explicitant davantage ces notions au plus près des activités mises en œuvre et pas seulement dans les objectifs généraux de la séquence, les enseignants novices seraient plus à même de repérer de possibles incohérences.

Enfin, la troisième question nous amène à considérer les relations que l'on peut établir entre le lexique présent dans les ressources, le type d'activités, les outils et le niveau scolaire des élèves. Nous faisons le constat que les activités de robotique, lorsqu'elles s'adressent à des élèves de niveau scolaire inférieur, sont plutôt orientées vers de la découverte. Les activités débranchées, elles, sont plutôt en lien avec des notions de déplacement. Les activités de programmation et de robotique, lorsqu'elles concernent des élèves de niveau scolaire supérieur, sont quant à elles les plus portées vers des notions informatiques. Ces résultats constituent autant de pistes de recherche pour la didactique de l'informatique et pour la formation des enseignants.

Recommandations et leçons tirées

Au niveau didactique, les conclusions de cette étude montrant une surreprésentation de termes informatiques dans les portions de ressources éloignées de la leçon mettent en lumière l'absence relative de notions informatiques dans le niveau de granularité fine des activités en classe. Cela nous amène à nous interroger sur l'explicitation des concepts informatiques durant les activités avec les élèves : comment les élèves s'approprient-ils ces concepts si le vocabulaire spécifique n'est que peu sollicité ? Quels liens peut-on établir entre apprentissages des élèves et langage, vocabulaire et notions ?

En ce qui concerne les connaissances professionnelles des enseignants pour enseigner l'informatique, nous estimons que l'explicitation des concepts constitue un enjeu de formation : comment rendre ces concepts actionnables par les enseignants dans leurs activités, afin qu'ils puissent y faire référence non seulement dans un contexte de connaissances institutionnelles, mais également au cœur des activités avec leurs élèves ?

En termes méthodologiques, les résultats de cette étude nous amènent à penser que les techniques issues de l'analyse de données textuelles sont porteuses d'un potentiel non négligeable et qu'il serait intéressant d'en explorer davantage les possibilités dans le domaine de la recherche en éducation et plus particulièrement en didactique de l'informatique. Il serait souhaitable de pouvoir les expérimenter sur d'autres supports textuels, comme dans l'analyse de transcriptions d'entretiens, de discours d'enseignants ou d'élèves en classe.

Références

Beaudouin, V. (2016). Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données, *Journées internationales d'Analyse statistique des Données Textuelle*, Nice, France, p. 17-27. https://hal.science/hal-01376938

Benzécri, J. P. (1973). *L'analyse des données, Tome 1 : La Taxinomie ; Tome 2 : L'analyse des correspondances,* Dunod.

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Kam Ho Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., (. . .) Wilson, C. D. (2019). « The refined consensus model of pedagogical content knowledge in science education », dans A. Hume, R. Cooper, et A. Borowski (dir.), *Repositioning Pedagogical Content Knowledge in Teachers' Knowledge for Teaching Science*, Springer, p. 77-94. https://doi.org/10.1007/978-981-13-5898-2_2

Cibois, P. (1990). Éclairer le vocabulaire des questions ouvertes par les questions fermées : Le tableau lexical des questions, *Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique*, vol. 26, n° 1, p. 12-21. https://doi.org/10.1177/075910639002600102

Drot-Delange, B., Parriaux, G., et Reffay, C. (2021). Futurs enseignants de l'école primaire : connaissances des stratégies d'enseignement, curriculaires et disciplinaires pour l'enseignement de la programmation, *Recherches en didactique des sciences et des technologies*, vol. 23, p. 55-76. https://doi.org/10.4000/rdst.3685

Kermen, I., et Izquierdo-Aymerich, M. (2017). Connaissances professionnelles didactiques des enseignants de sciences : un thème de recherche encore récent dans les recherches francophones, *Recherches en didactique des sciences et des technologies*, vol. 15, p. 9-32. https://doi.org/10.4000/rdst.1479

Lebart, L., Pincemin, B., et Poudat, C. (2019). *Analyse des données textuelles*, Presses de l'université du Québec. https://doi.org/10.2307/j.ctvq4bxws

Magnusson, S., Krajcik, J., et Borko, H. (1999). « Nature, sources, and development of pedagogical content knowledge for science teaching », dans *Examining pedagogical content knowledge*, Springer, p. 95-132. https://doi.org/10.1007/0-306-47217-1_4

Margolinas, C. (2014). Connaissance et savoir. Concepts didactiques et perspectives sociologiques ? *Revue française de pédagogie. Recherches en éducation*, vol. 188, p. 13-22. https://doi.org/10.4000/rfp.4530

Parriaux, G., Reffay, C., Drot-Delange, B., et Khaneboubi, M. (2023). « Teachers' knowledge in informatics—Exploring educational robotics resources through the lens of textual data analysis », dans J.-P. Pellet et G. Parriaux (dir.), *Informatics in Schools. Beyond Bits and Bytes : Nurturing Informatics Intelligence in Education*, ISSEP 2023, *Lecture Notes in Computer Science*, vol. 14 296, p. 126-138. https://doi.org/10.1007/978-3-031-44900-0_10

R Core Team, R. (2022). *R : A language and environment for statistical computing*.

Reffay, C., Parriaux, G., Drot-Delange, B., et Khaneboubi, M. (2023). « Robotics in primary education : A lexical analysis of teachers' resources across robots », dans T. Keane, C. Lewin, T. Brinda, et R. Bottino (dir.), *Towards a Collaborative Society through Creative Learning*, WCCE 2022, *IFIP Advances in Information and Communication Technology*, vol. 685. https://doi.org/10.1007/978-3-031-43393-1_-20

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte, *Cahiers de l'analyse des données*, vol. 8, n° 2, p. 187-198.

Shulman, L. S. (1986). Those who understand : Knowledge growth in teaching, *Educational researcher*, vol. 15, n° 2, p. 4-14. https://doi.org/10.30827/profesorado.v23i3.11230

Shulman, L.-S. (2007). Ceux qui comprennent. Le développement de la connaissance dans l'enseignement, *Éducation et didactique*, vol. 1, n° 1, p. 97-114. https://doi.org/10.4000/educationdidactique.121

Pour citer ce chapitre :

Parriaux, Gabriel, Reffay, Christophe, Drot-Delange, Béatrice, et Khaneboubi, Mehdi (2024). « Connaissances pour enseigner l'informatique : analyse textuelle de productions d'enseignants de l'école primaire », dans Cédric Fluckiger, Laetitia Boulc'h, Sandra Nogry et Christophe Reffay (dir.), *Enseigner, apprendre, former à l'informatique à l'école : regards croisés*, Université Paris Cité, p. 173-191. https://doi.org/10.53480/2024iecare09w