Analyse textuelle avec IRaMuTeQ et interprétations référentielles des programmes officiels de mathématiques en quatrième

Salone Jean-Jacques
EA 4671 ADEF, Aix-Marseille Université

Résumé
Une analyse lexicométrique avec le logiciel libre IRaMuTeQ est conduite sur les textes officiels du programme des mathématiques au collège en classe de quatrième. Elle fait apparaître l'existence de nombreuses références à des savoirs non disciplinaires. Au delà des autres disciplines scientifiques, toutes les disciplines enseignées sont concernées. En outre, non seulement cette ouverture écologique des savoirs porte sur le monde et le vivant, mais elle passe aussi par une prise en compte des pratiques personnelles des élèves dans la vie courante.

Mots-clés
Analyse textuelle multivariée, IRaMuTeQ, programmes officiels, mathématiques, références praxéologiques

Introduction

Au centre des systèmes didactiques élémentaires que sont les classes se trouvent des élèves, des professeurs et des savoirs à enseigner et apprendre. Ces savoirs, ou plutôt ces praxéologies, sont intimement liées aux agents du système car elles constituent tout ou partie de ce qu'elles savent déjà faire ou dire. Ainsi, dans une classe donnée, parmi les praxéologies relatives à la discipline enseignée qui sont en jeu, certaines sont routinières pour les élèves et d'autres sont problématiques. Plus généralement, la Théorie Anthropologique du Didactique postule que les institutions, comme le sont des personnes, une classe ou encore les noosphères du Ministère de l'Éducation Nationale en France, entretiennent des rapports aux savoirs qui fondent leurs équipements praxéologiques.
L'avancée du temps didactique dans la classe amène, comme réponses aux questions posées, des rencontres successives avec de nombreuses œuvres. Ces rencontres, certaines étant officiellement programmées, peuvent rester dans le champ disciplinaire concerné. Mais elles peuvent aussi s'en écarter, tendant alors à l'ouverture écologique du système aux autres disciplines, voire au monde. Ainsi, les rapports aux savoirs évoluent et modifient les équipements praxéologiques personnels sous l'influence des référencements praxéologiques réalisés dans la classe. Quelles sont alors ces sources de références non disciplinaires, c'est la question qui sera examinée ici à partir d'un texte noosphérique officiel, le Bulletin Officiel Spécial N°6 de 2008 (Ministère de l'Éducation Nationale, 2008), qui délimite les programmes de mathématiques au collège. Recommande t-il des ouvertures écologiques de la discipline ? Et si oui, quelles sources référentielles externes suggère-t-il ?
Après avoir rappelé les cadres théoriques et méthodologiques de l'étude, une analyse textuelle sera conduite pour étayer nos réponses à ces questions. Le logiciel libre et open source IRaMuTeQ sera employé pour cela.

Cadres théoriques et méthodologiques

Les personnes et plus généralement les institutions entretiennent des rapports aux savoirs qui se laissent observer par les actes, discursifs ou non, qu'elles réalisent et que leurs équipements praxéologiques permettent (Chevallard, 1988). L'évolution des rapports aux savoirs est induite par les actes de référencement, c'est-à-dire les mises en relation que les institutions font via leurs actes entre leurs propres savoirs et les savoirs en général. Plus formellement, si H est l'espace des institutions humaines et P l'espace des praxéologies, les rapports aux savoirs sont définis comme

des couples de l'espace produit HxP et les références comme des couples de l'espace $(HxP)^2$. Les niveaux de codétermination didactique (Chevallard, 2007) permettent alors de catégoriser les sources de référence en distinguant ce qui est propre à une discipline (niveaux inférieurs) de ce qui ne l'est pas (niveaux supérieurs), c'est-à-dire des références externes. L'existence de références externes dans un texte ou un discours didactique constitue une ouverture écologique des savoirs.
Ainsi le BOS6 (Ministère de l'éducation nationale, 2008) définit le rapport institutionnel de l'élève générique du collège aux savoirs mathématiques, ainsi qu'une partie du rapport personnel du professeur dans ce que sa praxéologie a justement de professionnel. Il est découpé en six grandes parties : une introduction communes aux disciplines scientifiques (mathématiques, sciences et vie de la Terre, sciences physiques, technologie), un préambule spécifique aux mathématiques du collège, et les quatre programmes, celui de la sixième, puis ceux de la cinquième, de la quatrième et de la troisième.
Ce texte est le patient d'une approche clinique (Leutenegger, 2009) de la problématique, chez qui les symptômes que sont les références vont être recherchés et interprétés. Il n'est pas analysé dans son intégralité, mais à partir de deux extraits : un premier corpus est constitué du programme de quatrième seul, et un corpus élargi contient en plus l'introduction et le préambule. Ces corpus sont formatés pour pouvoir être lus par le logiciel IRaMuTeQ, Version 0.6 alpha 3, conçu par Ratinaud P. et Dejean S. (Ratinaud, 2012 ou Ratinaud et Dejean, 2009) : déclaration de variables étoilées et de modalités, suppression des numéros de pages et suppression des titres de tableaux « Connaissances » « Capacités » « Commentaires » « Objectifs ». IRaMuTeQ poursuit ce formatage en ne conservant que certains caractères.
Trois algorithmes sont lancés sur chacun de ces deux corpus : une Classification Descendante Hiérarchique (CDH), une Analyse Factorielle de Correspondances (AFC) et une Analyse Des Similitudes (ADS). Des exemples d'analyses lexicométriques avec ces algorithmes sont proposés, en sciences de gestion, par Gavard-Perret, Gonzalez, Helme-Guizon, Labbé, Marchand et Reinert (2007).
La CDH et l'AFC proposent une approche globale du corpus. Après partitionnement de celui-ci, la CDH identifie des classes statistiquement indépendantes de mots (de formes). Ces classes sont interprétables grâce à leurs profils, qui sont caractérisés par des formes spécifiques corrélées entre elles. La CDH résume cela par un dendrogramme.
L'AFC, basée sur des calculs d'inertie du nuage de mots que constitue un corpus, fait davantage apparaître les oppositions ou rapprochement. Elle détermine pour cela des facteurs (des espaces propres de la matrice d'inertie) sur lesquels les formes se distribuent. À la notion d'appartenance à une classe se substitue ainsi celle de distance à un axe d'inertie. Les AFC proposées ici sont réalisées après lemmatisation et sont doubles. Leurs représentations graphiques du nuage de point sont bidimensionnelles, dans l'hyperplan définit par les deux premiers facteurs.
L'ADS envisage les corpus d'une façon complètement différente. L'approche est davantage locale, reposant sur des propriétés de connexité du corpus. Elle aboutit à une représentation graphique en arbre (maximal valué et connexe), où les nœuds sont les formes, et où il est possible de faire apparaître des communautés lexicales. Cet algorithme a tendance à renforcer les relations de voisinage entre les formes.

Analyse et interprétation du programme de quatrième seul

L'analyse statistique distingue 631 formes parmi 3073 occurrences, dont 297 hapax.
La forme active lemmatisée d'effectif maximum est la forme « calcul », avec 34 occurrences.
Puis viennent les formes « nombre » (31 oc.), « utiliser » (27 oc.), « relatif » (23 oc.) et « triangle » (20 oc.).
Cette hiérarchisation fréquentielle des formes actives est caractéristique de la discipline d'appartenance de ce texte officiel : c'est un texte de mathématiques, même si la forme « mathématiques » n'en est qu'un hapax. Deux des quatre domaines de détermination du BOS6 sont présents, celui des « nombres et calculs » et celui de la « géométrie », ce dernier étant instancié par la forme « triangle ».

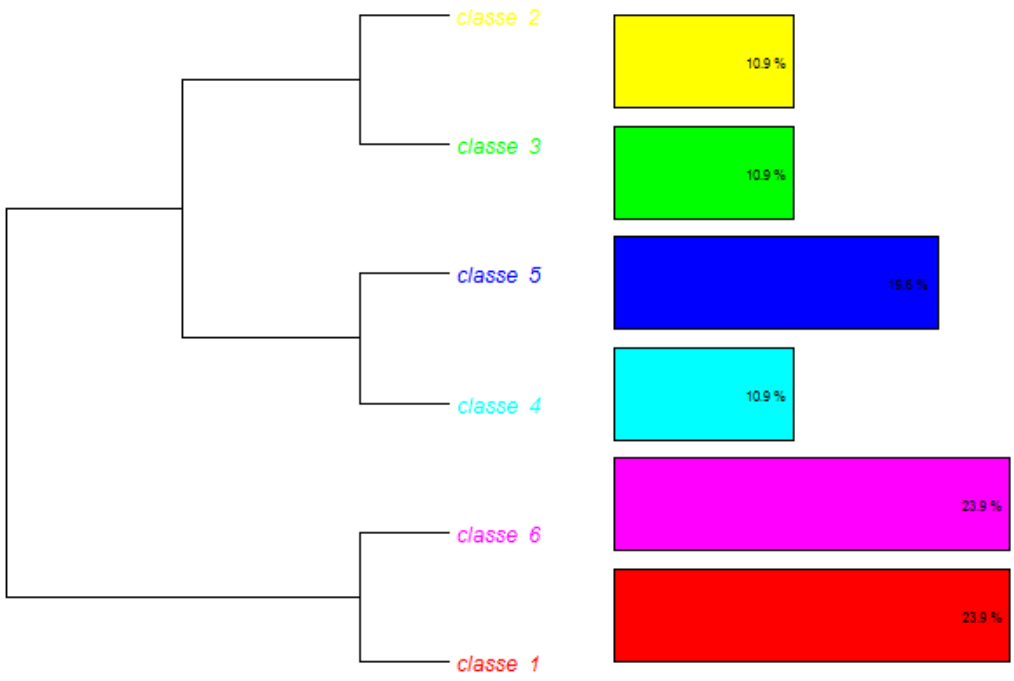

Figure 1 : Classification Descendante Hiérarchique du programme de mathématiques en quatrième par IRaMuTeQ

La CDH (fig. 1) distingue 6 classes de formes sur les 54,12% de segments de textes classés.
Les classes 1 et 6 sont les plus grandes avec toutes les deux 23,9% des formes. Leur regroupement constitue l'une des 2 branches du dendrogramme.
Les 4 classes restantes dans la deuxième branche sont hiérarchisées en 2 sous-branches, classes 4 et 5 (30,5% des formes) et classes 2 et 3 (21,8% des formes).
La classe 1 est caractérisée 12 formes actives :
« espace », « figure », « géométrie », « géométrique », « aire », « réduction », « agrandissement », « objet », « propriété », « relation », « symétrie » et « représentation ».
La classe 6 est caractérisée 13 formes actives :
« cercle », « rectangle », « triangle », « bissectrice », « caractériser », « angle », « côté », « droite », « Pythagore », « longueur », « donner », « construction » et « théorème ».
Ce monde lexical renvoie de façon évidente au domaine de la géométrie. Mais qu'est-ce qui différencie ces deux classes ?. La classe 6 est surtout composée d'objets, géométriques ou technologiques, tandis que la classe 1 regroupe des relations, conceptuelles ou opérationnelles, entre ces objets.
Les classes 4 et 5 contiennent respectivement 5 et 15 formes actives, les plus caractéristiques étant étant « décimal », « nombre », « écriture », « forme » et « relatif » pour la première, « exemple », « nombre », « notation », « parenthèse » et « numérique » pour la seconde. Le monde lexical de cet embranchement terminal correspond logiquement au deuxième domaine mathématique déjà repéré par l'analyse statistique, celui des « nombre et calculs ». La distinction entre objets et relations semble encore opérer.
Les classes 2 et 3 contiennent respectivement 7 et 10 formes actives, dont « proportionnalité », « thème », « convergence », « mettre » et « calculer » pour la première, « littéral », « utilisation », « expression », « résolution » et « calcul » pour la seconde. Cet embranchement terminal semble ne plus relever spécifiquement du monde mathématique, mais au contraire s'ouvrir vers les autres disciplines.
L'AFC (fig. 2) conforte et affine ces interprétations.

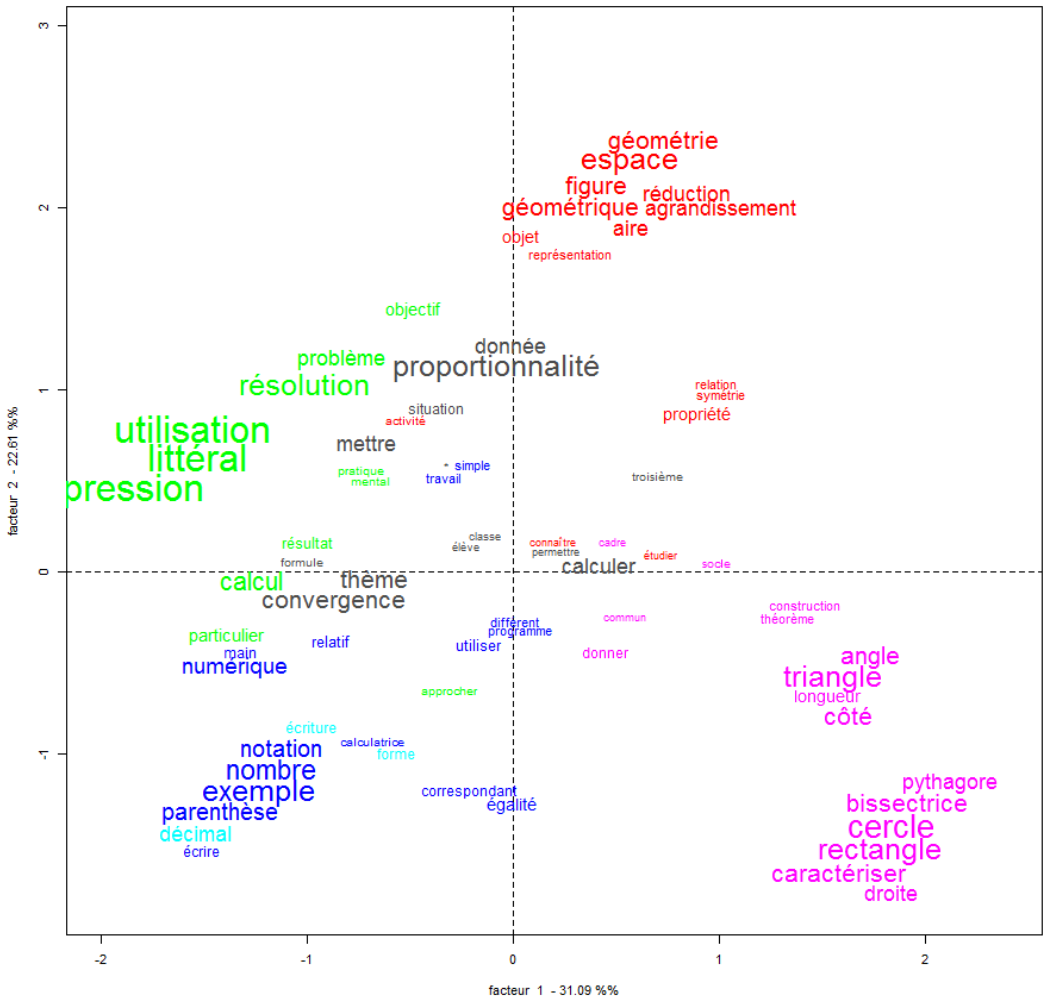

Figure 2 : Analyse Factorielle des Correspondances du programme de mathématiques en quatrième par IRaMuTeQ

Le premier facteur (31,09% de la masse du corpus), sépare nettement les classes 3,4,5 (abscisses négatives) des classes 1 et 6 (abscisses positives). Il retrouve la bipartition de la discipline en domaine géométrique et domaine numérique. La classe 2 est centrée pour ce facteur.

Le deuxième facteur (22,61%) assume davantage la distinction entre les objets et leurs relations . La classe 1 (ordonnée positive) y est nettement séparée des classes 4, 5 et 6 (ordonnées négatives), tandis que les classes 2 et 3 y apparaissent plus centrées.

La combinaison de ces deux facteurs fait que sur le graphique cinq zones apparaissent dans cette projection bidimensionnelle du corpus textuel :

- une zone à coordonnées négatives, en bas à gauche, regroupant les classes 4 et 5, relatives aux objets du domaine numérique,
- une zone à coordonnées positives, en haut à droite, qui isole la classe 1, relative aux relations géométriques,
- une zone à abscisses positives et ordonnées négatives, en bas à droite, où l'on trouve la classe 6 des objets géométriques,
- une zone centrale occupée essentiellement par la classe 2,
- une zone à abscisses négatives et ordonnées faiblement positives, en haut à gauche, où se localise la classe 3.

La place centrale de la classe 2 souligne sa position praxéologique intermédiaire entre les deux domaines mathématiques. Les formes « proportionnalité » et « calculer » renvoient ainsi l'une à un concept fondamental du programme de quatrième et l'autre à un genre de tâche majeur. Toutes deux évoquent les techniques fractionnaires étudiées à ce niveau scolaire, que ce soit pour elles-mêmes

ou comme outils de techniques géométriques autour du théorème de Thalès ou de la trigonométrie par exemple. Mais la classe 2 ne se contente pas de relier les domaines de la géométrie et du numérique par les secteurs d'étude particuliers que sont les fractions ou les triangles proportionnels. Elle renferme aussi les formes « thème » et « convergence » qui rappellent les nombreux « thèmes de convergence » que propose le texte officiel avec les autres disciplines. La classe 2 marque donc une ouverture écologique des programmes vers des praxéologies non mathématiques, à des niveaux de codétermination supérieurs.
Ce phénomène se retrouve dans la classe 3. Proche de la classe 2, sur le graphique et aussi par son lexique (avec « calcul » dans l'une et « calculer » dans l'autre), elle s'en distingue par un éloignement plus marqué avec les autres classes. Les formes « résolution » et « problèmes » y joue un rôle important et traduisent elles aussi une certaine ouverture des références praxéologiques.
Ainsi, les classes 2 et 3 semblent s'ouvrir vers des niveaux de codétermination didactique supérieurs, au delà de la discipline et de ses domaines géométrique ou numérique, vers des praxéologies plus générales.
L'ADS (fig. 3) va affiner encore un peu plus cette interprétation.

Figure 3 : Analyse Des Similitudes du programme de mathématiques en quatrième par IRaMuTeQ
À première vue, l'arbre se compose d'une branche principale dont le squelette est une suite de 7 formes : « situation », « calcul », « nombre », « relatif », « utiliser », « propriété », « triangle ». Les deux domaines des mathématique qui étaient identifiés dans l'AFC, la géométrie avec les classes 1 et 6, et le numérique avec les classes 4 et 5, deviennent dans l'ADS des communautés lexicales,

autour des formes « triangle » et « propriété » pour le premier, des formes « nombre » et « calcul » pour le second. Ils sont séparés par deux autres communautés : une autour de la forme « utiliser », l'autre autour de la forme « relatif ». En outre, ces deux dernières formes et la forme « problème » sous-tendent un triangle qui permet de relier la branche principale à une nouvelle communauté lexicale autour de la dernière de ces formes. L'interprétation conduit à retrouver là les classes 2 et 3 de l'AFC. Un examen plus détaillé fait apparaître que la communauté de « problème » se poursuit en deux branches autour de formes évoquant les programmes officiels, comme « socle » ou « objectif », puis autour de formes plus ouvertes sur le monde non mathématique, comme « représentations » ou « espace ». Cette tendance s'observe également dans la communauté de « utiliser » qui se termine sur des formes comme « convergence », « particulier » et « sentir ».
Repartant dans la branche principale de l'arbre du côté du monde numérique, on remarque là aussi un long prolongement de celle-ci vers des communautés de moins en moins spécifiques des mathématiques. Une boucle apparaît ainsi dans l'arbre autour de formes comme « situation » ou « élève », boucle dont un appendice renferme des formes qui n'avaient pas été retenues par l'AFC, les formes « vie », « courante » et « discipline ».
L'ADS semble donc étayer l'hypothèse de l'existence d'une ouverture écologique du programme officiel des mathématiques en quatrième vers des praxéologies de niveaux de codétermination supérieurs, impliquant le monde extérieur qui est lui aussi susceptible de fournir des sources de références. L'adjonction, au programme officiel de la classe de quatrième, de l'introduction et du préambule va amplifier ce phénomène.

Analyse et interprétation du corpus élargi

La CDH distingue 7 classes de formes sur les 63,66% de segments de textes classés.

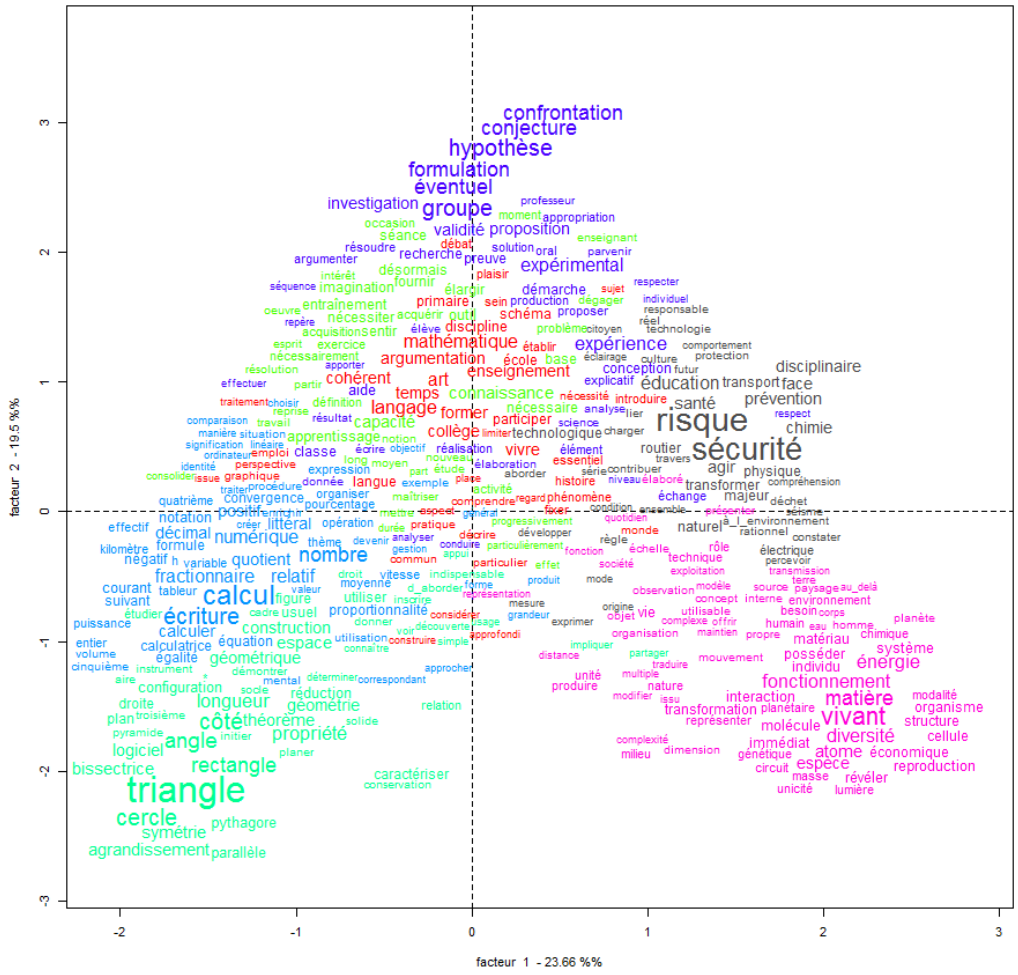

Figure 4 : Analyse Factorielle des Correspondances du corpus élargi par IRaMuTeQ

Le premier facteur de l'AFC (fig. 4) sépare nettement la zone des mathématiques, constituée des classes 4 (la géométrie) et 5 (le numérique), d'une zone explicitement relative aux praxéologies non mathématiques, caractéristique de la classe 7 et autour de formes comme « vivant » et « matière ». Cette zone est essentiellement apportée par la partie introduction du corpus, qui amplifie donc le phénomène d'ouverture écologique de la discipline à des praxéologies extérieures. Le graphe étoilé (fig. 5) à partir de la forme « vivant » souligne son ampleur.

Figure 5 : Graphe étoilé de la forme « vivant »

Le concordancier de la forme « vie » montre que ce mot est intriqué dans des expressions comme « vie courante » ou « vie quotidienne » :

« … au niveau des exigibles du socle commun, toute technicité est exclue puisque dans l'esprit général du socle, on se limite à des problèmes simples proches de la vie courante... »

« … le traitement de ces problèmes permet de souligner le sens, l'intérêt, la portée des connaissances mathématiques, que ce soit dans d'autres disciplines ou dans la vie quotidienne... »

Bien au delà de la simple question de l'élargissement des mathématiques aux proxéologies scientifiques, il s'agit donc d'intégrer les rapports personnels des élèves aux savoirs dans l'action conjointe de professeur et de la classe (Sensevy et Mercier, 2007).

Le deuxième facteur oppose ces trois zones à une troisième constituée par la classe 6, et caractérisée par des formes comme « hypothèse », « conjecture », « formulation », « investigation » ou « validité ». Il semble qu'est retrouvée là, de façon encore plus exacerbée, la distinction entre formes relatives à des objets et formes relatives à des concepts.

Les trois classes restantes, 1, 2 et 3, sont centrées sur le graphique et se mêlent aux autres. Peu discriminées par l'AFC, elles hébergent des formes non spécifiques du lexique mathématique, comme « langage », « art » et « collège » pour la classe 1, comme « risque », « sécurité » et « santé » pour la classe 2, et comme « connaissance », « capacité » et « apprentissage » pour la classe 3. Le préambule en particulier est caractéristique de la classe 3. Ces trois classes

s'interprètent elles aussi comme des ouvertures écologiques. Elles sont encore plus éloignées des praxéologies mathématiques que ne l'étaient les références scientifiques précédemment identifiées. Elles ouvre le système didactique sur le monde et le vivant, via les arts, les langues, la santé et la sécurité.
L'ADS (fig. 6) fait encore mieux apparaître ces cheminements praxéologiques.

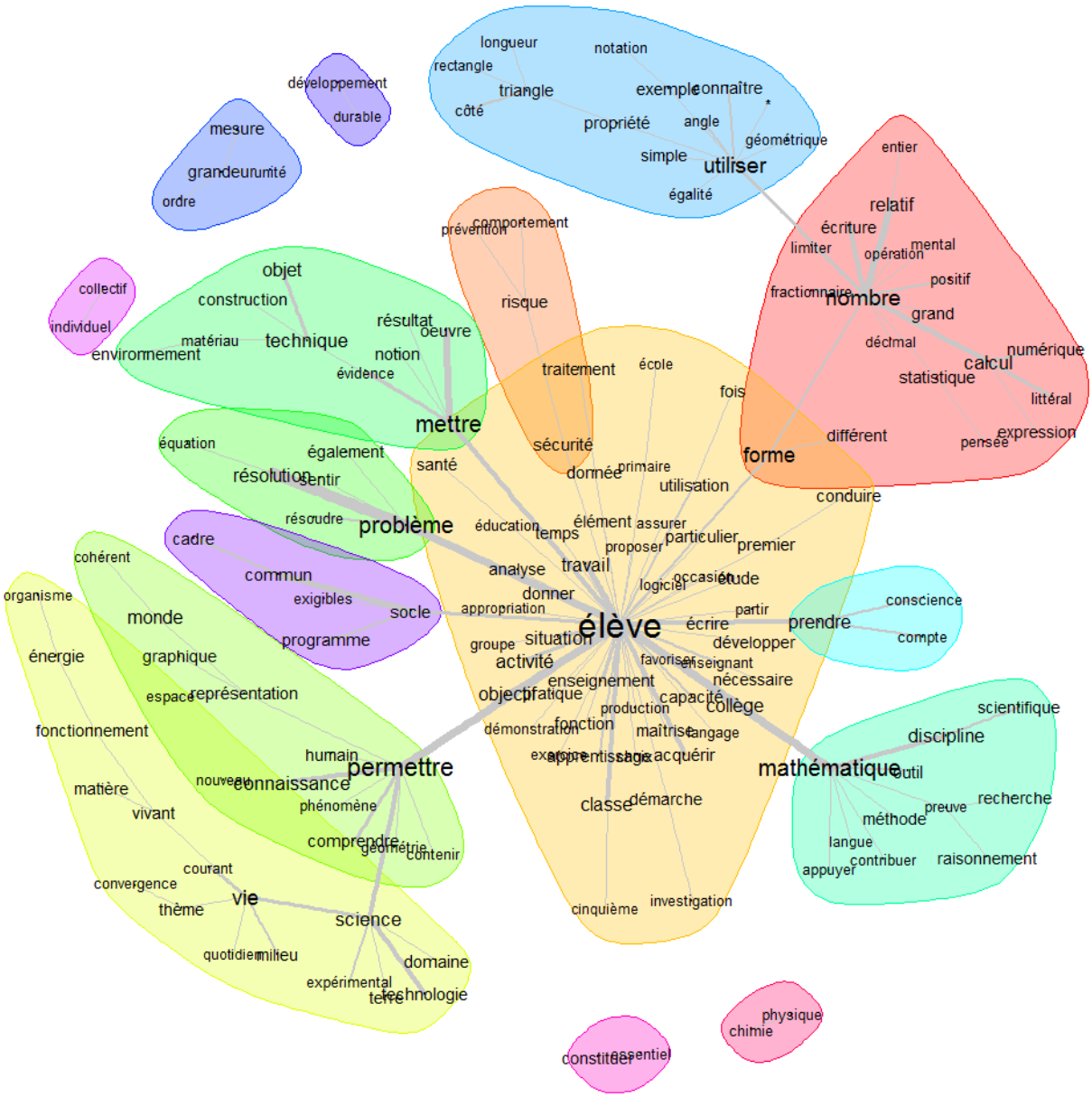

Figure 6 : Analyse Des Similitudes du programme de mathématiques en quatrième par IRaMuTeQ
Une communauté lexicale est à la racine de l'arbre, constituée autour de la forme « élève ». Trois branches principales se développent à partir de là. Une branche retrouve le lexique mathématique, d'abord avec une communauté autour de la forme « nombre », puis, via la forme « utiliser », avec une communauté géométrique instanciée par les formes « triangle » et « angle ». La forme « mathématique » se situe au centre d'une autre communauté sur une nouvelle branche, et son monde lexical renvoie au monde des concepts, avec des formes associées comme « raisonnement » ou « recherche ». On retrouve la classe 6.
Dans la troisième branche c'est la classe 7 des ouvertures écologiques au monde extérieur qui apparaît. Le forme « vie » y constitue une communauté lexicale terminale que l'on atteint via la communauté de la forme « permettre ». Cette dernière communauté héberge des formes comme « connaissance », « comprendre » ou « représentation » qui souligne son caractère sémantiquement

intermédiaire entre la praxéologie mathématique de l'élève générique et celles des autres mondes socio-culturels. D'autres formes que « vie » ou « vivant » caractérisent ces références externes, comme les formes « monde », « science » ou « Terre ».
Des branches secondaires partent encore de la communauté de « élève ». L'une retrouve le cadre contractuel du système didactique, avec les formes « programme », « socle », « commun » ou « exigible », une autre est relative à la « résolution » de « problèmes », une autre à la mise (forme « mettre ») en « œuvre » de « techniques » ou de « notions », et une dernière à la question de la « prévention » des « risques » et à la « sécurité ».
Enfin, d'autres petites communautés lexicales gravitent à la périphérie, dont deux sont remarquables quant aux ouvertures écologiques qu'elles suggèrent : la communauté autour des formes « développement » et « durable », et celle autour des formes « physique » et « chimie ».

Conclusion :

Les analyses textuelles conduites sur le BOS6 autour du programme de mathématiques en classe de quatrième du collège confirment l'existence d'ouvertures écologiques des savoirs dans ces textes officiels. Au delà des ouvertures aux autres disciplines scientifiques (Sciences physiques, Sciences et Vie de la Terre, Technologie) ou non scientifiques (Arts, langues) que les thèmes de convergence, entre autres, proposent aux professeurs, des références externes plus générales au monde et au vivant sont suggérées. Il s'agit de laisser une place aux praxéologies personnelles des élèves dans la construction des savoirs en classe, en se référant à des problèmes et des situations de leur vie quotidienne.
Qu'en est-il chez d'autres patients que les textes officiels ? La question est à l'étude, avec des analyses conduites sur les manuels scolaires agréés ou sur des transcriptions de séances d'enseignement réelles. Les premiers résultats tendent à confirmer l'existence de ce phénomène d'ouverture écologique quel que soit le corpus analysé, mais avec des profils très variables surtout lorsqu'il s'agit de transcription de séances. Dans ce dernier cas, l'équipement praxéologique personnel de l'enseignant est un paramètre vraisemblablement prépondérant.
Enfin une approche comparatiste est envisageable qui permettrait de détecter d'éventuelles spécificités disciplinaires, dans les textes ou dans les pratiques enseignantes.

Bibliographie

Chevallard, Y. (1988). *Esquisse d'une théorie formelle du didactique*. In C. Laborde (Ed.), *Actes du premier colloque franco-allemand de didactique des mathématiques et de l'informatique* (pp. 97-106). Grenoble : La Pensée Sauvage. Disponible sur internet à l'adresse : http://yves.chevallard.free.fr/spip/spip/IMG/pdf/Esquisse_d_une_theorie_formelle_du_didactique.pdf.

Chevallard, Y. (2007). *Passé et présent de la théorie anthropologique du didactique*. In L. Ruiz-Higueras, A. Estepa & F. Javier Garcia (Eds.), *Sociedad Escuela y Mathematicas : aportaniones de la Teoria Antropologica de la Didactico* (pp 705-746). Baeza (Espagne): Universidad de Jaen. Disponible sur internet à l'adresse : http://yves.chevallard.free.fr/spip/spip/IMG/pdf/Passe_et_present_de_la_TAD-2.pdf.

Gavard-Perret, ML., Gonzalez, C., Helme-Guizon, A., Labbé, S., Marchand, P. & Reinert, M. (2007). *Analyse statistique de données textuelles en sciences de gestion*. In C. Gauzente & D. Peyrat-Guillard (Coord.). Paris : éditions EMS, coll. Questions de société.

Leutenegger, F. (2009). *Le temps d'instruire*. Berne : Peter Lang.

Ministère de l'éducation nationale. (2008). *Bulletin officiel spécial n°6 du 28 août 2008, programmes de l'enseignement de mathématiques.* Disponible sur internet à l'adresse : http://cache.media.education.gouv.fr/file/special_6/52/5/Programme_math_33525.pdf.

Ratinaud, P. (2012). *Analyse Automatique de Textes*. Disponible sur internet à l'adresse :

http://repere.no-ip.org/Members/pratinaud/informatique/aat.pdf.

Ratinaud, P. & Déjean, S. (2009). *IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre*. Disponible sur internet à l'adresse : http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf.

Sensevy, G. & Mercier, A. (2007). *Agir ensemble : l'action didactique conjointe*. In G. Sensevy & A. Mercier (Eds.), *Agir ensemble:l'action didactique conjointe du professeur et des élèves* (pp. 187-211). Rennes : Presses Universitaires de Rennes.