L'analyse de similitude pour modéliser les CHD

Lucie Loubère

L'analyse de similitude pour modéliser les CHD

Lucie Loubère

Université de Toulouse – France

Abstract

The corpus analysis can be seen (among others) by down hierarchical classification type Reinnert , but also by analyzing similarities. If these two approaches are often used in parallel to the complementary elements they present, they don't offer the combination protocol. We propose here a working farm close to the AFC after the CHD : a similarity analysis on the shapes highlighted in the classification taking the color code of the classes. This procedure was tested on the forum entries on the National Consensus on digital education . This corpus presented in classes 7 shows in the merger of similarity analyzes recent modeling of inter-class relationships .

Résumé

L'analyse de corpus peut être abordée (entre autre) par classification hiérarchique descendante de type Reinert, mais également par analyse de similitudes. Si ces deux approches sont souvent exploitées parallèlement pour la complémentarité des éléments qu'elles présentent, elles n'offrent pas de protocole les associant. Nous proposons ici un travail proche de l'exploitation de l'AFC issue de la CHD : une analyse de similitude sur les formes mises en relief lors de la classification reprenant le code couleur des classes. Cette procédure a été testée sur les participations du forum portant sur la concertation nationale sur le numérique éducatif. Ce corpus se décompose en 7 classes, la fusion de leur analyse de similitude fait ressortir la modélisation des liens inter-classes.

Key words : analyse de similitudes, classification hiérarchique descendantes,

1. Introduction

L'analyse de corpus textuels par classification hiérarchique descendante de type Reinert permet de mettre à jour les mondes lexicaux (Reinert, 1983). Ces « traces possibles de nos contenus d'activités » (Reinert, 2008) constituées par les formes pleines permettent de voir les différents discours se construisant dans l'énonciation. Ce choix d'étude exploratoire nous amène trop vite à une vision scindée des corpus, où chaque type de discours serait séparé lexicalement des autres. Or certaines classes peuvent partager des formes significativement présentes. Le travail présenté ici part de l'utilité de ces « pont lexicaux », ces formes pleines présentes dans les profils de plusieurs classes qui dans une étude exploratoire peuvent être difficiles à repérer. Pour modéliser leur fonction architecturale dans le discours nous allons à partir des analyses de similitudes de chaque classe d'une CHD, fusionner les graphes en affectant à chaque sommet la couleur de la classe sur laquelle il est le plus représentatif. Nous prendrons comme exemple d'application le contenu du forum sur la concertation nationale sur le numérique éducatif.

2. Méthodologie

Les outils lexicométriques utilisés dans notre méthodologie étant déjà longuement documentés dans les actes des JADT précédents, nous nous permettrons une description succincte, préférant focaliser notre présentations sur leurs particularités et complémentarités. En dernier point, nous mettrons en avant la démarche spécifique employée ici.

2.1 La classification hiérarchique descendante (CHD)

Le modèle de CHD employé dans notre travail est la classification de type Reinert proposée par le logiciel Iramuteq. Cette classification implémentée pour la première fois dans le logiciel Alceste® (Reinert, 1983) permet de mettre en avant les mondes lexicaux. Ces structures du discours partent du principe que l'énoncé est un point de vue dépendant du sujet mais aussi de son activité et son contexte, où « l*e vocabulaire d'un énoncé particulier [est considéré] comme une trace pertinente de ce « point de vue » il est à la fois la trace d'un lieu référentiel et d'une activité cohérente du sujet-énonciateur. Nous appelons mondes lexicaux, les traces les plus prégnantes de ces activités dans le lexique*. » (Reinert, 1993)

- après lemmatisation les textes sont segmentés, puis la ponctuation est supprimée.
- à partir de ce matériau est construit un tableau à double entrée répertoriant la présence ou absence dans les segments des formes pleines retenues.
- sur ce tableau est effectuée une série de bi-partitions reposant sur une analyse factorielle des correspondances.

2.1 L'analyse de similitude

Ce modèle issue de la théorie des graphes (Flament, 1962, 1981; Vergès & Bouriche, 2001) représente la structure d'un corpus par la schématisation de ces relations, permettant ainsi de faire ressortir les liens des formes dans les segments de textes (Marchand & Ratinaud, 2012).

Nous utiliserons ici les arbres maximum, constitués des 2 liens les plus forts de chaque clique (relation entre 3 formes), ils permettent d'avoir un « graphe connexe et sans cycle » (Degenne & Vergès, 1973) modélisant le plus simplement le texte tout en conservant son architecture principale.

2.1 Le procédé employé ici

Après avoir effectué une CHD sur le corpus, nous exécutons sur chaque classe une analyse de similitude sur les formes actives significatives de cette dernière (formes présentes dans les profils), afin de comparer toute les classes quelles que soient leur taille, nous utilisons l'indice « pourcentage de cooccurrence » pour effectuer ces analyses de similitude. Enfin, nous joignons ces graphes en les assemblant par sommet de même label et en leur assignant la taille la plus élevée des 2 graphes, et la couleur de la classe où il est le plus représenté.

2. Mise en application

2.1 Le corpus

Le corpus étudié est issu du forum sur la concertation nationale sur le numérique éducatif (http://ecolenumerique.education.gouv.fr/). Cette concertation effectuée entre Janvier et Février 2015 ouverte à toute personne proposait une réflexion autour de 5 thèmes :

- le numérique, les apprentissages et la réussite de tous les élèves (380 articles)
- le numérique, renouvellement et diversifications des pratiques pédagogiques et éducatives (367 articles)
- le numérique et les compétences de demain (165 articles)
- le numérique et la réduction des inégalités (127 articles)

- le numérique, un facteur d'ouverture de l'école à son territoire et à son environnement (63 articles)

Les contributions étaient présentées dans leur intégralité sur une page d'index pour chaque thématique. Chacun de ces posts permettait d'ouvrir une page contenant l'article lui même et les réponses lui étant assignées.

Nous avons à l'aide du logiciel Gromoteur[1] récolté les contenus des articles, ainsi que leurs réponses. Cette démarche s'est basée sur un import récursif à partir de la page d'index de chaque thématique.

2.2 Les résultats de la CHD

Nous avons choisi d'effectuer une CHD sur l'ensemble du corpus toute thématique confondue. La présence de contributions aux contenus identiques, mais présentes dans différentes thématiques nous a forcé à augmenter le nombre de classes terminales demandées, pour avoir accès aux mondes lexicaux présents dans notre corpus. Le dendrogramme présent sur l'illustration 1 est donc issue d'une CHD avec 37 classes terminales en phase 1 et un seuil minimal de 100 segments de texte par classe.

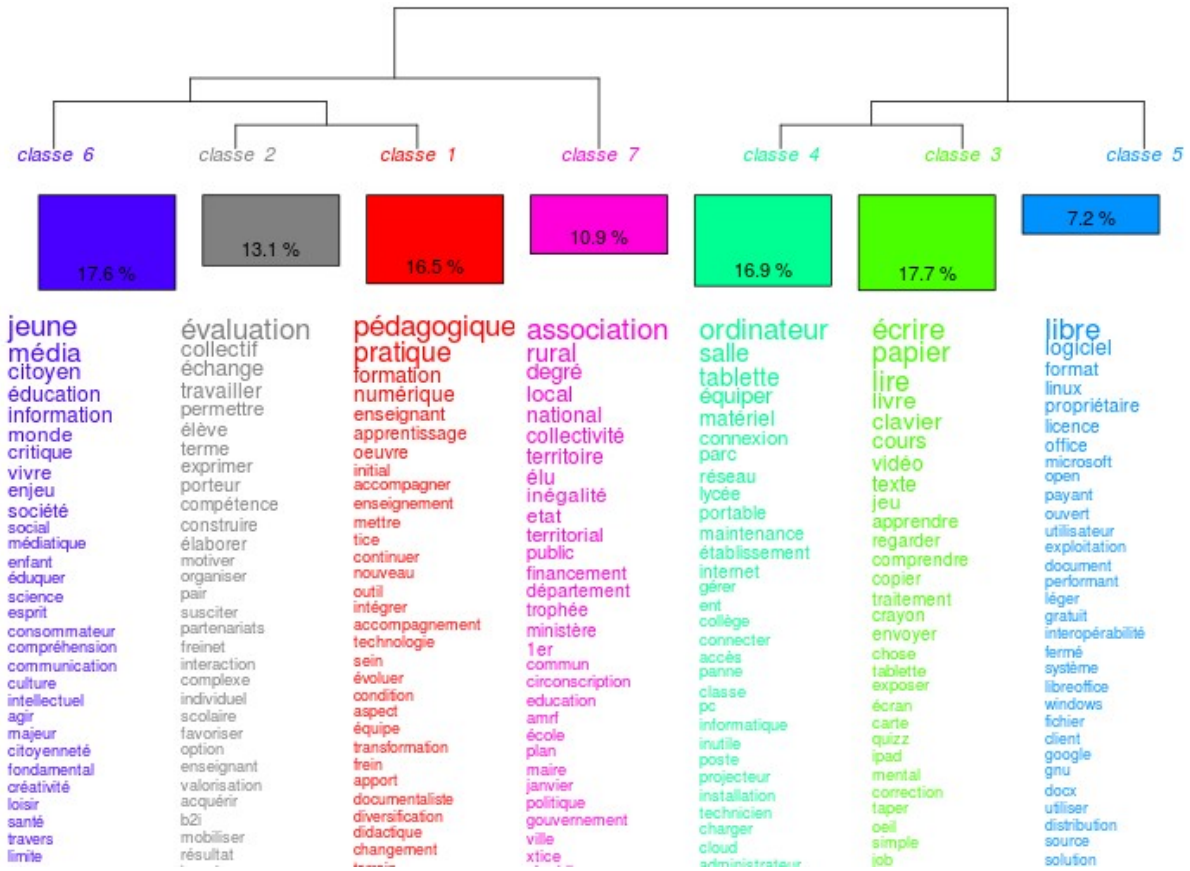

Illustration 1: Dendrogramme de la classification, taille des classes et liste des mots les plus caractéristiques de chaque classes (par ordre décroissant des chi2 de liaison aux classe)

[1]http://gromoteur.ilpga.fr/

Nous pouvons observer un découpage en deux grandes thématiques[2], d'une part les outils avec les classes 4, 3 et 5 opposés aux moyens mis en place avec les classes 6, 2, 1 et 7. Voici succinctement le contenu de ces classes.

- Les classes portant sur les outils :

L'équipement informatique des établissements scolaire est abordé dans la classe 5, nous retrouvons ici les discours généraux sur les équipements informatiques, l'infrastructure réseau et les moyens humains alloués à ce secteur. La question du choix de solutions logicielles libres ou propriétaires est plus particulièrement développée.

> *et puissent avoir le libre choix de la configuration en fonction des capacités réseau des matériels disponibles et des coûts logiciels associés (dont le coût des licences et participation financière développement/maintenance des solutions libres utilisées)*

L'opposition entre la proposition d'investissement dans de nouveaux équipements à destination des élèves (équiper les élèves de 5ème de tablettes individuelles) et le manque de moyen humain et d'équipement actuel des établissements scolaire est débattue dans la classe 4. La justification de cette dépense est mise en cause face aux faiblesses du numérique dans les établissements.

> *Expérimentation tablettes, il me semble qu'avec les mêmes moyens nous devrions parvenir à équiper toutes les salles de classe d'une connexion internet d'un vidéoprojecteur et d'un ordinateur portable.*
>
> *Un administrateur réseau par établissement pour gérer la fiabilité de la connexion haut débit nécessaire aux visio-conférences et aux accès multiples à internet, si le réseau tombe toutes les deux heurs les investissement colossaux en matériels types tablettes n'auront servi à rien.*

L'affrontement entre l'acquisition des savoirs fondamentaux et les besoins numériques est mis en perspective dans la classe 3. Ici nous relevons une argumentation prônant le besoins de familiarité avec l'outil dés les premiers apprentissages, à laquelle s'oppose une vision prioritaire des savoirs fondamentaux.

> *Si à l'école primaire ils utilisaient cet appareil pour lire et écrire à la place de la feuille de papier et du stylo on ferait une grande avancée cela permettrait de l'utiliser pour faire des calculs*
>
> *J'ai peur que notre pays s'aligne sur les États-Unis et oublie des compétences basiques lire écrire au profit du numérique, en effet aux États-Unis il devient commun d'apprendre aux élèves à taper sur un ordinateur plutôt que d'apprendre à écrire sur une feuille.*

- À ces classes, s'opposent les discours portés sur les moyens mis en places :

La question du financement de l'éducation et donc du numérique éducatif est soulevée dans la classe 7. Portée de façon significative dans la thématique sur la réduction des inégalités, nous ne trouvons pas un discours centré sur le bénéfice potentiel de ces technologies pour l'ensemble du public, mais décrivant l'inégalité de moyen d'une ville à l'autre. Le découpage institutionnel et administratif des établissements scolaires est décrit ici comme source d'inégalité entre zone rurale et urbaine, et accentuant la fracture numérique.

[2]La classe 8 étant constituée d'éléments architecturaux du site internet, nous ne la prenons pas en compte dans notre étude.

> *Les inégalités crées entre les communes pour le 1er degré et entre les départements pour le 2e degré risquent en effet de s'aggraver.*
>
> *Dans le premier degrés l'échelon territorial de fonctionnement est le département c'est aussi l'échelon qui semble le plus pertinent en matière d'équipement pour arriver à une unité pour éviter les disparités liées à la richesse des communes*

L'enseignement de l'éducation aux médias est l'objet de la classe 6. À l'instar des besoins de formation aux outils informatique que développait cette classe, ici c'est le versant « société de l'information » qui représente l'obligation de sensibilisation. Au delà du périmètre d'une discipline c'est un projet global porté par les enseignants documentalistes qui est revendiqué.

> *Une éducation à l'actualité critique et distanciée, orientée vers une compréhension du monde et de ses enjeux doit trouver ici les points d'appui efficaces dans une finalité de formation citoyenne des jeune ouvrant l'école aux savoir vivants et mouvants à la culture d'aujourd'hui*

Le discours présent dans la classe 2 est centré sur l'ingénierie pédagogique. Significative de la thématique sur le renouvellement des pratiques pédagogiques, nous observons ici la description des différentes architectures de cours que devrait rendre possible le numérique.

> *La conception pédagogique le learning design aussi traduit par conception de l'apprentissage s'est développée comme un moyen d'aider les enseignants à faire des choix éclairés en termes de création d'interventions d'apprentissage qui sont pédagogiquement efficaces et utilisent efficacement les nouvelles technologies*

Enfin la classe 1 est le discours concernant la formation des enseignants. Qu'elle soit initiale, continue, institutionnelle ou de leur propre chef, elle est présentée comme un pré-requis indispensable à la mutation du système éducatif.

> *Propositions pour contribuer au renouvellement et à la diversification des pratiques pédagogiques : sans formation initiale et continue il n'y aura pas de révolution numérique les enseignants acteurs de terrain sont les réel leviers des changements*

2.1.1 L'AFC

Le dendrogramme de la CHD modélise le découpage factoriel des classes, mais n'est pas en mesure de révéler le positionnement de son lexique. Pour étudier ces tendances, nous avons habituellement recours à l'AFC issue des classes de cette CHD proposée par Iramuteq. Nous pouvons ainsi observer les superpositions de classes, ou au contraire le détachement d'autres.

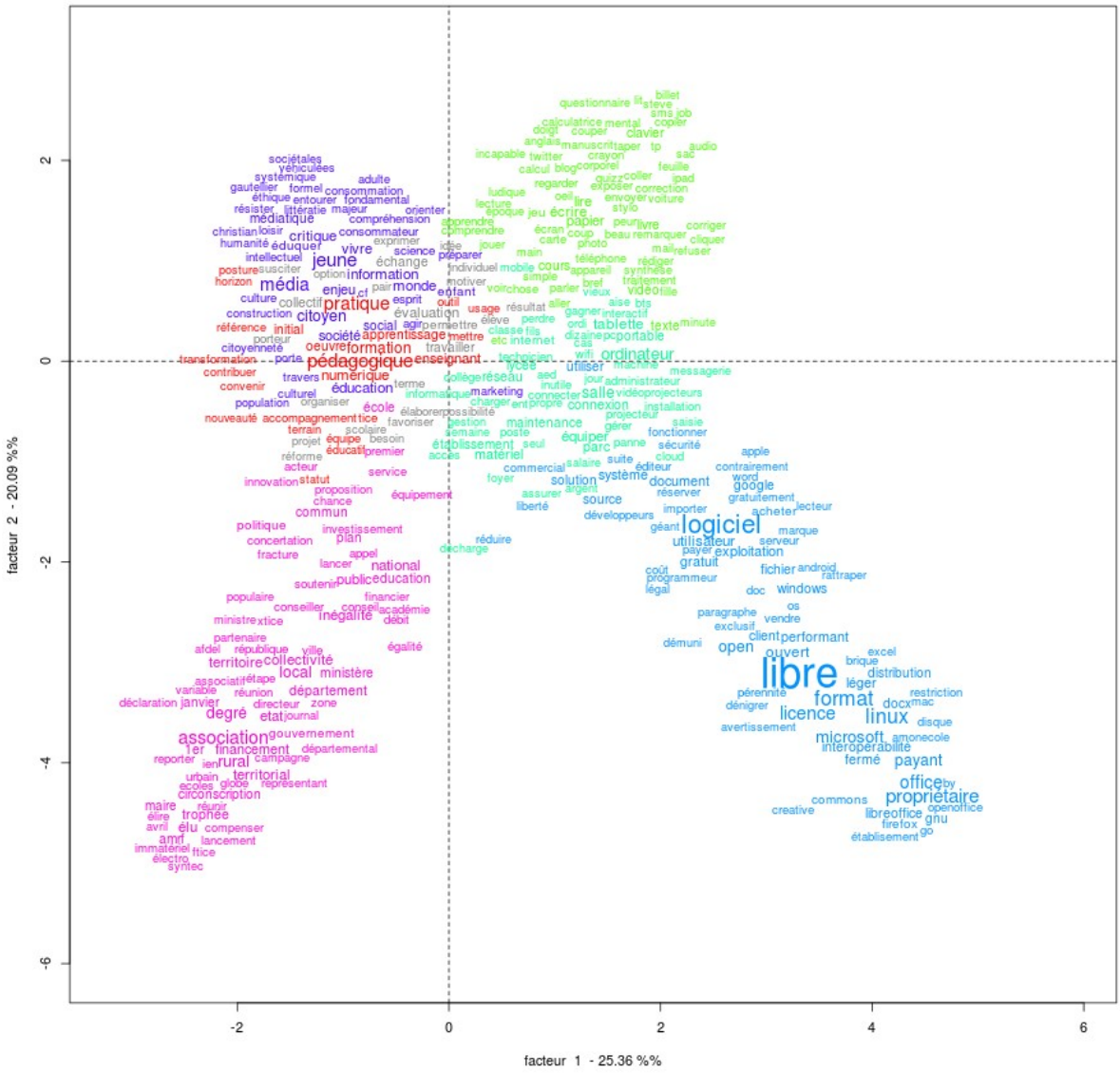

Illustration 2: Plan factoriel facteur 2 / facteur 3

L'illustration 2 fait par exemple ressortir un très net découpage des classes traitant des outils TICE (partie de droite) alors que dans le groupe traitant des moyens seule la classe concernant le financement des établissements scolaire se détache sur le facteur 2.

En effet, les classes 3, 4 et 5 (vert, bleu cyan et bleu) n'ont que peu de mots se mélangeant sur le graphe, là où les classes 6, 2 et 1 (mauve, gris et rouge) sont superposées. Si cela nous informe sur leurs proximités, leurs structures internes, les relations entre elles et les liens qu'elles entretiennent par les formes communes ne sont pas lisibles dans cette représentation graphique.

2.1.2 La fusion d'analyse de similitudes

Le procédé que nous proposons dans ce travail est la fusion des analyses de similitudes de chaque classe effectuées sur les termes présents dan les profils, avec l'attribution de la couleur de la classe dans lequel il occupe une plus grande place. Le graphe a été produit par la

fonction « merge graph » présente sur Iramuteq[3], le graphique quant à lui à été élaboré à partir du logiciel Gephi[4].

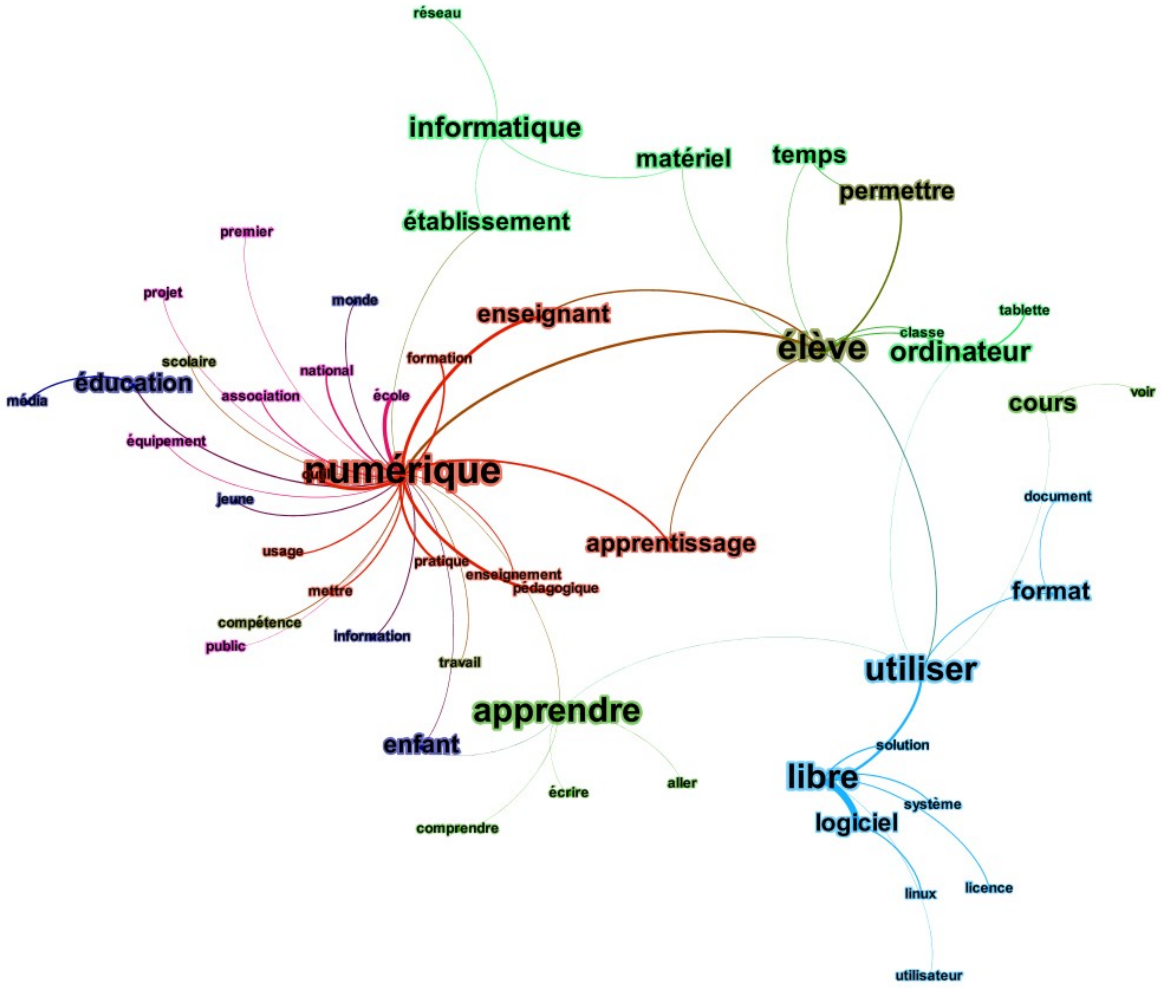

Illustration 3: Fusion des graphes de similitude (sur graphe contenant les 25 formes actives les plus présentes par classe)

La schématisation des liens, nous montre ici une liaison entre les classes sur les moyens de mise en place (rouge, rose et bleu) et les outils (vert, vert clair et bleu cyan) par la classe sur

[3]Version en développement accessible sur le dépôt : http://www.iramuteq.org/git/iramuteq déposée le 10/06/2015

[4]https://gephi.org/

l'ingénierie pédagogique (gris). Seuls les mots *élève* et *permettre* sont encore inscrits dans cette thématique qui représente pourtant 13 % du discours classé.

Nous pouvons également observer que les classes sur l'éducation aux médias (bleu), le financement des établissements (rose), ou encore les équipements de ces derniers (bleu cyan) constituent les discours périphériques (ils n'ont de liens qu'avec une autre classe) ; alors que les classes sur la formation des enseignants (rouge), l'ingénierie pédagogique (gris), la polémique sur l'utilité du numérique face aux enseignement fondamentaux (vert), l'état des équipements actuels (turquoise) sont des classes partageant des ponts lexicaux avec plusieurs classes, elles représentent ainsi l'articulation des discours. Nous pouvons également observer une « boucle » reliant 4 des 7 classes sur les termes *utiliser, élèves, enseignant, numérique et apprendre*.

Ce graphique fait donc ressortir un raisonnement où les thématiques des besoins de l'éducation aux médias (bleu), nécessitent une ingénierie pédagogique (grise), avec une formation des enseignants (rouge), pour légitimer et développer l'emploi du numérique dans l'éducation.

2.1.3 Les limites du procédé

La première limite est illustrative L'analyse des classes de discours dans leur intégralité nécessite un affichage très étendu, or pour que le graphe reste lisible sur un format papier, seul un petit nombre de mots peuvent être pris en compte, il faut donc limiter les formes dès la création des graphes de classe, et affecter le même filtre à chaque classe. Ce choix va impliquer un affaiblissement supplémentaire des classes les plus petites. De plus si ces dernières ont une densité plus faible elles peuvent presque disparaître du graphe.

Le fonctionnement de la CHD en étudiant les cooccurrences dans les segments de textes permet de lever une grande part d'ambiguïté des termes. Ainsi une forme pouvant porter plusieurs significations peut se retrouver dans plusieurs classes qui relèvent de ces différents sens. Le protocole que nous montrons ici en fusionnant les sommets de mêmes labels fait réapparaître cette ambiguïté.

3 Conclusion

Le protocole que nous montrons ici nous a permis de mettre en relief des articulations entre classes reposant sur les ponts lexicaux. Au delà du positionnement géographique, les liens proposés par l'analyse de similitude permettent une modélisation supplémentaire de l'articulation des discours. Bien que suivant le découpage du dendrogramme, le graphe obtenu nous permet de mettre en relation des classes séparées initialement par les premiers facteurs, permettant ainsi après lecture d'une CHD de relier les mondes lexicaux entre eux, quel que soit leur moment d'apparition dans la classification.

Référence

Degenne, A., & Vergès, P. (1973). Introduction à l'analyse de similitude. *Revue française de sociologie*, 471-512.

Flament, C. (1962). L'analyse de similitude, (4), 63-97.

Flament, C. (1981). L'analyse de similitude: une technique pour les recherches sur les représentations sociales. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*.

Marchand, P., & Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes*

des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT, *2012*, 687-699.

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

Reinert, M. (1993). Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, *66*, 5-39.

Reinert, M. (2008). Mondes lexicaux stabilisés et analyse statistique de discours. *9è JADT*.

Vergès, P., & Bouriche, B. (2001). L'analyse des données par les graphes de similitude. *Sciences humaines*.