Le traitement des TICE dans les discours politiques et dans la presse

Lucie Loubère

Lerass – lucie.loubere@iut-tlse3.fr

Abstract

The very existence of databases in addition to the ability to scan press materiel now allows researchers to access very large corpuses. However the availability of these databases and their number in constant evolution raise the problem of the thematic relevance of the selected articles. This study proposes to handle two corpus by using the software Iramuteq (P. Ratinaud, Lerass, University of Toulouse). By the means of the descendant hierarchical clustering, following to the Reinert method, we will examine the themes developed throughout the corpus, in order to extract the segment of texts for our topic. The sub-corpus obtained in this manner will then allow to identify, via a similarity analysis, the organization of the discourse in the targeted thematic.

Résumé

La numérisation des documents de presse et l'existence de base de données permettent aujourd'hui aux chercheurs d'accéder à des corpus de plus en plus grands. Cependant la facilité d'accès à ces données et leur nombre en constante évolution posent le problème de la pertinence thématique des articles sélectionnés. Nous proposons ici une étude sur la manipulation de deux corpus à l'aide du logiciel Iramuteq (P. Ratinaud, Lerass, Université de Toulouse). La Classification Hiérarchique Descendante (CHD), selon la méthode Reinert, nous permettra d'étudier les thèmes développés dans l'ensemble des corpus, afin d'extraire les segments de texte ciblant notre sujet. Les analyses menées sur les sous-corpus obtenus permettront alors d'identifier l'organisation des discours sur la thématique ciblée.

Mots-clés : Classification Hiérarchique Descendante, méthode Reinert, gros corpus, Iramuteq, TICE

1. Introduction

La numérisation et le stockage d'écrits dans des bases de données proposent de nouvelles perspectives à la recherche en permettant d'élargir la quantité de sources. Cependant, cette ouverture crée de nouveaux problèmes dans la sélection des textes étudiés. La base de données d'articles de presse par exemple, nous permet aisément de récolter de nombreux articles sur les soixante dernières années. Les requêtes s'exerçant par recherche en texte intégral, et malgré les différents filtres disponibles (région géographique, date, thématique, source...), le risque d'atteindre des documents sans réel rapport avec notre recherche persiste. De plus, l'obtention d'une grande quantité de textes issus de la presse rend difficile leur comparaison à d'autres corpus issus de sources moins abondantes, mais toutes aussi pertinentes (les textes institutionnels par exemple). Les travaux que nous allons présenter sont fondés sur l'utilisation de la Classification Hiérarchique Descendante (CHD) de type Reinert (1983) sur deux corpus de tailles et de types différents. Cette méthode nous permettra de prélever les éléments significatifs à notre thématique de recherche, puis de dégager les principaux thèmes abordés. Pour mener cette recherche, nous nous baserons sur les textes traitant des Technologies de l'Information et de la Communication dans l'Enseignement (TICE), à partir des articles proposés dans la base de données Factiva®, et des discours politiques disponibles sur le site Vie Publique®.

2. Problématique

Les Technologies de l'Information et de la Communication pour l'Enseignement sont un sujet récurrent dans les discours politiques et médiatiques. L'étude et la comparaison des types de discours portés sur ce sujet dans ces deux sources posent plusieurs difficultés. La première étant la sélection des données issues de la presse. Une sélection arbitraire des sources journalistiques impliquerait un important biais méthodologique, Ce sujet pouvant être traité de façon différente entre les divers types de presse (presse régionale/nationale, économique, spécialiste...). L'élargissement de ce panel rédactionnel augmente de façon exponentielle des documents traités. A la question sur la quantité de documents, s'ajoute celle de la qualité. Comme le montrent les travaux d'Iyengar (1991), les médias porteraient un discours de type épisodique, basé sur des moments concrets de l'actualité nationale ou régionale, impliquant une diversité de sujets traités (rentrée scolaire, parution de nouvelles méthodes didactiques, nouvelles orientations ministérielles...). Cette caractéristique impose pour accéder à l'échantillon le plus complet possible, d'établir des requêtes basées sur des mots-clés vastes. Ce choix méthodologique génère une grande quantité de textes ne comportant que quelques lignes sur le numérique, il est donc nécessaire de gérer le « bruit ainsi généré ». A ces impératifs propres aux discours médiatiques et institutionnels, s'ajoute la difficulté de confronter ce corpus à celui des textes politiques. Ces données étant quantitativement beaucoup moins importantes, elles ne peuvent intégrer le premier corpus sur une analyse quantitative, la variation d'effectif, ferait disparaître les spécificités du dernier corpus. Afin de confronter ces deux matériaux, nous proposons à l'aide du logiciel *Iramuteq* un processus d'extraction dans nos corpus des segments portant un discours sur le numérique. Puis, après étude des thématiques développées dans chaque type de discours, une confrontation des deux sous-corpus.

3. Les corpus

Le premier corpus exploité dans cette étude, est constitué des 147 discours politiques, issus du site « vie publique », contenant le terme « numérique », et la chaîne « ministre de l'éducation ». Cette base de données ministérielle est éditée par la Direction de l'information légale et administrative elle « *se propose de faciliter l'accès des internautes aux ressources et données utiles pour appréhender les grands sujets qui animent le débat public* ». Il est ainsi possible de consulter plus de 14 000 rapports, discours et débats publics depuis 1947. Par souci d'efficacité lors du traitement informatique de ce corpus, un nettoyage manuel a été effectué pour enlever les didascalies (notamment lors des interviews), ainsi que les trois discours présentant la liste du gouvernement. Le second corpus, extrait de la base de données Factiva®, correspond aux articles de presse de 1999 à 2013 parus dans la presse française[1] contenant le terme « numérique », et indexés dans la thématique éducation. Ces requêtes nous ont retourné 8995 articles, sur lesquels aucune modification manuelle n'a été apportée.

4. Les classifications sur les corpus complets

4.1 Le choix de la méthode Reinert

Notre objectif étant d'extraire à partir d'un grand nombre d'articles les épisodes portant plus spécifiquement sur notre objet d'étude. Cette sélection, nous permettra de porter une analyse plus précise sur l'organisation des discours. Pour effectuer cette manipulation, notre choix

[1] Soumis à l'abonnement Factiva®

s'est porté sur la méthode Reinert. Cette méthode implémentée initialement dans le logiciel Alceste®, est basée sur le découpage des textes en segments (une moyenne de 40 formes actives dans nos expérimentations), sur lequel est faite une classification. Cette particularité, permet dans le premier temps de notre méthodologie d'extraire un contenu suffisamment structuré pour subir d'autres traitements par la suite. Au-delà de cette manœuvre d'extraction, la conservation de segments nous permet en gardant un contexte d'analyser les formes ambigües. Notre travail étant porté sur la méthodologie d'extraction de sous-corpus, nous ne détaillerons pas les thématiques abordées dans les discours, nous ne feront qu'énumérer ces dernières et leurs caractéristiques principales afin de prendre conscience de la diversité des thèmes abordés dans nos corpus.

4.2. Le corpus composé des discours politiques

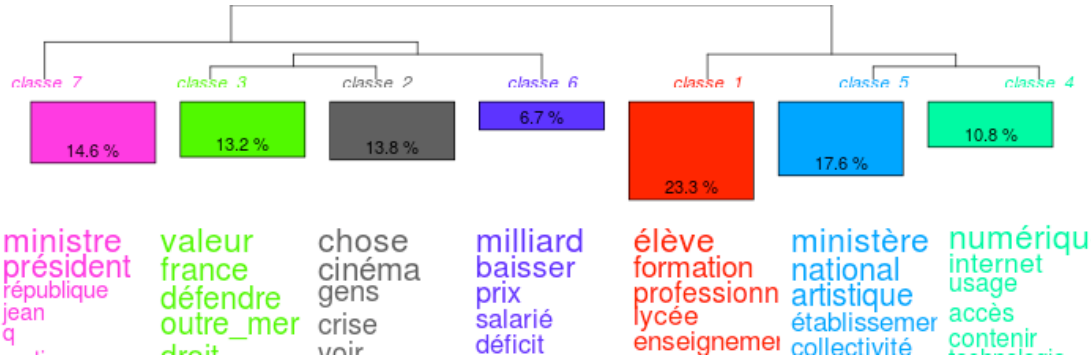

Figure 1. Dendogramme de la CHD effectuée sur les textes politiques

La CHD effectuée ici est une classification simple sur segments de textes, avec 10 classes en fin de première phase, et un seuil minimal de 13 segments par classe. Le résultat obtenu classe 1 258 segments sur 12 884 soit 99,80 % du corpus en 7 classes, dont une seule est caractérisée par la présence prononcée du terme « Numérique ». Nous retrouvons comme classe de discours :

- **Une classe basée sur un discours institutionnel :** centrée autour des fonctions politiques, et des noms de personnalités, elle est constituée des interventions situant le discours dans le contexte politique.
- **Le discours économique** : le numérique éducatif est une source d'investissement de la part des organismes politiques (du ministère aux collectivités territoriales), il est donc un enjeu dans l'économie. Se croisent ici un discours sur les investissements (Milliard, Investissement, prix...), et sur l'économie générale (croissance, TVA...).
- **Le discours sur le cinéma** : les ministères de l'éducation et de la culture ayant mis en place plusieurs partenariats, sont présents dans ce corpus des textes mêlant les deux domaines (Cinémas, Voir, Film...).
- **Le discours sur les valeurs de la république** : le discours porté ici est centré sur les valeurs de la République (Valeur, France, Droit, Paix...).
- **Le discours centré sur l'enseignement** : nous retrouvons ici le lexique propre à l'enseignement (Élève, Formation, Lycée...).
- **Le discours sur le partenariat école/culture : c**ette classe correspond à un partenariat entre le ministère de la culture et de l'éducation nationale. Cette classe est

marquée par un lexique spécifique des Arts (Artistique, Danse, Culturel...) auquel est associé celui du champ scolaire (Éducation, Université, Éducatif...), les deux étant reliés par les termes du partenariat (Projet, Équipe, Associer...).

- **Le discours sur le numérique** : ce discours est composé par le vocabulaire informatique (Numérique, Internet, Informatique...) associé à celui des usages (Usages, Accès, Ressource Information, Communication...). C'est dans cette classe que nous retrouverons les segments portant sur notre sujet, il ne représente que 10.8% du corpus classé.

4.3. Le corpus composé des articles de presse

Figure 2. Dendogramme de la CHD effectuée sur les articles de presse

La CHD effectuée ici est une CHD simple sur segments de textes avec 10 classes en fin de première phase, 107 874 segments de texte ont été classés sur les 108 156 contenus dans le corpus, soit 99.74 %. Nous retrouvons comme classe de discours :

- **Le discours sur l'enseignement supérieur** : l'enseignement supérieur étant spécifique, il porte un vocabulaire qui lui est propre, la classification fait donc ressortir ce lexique (Université, recherche, supérieur...).
- **Le discours sur les institutions** : il représente, comme dans le corpus précédent, la mise en contexte des articles par la citation des personnalités. Nous y retrouvons de façon caractéristique les termes (Président, Général, Ministre, Peillon, Hollande...).
- **Les adresse web** : ici sont classés tous les segments contenant les adresses internet (http, www, …).
- **Le discours sur le numérique** : c'est dans cette classe que sont regroupés les segments propres à notre thématique, composée de 19349 segments sur les 93145 classés, elle ne représente que 20,77 % du corpus analysé. Sont présents de façon significative les termes spécifiques au matériel informatique (Numérique, Ordinateur, Internet...), les verbes propres aux usages (Utiliser, Connecter, Disposer...), et enfin, le vocabulaire de gestion de projet (Équiper, Doter, Installer...).
- **Le discours sur la structure scolaire** : nous retrouvons le discours sur les établissements et la structure scolaire (Bac, Lycée, Collège...).
- **Les discours sur le travail pédagogique** : constitué de deux classes, nous y trouverons les termes liés à l'enseignement (généralité et disciplinaire) (compétence, Langue, Formation...) associés à celui de la difficulté scolaire (Difficulté, Accompagnement, Échec...).

- **Le discours sur les journées événementielles** : dans cette classe est porté le discours sur l'actualité événementielle de l'établissement scolaire (Photo, Fête, Loto, Porte, Accueillir, Rentrer, Effectif...).

Ce temps de notre étude étant réservé à isoler les discours spécifiques au numérique dans chaque corpus, cette première étude, en restant sur un faible nombre de classes ne nous permet pas de porter une analyse plus fine des thèmes abordés.

Le parcours des profils de classe que nous venons d'effectuer, aussi succinct soit-il, nous permet d'appréhender la variété de thématiques traitées dans notre corpus, ainsi que le peu de proximité qu'elles partagent avec les classes nous intéressant.

5. Les CHD sur chaque sous-corpus « numérique »

Sur chacune des deux classes extraites nous avons à nouveau exécuté une CHD, de façon à séparer les différents sujets traités dans ces textes.

5.1. Les discours politiques

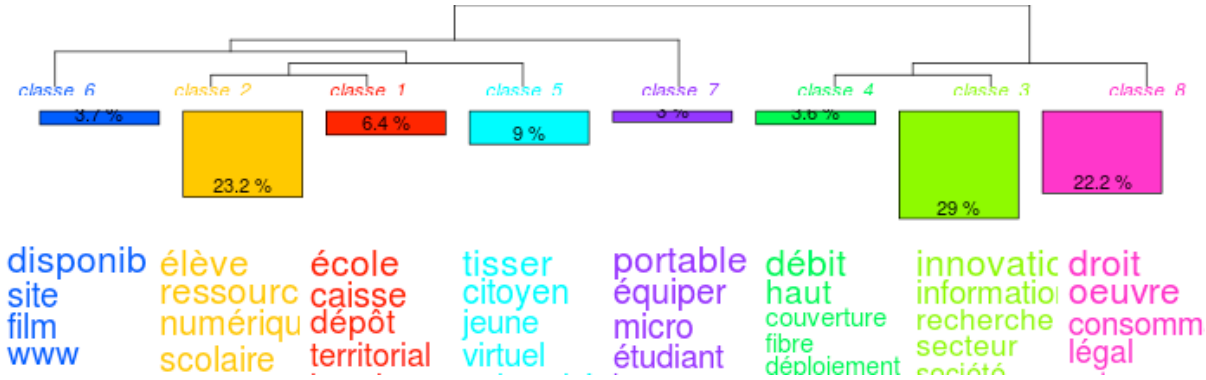

Figure 3. Dendrogramme de la CHD effectuée sur le sous corpus issu des discours politique

Nous avons effectué une CHD avec 15 classes en phase 1, le résultat obtenu classe 1695 segments de texte sur les 1701 que contenait le sous corpus soit 99,65 % en 8 classes.

5.1.1. La classe sur les ressources mises à disposition

Cette classe de discours est centrée sur le réseau de ressources mis en place par le ministère, sont donc surreprésentés les adresses internet, ainsi que le vocabulaire de l'accessibilité (Disponible, Portail, Accessible...). Le tout relié aux termes Pédagogie, Académie, Discipline.

Figure 4. Wordcloud ressources mises à disposition

5.1.2. La classe sur les ressources employées

Nous retrouvons ici les ressources numériques utilisées en situations d'enseignement (Élèves, Enseignant, Pédagogique...)

Figure 5. Wordcloud ressources employées

5.1.3. La classe sur l'école primaire

Ce discours est marqué par les segments spécifiques à l'école primaire. Ce niveau d'enseignement étant financé par les communes, le développement des TICEs n'a pas suivi le même rythme que dans le secondaire. Nous retrouvons les termes marquant cette temporalité (Entrer, Rapidement, Lancer, Progressivement).

Figure 6. Wordcloud école primaire

5.1.4. La classe sur les risques du numérique

Cette classe est centrée sur les risques du numérique, et les besoins de formation qu'ils engagent. Nous retrouverons comme segments caractéristiques :

« c'est dans l'utilisation au **quotidien** de l'**informatique** et d'**internet** sous la **conduite** de leurs **enseignants** que les **jeunes** seront le mieux **formés** à l'**usage** raisonné responsable et **citoyen** de nouveaux **médias** » **ou encore** « en **effet** il est indispensable que les **jeunes** puissent exercer un **regard critique** devant le flux incessant d **images diffusées** par les nouveaux **médias** et les **sites** de partage tels que Google YouTube, Dailymotion, **etc** … »

5.1.5. La classe sur la dotation premier équipement

Ce discours porte le lexique des projets de dotation en premier équipement, ainsi ressortent les termes matériels (Portable, Micro, Wifi...), associés aux publics auxquels se destinaient les différentes campagnes (Étudiant, Foyer, Ménage...).

Figure 7. Wordcloud 1er équipement

5.1.6. La classe sur l'infrastructure connectique

Cette classe est constituée des segments traitant de la mise en place de l'infrastructure connectique. La progression du numérique dans les classes nécessitant une fiabilité de connexion, le thème de l'amélioration du débit est récurrent.

Figure 8. Wordcloud connectique

5.1.7. La classe sur la recherche et l'innovation

Ce discours traite de l'innovation dans le domaine des TICE, nous trouverons de façon significative les termes propres à la recherche et au développement (Innovation, Recherche, Ambition...). Ainsi que ceux spécifiques aux TIC (Information, Technologie, Communication...).

Figure 9. Wordcloud recherche innovation

5.1.8. La classe sur les droits d'auteur

Cette dernière classe est composée des segments de texte portant sur le domaine légal des TICE, et plus spécifiquement celui des droits ; d'auteur et de l'exception pédagogique. Les termes significativement présents sont du domaine légal (Légal, Loi, Contrefaçon...), mais également du monde industriel (Consommateur, Industrie, Rémunération...).

Figure 10. Wordcloud droit d'auteur

5.2. Les articles de presse

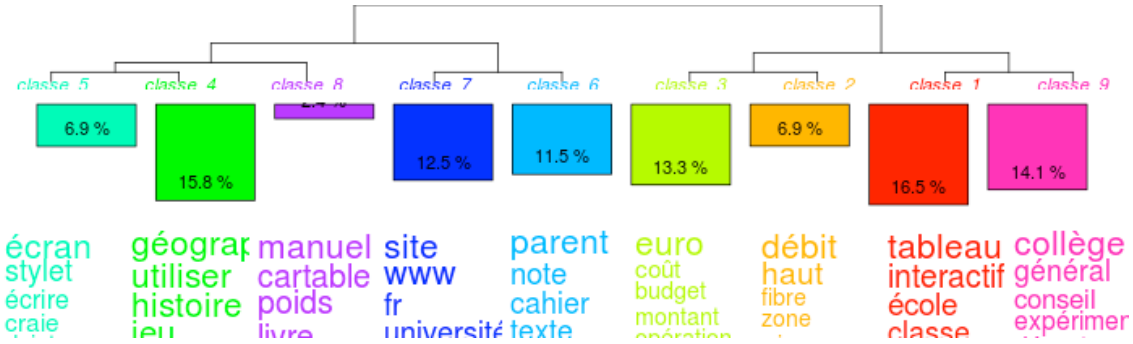

Figure 11. Dendrogramme de la CHD effectuée sur le sous corpus issu des articles de presse

Nous avons effectué une CHD avec 15 classes en fin de première phase, le résultat a classé 19 273 segments sur les 19 345 du sous-corpus, soit 99,63 %.

5.2.1. La classe sur la numérisation de l'écriture

Cette classe est centrée sur les termes spécifiques de la migration de l'écriture manuelle à l'utilisation d'outils numériques. Nous retrouvons donc un discours propre à l'écriture traditionnelle (Écrire, Doigt, Stylo...) auxquels sont juxtaposés ceux propres à l'écriture numérique (Écran, Stylet, Souris...). Nous notons également la présence du vocabulaire de transition (Remplacer, Modifier,...).

Figure 12. Wordcloud numérisation écriture

5.2.2. La classe Disciplinaire

Ce discours gravite autour des disciplines enseignées (Histoire, Géographie, Maths...) ainsi que le vocabulaire propre à l'enseignement (Cours, Apprendre, Professeur...) et aux ressources pédagogiques (Image, Livre, Vidéo...). Enfin, nous notons la présence de nombreux termes valorisants (Ludique, Enthousiasme, Enrichir...).

Figure 13. Wordcloud disciplinaire

4.2.3. La classe sur les manuels numériques

C'est le discours portant sur les manuels numériques. Ainsi, sont surreprésentées les formes propres aux manuels scolaire (Livre, Papier, Éditeur...), et numérique (CD-Rom, Numérisation...). Ces termes sont complétés par un vocabulaire portant sur l'allègement des cartables (Poids, Kg, Alléger...).

Figure 14. Wordcloud manuels numériques

5.2.4. La classe sur les ressources disponibles

Cette classe de discours présente les ressources officielles disponibles, sont associés les termes des sites web (Site, Internet, Web...) et le champ lexical des ressources (Information, Recherche, Bibliothèque...). A ces deux catégories, s'ajoute celle du partage (Contenir, Disponible, Gratuitement...).

Figure 15. Wordcloud ressource disponibles

5.2.5. La classe échange famille/établissement

Dans cette classe de discours, nous retrouvons les pratiques concernant le lien entre parents et établissements (Parent, Communication...). L'outil majeur pour ces pratiques étant l'ENT, pouvons voi les services proposés par ce dernier (Cahier de texte, ENT, Agenda...). Le vocabulaire lié à la communication est très présent dans de cette classe (Communication, Communiquer, Consulter...). Enfin, ces outils offrant un espace personnel aux usagers, le vocabulaire de la sécurité est surreprésenté (Code, Sécurisé, Identifier...).

Figure 16. Wordcloud famille/établissement

5.2.6. La classe sur l'investissement

Nous retrouvons ici tout le discours économique et budgétaire (Euro, Coût, Montant...), mais aussi un champ lexical de la dépense (Investissement Dépenser, Subventionner...).

Figure 17. Wordcloud investissement

5.2.7. La classe sur la connectique

C'est le discours sur les investissements faits sur la mise en place de connexion à haut débit (Débit, Haut, Fibre...). Les termes propres au marché numérique sont également présents (Orange, Galaxy, Ipad...).

Figure 18. Wordcloud connectique

5.2.8. La classe Équipements

Sur cette classe est présenté le discours sur les investissements pour les équipements (Tableau, Interactif, Ordinateur, Projecteur...), associés aux verbes propres à l'équipement (Équiper, Doter, installer...).Nous notons également un aspect évolutif, et une notion de remise à niveau (Compléter, Performant, Obsolète...).

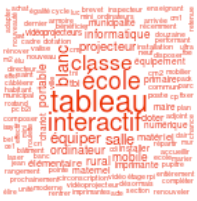

Figure 19. Wordcloud équipement

5.2.9. La classe sur les collectivités

Le champ éducatif est du ressort des collectivités territoriales, ainsi, nous retrouvons ici les différents partenaires (Conseil, Général, Région...).

Figure 20. Wordcloud collectivités

6. La confrontation des deux sous corpus par Analyse factorielle de correspondance (AFC)

Sur cette partie, nous avons regroupé les deux sous-corpus, en affectant à chaque segment une variable illustrative contenant le corpus d'origine, et leur classe d'appartenance. Nous avons également modifié le codage de l'année, de façon à faire ressortir la source des segments sur cette donnée.

6.1. L'AFC selon les classes

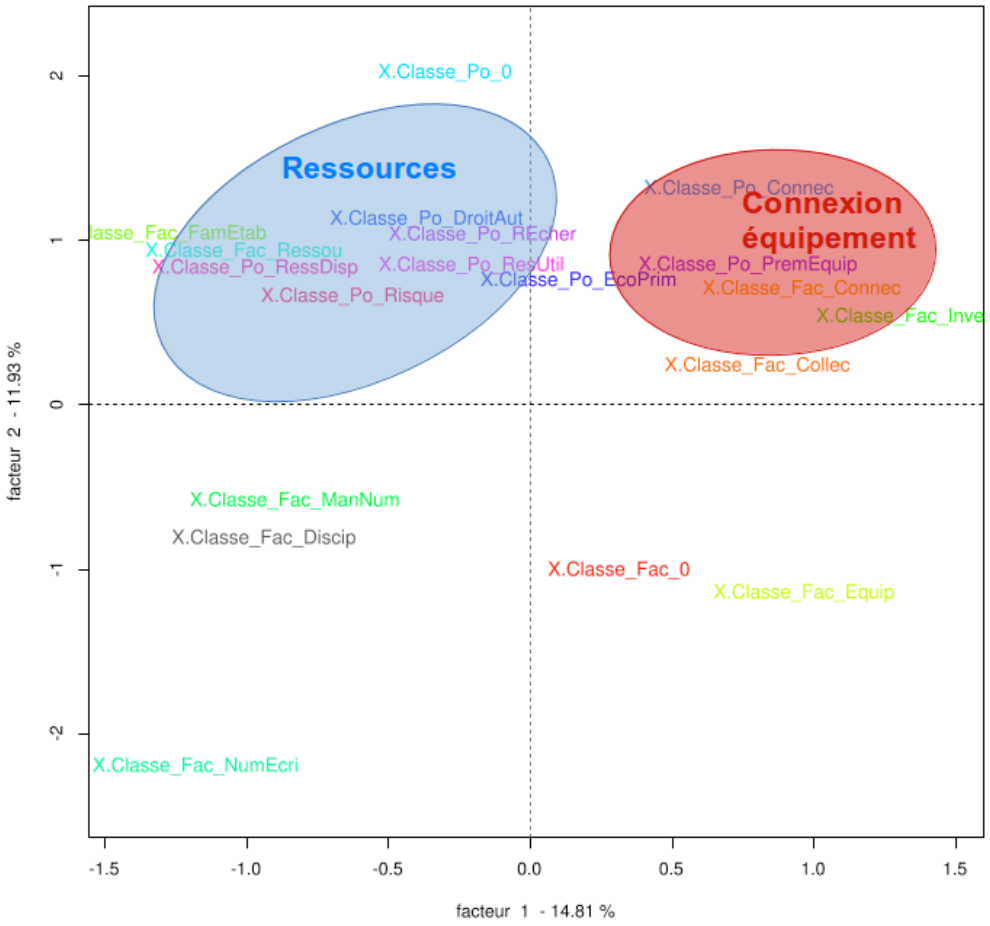

Figure 21. AFC par corpus et classes

D'une façon générale sur les deux premiers facteurs, les classes provenant des discours politiques s'entremêlent avec les classes des articles de presse (partie supérieure du graphique), cependant 4 des 9 thématiques portées par le sous-corpus provenant de Factiva sont isolées sur des coordonnées négatives du second facteur. Sur ces discours, nous retrouvons les thématiques sur le quotidien des élèves (Manuel numérique et la numérisation de l'écrit) ainsi que le traitement des équipements (qui partage cependant des coordonnées très proches sur le premier facteur avec la classe sur les équipements des textes politiques) et la classe sur les disciplines scolaires. Nous relevons plus spécifiquement, la forte proximité des classes concernant les ressources numériques, que ce soit dans les articles de presse, ou les discours politiques, et de façon plus atténuée, mais qui sont regroupés sur le même quart du graphique, les discours sur la connectivité.

6.2. L'AFC selon les années

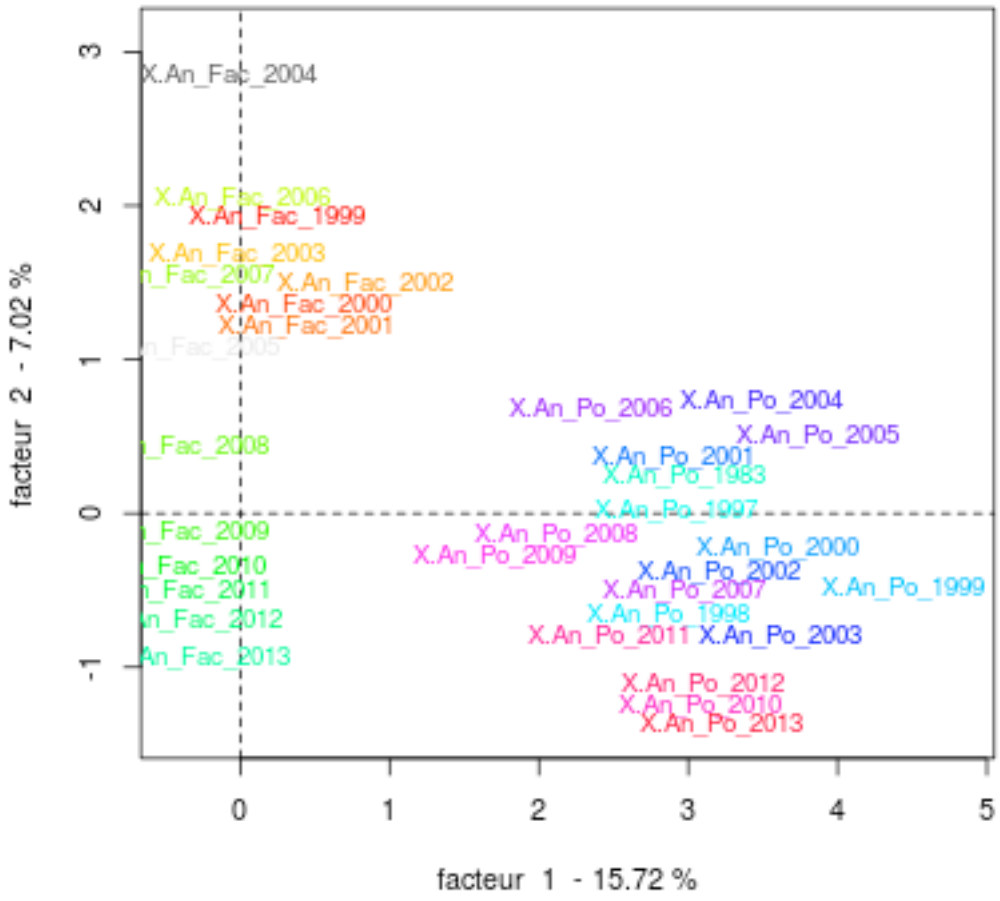

Figure 22. AFC par corpus et année de parution

Cette analyse factorielle nous montre une rupture sur le premier facteur entre les discours politiques et les articles de presse. Le second facteur marque une évolution au fil des années avec pour le corpus de Factiva® un enchaînement régulier de 2008 à 2013. Nous retrouvons de façon plus instable la même tendance chronologique dans les textes politiques, avec un décalage. Ces variations peuvent s'expliquer par la rapidité avec laquelle évoluent les technologies numériques et leur vocabulaire. De plus, un décalage temporel persiste entre les textes institutionnels, et la mise en place des projets qui feront l'actualité. Une étude plus détaillée sur la séquentialité de ces corpus devrait être développée.

7. Conclusion et perspectives

L'étude parallèle de ces deux corpus nous permet lors de la première phase de l'analyse de montrer que le sujet du numérique, est pour nos deux catégories de données inclus dans une multitude d'autres thématiques. L'analyse des corpus entiers, si elle apporte de nombreux éléments de contextualisation, produit trop de bruit pour une analyse lexicométrique.

La seconde partie de l'analyse, en ne ciblant que les passages portant sur le numérique, nous permet de façon beaucoup plus fine de dresser un tableau des différents discours portés sur le numérique. Ces 17 classes étudiées nous apparaissent dans des domaines relevant de la pédagogie, du matériel, des infrastructures, mais aussi de la logistique, des partenariats, de l'économie ou encore de la législation.

A l'aide de l'AFC sur les corpus et classes d'appartenance, nous avons pu observer de forts rapprochements entre les classes centrées sur les ressources pédagogique, et les connectiques quelque soit l'origine des textes et des classes propres aux articles de presse. L'AFC menée sur les années de parution des textes a quant à elle mis en relief une évolution chronologique des discours.

Cette étude exploratoire nous a permis d'appréhender le traitement du numérique éducatif dans la presse et les discours politiques. Au-delà de la description de ces thématiques, les AFC ont mis en évidence des liens entre des catégories de discours portés par ces deux sources. Une étude plus approfondie sur la séquentialité de ces discours et sur l'enchaînement des thèmes développés dans chaque corpus pourrait être envisagée.

Références

Iyengar S. (1991). *Is anyone responsible?: how television frames political issues* (Vol. 1-1). Chicago Ill., Etats-Unis.

Marty E., Marchand P. et Ratinaud P. (2013). Les médias et l'opinion: éléments théoriques et méthodologiques pour une analyse du débat sur l'identité nationale. *Bulletin de méthodologie sociologique*, *117*(1) : 46–60.

Ratinaud P. et Marchand P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux"?: analyse du "CableGate" avec IRaMuTeQ. In *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles* (pp. 835–844). Presented at the 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012, Liège, Belgique. Retrieved from http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications

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