FABIEN EMPRIN

LES APPORTS D'UNE ANALYSE STATISTIQUE DES DONNÉES TEXTUELLES POUR LES RECHERCHES EN DIDACTIQUE : L'EXEMPLE DE LA MÉTHODE REINERT

Abstract. How statistical analysis of texts can help research in didactics: the example of Reinert's method. In this paper, we analyse statistical analysis of texts' potential for researchers in didactics. We focus in Reinert's method. Research methodologies lead us to analyze corpus with large amount of text. Without making it the only tool of analysis we show on two specific examples how this statistical treatment makes it possible to access corpuses hardly accessible otherwise or to make hypotheses facilitate then the human analysis. First we present the foundations and techniques related to this kind of analysis and then we develop the analysis around two examples: the processing of exchanges between trainers and trainees through a digital portfolio and the analysis of the mathematics primary school curricula in France since 1976.

Résumé. Dans cet article nous analysons les potentialités offertes aux didacticiens par l'analyse statistique des données textuelles au moyen de la méthode Reinert. Les méthodologies de recherche nous amènent en effet à analyser des corpus comportant de grandes quantités de texte. Sans en faire l'unique outil d'analyse, nous montrons sur deux exemples spécifiques comment ce traitement statistique permet d'accéder à des corpus difficilement accessibles sinon ou d'émettre des hypothèses pour faciliter ensuite l'analyse « manuelle ». Nous commençons par exposer les fondements et les techniques liés à ce type d'analyse puis nous développons l'analyse autour de deux exemples : le traitement des échanges formateurs/stagiaires au travers d'un portfolio numérique et l'analyse des programmes de mathématiques de l'école primaire en France depuis 1976.

Mots clés. Statistiques textuelles, Recherches en didactique, Mondes lexicaux,

Introduction

L'arrivée depuis des dizaines d'années des outils numériques : documents numériques, échanges par courriels, MOOC, plateformes de cours en ligne, blogues, espaces de dialogues… rend accessible aux chercheurs en didactique une masse de données qui n'existait pas auparavant. En effet, là où il fallait enregistrer puis transcrire les échanges entre un étudiant et son formateur, lire et retranscrire les annotations des enseignants sur les copies, une partie importante des échanges transite maintenant par des canaux numériques. Ils sont facilement accessibles sous réserve des autorisations d'usage. Ces données, pour beaucoup textuelles, sont un « gisement » d'informations nouvelles pour les recherches en didactiques.

ANNALES de DIDACTIQUE et de SCIENCES COGNITIVES, volume 23, p. 179 - 200.

De plus, qu'il fasse des analyses de cas ou des analyses à grande échelle, le chercheur peut se trouver également confronté à des corpus de données comportant de grandes quantités de texte. Si cela peut sembler une évidence pour les études à grande échelle, notamment lorsqu'elles comportent des questionnaires avec des champs de saisie de texte libre, cela peut apparaître plus surprenant pour les études de cas. Pourtant, ces études amènent le chercheur à enregistrer des temps de classes ou des entretiens qui, une fois transcrits, forment un corpus textuel parfois encore plus volumineux que pour des études à plus grande échelle. Par exemple (Labbé & Labbé, 2012, p. 5) qui analysent les questions ouvertes des sondages d'opinion, dénombrent entre 86 et 108 mots par répondant à ce type de question. Pour obtenir un corpus de 78 813 mots, l'échantillon de départ était de 1 010 personnes (727 ayant accepté de répondre à la question ouverte). Par comparaison nous avons obtenu 79 298 mots en transcrivant les paroles des formateurs et des stagiaires durant 5 formations (Emprin, 2007).

Comment alors exploiter ces données textuelles ?

Nous avons cherché à savoir comment étaient analysées les données textuelles dans les recherches en didactique des mathématiques. En nous basant sur le classement des revues établi par la European Society for Research in Mathematics Education (Toerner & Arzarello, 2012, p. 52‑53) nous avons consulté les articles publiés dans les sept revues de catégorie A* et A. Une recherche selon les mots clefs « textual analysis » nous a permis d'extraire quarante-sept articles. La consultation des résumés et des bibliographies fait apparaître que les textes y sont traités par des analyses de contenus, des classifications basées sur des analyses grammaticales, syntaxiques, sémiotiques, pédagogiques, sociologiques, psychanalytiques, des tâches mathématiques ou des mathématiques enseignées. Toutes ces méthodes sont basées sur la lecture et la catégorisation par un chercheur ce qui les rend très coûteuses quand le corpus de données est très grand. En tant que chercheur en didactique, nous avons été confrontés à ces types de corpus et nous avons utilisé des systèmes d'analyse et de classifications du corpus basés sur des lectures successives. Les biais d'interprétation des textes liés au lecteur peuvent être réduits par un codage en double aveugle, mais il rend le travail d'analyse encore plus coûteux.

Pour résoudre ces difficultés, des réponses existent dans d'autres champs de recherche comme celui de l'analyse en management stratégique :

« L'Analyse de Données Textuelles (A.D.T.) regroupe les méthodes qui visent à découvrir l'information « essentielle » contenue dans un texte, et le foisonnement de nouveaux outils auquel on peut assister aujourd'hui est à la conjonction de deux demandes différentes :

- d'une part une demande des entreprises, qui peuvent aujourd'hui collecter très facilement une grande quantité de textes avec Internet (articles, brevets, dépêches, rapports, études, mais aussi courriels, messages de forums, enquêtes clients, fiches de centres d'appel, descriptifs de produits...). […]

- et d'autre part une demande des chercheurs, qui ont besoin d'une alternative est à de traditionnelles analyses de contenu jugée trop subjective, soit à de simples analyses par mots-clés jugées trop pauvres (Bournois et al., 2002). » (Fallery & Rodhain, 2007, p. 2)

Ce type d'analyse existe également dans les champs de la politique ou la littérature. (Mayaffre, 2005) en fait le constat : « Depuis sa constitution à la fin des années 1960, la lexicométrie politique a connu en France une heure de gloire pour aujourd'hui s'essouffler. […]. Dans le même temps, notons que le développement d'une lexicométrie littéraire, dont on date les premiers balbutiements dès l'après-guerre avec les travaux de Busa (voir l'historique dans Busa, 1998) puis de Guiraud (Guiraud, 1954), a connu le même désenchantement […] ».

Nous avons alors souhaité confirmer que ce type de méthodes était peu utilisé en didactique et dans le domaine des recherches en éducation en général par une recherche dans la base de données ERIC (Education Resources Information Center[1]).

- Les mots clés « analyse » « données » « textuelles » renvoient quatre réponses dans le champ de la littérature et de la linguistique ;
- « textual data analysis » renvoie cinq réponses dont deux sont en relation avec des problématiques d'éducation ;
- « lexicométrie » en anglais et en français donne trois résultats, dont un en relation avec l'éducation ;
- « text mining » permet d'obtenir 196 résultats dont seulement 17 sont classés dans les recherches en éducation.

Ces éléments nous amènent à nous demander si les outils statistiques d'analyse de données textuelles permettraient d'accompagner le travail du chercheur en didactique, d'analyser des corpus inaccessibles en raison de leur taille et de fiabiliser les analyses par lecture.

Une méthode a particulièrement retenu notre attention : la détermination de mondes lexicaux au moyen de l'analyse statistique des cooccurrences (Reinert, 2007, 1986), elle-même basée sur l'approche fréquentiste ou l'analyse géométrique

[1] https://eric.ed.gov

de données (Benzécri, 1973 ; Benzécri & Benzecri, 1984). Cette méthode est implémentée à l'origine dans le logiciel ALCESTE[2] (Analyse des Lexèmes Cooccurrents dans les Énoncés Simplifiés d'un Texte), mais également dans le logiciel IRaMuTeQ[3] (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires). Nous privilégions ici ce dernier, car il est libre, multilingue et possède des versions pour les trois grands types de systèmes d'exploitation (Windows®, Mac OS® et Linux).

Nous examinons donc, dans cet article, les apports méthodologiques de ce type d'analyse statistique des données textuelles aux recherches en didactique en prenant comme appui la didactique des mathématiques.

Après une description de la méthode statistique utilisée, nous en explorons les potentialités au travers de deux usages : traiter des échanges entre formateurs et stagiaires et analyser des textes des programmes d'enseignement des mathématiques à l'école primaire française depuis les années 1970.

1. De l'analyse statistique à la détermination de mondes lexicaux

1.1. Posture de recherche

Dans notre approche il ne s'agit pas de réduire la méthode d'analyse des données textuelles à un traitement informatique, mais bien d'intégrer dans nos méthodes des résultats statistiques. Cela ne nous dispense pas d'avoir un regard sur les points de vue épistémologiques sous-jacents à ces traitements, sans pour autant en faire nos propres fondements. Les résultats statistiques sont pris ici comme un indicateur parmi d'autres. Nous considérons que nous pouvons interpréter les informations statistiques comme la présence ou l'absence significative d'un terme dans un corpus ou les cooccurrences (des présences conjointes de plusieurs termes dans des portions de textes définies) significatives avec nos cadres habituels comme la double approche (Robert, 1999) pour les pratiques enseignantes, les genèses instrumentales (Rabardel, 1995) pour analyser l'introduction d'un artefact numérique ou non. Cette démarche est bien dans la lignée des recherches sur le langage par exemple géométrique (Barrera Curin, Bulf & Venant, 2016, p. 73) qui combinent analyse didactique avec d'autres types d'approches comme l'analyse sémantique, discursive.

---

[2] CNRS-PRINTEMPS.

[3] Développé au sein du Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales de l'Université de Toulouse 3 – Paul Sabatier (LERASS) au sein de l'équipe REPERE (Représentations et Engagements Professionnels, leurs Evolutions : Recherches et Expertises) du CREFI-T (Centre de Recherche en Éducation, Formation et Insertion de Toulouse)

1.2. Premier traitement des textes et statistiques textuelles

Le traitement statistique des corpus textuels par un « simple » comptage des formes graphiques des mots peut apporter un certain nombre d'informations de base, mais il confronte l'utilisateur à des problèmes liés notamment aux variations de chaque mot : singulier, pluriel, masculin, féminin. Par exemple si l'on conteste l'usage du mot « formateur » il faut regarder « formateur », « formatrice », « formateurs », « formatrices » et parfois « maître-formateur »… Les logiciels comme IRaMuTeQ proposent un premier traitement permettant de simplifier le corpus : la lemmatisation « est une autre technique de normalisation : pour chaque forme déclinée d'un mot dans un document ou une recherche est identifiée une forme de base : le lemme. Les bénéfices de la lemmatisation sont les mêmes que pour la racinisation. De plus, en utilisant les formes de base des mots, le chercheur peut faire correspondre une clef de recherche exacte à une clé indexée » (traduit[4] de Korenius et al., 2004).

Le procédé de lemmatisation est une réduction à une forme canonique de la multitude de formes (flexions ou conjugaisons). Son utilisation permet de faire apparaître des occurrences d'utilisation de mots quels que soient la variante : masculin, féminin, pluriel et le temps pour les verbes.

Comme toute réduction elle masque un certain nombre d'informations (sans toutefois les perdre complètement, nous le voyons plus loin avec IRaMuTeQ), mais elle rend l'extraction de statistiques signifiantes plus aisée.

Lors du cours de statistiques pour la recherche dans le master MEEF (métiers de l'enseignement, de l'éducation et de la formation) pour les professeurs d'Éducation Physique et Sportive (EPS), nous leur avons fait soumettre à IRaMuTeQ les programmes d'enseignement de l'EPS au collège et au lycée en vigueur en 2014. Le tableau 1 fournit la liste des douze premiers lemmes classés par occurrence décroissante.

---

[4] "is another normalization technique: for each inflected word form in a document or request, its basic form, the lemma, is identified. The benefits of lemmatization are the same as in stemming. In addition, when basic word forms are used, the searcher may match an exact search key to an exact index key" (Korenius, Laurikkala, Järvelin & Juhola, 2004).

Lemme	occurences	Types
niveau	196	nom
compétence	103	nom
élève	88	nom
activité	88	nom
physique	80	adj
enseignement	76	nom
eps	73	nr
projet	57	nom
performance	56	nom
réaliser	55	ver
meilleur	49	adj
pratique	48	adj

Tableau 1. les douze lemmes les plus occurrents dans le corpus : forme, nombre d'occurrences, nature grammaticale.

Ce corpus comporte au total 15 111 occurrences et 1 633 lemmes ont été détectés à partir de 2 156 formes.

Ce qui a surpris des étudiants stagiaires, c'est la présence en premier du mot « niveau », mais surtout le classement du mot « performance » en dixième place avant « pratique » par exemple. Ils ont donc cherché à savoir comment ce mot était utilisé en accédant à ce qui est nommé dans IRaMuTeQ c'est-à-dire le contexte d'utilisation du mot dans le corpus. On peut y lire : « niveau 4 pour produire la meilleure *performance*, se préparer et récupérer efficacement… ». Sans entrer dans une analyse plus approfondie, le traitement donne une information qui n'était pas accessible aux étudiants sinon : un lecteur lit plus de fois le mot *performance* que le mot *pratique* quand il lit le programme d'EPS du collège. L'analyse doit être approfondie, mais il révèle des éléments que la lecture seule n'avait pas mis en évidence.

De même lorsque nous avons commencé à travailler avec des collègues spécialistes en relations internationales sur la comparaison des discours de l'état de l'union de Georges W. Bush et Barack Obama (Baillat, Emprin & Ramel, 2016) nous avons commencé par analyser les comptages de mots, de lemmes, de phrases, de mots par phrase dans les textes pour vérifier s'il existait des différences dans la façon de s'exprimer des deux présidents. Est-ce que les phrases utilisées par Georges W. Bush sont moins élaborées (nombre de mots, richesse du vocabulaire utilisé…) que

celles de Barack Obama ? Il s'avère que les phrases des discours de G. W. Bush font en moyenne 18,2 mots alors que celle de B. Obama en font 17. En revanche les discours du premier sont plus courts et moins riches en lemmes (en moyenne 4 900 mots et 3 526 lemmes contre 7 000 et 4 592 lemmes pour B. Obama). Ces premiers éléments ne permettent pas à eux seuls de construire une analyse, mais ils fournissent là encore des données objectives permettant de commencer la réflexion en caractérisant le corpus de données et de mettre à distance certains *a priori* (l'image de G. W. Bush par rapport à celle de B. Obama laissait penser à des phrases plus simples donc plus courtes pour le premier).

1.3. Les unités de contexte et les cooccurrences

Ces éléments statistiques aident également l'utilisateur à définir les unités de contexte : « Nous entendons par « unités de contexte » (u.c.), tout segment de texte pouvant servir de support à l'étude des représentations envisagées. Généralement la segmentation du corpus en u.c. suit la segmentation « naturelle » de sens : proposition, phrases, paragraphes, réponses, etc. » (Reinert, 1986). La méthode de traitement que nous utilisons nécessite de segmenter le texte en unités d'une taille préétablie, plus la taille de ces unités est proche de la réalité du texte plus les résultats fournis pourront être significatifs. Si on considère qu'un paragraphe correspond à une idée et que pour les discours de l'état de l'union il y a en moyenne quatre phrases par paragraphe avec une taille de phrase moyenne de 18 mots, on peut faire le choix de paramétrer les unités de contexte à 70 mots. Pour un dialogue en classe où les échanges font en moyenne 4 phrases de 10 mots on choisira une taille d'unités de contexte de 40 mots.

La définition de la taille des unités de contexte est d'autant plus importante que pour aller plus loin dans l'analyse des textes, le concept de *cooccurrence* est mobilisé. Il s'agit de compter le nombre de fois où des mots apparaissent ensemble, dans la même unité contextuelle. Une hypothèse de base de ce travail est que « pour pouvoir énoncer, le sujet énonçant doit se représenter ce qu'il va dire dans un certain espace mental (qui lui sert de "référence"). Le choix de cet espace référentiel, de ce "lieu" – qui ne dépend pas forcement d'une opération consciente – implique le choix d'un type d'objet : il implique, par là même, un type de vocabulaire. En conséquence, l'étude statistique de la distribution du vocabulaire dans les différents énoncés d'un corpus doit permettre une discrimination de ce vocabulaire révélatrice des différents choix référentiels effectués par l'énonciateur. » (Reinert, 1993, p. 2)

La méthode Alceste consiste à redécouper le corpus en unités de contexte élémentaires (UCE), plus petites que l'unité de contexte, fournies par les textes (comme les phrases) et établir un tableau de comptage des cooccurrences entre les mots lexicaux (que l'auteur nomme à l'époque de la création de la méthode mots

pleins), c'est-à-dire à l'exclusion des mots grammaticaux comme les prépositions, les pronoms (appelés par l'auteur mots outils). En effet « les mots pleins ont cependant la faculté de susciter des contenus en tant qu'ils stabilisent nos visions du monde, mais le contenu n'est pas dans le mot ; il est dans l'acte, dont le mot est une trace. » (Reinert, 2008, p. 982). Ce tableau que l'on peut imaginer avec en ligne les mots lexicaux lemmatisés et en colonne les UCE est ensuite traité au moyen d'une analyse factorielle des correspondances (AFC) illustrée figure 1.

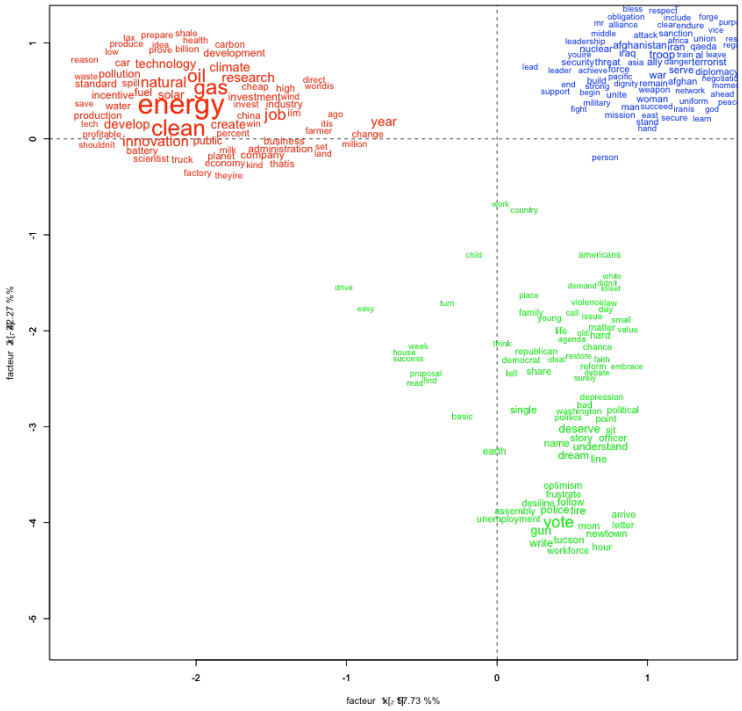

Figure 1. Représentation de l'AFC des cooccurrences des mots lexicaux lemmatisés dans les discours de l'état de l'union de B. Obama

Par essence, l'AFC permet d'identifier si des UCE se ressemblent du point de vue de la distribution des mots. Nous ne décrivons pas en détail le procédé de calcul des AFC, mais uniquement ses spécificités ayant un impact sur l'analyse finale. Tout d'abord le tableau des UCE n'est pas traité directement, mais en relation avec un tableau équiprobable par la métrique du χ^2 normalisé. En conséquence ce ne sont pas les cooccurrences les plus fréquentes qui sont mises en avant, mais celles qui se distinguent le plus d'une répartition aléatoire. On recherche ensuite la factorisation optimale de la matrice, c'est-à-dire que le premier axe factoriel est celui qui représente le plus (en termes de χ^2) les données de la matrice initiale, puis le second axe factoriel... On parle alors, pour chaque axe, du pourcentage de

variance exprimé. Ainsi en représentant les données dans un espace à deux dimensions suivant les deux premiers axes factoriels, ce graphique exprimera un certain pourcentage des données initiales.

1.4. Classification hiérarchique descendante et mondes lexicaux

Ensuite est effectuée une classification hiérarchique descendante :

« 1. une analyse factorielle des correspondances (A.F.C.) est menée sur le tableau, puis pour toutes les partitions possibles le long du 1er facteur de l'AFC, l'inertie inter-classe est calculée. Une première coupure intervient pour la partition qui maximise l'inertie inter-classe.

2. Chaque unité du tableau est permutée d'une classe à l'autre et l'inertie inter-classe est recalculée. Si celle-ci est supérieure à l'inertie inter-classe précédente, la permutation est conservée. Cette partie de l'algorithme boucle jusqu'à ce qu'aucune permutation n'augmente l'inertie inter-classe.

3. Les formes spécifiques d'une classe (au sens du χ^2) sont retirées de l'autre. » (Ratinaud & Marchand, 2012, p. 837).

Les classes ainsi obtenues, si elles résistent à des variations de découpage des UCE, révèlent des mondes lexicaux : « Ce modèle simplifié de représentation statistique d'un discours suffit à mettre en évidence, du moins dans l'analyse de certains corpus, une tendance du vocabulaire à se distribuer dans des mondes lexicaux stabilisés… » (Reinert, 2008, p. 982).

C'est cette méthode qui est implémentée et adaptée dans IRaMuTeQ, notamment au niveau de construction des matrices qui, dans le cadre de ce type d'analyse, sont principalement constituées de zéros.

« Par ailleurs, les adaptations réalisées sur l'algorithme de classification rendent possible l'utilisation de la méthode ALCESTE classique (avec la double classification sur u.c.) sur des corpus de plusieurs dizaines de millions d'occurrences, tout en travaillant sur un nombre important de formes pleines. La procédure que nous avons suivie pourrait également permettre d'améliorer la recherche de stabilité dans l'analyse des gros corpus en utilisant les classifications sur les formes moins fréquentes pour préciser le contour des classes obtenues sur les classifications des formes fréquentes. » (Ratinaud & Marchand, 2012, p. 844)

1.5. Les outils pour le chercheur

Les textes soumis à l'analyse peuvent être de différentes natures : ils peuvent provenir d'un même auteur (discours, texte d'un auteur), de plusieurs auteurs (Bush *vs* Obama, échange sur une plateforme en ligne, transcription des séances de formation) ou être une œuvre collégiale (les programmes…). Les textes soumis

peuvent aussi appartenir à différents contextes (blogues différents, différentes années, dates…). Il est important que l'analyse puisse tenir compte de ces paramètres sans pour autant qu'ils influent sur le traitement statistique. Les lignes étoilées d'IRaMuTeQ permettent de réattribuer, *a posteriori*, ces paramètres aux différents mondes lexicaux.

Au début de chaque portion de texte soumise, le chercheur introduit une ligne étoilée « **** » et indique les paramètres, par exemple le sexe du locuteur : sexe_F ou sexe_M, son nom loc_Arne ou l'année de production annee_2009. Ces paramètres sont définis par le chercheur, de même que les valeurs possibles, celles qui suivent le « _ ». Ces lignes ne sont pas incluses dans l'analyse factorielle, mais elles sont conservées pour la restitution. Elles permettent donc de vérifier des hypothèses et attribuer certains mondes lexicaux à l'un ou l'autre des paramètres. Dans le travail sur les discours de l'état de l'union, nous avons dans un premier temps soumis l'ensemble des textes en choisissant les paramètres suivants discours_bush/discours_Obama et annee_xxxx. Notre première question était de savoir si les mondes lexicaux pouvaient être communs aux deux présidents et si certaines années (par exemple 2011, suite au 11 septembre) pouvaient marquer des changements dans les discours. Le premier résultat de l'analyse par IRaMuTeQ a été de mettre en évidence deux classes stables illustrées par les deux premières branches du dendogrammes (figure 2) : l'une attribuée à G.W. Bush de façon significative (en termes de χ^2 comme le montre la copie d'écran du logiciel en figure 3) et l'autre à B. Obama.

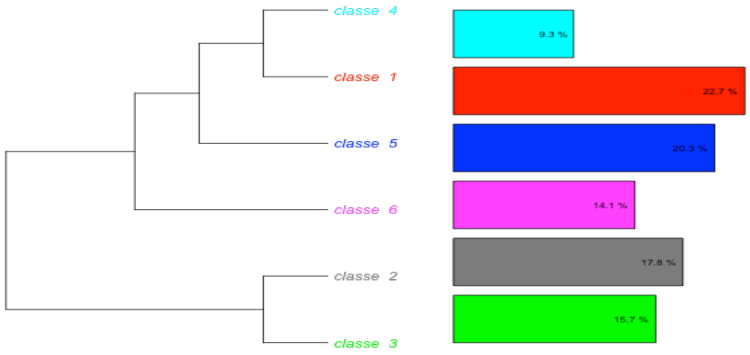

Figure 2. Dendrogramme présentant l'analyse IRaMuTeQ pour les discours de l'état de l'union de G. W. Bush et B. Obama. La première branche principale est révélatrice des discours de B. Obama et la seconde de G.W. Bush

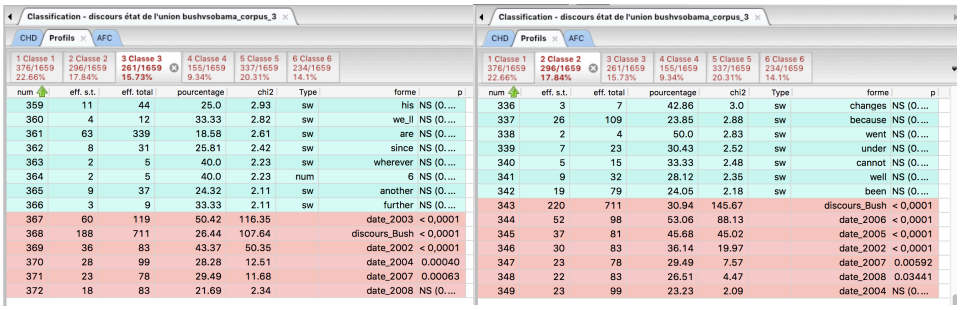

Figure 3. Copie d'écran des tableaux fournis par IRaMuTeQ montrant la significativité en termes de χ^2 de la modalité Bush pour la variable discours pour les classes 2 et 3

La thématique des « États voyous » (rogue nations) est réactivée dans le discours de G. W. Bush après le 11 septembre 2001. Alors même que ce terme n'apparaît qu'une seule fois dans le corpus « on le retrouve cependant en filigrane des discours de Bush Jr. grâce aux nombreuses cooccurrences entre ces pays et un vocabulaire associé à la menace et à la défense de la démocratie » (Baillat et al., 2016, p. 234). Le traitement des lignes de paramètres a mis en évidence que les mondes lexicaux des deux présidents étaient distincts au regard de cette méthode et que pour aller plus loin il nous fallait reprendre le traitement séparé des deux corpus, ce que nous ne détaillons pas ici.

Nous avons précisé que les tableaux de contingences ne prenaient en compte que les mots lexicaux. Cela ne veut pas dire que les formes supplémentaires, les mots grammaticaux soient perdus. Comme pour les lignes étoilées, elles sont réintégrées à l'analyse et peuvent être interprétées. Par exemple, dans l'analyse des discours de B. Obama une classe fait apparaître de façon cooccurrente les mots « school », « teacher », « child », « education ». Nous donnons dans le tableau 2 un extrait du tableau des χ^2 ordonné suivant la classe 3. Ce qui est intéressant c'est que cette classe est caractérisée par la présence des formes « every » et « each ». Le fait de pouvoir intégrer *a posteriori* ces formes supplémentaires sans influencer le traitement statistique permet de compléter l'analyse.

À chacune des étapes, le logiciel permet de retrouver les UCE où se trouvent les mots ou les cooccurrences. Cet outil, nommé *concordonancier* et illustré par la capture d'écran en figure 4, permet en permanence au chercheur de revenir au texte et donc à l'acception dans laquelle un terme donné est utilisé.

Comme dans toute méthode statistique le logiciel fournit également des données calculées permettant d'analyser les résultats. Le pourcentage de segments classés est un de ces paramètres, qui permet de savoir quelle part du corpus est effectivement prise en compte dans la classification.

D'autres outils sont également disponibles pour permettre de représenter, de façon graphique, les résultats obtenus : les nuages de mots, les représentations graphiques de AFC… Nous en donnons quelques illustrations dans les exemples qui suivent.

	classe 1	classe 2	classe 3	classe 4	classe 5
school	-2,97	-8,327	118,959	-1,662	-13,78
teacher	-4,314	-5,039	102,578	-2,273	-10,124
college	-3,891	-4,545	91,356	-1,884	-9,131
child	-1,412	-2,164	75,582	-4,754	-8,273
student	-0,52	-4,545	68,606	-1,884	-9,131
education	-1,715	-6,54	54,558	1,636	-13,139
…					
every	-0,626	1,278	14,883	-2,7	-4,077
…					
each	-4,314	-0,296	11,879	-0,839	0,023
…					
possible	0,129	-3,566	10,944	-0,139	-0,989

Tableau 2. Tableau des χ^2 ordonné suivant la classe 3 pour l'analyse des discours de B. Obama

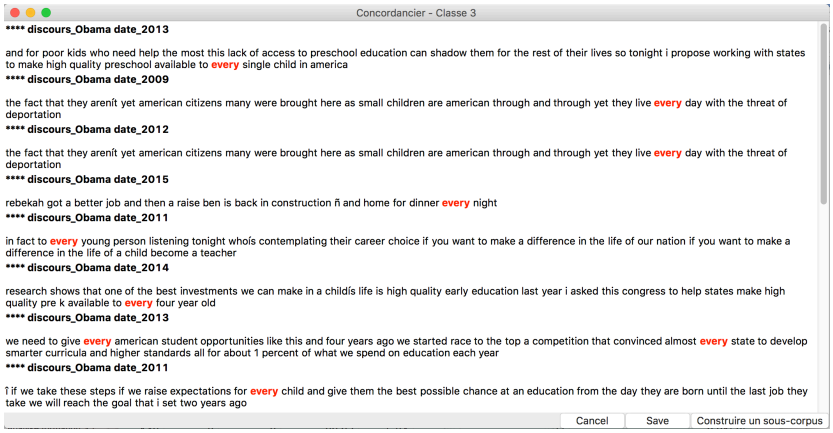

Figure 4. Copie d'écran du concordancier pour la forme supplémentaire « every » dans l'analyse des discours de B. Obama

1.6. Conclusions sur la méthode Alceste

La méthode Alceste implémentée dans le logiciel IRaMuTeQ est donc un outil pour mettre en évidence des mondes lexicaux stabilisés dans nos corpus de données textuelles. Cette mise en évidence n'est pas une fin en soi, mais elle vise à donner des éléments de réflexion au chercheur en s'appuyant sur les traces purement langagières laissées dans le texte par les locuteurs. Par ailleurs, IRaMuTeQ fournit

des outils complémentaires permettant d'approfondir l'analyse qualitative des données ainsi obtenues.

2. Deux exemples pour mettre en évidence les potentialités de ces outils statistiques

Nous prenons maintenant deux exemples issus de nos travaux pour illustrer comment l'usage de ce type d'outil peut contribuer au travail du didacticien.

Le premier consiste à traiter un corpus de grande taille pour lequel nous disposons de nombreux paramètres. Il s'agit d'échanges entre les formateurs et les stagiaires autour d'un portfolio centré sur l'évaluation des compétences professionnelles. Nous disposons donc de plusieurs paramètres comme l'année scolaire, la discipline dans laquelle le travail a été rendu, le site de formation du stagiaire. Le traitement statistique nous permet, dans ce cas, de répondre à des questions de significativité des résultats obtenus.

Le second part d'un corpus déjà existant : les textes des programmes officiels. Il permet de montrer comment l'analyse par des statistiques textuelles peut mettre en évidence des questions *a priori*.

2.1 Le portfolio numérique : analyse d'une grande quantité de texte et croisement des méthodologies

Dans une recherche menée en 2011, notre objectif était de comprendre l'évolution des usages d'un portfolio numérique au sein de l'IUFM de Reims Champagne Ardenne (Connan & Emprin, 2011). Pour accéder aux usages, nous disposons de deux types de traces :

- des enquêtes de l'observatoire des formations, structure organisant l'évaluation des formations par les étudiants et les formateurs ;
- les échanges au travers du carnet de bord informatisé (CBI) : le portfolio numérique, passage obligé de toutes les évaluations des fonctionnaires stagiaires.

Ces derniers devaient faire la preuve, par le dépôt de traces (textes d'analyse de séances, travaux d'élève…) de l'acquisition des compétences professionnelles attendues en fin de formation initiale. Ces dépôts donnaient lieu à des échanges entre stagiaires et formateurs autour des compétences acquises et des travaux déposés. Nous revenons, ici, plus en détail sur l'analyse de ce dernier corpus comportant 6 823 textes (433 870 mots, dont 16 489 mots distincts) écrits par les formateurs et les formés pendant quatre années (de septembre 2006 à août 2010). Au départ nous avions 12 255 échanges desquels ont été uniquement supprimés les échanges quasi-vides : « merci », « bien reçu », « vu », « laisser votre message ici »… pour obtenir les 6 823 textes finalement analysés.

Le traitement statistique a d'abord l'intérêt de pouvoir traiter rapidement ces textes, mais également de déterminer si certains phénomènes sont attachés à des paramètres. Les questions que nous nous posions étaient en premier lieu de savoir si les échanges avaient évolué dans leur contenu (échanges techniques, administratifs, évaluatifs, formatifs…) entre 2006 et 2010 mettant ainsi en évidence une forme d'évolution des usages de l'artefact numérique. Une autre question portait sur les spécificités disciplinaires ou locales : est-ce que certains usages étaient révélateurs d'une spécificité disciplinaire ou de pratiques liées à un site de formation ? Les textes ont été indexés par leur année universitaire d'émission, la discipline du formateur, le site de formation du stagiaire et du formateur. Le traitement informatique permet alors de mettre en évidence des présences statistiquement significatives d'échanges provenant d'un paramètre ou de l'autre.

Le traitement par le logiciel ALCESTE identifie quatre classes stables, en hiérarchisant 8 213 unités de contexte élémentaire (UCE) avec 1 371 formes analysées qui correspondent à 93 % des UCE classées (figure 5). Le premier résultat est que l'année n'apparaît de façon significative dans aucune des classes. L'hypothèse d'une évolution de la nature des échanges au cours des années n'est pas vérifiée par l'analyse des mondes lexicaux. Les autres paramètres tels que la discipline n'apparaissent qu'anecdotiquement au niveau des mondes lexicaux. Les échanges formateurs/stagiaires sont donc indépendants du site et de la discipline d'enseignement. Les classes sont interprétées par les chercheurs de la façon suivante :

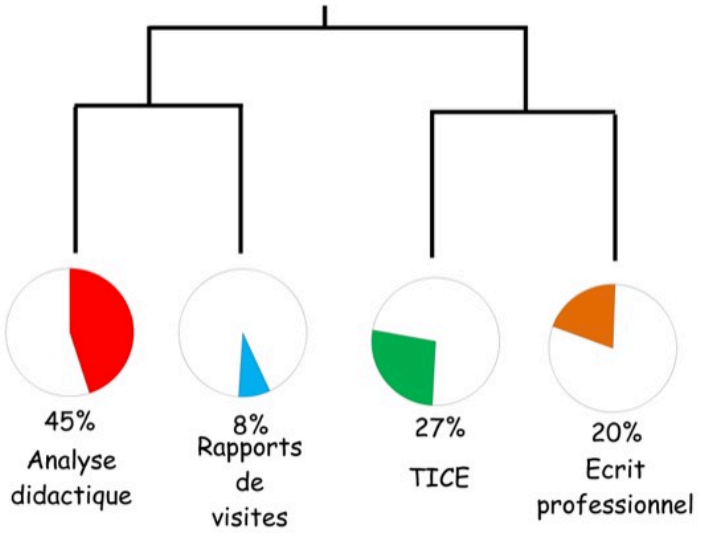

Figure 5. Analyse et interprétation des classes Alceste

L'analyse de ce corpus par la méthode Reinert met en évidence la correspondance entre les quatre mondes lexicaux et les quatre types de travaux que les stagiaires avaient à déposer sur les CBI. Ce résultat montre qu'il y a une forme de dissociation dans les différentes formes d'évaluation : l'analyse didactique demandée dans le cadre de la formation, les rapports de visites correspondant à l'évaluation du stage, l'évaluation des usages du numérique et l'écrit professionnel. Ce constat permet de contester les différences entre les échanges sur l'analyse de travaux didactiques (première catégorie) et ceux sur l'analyse des usages des technologies (troisième catégorie) : « La classe 3 regroupe les commentaires liés aux TICE ; on y voit apparaître des termes dédiés tels que « fichier », « FOAD » (Formation Ouverte A Distance), « mutualisation », « ressources » « informations » ainsi que des listes de compétences du C2i® 2e (dépôts effectués à la fois dans le cadre du référentiel métier et de celui du C2i® 2e). Les absences significatives sont celles des termes de la classe 1, par exemple « élève » ou « séance » ainsi que celles des adjectifs en général. Il y a donc une véritable singularité. » (Connan & Emprin, 2011, p. 14)

Une analyse de la classe 2 montre qu'elle est constituée d'échanges réduits (ce que confirme son poids de 8 %). Dans la mesure où un rapport détaillé est associé à ces échanges, on peut supposer que des éléments plus formatifs y sont indiqués. Le dialogue sur la plateforme entre le formateur et le stagiaire est donc réduit.

Ce premier exemple montre le rôle de ce type d'analyse :

- la mise en évidence d'une dissociation des différentes formes d'évaluation du point de vue des échanges formateurs/stagiaires. Elle permet de donner des éléments de discussion pragmatiques et dépersonnalisés pour engager une réflexion par exemple sur l'évaluation de la formation des fonctionnaires stagiaires.
- L'hypothèse d'évolution des pratiques n'est pas confirmée en termes de transformation des mondes lexicaux. Si l'on veut regarder l'évolution des pratiques, il ne faut donc vraisemblablement pas la chercher dans les discours formateurs/formés, mais dans la perception de l'outil. Le travail sur les questionnaires (évaluation des formations par les stagiaires et les formateurs) montrera, par exemple, que l'outil est de plus en plus perçu comme simple d'utilisation, facilitateur d'échanges entre formateurs et stagiaires, mais également comme outil d'évaluation et non pas de formation. Ce dernier point met en évidence l'importance de croiser les méthodologies d'analyses.

La méthode Reinert a été utilisée seule, sans lecture complète du corpus par un chercheur, mais elle a permis de mettre en évidence des faits statistiques utiles pour mieux comprendre les pratiques liées au portfolio. Elle a également invalidé certaines hypothèses de recherche comme la présence de spécificité locale.

Le second exemple choisi illustre un usage plus exploratoire d'un corpus textuel. Il ne s'agit pas cette fois de répondre à une question, mais d'explorer les données à la recherche d'hypothèses.

2.2. Traiter un corpus et émettre des hypothèses : analyse des programmes officiels

Nous avons été sollicités en 2015 par le conseil supérieur des programmes (CSP), comme beaucoup d'autres chercheurs en didactiques, pour contribuer à la réflexion sur les programmes de 2016[5]. Les questions posées portaient notamment sur une analyse des programmes de 2002 et de 2008. Cette demande nous a incité à regarder si, d'un point de vue lexical les programmes des différentes années présentaient des singularités. Après la parution des programmes, nous avons continué notre analyse en partant des programmes de 1976 jusqu'à ceux actuels (mis en œuvre à la rentrée 2016). Le choix de commencer l'analyse en 1976 correspond au fait que ce sont les premiers programmes post « réforme des mathématiques modernes ». Cette dernière marquait une rupture telle que nous avons fait le choix de ne pas l'inclure dans l'analyse.

Notre corpus comporte donc la partie mathématique des programmes officiels de 2016, 2008, 2002, 1995, 1985, 1980, 1976 à 1978 (programmes par année). Ce corpus comporte 48715 mots.

Nous avons paramétré notre corpus par l'année d'émission des programmes et le cycle concerné. Ce concept de cycle est anachronique en ce qui concerne les programmes des années 1970 ; de plus il a changé dans les derniers programmes officiels : le cycle 2 intègre maintenant le CE2 et le cycle 3 la 6ème. Comme les programmes des années 2000 sont décrits par cycle, nous avons conservé le paramétrage en cycle sans pouvoir extraire le CE2 du cycle 3. En revanche pour les programmes libellés par année nous avons associé le CE2 au cycle 2. Nous devons donc en tenir compte dans l'analyse.

Le traitement est réalisé par la méthode Reinert implémentée dans IRaMuTeQ : une double classification sur RST (regroupement de segments de textes), permet de classer 58 % des formes et fait apparaître 6 classes décrites en figure 6.

---

[5] A consulter sur la page http://www.education.gouv.fr/cid82307/le-conseil-superieur-des-programmes-contributions-des-experts-sollicites-par-les-groupes-charges-de-l-elaboration-des-projets-de-programmes.html

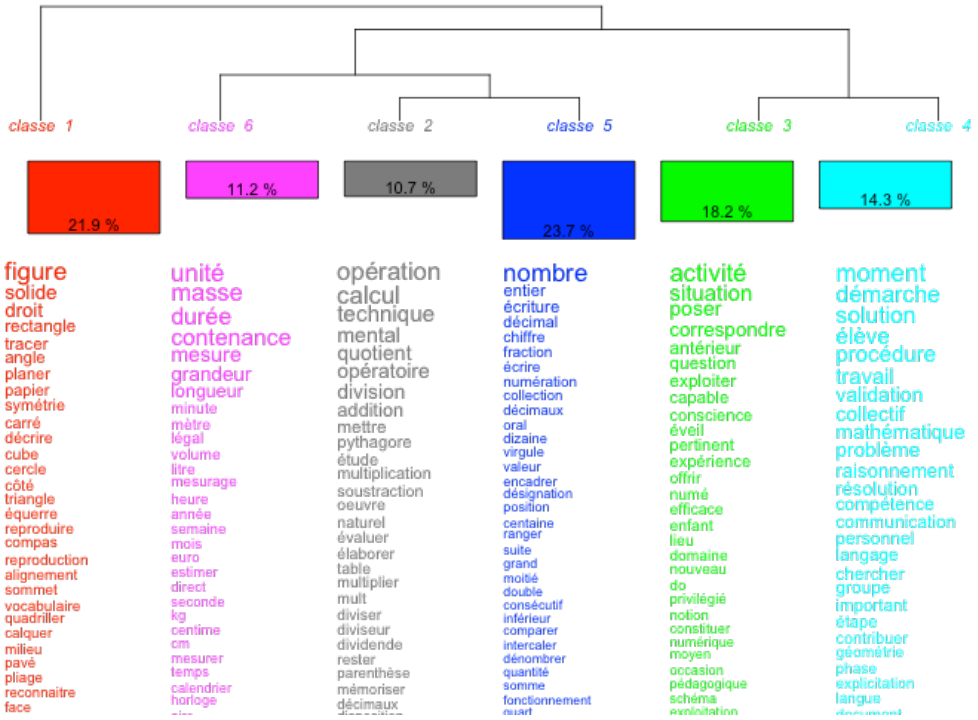

Figure 6. Dendrogramme des classes dans l'analyse des programmes d'enseignement des mathématiques à l'école primaire depuis 1976

Nous avons réalisé une analyse des classes en nous basant sur les listes de mots cooccurrentes, sur les paramètres significatifs dans les classes et les formes supplémentaires qui n'apparaissent pas dans le dendrogramme.

Des hypothèses peuvent être émises suite à cette analyse (58 % des UCE - unités de contexte élémentaire- traitées) :

- Il semble y avoir une spécificité du vocabulaire employé pour parler de géométrie. Ce monde lexical (classe 1) se distingue des autres classes. Cette classe est présente de façon significative dans les programmes de 2008 et de 2016.
- Les programmes de 2002 se distinguent par l'emploi d'un vocabulaire autour des démarches d'apprentissages, des procédures et des raisonnements des élèves.
- Les programmes de 76 à 80 insistent sur les apprentissages des techniques opératoires et les savoirs, mais également sur les activités dans lesquelles les mathématiques apparaissent.
- Les programmes de 2016 sont spécifiques par l'importance portée aux apprentissages liés à la mesure.

-

Niv.	Classe	Mode lexical	Programmes	Formes supplémentaires
	Classe 1	la géométrie	2008 et 2016, Cycle 2	
	Classe 6	la mesure	2016, cycle 2	
	Classe 2	opération et techniques de calcul	1976/78 et 1980	Savoir
	Classe 5	nombre en écriture décimal et travail sur la numération	2002, 2008, cycle 3	
	Classe 3	activités (manuelles, d'éveil…), situations (conçues par l'enseignant)	1976/78 et 1980 1995 et cycle 3	Devoir, permettre (aux enfants de développer des attitudes de recherche), pouvoir
	Classe 4	démarche, procédure, résolution	2002	

Tableau 3. Interprétation des mondes lexicaux issus de l'analyse des programmes d'enseignement des mathématiques à l'école primaire depuis 1976, suivant la classification hiérarchique

- Le cycle 1 n'apparaît pas spécifiquement dans les classes, mais les cycles 2 et 3 sont caractéristiques de certaines. S'il n'est pas étonnant, par exemple, pour que classe 5 où apparaît le vocabulaire spécifique aux nombres décimaux soit spécifique du cycle 3 il est plus surprenant que cycle 2 soit significativement présent dans la classe 1, centrée sur la géométrie. Cette faible particularité des cycles peut être intéressante à questionner alors même que les programmes des différents cycles peuvent être rédigés par des « équipes » différentes. Par ailleurs lorsque l'on analyse les programmes année par année, le vocabulaire autour du cycle 1 ressort comme spécifique avec l'usage du mot *enfant* à la place du mot *élève*. Là encore, il faut tenir compte du fait que seulement 58 % du corpus est interprété ici.

Le logiciel IRaMuTeQ permet donc de mettre en évidence des phénomènes liés spécifiquement au langage employé dans la rédaction de programmes. Ces phénomènes permettent de construire des hypothèses pouvant ensuite être traitées par d'autres méthodologies et croisées avec d'autres entrées. Ainsi l'analyse des manuels scolaires pourrait être révélatrice de la réception de ces programmes et de leur interprétation par les auteurs de ces outils. L'usage que nous avons fait, ici, du traitement des données textuelles est rendu rentable par la mise à disposition, au format numérique d'une grande quantité de textes.

3. Conclusions sur les potentialités et les limites de l'analyse statistique.

En constatant que les méthodes d'analyse de données textuelles ne sont que très peu utilisées dans les recherches en didactique alors qu'elles le sont dans d'autres champs, nous avons voulu explorer une méthode particulière d'analyse statistique des données textuelles pour en dégager les potentialités et les limites. En partant de deux exemples détaillés nous avons identifié deux fonctions : vérifier des hypothèses émises au préalable grâce a des données pragmatiques et avoir une approche inductive de l'analyse d'un corpus.

La première limite est liée aux fondements épistémologiques mêmes de la méthode. Il faut admettre que le ou les locuteurs laissent des traces de leurs intentions au travers du choix des mots utilisés : quand l'enseignant fait la classe, que le formateur interagit avec les stagiaires, lors de la production d'écrits professionnels, la réponse à une enquête ou la production d'un écrit officiel. Nous résolvons ce problème en ne faisant pas du traitement statistique des cooccurrences le fondement de notre analyse, mais un fait statistique à analyser, et à comprendre aux moyens d'autres cadres théoriques tels que la double approche (Robert et Rogalski, 2002) pour l'analyse des pratiques enseignantes.

Une seconde limite est la méthode statistique elle-même qui perd une partie de l'information en raison de la factorisation et de la projection d'un nuage de points dans un espace à n dimensions (les n axes factoriel) sur une espace à deux dimensions ainsi que de l'utilisation de formes réduites lemmatisées. Il faut donc tenir compte de cette perte d'information dans l'analyse.

La méthode de découpage du corpus, par des segments de longueur fixe, rend également le découpage des UCE potentiellement risqué, notamment quand de grandes unités de contextes sont mêlées à des unités plus petites. Le choix de la taille des UCE est alors important et doit être manié avec précaution. En modifiant ce paramètre, il n'est pas rare de perdre une part importante du corpus dans la classification.

Les exemples pris montrent également l'importance de croiser les méthodologies pour répondre aux questions de recherche posées. Dans le cas du travail sur le

portfolio, le questionnaire permet d'accéder à des éléments subjectifs : la perception des stagiaires sur l'outil, ou plus précisément la façon dont les stagiaires veulent informer l'institution de leur perception de l'outil. Les stagiaires et les formateurs ressentent l'outil comme de plus en plus simple au fil des années. Si cela peut être facilement expliqué pour les formateurs qui majoritairement restent les mêmes tous les ans, les stagiaires eux changent. Leurs perceptions de la simplicité de l'artefact sont donc vraisemblablement liées aux interactions avec des formateurs de plus en plus aguerris dans l'usage du portfolio.

L'analyse des programmes permet d'émettre des hypothèses tirées non pas de présupposés ou d'impressions, mais d'éléments objectifs. En effet, le logiciel « ne sait pas lire », il traite les mots sans relation avec leur sens et lorsqu'il montre par exemple que les programmes de 2002 sont caractérisés par un monde lexical lié à la démarche et aux procédures des élèves il le fait sans aucun préjugé sur le contexte. Lorsque l'on travaille sur les cooccurrences, il faut être conscient que le logiciel classe dans la même catégorie un texte qui dirait : « il faut prendre en compte les procédures des élèves » et « il ne faut pas prendre en compte les procédures des élèves ». Le traitement statistique montre que les deux textes sont liés par le fait qu'ils parlent des « procédures », des « élèves » et de « prise en compte ». D'autres logiciels de lexicométrie ou de textométrie fonctionnant avec d'autres modèles d'analyse pourraient fournir une analyse croisée pertinente.

Le traitement statistique ouvre donc la possibilité au chercheur d'accéder à des corpus nouveaux comportant de grandes quantités de textes. Une entrée progressive dans ces corpus est alors possible en se concentrant sur des mondes lexicaux qui émergent.

Bibliographie

BAILLAT A., EMPRIN F. & RAMEL F. (2016), chapitre 12 – des mots et des discours. in *Méthodes de recherche en relations internationales* (p. 227-246). Presses de Sciences Po.

BARRERA CURIN, R., BULF, C., & VENANT, F. (2016). Didactique, sémantique et métaphores : analyse de langages en classe de géométrie. Annales de didactique et de sciences cognitives, **21**, p. 73‑78.

BENZECRI, J.-P. & BENZECRI, F. (1984), *Practique de lanalyse des donnes. Analyse des correspondances and classification.* Exposé élémentaire. 2e éd. Paris : Dunod.

BENZECRI, J.-P. (1973), *L'analyse des données* (Vol. 2). Paris : Dunod.

BOURNOIS, F., POINT, S., VOYNNET-FOURBOUL, C. (2002), L'analyse de données qualitatives assistée par ordinateur : une évaluation, *Revue française de Gestion*,

137.

BUSA, R. (1998). Dernières réflexions sur la statistique textuelle, in S. Mellet (éd.), *JADT 1998, 4e Journées internationales d'analyse des données textuelles,* UNSA-CNRS, Nice, p. 179-183.

CONNAN, P.-Y. & EMPRIN, F. (2011), Le portfolio numérique : quelles évolutions des usages et des représentations chez les formateurs d'enseignants ? *Revue Sticef. org*, **14**, 10.

EMPRIN, F. (2007, décembre 14), *Formation initiale et continue pour l'enseignement des mathématiques avec les TICE : cadre d'analyse des formations et ingénierie didactique.* Université Paris-Diderot - Paris VII. Consulté à l'adresse http://tel.archives-ouvertes.fr/tel-00199005

FALLERY, B., & RODHAIN, F. (2007). Quatre approches pour l'analyse de données textuelles: lexicale, linguistique, cognitive, thématique. In *XVI ème Conférence de l'Association Internationale de Management Stratégique AIM*S (pp. 1-16). AIMS.

GUIRAUD, P. (1954). *Les caractères statistiques du vocabulaire.* Presses universitaires de France.

KORENIUS, T., LAURIKKALA, J., JÄRVELIN, K. & JUHOLA, M. (2004), Stemming and lemmatization in the clustering of finnish text documents (p. 625 -633). *Proceedings of the thirteenth ACM international conference on Information and knowledge management,* ACM.

LABBE, C. & LABBE, D. (2012), *Analyser les questions ouvertes dans les sondages. In Comment convaincre ? Analyse scientifique de la campagne électorale 2012*. Grenoble.

MAYAFFRE, D. (2005), *De la lexicométrie à la logométrie*. Astrolabe, 1-11.

RABARDEL, P. (1995). *Les hommes et les technologies; approche cognitive des instruments contemporains.* Armand Colin, pp. 239. <hal-01017462>.

RATINAUD, P. & MARCHAND, P. (2012), Application de la méthode ALCESTE à de « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRaMuTeQ. *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles*, 835 -844.

REINERT, M. (1986), Un logiciel d'analyse lexicale. *Les Cahiers de l'analyse des données*, **11(4),** 471-481.

REINERT, M. (1993), Quelques problèmes méthodologiques posés par l'analyse de tableaux 'Énonces x Vocabulaire'. *Actes des secondes journées internationales d'analyse statistique des données textuelles*. Montpellier, 21-22.

REINERT, M. (2007), Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. *Langage et société,* **3***,* 189-202.

REINERT, M. (2008), Mondes lexicaux stabilisés et analyse statistique de discours. *Actes de la JADT 2008*, 981-993.

ROBERT, A. (1999). Recherches didactiques sur la formation professionnelle des enseignants de mathématiques du second degré et leurs pratiques en classe. *Didaskalia, 15,* 123-157.

ROBERT, A. & ROGALSKI, J. (2002). « Le système complexe et cohérent des pratiques des enseignants de mathématiques : une double approche », *Canadian Journal of Science, Mathematics and technology Education* (La revue canadienne de l'enseignement des sciences, des mathématiques et des technologies), **2(4)**, p. 505-528.

TOERNER, G., & ARZARELLO, F. (2012). Grading mathematics education research journals. *Newsletter of the European Mathematical Society, 86,* 52–54.

FABIEN EMPRIN
Laboratoire CEREP
Université de Reims Champagne Ardenne
fabien.emprin@univ-reims.fr