Déconstruction et reconstruction de corpus... À la recherche de la pertinence et du contexte

Lucie Loubère

Déconstruction et reconstruction de corpus... À la recherche de la pertinence et du contexte

Lucie Loubère[1]

[1]Lerass Université de Toulouse – lucie.loubere@iut-tlse3.fr

Abstract

Faced with corpora of large sets of texts, we propose a method of selection, based on the identification of segments of texts relevant to a topic by successive classification, then recomposition of the corpus with all the texts having at least one relevant segment . This approach makes it possible to preserve the contextualizations and narrative discourses surrounding a theme while excluding off-topic texts.

Résumé

Face aux corpus constitués de grands ensembles de textes, nous proposons une méthode de sélection, basée sur l'identification de segments de textes pertinents à une thématique par classification successive, puis recomposition du corpus avec l'intégralité des textes ayant au moins un segment pertinent. Cette démarche permet ainsi de conserver les contextualisations et discours narratifs entourant une thématique tout en excluant les textes hors-sujet.

Keywords: Big corpus, Reinert classification, Iramuteq

1. Introduction

La multiplication d'outils d'extraction de contenus numériques ou l'abonnement des universités aux bases de données de presse, sont autant de raisons favorisant la création de corpus de grande taille. À ces facilités grandissantes s'opposent de nouvelles difficultés. L'hétérogénéité des contenus mis à disposition par une communauté, les algorithmes de recherche de bases de données, ou simplement les limites d'ambiguïté de requêtes génèrent de nombreux bruits à nos corpus. Nous proposerons ici une méthode s'appuyant sur une identification de contenu par classification successive (Ratinaud et Marchand, 2015), puis une régénération du corpus par concaténation de l'intégralité des articles contenant au moins un segment de texte (ST) dan le matériel identifié comme pertinent.

2. Problématique

La sélection de corpus par classifications successives, en utilisant comme unité le segment de texte, permet d'obtenir un sous corpus pertinent avec une thématique (Loubère, 2014; Ratinaud et Marchand, 2015). Cependant, lorsque le corpus de départ est constitué de textes au contenu narratif structuré et délimité (article de presse, blog, argumentaires dans une concertation…) ce processus peut supprimer les éléments périphériques au thème étudié. Ces contenus restent portant pertinents pour la compréhension de l'objet d'étude, mais peuvent être classés avec le bruit des textes hors sujet dès les premières étapes de sélection. L'objectif de cettte méthode est donc d'exclure le bruit de textes hors-sujets tout en conservant le contexte d'évocation de la thématique principale.

3. Méthodologie

Le processus proposé ici se décompose en trois étapes :

1. Numérotation des textes par un identifiant en méthadonnée
2. Extractions des segments de textes propres à notre thématique par classifications successives. Cette étape repose sur la classification hiérarchique descendante (CHD) de type Reinert (Reinert, 1983) proposée par le logiciel Iramuteq (Ratinaud, 2009). En permettant de faire émerger les mondes lexicaux, ce traitement nous permet de sélectionner les segments concernant notre thématique, puis de les re-soumettre à une CHD afin de préciser le corpus. Cette étape est reconduite jusqu'à avoir une classification dont toutes les classes concernent la thématique étudiée.
3. Re-composition du corpus par concaténation des articles apparaissant au moins une fois dans l'extraction finale de l'étape 2

4. Exemple empirique

Dans les parties qui suivront, nous présenterons une mise en application de cette méthode sur un corpus utilisé lors de notre thèse (Loubère, 2018). Il est constitué d'une extraction d'article de presses quotidiennes nationales (libération, l'humanité, le monde, la croix, le figaro) portant sur la thématique du numérique éducatif du 01/01/2000 au 31/12/2014. Afin de couvrir le plus d'informations possible la requête exécutée sur la base de donnée d'Europresse retournait tous les articles contenant au moins un terme éducatif dans la liste : collège, lycée, école, éducation et au moins un terme numérique dans la liste : numérique, informatique, multimédia, TICE.

4.1. Les classifications successives

Cette extraction retourna 18 804 articles, auxquels nous avons retiré 875 doublons. LE corpus exploité ici est donc constitué de 17 929 articles représentant 450 815 segments de textes, sur lesquels nous avons apposé en méthadonnée le numéro de l'article source. Nous allons présenter ici les classifications successives

Nous avons effectué une CHD de 20 classes en phase 1 et un minimum de 1000 ST par classe, nous obtenons 16 classes représentant 99,72 % du corpus. Le résultat obtenu est présenté sur le dendrogramme en illustration 1

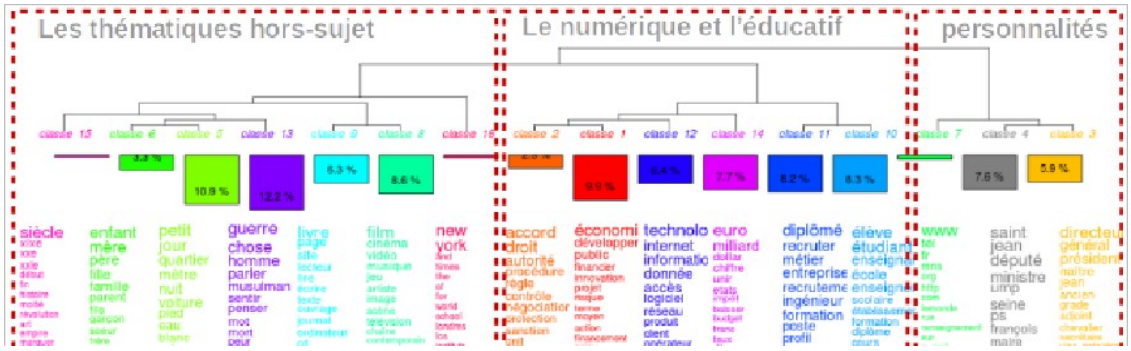

Illustration 1 : dendrogramme de la première CHD

Ce premier découpage montre une séparation en 3 blocs. Le premier est composé des personnalités publiques, le second est composé par des thématiques extérieures à notre sujet. En effet, de nombreux articles contiennent les termes de notre requête sans être pour autant dans le domaine éducatif (ou numérique). Ainsi, les classes 9 et 8 regroupent les actualités ou dossiers portant dans le domaine de la culture. Nous citerons comme exemple non exhaustif d'article de ce domaine un article du journal Le monde commentant les sorties cinématographiques dans lequel nous relèverons « les enfants privés d'école jouant dans les rues », et pour un autre film « les décors numériques ». Nous retrouvons sur le même principe les classes 6, 5 et 13 traitant des conflits armés détruisant les lycées et relatant une infériorité numérique.Enfin, le troisième bloc présente une classe centrée sur le numérique (classe 12), deux classes centrées sur l'éducatif (11 et 10) et deux classes sur l'aspect législatif et économique (classes 1 et 2). Afin de pouvoir affiner ces thématiques et les possibles interactions, nous avons choisi de conserver le bloc entier, soit les segments composant les classes 1, 2, 10, 11, 12 et 14.

L'export précédent nous a permis d'obtenir 194 966 segments de texte sur lesquels nous avons effectué une deuxième CHD de 15 classes en phase 1 et seuil minimal de 100 ST. Nous obtenons 14 classes portant sur 99,97 % des segments. Le résultat est présenté en illustration 2.

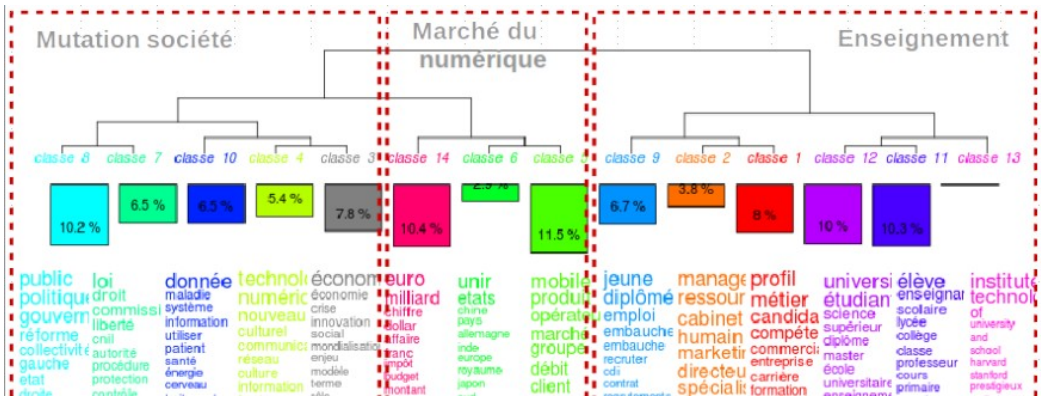

Illustration 2 : dendrogramme de la deuxième CHD

Ce deuxième découpage reprend une structure en trois groupes. Ici, nous relevons le contexte économique du marché du numérique (classe 14, 5 et 6). Le second bloc (classe 4, 3, 7, 8, 10) est constitué des différents discours témoins de la numérisation de la société. Le troisième groupe séparé du reste du corpus par le premier facteur est centré sur-le-champ éducatif. Les trois premières classes à se détacher partagent un discours sur l'après-formation et le recrutement (classes 9, 2 et 1). La classe 11 constituant 10,3 % du corpus est entrée sur l'éducation primaire et secondaire, alors que la classe 12 porte sur l'enseignement supérieur et la recherche. Notre étude portant sur le système scolaire secondaire, nous ne conserverons que la classe 11 pour l'étape suivante.

L'export de cette dernière constitue un corpus de 20 167 segments de texte sur lesquels nous avons effectué une CHD de 15 classes en phase 1 et un minimum de 100 ST par classe. Nous obtenons 8 classes rapportant 99,22 % des segments.
Ce dendrogramme, structuré en deux blocs, nous montre une séparation entre un discours

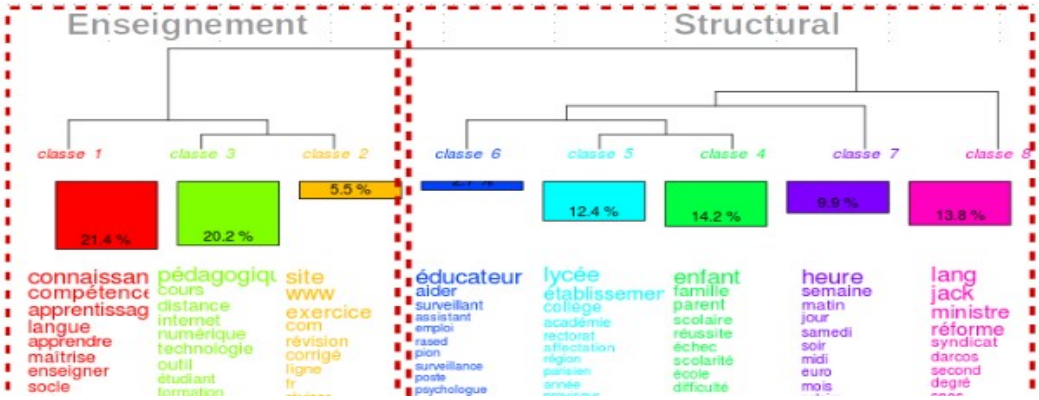

Illustration 3: dendrogramme de la troisième CHD

centré sur l'aspect structurel de l'éducation (classes 8, 6, 4, 3) et celui traitant de l'enseignement (classes 2, 1, 5, 7).

Dans la partie structurale nous retrouverons les segments de texte traitant des réformes sous un angle gouvernemental (classe 8), suivie de tout le discours se regroupant des aspects temporels, comme le temps de travail mais également les rythmes scolaires (classe 6). La classe 3 constitue un discours sociologique sur l'éducation, nous y retrouvons de nombreuses statistiques étudiant les répartitions sociales dans les différents cursus. Enfin, la classe 4 traite des établissements scolaires dans leurs diversités.

Les autres classes portent toutes sur le domaine pédagogique : la classe 7 concerne les contenus d'enseignement. La classe 5 traite de la mise en place d'outil numérique parascolaire (jeux éducatifs, fiche de révision) alors que la classe 2 est centrée sur la mise en place de formations à distance. Enfin, la classe 1 est le discours portant sur le numérique dans l'éducation, les mots clés employés dans notre requête y sont tous surreprésentés. Nous ne conserverons dons que les segments composant cette classe.

L'extraction de cette dernière classe nous permet d'obtenir 2072 segments sur lesquels nous avons effectué une CHD de 20 classes en phase 1 avec un seuil de 100 ST par classe. Cette

classification nous a montré une réelle stabilité de la thématique. En effet, les 8 classes exposée portent chacune sur un aspect du numérique éducatif.

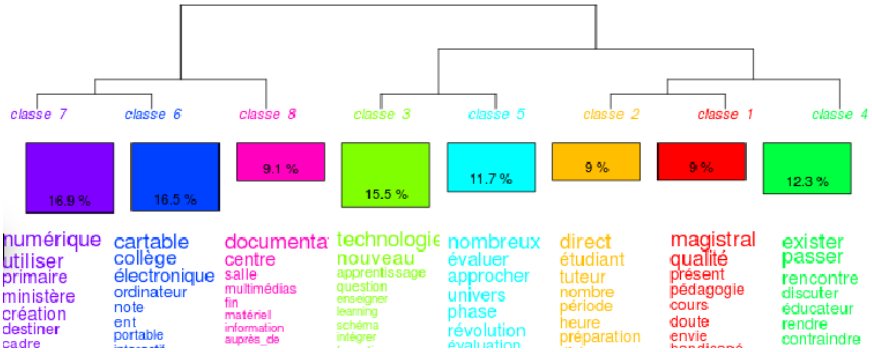

Illustration 4 : dendrogramme de la quatrième CHD

4.2. Classification du corpus recomposé

Le corpus recomposé des 2902 articles contenant au moins un segment de texte dans la classe 1 de la troisième CHD est constitué de 72460 segments. Une CHD de 20 classes en phase 1 et un minimum de 800 ST par classe nous donne le dendrograme suivant :

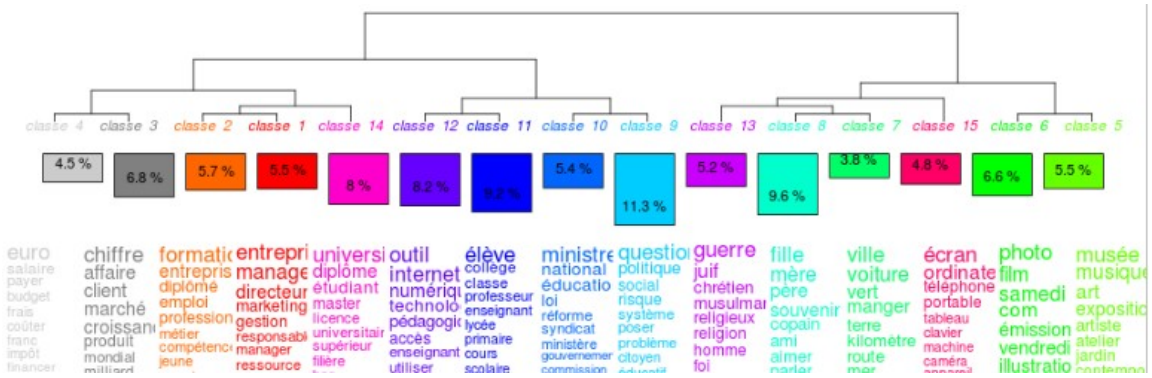

Illustration 5 : dendrogramme de la CHD sur le corpus recomposé

Nous y retrouvons donc au-delà de discours sur l'utilisation du numérique dans les établissements, un discours sur l'économie reflétant le marché du numérique éducatif et les frais engendrés par les dotations des établissements. Un discours à la frontière de la culture et de l'éducation, avec les formations de ces domaines empreinte de numérique. Mais également un discours sur l'actualité géopolitique mondiale contextualisant des initiatives où le numérique apporte des solutions éducatives lors de ségrégation ethniques, ou éloignements géographiques. Tous ces mondes lexicaux constituent des éléments du discours social sur notre sujet, qu'une étude réduite aux segments ciblés lors des CHD successives ne permettrait pas d'explorer.

5. Conclusion

Le principe des CHD successives, s'il nous permet d'accéder finement aux segments contenant le discours sur le numérique éducatif, nous éloigne d'une compréhension globale du sujet. En effet, interroger les bases de données de presse sur une longue période et une sélection de presse généraliste apporte une quantité importante de documents hors contexte.

Ces données portent des éléments contextuels communs avec les articles traitant de notre sujet (personnalités politiques, discours économique…), la proximité lexicale des segments de ces champs structure les classes de discours communes aux articles portant sur notre sujet ou non. Cette hétérogénéité associée à l'insécurité d'un grand ensemble (Geffroy et Lafon, 1982) nous empêchant une connaissance du corpus antérieure à l'analyse lexicométrique conduit « à tracer un peu trop vite une autoroute » (Geffroy et Lafon, 1982, p. 140) jusqu'à notre classe 1 finale. Ce phénomène questionne la constitution d'un corpus sur une dimension architextuelle, alors même que l'outil de classification utilisé ici joue sur un niveau intertextuel et cotextuel (Rastier, 2015), rapprochant des passages de textes en fonction de leur structure lexicale. La présence de textes aux sujets hétéroclites fait ressortir de façon précoce des thématiques indépendamment de leur hypothétique poids dans le corpus qu'aurait constitué une sélection de textes centrés sur notre sujet. Ainsi, les segments traitant de sujets de politique générale ou exposant le contexte social d'un pays dans les articles traitant du numérique éducatif sont classés avec ceux des articles hors sujets. Cette difficulté éloigne le chercheur de la compréhension d'un discours. La démarche que nous venons de présenter nous permet de se rapprocher d'un positionnement de textomètre (Pincemin, 2012), sélectionnant les segments pertinent par une démarche inductive, mais en conservant l'unité sématique du texte dans la construction du corpus final.

6. Bibliography

Geffroy, A., & Lafon, P. (1982). L'insécurité dans les grands ensembles. Aperçu critique sur le vocabulaire français de 1789 à nos jours d'Etienne Brunet. Mots, 5(1), 129-141.

Loubère, L. (2014). Le traitement des TICE dans les discours politiques et dans la presse. In Présenté à 12èmes Journées internationales d'Analyse statistique des Données Textuelles.

Pincemin, B. (2012). Sémantique interprétative et textométrie. Texto! Textes et Cultures, 17(3), 1-21.

Rastier, F. (2015). Arts et sciences du texte. Paris: Presses universitaires de France.

Ratinaud, P. (2009). IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires. Consulté à l'adresse http://www.iramuteq.org

Ratinaud, P., & Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). Mots. Les langages du politique, (2), 57-77.

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.