Lexicométrie pour l'analyse qualitative : Pourquoi et comment résoudre le paradoxe ?

Christophe Lejeune[1] et Aurélien Bénel[2]

[1] Université de Liège – christophe.lejeune@ulg.ac.be

[2] Université de technologie de Troyes – aurelien.benel@utt.fr

Abstract

Using text statistics for qualitative analysis is surely a paradox – counting words in an approach where "numbers doesn't count". Hence, we were stunned by the repeated expectation of text mining features from the users of our qualitative analysis platform. Reluctant at first to methodological hybridizations that could be seen as flawed, we then had to admit that even the founders of qualitative analysis envisioned qualitative ways to use statistics. Counts cannot be regarded as proofs by qualitative analyzers, but as clues for an interpretation, as incentives to look closer at the materials. To achieve this goal, we designed a new visualization of text statistics, displayed *in* the text without any numbers nor tables.

Résumé

Pour la recherche qualitative, les nombres ne comptent pas. Dans une telle démarche, faire appel à la statistique textuelle devrait relever du paradoxe. Nous avons pourtant décidé d'intégrer de tels comptages dans la plateforme d'analyse qualitative que nous concevons. L'enjeu était d'assigner une fonction heuristique (et non plus probatoire) aux comptages, afin de les mettre au service de la perspective initiale, sans la dévoyer. Pour le qualitativiste, les mots rares, les spécificités et les segments répétés sont autant de pistes d'interprétation. Restait à trouver un mode de visualisation, sans tableaux ni chiffres, adapté à une approche visant davantage à l'immersion dans les textes qu'à leur objectivation.

Mots-clés : recherche qualitative, lexicométrie, segments répétés, spécificités, visualisation.

1. Introduction

La plateforme Hypertopic rassemble des outils dédiés à l'analyse qualitative de matériaux textuels et photographiques. Cet article se concentre sur les outils textuels. Ces derniers combinent un mode d'annotation "manuel" et un mode d'annotation semi-automatique, par mots-clés (Bénel *et al.*, 2010). Dans les deux cas, il s'agit d'assister une méthode essentiellement qualitative. *A priori*, nos outils ne s'inscrivent donc pas dans l'analyse statistique de données textuelles. Pourtant, nous y avons intégré certains calculs lexicométriques.

Nous exposons tout d'abord en quoi la combinaison de la recherche qualitative et de la lexicométrie peut sembler paradoxale. Nous montrons comment dépasser ce paradoxe en affichant les informations lexicométriques sans chiffres ni tableaux. Le mode de calcul des mots rares, des spécificités et des segments répétés est ensuite discuté sur base d'exemples issus d'une re-

cherche qualitative sur les représentations sociales du handicap [1]. Compte tenu que les corpus analysés sont partagés en ligne et que leur analyse débute dès le début de leur constitution, leur mise à jour est susceptible d'être continue. Nous montrons donc comment l'implémentation de ces calculs gère finement l'impact de ces mises à jour.

2. Traversée et dépassement du paradoxe

2.1. Recherche qualitative

Toutes les sciences humaines et sociales collectent des matériaux empiriques, qu'elles entendent analyser. Elles y distinguent les matériaux dits quantitatifs des matériaux dit qualitatifs. Ainsi, le matériau collecté au moyen de questionnaires est-il dit quantitatif et le matériau issu d'entretiens ou d'observations est-il considéré qualitatif (Pires, 1987). La différence de matériau ne suffit cependant pas à distinguer les approches quantitative et qualitative (Charlier et Moens, 2006).

Plus que la nature du matériau, ce sont les techniques impliquées dans son traitement qui définissent une approche (Paquay *et al.*, 2006). Ainsi l'approche qualitative repose sur l'explicitation, en profondeur, des significations impliquées dans les formulations du matériau (Olivier de Sardan, 2008). La visée y est avant tout compréhensive, avant d'être explicative, même si, comme l'affirme Weber (1995), une bonne compréhension d'un phénomène permet de l'expliquer. La question des causes, donc de l'identification de variables dont il faut vérifier la corrélation (statistique) passe au second plan. L'approche qualitative privilégie donc un traitement en profondeur d'un corpus empirique de petite taille, à un traitement extensif de large corpus. La question du comptage lui est par essence étrangère.

2.2. Traitements quantitatifs de données textuelles

Un matériau textuel (qualitatif) peut bien entendu faire l'objet d'une analyse quantitative. C'est ce que font notamment l'analyse de contenu, la lexicométrie et le traitement automatique des langues (TAL). En ayant recours à des traitements statistiques, elles embarquent un ensemble cohérent de postulats épistémologiques : (1) le nombre compte ; (2) la représentativité détermine la validité ; (3) la force des corrélations et la significativité des différences se mesurent au moyen de tests standardisés par la statistique (Atifi *et al.*, 2006). Ces règles permettent l'élaboration d'analyses robustes et éprouvées dans un cadre quantitatif.

2.3. Un certain purisme épistémologique

Lorsque nous avons entamé la conception d'une plateforme collaborative d'analyse de textes (Zacklad *et al.*, 2007), nous étions résolus à n'intégrer aucune fonctionnalité reposant sur des comptages. Cette décision provient initialement d'une volonté de concevoir un outil résolument qualitatif, un outil utilisable, donc, dans une perspective selon laquelle le nombre ne compte pas. La conception de nos outils s'inscrit ainsi dans une volonté de rigueur épistémologique, assumant une cohérence voire un certain purisme épistémologique. Ce purisme proscrit donc tout comptage. Dans cette optique, la conjugaison de la statistique textuelle et de l'approche qualitative n'est ni envisageable, ni souhaitable. La simple mention d'une telle conjugaison relève du paradoxe.

1. Les auteurs remercient Sébastien Fontaine, pour sa collaboration dans le cadre de cette recherche, ainsi que Marine Leleu et Coralie Darcis, qui ont conduit les entretiens apparaissant dans les exemples. Merci également à Gaëlle Lortal qui a relu une version antérieure de ce texte.

2.4. Retour des utilisateurs

De leur côté, les utilisateurs de logiciels d'analyse textuelle sont habitués aux mesures statistiques. Il s'agit là de fonctionnalités "attendues". Plusieurs de nos utilisateurs se sont étonnés de leur absence. Certains nous ont réclamé leur intégration. Interpellés par ces requêtes, nous avons d'abord adopté une posture pédagogique. Nous avons déplacé la discussion autour des outils à une question de méthode, ce qui nous apparaît comme le plan le plus pertinent pour les questions de ce type.

Notre attitude fut alors corrective. Nous avons rappelé à nos utilisateurs que la recherche qualitative n'avait pas besoin de chiffres et ce, par conception. Adjoindre des comptages risquerait de la dénaturer, de produire des analyses quasi-qualitatives (Paillé, 2006). De telles analyses existent (Lejeune, 2010). Elles sont même courantes. En recourant aux comptages, elles souscrivent aux postulats quantitatifs (parfois sans même que le chercheur ne s'en rende compte) sans néanmoins se donner les moyens de quantifier de manière rigoureuse ce qui doit l'être. Pourtant, avec l'analyse de contenu, il existe depuis les années cinquante un exemple de définition rigoureuse des quantifications sur lesquelles appuyer une analyse (quantitative) de textes (Berelson, 1952). Les analyses quasi-qualitatives n'appliquent cependant pas les règles de méthode présidant à la réalisation d'une analyse de contenu. Par ailleurs, reposant sur des quantifications (mêmes floues) comme "tous les", "la plupart de" ou "en général", les analyses quasi-qualitatives quittent le domaine strict de la recherche qualitative. Notre ambition n'était cependant ni de concevoir un outil d'analyse de contenu (il en existe déjà par ailleurs) ni d'offrir une fonctionnalité risquant d'induire en erreur nos usagers.

2.5. Conversion, trahison et continuité

Les requêtes des utilisateurs furent d'abord temporisées. Parallèlement, cependant, les discussions avec les utilisateurs (qui sont, comme nous, des chercheurs) initièrent une réflexion sur la façon dont un comptage peut effectivement être mobilisé dans une recherche qualitative (Bénel, 2006). Et sur une façon de visualiser de tels comptages qui ne souscrirait pas à la sémantique visuelle du tableau de chiffres (inscription essentiellement quantitative).

C'est, en définitive, l'usage et la pratique qui ont tranché. Notre plateforme sert en fait à partager des textes, à les annoter et à partager ces annotations. Plus précisément, nos outils proposent deux modes d'annotation des textes analysés (Bénel *et al.*, 2010). La première repose sur une annotation libre : l'usager sélectionne des passages textuels à la souris et leur associe une étiquette. La deuxième est semi-automatique : l'usager repère, dans le texte, des mots-clés ou des expressions-clés et leur assigne une fonction de marqueur (Lejeune, 2008). Dès lors, tous les passages qui comprennent le mot ou l'expression considérée sont automatiquement associés à l'étiquette choisie par le chercheur. Or, dans ce deuxième cas de figure, certains utilisateurs considèrent que certains mots rares, fréquents ou certaines expressions (composées de plusieurs mots) sont des marqueurs particuliers à certains domaines qu'il est pertinent de relever. Au fond, dans le vocabulaire de la lexicométrie, les mots rares, les spécificités et les segments répétés se révèlent de bons candidats marqueurs. Tous les hapax, toutes les spécificités et tous les segments répétés ne deviendront évidemment pas des marqueurs. Mais leur identification peut, en première analyse, suggérer quelques pistes heuristiques à l'analyste.

Envisager les mesures lexicométriques selon leur apport heuristique n'est pas nouveau. C'est précisément l'inscription épistémologique que Max Reinert (1983) donne à ce type d'outils. Une telle conception ne confère pas un rôle probatoire à la mesure quantitative. La fréquence n'est pas le résultat. Elle offre une prise à l'analyste, qui s'en empare (ou non) comme d'un marqueur possible.

3. Le texte dans tous ses états

La lexicométrie s'est fait une spécialité de trier les unités lexicales en fonction de leur fréquence. Les résultats de ces tris sont typiquement affichés dans des tableaux d'occurrences. Lus à partir du haut, ces tableaux permettent de prendre connaissance des formes les plus fréquentes d'un corpus. L'analyse, par Sébastien Fontaine, d'entretiens sur les représentations sociales du handicap permet d'illustrer ce que deviennent ces comptages dans une recherche qualitative. Les formes qui apparaissent le plus fréquemment dans ces entretiens sont “il", “est", “de", “que" et “je". On y reconnaît les pronoms typiques de la langue orale ainsi que les mots pivots qui, figurant dans son guide d'entretien, ont été mentionnés par le chercheur et sont repris, de manière récurrente, par les informateurs. Du point de vue de Sébastien Fontaine, ces informations se révèlent relativement triviales. Ces formes ne constituent pas, pour lui, l'information la plus heuristique.

Par ailleurs, la présentation en tableau ne convient pas à une pratique ne jurant que par le travail sur des sources textuelles originales. Épistémologiquement, le tableau n'est pas un intermédiaire valide pour les qualitativistes. Nous avons résolu cet apparent paradoxe en concevant un mode d'affichage de mesures lexicométriques dans le texte lui-même, en jouant sur des contrastes de gris. Un tel mode de visualisation reste compatible avec les options épistémologiques des qualitativistes ; le contact à la source reste constant, tout en lui adjoignant une information produite de manière quantitative, proposée à titre de suggestion ou de piste pour le parcours interprétatif. C'est donc en nuances de gris que s'affichent les mots rares, les spécificités et les segments répétés.

3.1. Raretés

Les mots qui n'apparaissent qu'une fois (appelés *hapax*) sont susceptibles de délivrer des pistes heuristiques plus fécondes que les fréquences brutes. Cet intérêt pour les formes rares correspond d'ailleurs à la lecture des tableaux de fréquences lexicométriques à partir du bas.

Soit m un mot, C un corpus et $\underset{m,C}{\Omega}$ l'ensemble des occurrences de m dans C, on pourrait définir la rareté de la manière suivante :

$$(\forall m)(\forall C)\ \ rarete(m,C) = \frac{1}{card(\underset{m,C}{\Omega})} \tag{1}$$

Cette mesure de la rareté permet d'identifier les mots les moins fréquents. Quand elle est utilisée pour visualiser un texte, cette définition, aussi “naïve" qu'elle soit, comporte déjà un intérêt heuristique (cf. figure 1). Dans la recherche sus-mentionnée, des entretiens ont été conduits auprès de différents informateurs vivant au quotidien avec une personne handicapée. Les deux premiers entretiens réalisés pour cette recherche portent sur l'expérience de deux informatrices

différentes. Le mari de la première est sourd ; le frère de la deuxième est atteint du syndrome X fragile (un handicap mental comparable à l'autisme). La figure 1 montre les résultats du calcul de la rareté pour le deuxième entretien. Vu la taille restreinte des corpus au début de la recherche, les mots rares (en noir) n'apparaissent qu'une seule fois (ce sont des *hapax*). À la lecture du passage considéré, l'analyste comprend aisément que la rareté de qualification comme "bête" ou "limité" traduit la préoccupation de l'informatrice d'éviter la stigmatisation de la maladie mentale. L'évitement (partiel) de ce type de qualification traduit donc un enjeu capital pour l'informatrice, qui recouvre également une problématisation classique pour la sociologie du handicap (Goffman, 1975).

Figure 1 – Affichage des mots rares en nuances de gris (copie d'écran de Cassandre*)*

Dès ce premier exemple, on peut noter un certain nombre de choix de conception : les mesures globales sont faites sur un corpus et non sur l'ensemble de la base, quant aux "poids" des mots, ils sont normalisés par rapport à leur maximum dans le document.

Les raisons de ces choix sont multiples : elles sont informatiques (réduire les temps de calcul et de téléchargement), ergonomiques (obtenir une palette de niveaux de gris allant jusqu'au noir), linguistiques (ne pas faire l'hypothèse de l'existence de la "langue") et méthodologiques (faciliter la reproductibilité sur un corpus).

Les résultats de ce calcul de la rareté sont instructifs. Cette mesure est cependant adaptée pour des corpus de taille réduite. Or, en recherche qualitative, la constitution et l'analyse de ses corpus se réalisent parallèlement. Autrement dit, le calcul de la rareté est particulièrement adapté à des corpus "naissants". Lorsque le corpus s'étoffe, le rapport des fréquences des mots du texte à l'ensemble du corpus bascule. La mesure des spécificités (présentée dans la section suivante) offre alors des pistes plus pertinentes.

3.2. Spécificités

La fréquence brute dépend largement de la taille des corpus et des textes analysés. La lexicométrie a donc défini d'autres mesures permettant de dépasser les simples fréquences. Les *spécificités* permettent ainsi d'identifier si une unité lexicale est sur-représentée (ou sous-représentée) dans une portion du corpus par rapport au reste du corpus considéré.

En recherche d'information, par exemple, le $tf.idf$ (*term frequency, inverse document frequency*) est utilisé d'abord pour extraire d'un document les mots qui le caractérisent le plus par rapport au corpus, puis, en réponse à une requête, pour trier les documents en donnant plus

de poids aux mots de la requête les plus discriminants (Spärck Jones, 1972) et plus de poids aux documents utilisant ces mots fréquemment.

Pour notre visualisation des spécificités, nous nous sommes inspirés du $tf.idf$ mais en ne gardant que son principe et en simplifiant grandement sa formulation. En effet, notre visualisation étant celle d'un document et notre mesure étant déjà normalisée, nous avons supprimé tout ce qui était constant pour un document donné et tout ce qui participait à la normalisation du résultat. Ensuite, il ne nous restait donc plus qu'à diminuer l'impact du nombre d'occurrences.

Soit m un mot, d un document, C un corpus, $\underset{m,d}{\mathscr{O}}$ l'ensemble des occurrences de m dans d et $\underset{m,C}{\mathscr{D}}$ l'ensemble des documents de C contenant m, nous définissons la spécificité de la manière suivante :

$$(\forall m)(\forall d \in C)\ \ specificite(m,d) = \frac{\sqrt{card(\underset{m,d}{\mathscr{O}})}}{card(\underset{m,C}{\mathscr{D}})} \tag{2}$$

Figure 2 – Affichage des spécificités en nuances de gris (copie d'écran de Cassandre*)*

Pour cet exemple, la fenêtre s'ouvre sur l'entretien mené avec une personne dont l'époux est sourd. Lorsqu'il actionne le bouton des spécificités, l'analyste ne s'émeut pas outre mesure que s'affichent en gris foncé les mots "sourd", "surdité", "malentendant". Ces mots sont en effet spécifiques par rapport aux autres entretiens, qui ont été menés auprès de personnes vivant par exemple avec un malvoyant ou un autiste. Pour l'analyste, les spécificités relatives au thème de la surdité sont donc triviales. Les spécificités affichées sur la figure 2 l'intéressent bien plus : il y est question de "casquette" et de "vêtement à capuche". Ce type de vêtements correspond à l'uniforme porté par les ouvriers sur chantier pour se protéger des intempéries (le mari de l'informatrice est maçon). La question des couvre-chefs est une spécificité particulièrement pertinente pour cette recherche. Les capuches vont en effet subir un découpage autorisant un contact visuel constant avec les collègues. Cette opération constitue un aménagement de l'équipement standardisé et permet à la personne handicapée d'évoluer sans heurt dans le cadre de son travail.

La formule du calcul des spécificités circonscrit sa pertinence. En effet, la division par le nombre de textes perd de sa force lorsque le dénominateur est un petit nombre. Il en résulte que le calcul des spécificités est particulièrement adapté aux corpus d'une certaine taille. Pour analyser des corpus à géométrie variable, la mesure des mots rares et celle spécificités sont donc complémentaires. La deuxième peut prendre le relais de la première lorsque le corpus grandit.

3.3. Séquences

Pondérée par les spécificités, la fréquence peut fournir une information utile à l'analyste d'un corpus. Il arrive cependant que la récurrence porte moins sur une unité lexicale isolée que sur une expression ou un groupe de formes. La répétition d'une suite d'unités lexicales renvoie au repérage de ce que la littérature appelle, selon les domaines, les collocations, les polyformes, les segments répétés ou les n-grammes. Chacun de ces concepts renvoie à une tradition différente. Et, même au sein des différents domaines, chaque appellation renvoie à différentes définitions (Colson, 2010, 397).

Les phénomènes considérés renvoient à ce que la linguistique appelle *collocation*. Le rapprochement ne fonctionne toutefois que si l'on donne à ce concept une définition relativement lâche. Dans la mesure où elles se réfèrent au sens de la forme composée, les définitions arrêtées par les chercheurs en sémantique et en lexicographie se révèlent trop restrictives pour les repérages automatiques que nous envisageons ici.

La détection de telles séquences de données textuelles peut être opérée au moyen de techniques soit reposant sur des dictionnaires, soit n'impliquant aucune information linguistique (Khokhlova, 2008). Dans ce dernier cas, la statistique textuelle désigne sous l'expression "segments répétés" les suites de formes qui apparaissent de manière répétée dans un corpus de textes (Lebart et Salem, 1994). On parle par exemple de "bigramme" quand il s'agit de la succession répétée de deux unités. Cette notion de "bigramme" est relativement générique puisqu'elle peut aussi renvoyer au traitement de matériaux non textuels comme, par exemple, le repérage de récurrence dans les séquences d'ADN.

La plupart des techniques de détection de séquences répétées se concentrent sur les segments composés de deux unités lexicales (Heid, 2007, 1042) (cité par Colson, 2010). Le phénomène est d'ailleurs parfois décrit comme une cooccurrence linéaire (Halliday, 1966) (cité par Colson, 2010), entendue comme l'occurrence simultanée de deux mots qui se suivent. La généralisation de la collocation aux *n*-grammes (quel que soit *n*) est plus complexe qu'il n'y paraît. En effet, se posent le problème de la taille maximale des séquences à repérer ainsi que celui du repérage ou non des séquences incluses dans d'autres.

Dans notre cas, le fait d'afficher les résultats *sur* le texte d'origine, déplace le problème : les segments répétés seront en quelque sorte "empilés" les uns sur les autres. Dès lors, il ne s'agit plus de détecter des segments répétés de longueur indéterminée, mais des zones de taille indéterminée contenant des segments répétés de taille fixe (en nombre de mots). Le poids de chaque occurrence est ainsi déterminé en faisant glisser une fenêtre de longueur déterminée. Étant donné le grand nombre de mots d'un document habituellement contenus dans les bigrammes répétés d'un corpus, notre choix s'est porté sur les trigrammes.

Soit i l'indice d'un mot dans un document, d ce document, C son corpus et $\underset{[a,b,c],C}{\Omega}$ l'ensemble des occurrences de la séquence de mots abc dans C, nous définissons le degré de collocation de la manière suivante :

$$(\forall i)(\forall d \in C)\ collocation(i,d) = \max\left(\left\{\begin{array}{l} card\left(\underset{[m_{i-2},m_{i-1},m_i],C}{\Omega}\right) \\ card\left(\underset{[m_{i-1},m_i,m_{i+1}],C}{\Omega}\right) \\ card\left(\underset{[m_i,m_{i+1},m_{i+2}],C}{\Omega}\right) \end{array}\right\}\right) \quad (3)$$

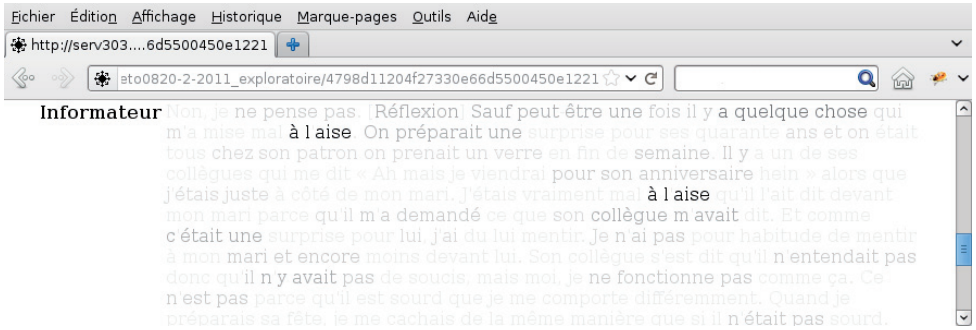

Figure 3 – Affichage des segments répétés en nuances de gris (copie d'écran de Cassandre*)*

La figure 3 est issue du même entretien que la figure 2. L'expression "à l'aise" apparaît saillante. D'un coup d'œil, l'analyste identifie que ce segment répété s'inscrit dans des expressions plus larges comme "mal à l'aise" ou "pas à l'aise" qui renvoient au vécu de la personne dans les situations où elle est confrontée à des asymétries de communication. C'est le cas, par exemple, lorsque ses enfants ou un collègue s'adresse oralement à elle en présence de son mari. De telles situations introduisent une asymétrie : le mari, sourd, ne perçoit qu'une partie de l'interaction. En outre, les interlocuteurs de l'épouse produisent cette asymétrie en connaissance de cause (ils savent que le mari est sourd). Cette configuration diffère donc radicalement de la maladresse qu'un inconnu rencontré en rue pourrait commettre, ne sachant pas le mari sourd. C'est donc moins la maladresse que l'asymétrie produite délibérément qui dérange l'épouse. Cet élément se révèle central pour la recherche considérée qui porte précisément, comme la plupart des recherches qualitatives, sur l'explicitation du vécu des acteurs.

4. Calcul au fil de l'eau

En recherche qualitative, la collecte et l'analyse ne sont pas envisagées comme des phases qui se succèdent (Glaser et Strauss, 2010, 136, 172). Au contraire, l'analyse est intégrée à la collecte ; elle l'accompagne et la guide (Glaser et Strauss, 2010, 138). Les deux activités ne sont donc pas menées l'une après l'autre (en séquence) mais sont concomitantes (en parallèle). En conséquence, le corpus de textes n'est pas fermé, mais évolue tout au long de son analyse. C'est pour cette raison que nous conservons la mesure de la rareté (utile en début de recherche) et de la spécificité (adaptée aux corpus "fournis"). Peu de systèmes sont cependant conçus pour gérer des corpus en cours de maturation. La plupart du temps, tous les calculs doivent être reconduits à chaque modification.

Pour implémenter de manière efficace des calculs lexicométriques, il est nécessaire de gérer finement l'impact des mises à jour du corpus : lors de l'ajout, du retrait ou de la modification d'un texte, il s'agit de conserver les résultats intermédiaires non impactés et de recalculer le reste.

Cette tâche est aujourd'hui grandement facilitée par les cadriciels (*frameworks*) inspirés du modèle *MapReduce* de Google (Dean et Ghemawat, 2004). En particulier, dans la base de données CouchDB, pour une *vue* donnée, chaque objet semi-structuré (ici un entretien découpé en tours de parole) est traité par une fonction *map* pour donner un ensemble de couples $(clef, valeur)$. Ces couples sont ensuite triés pour l'ensemble de la base. Enfin, une fonction *reduce* (qui doit être associative et commutative) permet d'agréger les valeurs d'une même clef. Un tel cadre de développement permet ainsi de garantir que la modification d'un objet n'aura d'impact que sur le résultat de la fonction *map* appliquée à cet objet et sur ceux de la fonction *reduce* qui en dépendent.

La première étape de notre implémentation a donc été de déterminer dans nos fonctions lexicométriques les calculs qui pourraient correspondre à ces contraintes. Il est apparu que le nombre d'occurrences d'un mot dans un document ($card(\underset{m,d}{\mathscr{O}})$) était calculable par une fonction *map* (cf. figure 4) et que, de plus, les réductions de ce résultat par somme et par comptage pouvaient être utilisées à d'autres endroits dans nos calculs :

$$\begin{aligned}(\forall m)(\forall d \in C)\ \ specificite(m,d) &= \frac{\sqrt{card(\underset{m,d}{\mathscr{O}})}}{card(\underset{m,C}{\mathscr{D}})} \\ &= \frac{\sqrt{card(\underset{m,d}{\mathscr{O}})}}{card(\{n|(\exists i \in C)\ n = card(\underset{m,i}{\mathscr{O}})\})}\end{aligned} \tag{4}$$

$$\begin{aligned}(\forall m)(\forall C)\ \ rarete(m,C) &= \frac{1}{card(\underset{m,C}{\Omega})} \\ &= \frac{1}{\sum_{d \in C} card(\underset{m,d}{\mathscr{O}})}\end{aligned} \tag{5}$$

La fonction *map* associe, pour chaque texte, chaque mot à une fréquence. La fonction *reduce* agrège, quant à elle, ces résultats intermédiaires et détermine le nombre de textes et la fréquence de chaque mot, pour l'ensemble du corpus. Ainsi, dans l'exemple [2] de la figure 4, à la sortie des fonctions *map* et *reduce*, les deux premiers textes du corpus ont produit 12 occurrences du mot “web". L'ajout d'un troisième document nécessite des calculs complémentaires pour que, “web" soit compté 31 fois. Cependant, ces calculs complémentaires (nœuds colorés dans la figure) peuvent s'appuyer sur les calculs précédents. Ainsi les 19 nouvelles occurrences viennent

2. Le corpus utilisé dans cet exemple provient du site d'actualités technologiques InternetActu.

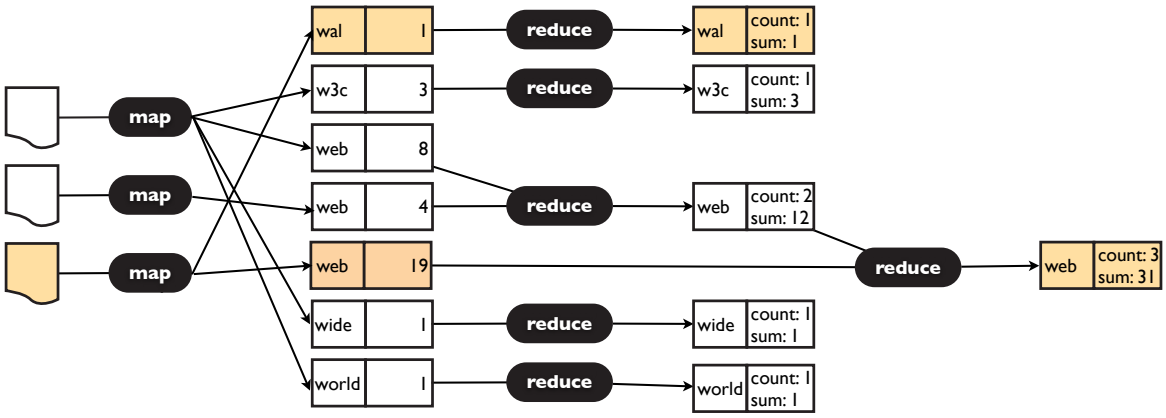

Figure 4 – MapReduce *appliqué à la lexicométrie : Impact de l'ajout d'un document*

s'ajouter aux 12 précédentes. Ensuite, à partir de ces résultats sont calculés les mesures de la rareté et de la spécificité. De la même manière, les segments répétés s'appuient sur le calcul par *MapReduce* des comptages de trigrammes. Cet exemple illustre l'économie calculatoire opérée, qui devient vitale lorsque les utilisateurs d'une même plateforme ne clôturent pas leurs corpus une fois pour toutes mais les construisent au fur et à mesure.

5. Conclusion

La conception de la plateforme Hypertopic d'analyse collaborative de données qualitatives nous a amenés à conduire deux innovations : la première concerne l'affichage des résultats des calculs lexicométriques ; la deuxième porte sur l'optimisation de ces calculs pour des corpus en constante évolution.

Initialement envisagée comme une entreprise exclusivement qualitative, notre plateforme excluait d'abord tout comptage. Pourtant, certains calculs peuvent servir de pistes pour l'annotation. Nous avons surmonté le paradoxe en présentant le résultat de calculs lexicométriques en nuances de gris, directement dans le texte. Un tel affichage est compatible avec l'exigence – déterminante en recherche qualitative – de "coller aux textes".

En outre, la volonté de soutenir des recherches qualitatives "en train de se faire" nous a amenés à tenir compte de corpus nécessairement évolutifs. Adossé au modèle *MapReduce*, notre système ne recalcule que ce qui est nécessaire. Les résultats intermédiaires non impactés ne sont, eux, pas modifiés. En conséquence, la plateforme n'est jamais surchargée, même si de nombreux chercheurs en sciences humaines et sociales se la partagent.

Références

Atifi, H., Lejeune, C., Ninova, G. et Zacklad, M. (2006). Méthodologie transdisciplinaire de gestion du corpus pour les disciplines de l'interaction : recherche de principes directeurs. *Actes du colloque international "Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation"*, XI(2):185–190.

Berelson, B. (1952). *Content Analysis in Communication Research*. The Free Press, Glencoe.

Bénel, A. (2006). Porphyry au pays des paestans : usages d'un outil d'analyse qualitative de documents par des étudiantes de maîtrise en iconographie grecque. *Actes du colloque international "Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation", Albi, juillet 2006*, XI(2):191–197.

Bénel, A., Lejeune, C. et Zhou, C. (2010). Éloge de l'hétérogénéité des structures d'analyse de texte. *Document numérique*, 13(2):41–56.

Charlier, J.-É. et Moens, F. (2006). *Observer, décrire, interpréter*. Institut National de Recherche Pédagogique, Bruxelles.

Colson, J.-P. (2010). Automatic extraction of collocations : a new web-based method. *In* Bolasco, S., Chiari, I. et Giuliano, L., éditeurs : *Statistical Analysis of Textual Data. Proceedings of* 10^{th} *International Conference Journées internationales d'Analyse statistique des Données Textuelles*, pages 397–408, Milan. Edizioni Universitarie di Lettere Economia Diritto.

Dean, J. et Ghemawat, S. (2004). MapReduce : Simplified data processing on large clusters. *In Proceedings of the Sixth Symposium on Operating System Design and Implementation*.

Glaser, B. G. et Strauss, A. L. (2010). *La découverte de la théorie ancrée. Stratégies pour la recherche qualitative*. Armand Colin, Paris.

Goffman, E. (1975). *Stigmates. Les usages sociaux des handicaps*. Minuit, Paris.

Halliday, M. A. K. (2002 (1966)). Lexis as a linguistic level. *In On Grammar*, pages 158–172. Longman, London.

Heid, U. (2007). Computational linguistic aspects of phraseology. *In* Burger, H., Dobrovol'skij, D., Kühn, P. et Norrick, N. R., éditeurs : *Phraseology. An international handbook*, pages 1036–1044. Mouton de Gruyter, Berlin.

Heiden, S. et Pincemin, B., éditeurs (2008). *Actes des* 9^{es} *Journées internationales d'Analyse statistique des Données Textuelles*, Lyon. Presses universitaires de Lyon.

Khokhlova, M. (2008). Extracting collocations in russian : Statistics vs. dictionary. *In* Heiden et Pincemin (2008), pages 613–624.

Lebart, L. et Salem, A. (1994). *Statistique textuelle*. Dunod, Paris.

Lejeune, C. (2008). Au fil de l'interprétation. L'apport des registres aux logiciels d'analyse qualitative. *Revue Suisse de Sociologie*, 34(3):593–603.

Lejeune, C. (2010). Cassandre, un outil pour construire, confronter et expliciter les interprétations. *In* Beauvais, M. et Clénet, J., éditeurs : *Actes du 2ème colloque international francophone sur les méthodes qualitatives*.

Olivier de Sardan, J.-P. (2008). *La rigueur du qualitatif. Les contraintes empiriques de l'interprétation socio-anthropologique*. Academia Bruylant, Louvain-la-Neuve.

Paillé, P. (2006). Lumières et flammes autour de ma petite histoire de la recherche qualitative. *Recherches Qualitatives*, 26(1):139–153.

Paquay, L., Crahay, M. et Ketele, J.-M. D. (2006). *L'analyse qualitative en éducation : Des pratiques de recherche aux critères de qualité, Hommage à Michael Huberman*. De Boeck, Bruxelles.

Pires, A. (1987). Deux thèses erronées sur les lettres et les chiffres. *Cahiers de recherche sociologique*, 5(2):85–105.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. *Cahiers de l'Analyse des Données*, 3:187–198.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Weber, M. (1995). *Économie et société. Les catégories de la sociologie*. Presses Pocket [Plon, 1971], Paris.

Zacklad, M., Cahier, J.-P., Zaher, L., Bénel, A., Lejeune, C. et Zhou, C. (2007). Hypertopic : une métasémiotique et un protocole pour le web socio-sémantique. *In* Trichet, F., éditeur : *Actes des 18e Journées Francophones d'Ingénierie des Connaissances*, pages 217–228, Grenoble. Cepadues.