L'analyse statistique des données langagières pour une approche de l'activité cognitive des collégiens en situation de recherche d'information

Leila El Allouche

Université de Nice Sophia Antipolis – France

Abstract

Using an information search engine on the Internet is now a regular activity, recommended by learning institutions. Whether it is centered on the technique or on the human, the analysis of the research activity is booming, which is making traces of activities a privileged and coveted object of observation. However if the control of the research activity is now the foundation of the recommendation industry, it also causes the failure of educational initiatives. Indeed, middle school students use search robots with little discernment, and they ignore their indexing and referencing techniques. Our empirical observation of activity traces on a body of middle school students shows a succession of tasks that are repeated in a successive and predictable manner. These behavioral canons constitute an obstacle to the rational strategies, but also question the symbolic activity they deploy. Thus, our work focuses on the difficulty of interpreting the search engines behavioral rules developed by middle school students based only on the visible and partial data of the activity. We are showing that the statistical analysis of the language production can allow to reintroduce the symbolic activity of middle school students.

Résumé

Utiliser un moteur de recherche d'information sur Internet est désormais une activité régulière, recommandée par l'école. Qu'elle soit centrée sur la technique ou sur l'humain, l'analyse de l'activité de recherche est en plein essor, faisant des traces d'activité un objet d'observation privilégié et convoité. Mais si le contrôle de l'activité de recherche constitue aujourd'hui les bases de l'industrie de la recommandation, il cause aussi la ruine des entreprises éducatives. En effet, les collégiens utilisent avec peu de discernement les robots de recherche dont ils ignorent les techniques d'indexation et de référencement. Notre observation empirique des traces d'activité d'un corpus de collégiens montre un enchainement de tâches qui se répètent de manière successive et prévisible. Ces grammaires d'usage constituent un obstacle aux stratégies rationnelles de l'activité mais interrogent aussi l'activité symbolique qu'elles déploient. Ainsi notre travail porte sur la difficulté d'interprétation des grammaires d'usage des moteurs de recherche élaborées par les collégiens à partir des seules données visibles et partielles de l'activité. Nous montrons que l'analyse statistique de la production langagière peut permettre de réintroduire l'activité symbolique des collégiens.

Key words :
Recherche Internet, Activité située, Activité cognitive, Interaction homme-machine, Données langagières, Pratiques juveniles, Culture numérique.

1. Introduction

L'intégration des moteurs de recherche dans l'institution scolaire, en recommandant l'usage des technologies cognitives, répondait aux promesses des nouvelles pédagogies et aux attentes des effets bénéfiques pour l'apprentissage. L'interdisciplinarité et la pédagogie documentaire ont ainsi porté la recherche documentaire puis la recherche sur le web au rang des activités éducatives, elles ont fait de la recherche d'information une activité cognitive et du moteur de

recherche un OPC[1] (Depover et al, 2007). Si l'usage des robots de recherche est désormais bien implanté, force est de constater que l'activité de recherche des collégiens échappe pourtant aux attentes du corps enseignant (Ravenstein et al, 2007). En pratiquant la recherche sur le web, les adolescents développent une activité cognitive mais qui reste encore mal délimitée. Un vide pour l'institution qui s'interroge alors sur ses modalités d'avenir, en quête de nouveaux formats de formation, voire de nouvelles disciplines d'enseignement. La loi d'orientation et de programmation pour la refondation de l'école de la République du 8 juillet 2013, tente d'y apporter quelques réponses[2].

Si l'activité cognitive des adolescents en situation de recherche d'information demeure floue, les entreprises de contrôle de l'activité sont pourtant juteuses. Devenue une préoccupation majeure, la maîtrise des comportements n'intéresse pas seulement les communautés éducatives, aucun domaine n'y échappe, il est vrai que l'enjeu est de taille, notamment depuis l'apparition de l'industrie de la recommandation. D'où l'intérêt partagé pour la récupération des traces. La question de la trace et de la traçabilité est en effet au cœur même du fondement du développement aussi bien du web, des outils que des usages (Ertzscheid et al., 2013). Pourtant l'activité ne se réduit pas à ses traces perceptibles sur le web.

Le projet initial de Vannevar Bush d'instituer une machine capable de mémoriser les liens de consultation des documents, le memex, plaçait la machine au service de l'activité humaine. Cette idée se dissout aujourd'hui derrière la toute puissance affichée des géants du web. Pourtant si les liens hypertextes ont transformé l'activité cognitive de l'internaute, « l'apprenaute » rappelle le rôle de l'homme dans l'activité (Simonian, 2014). L'usage d'un objet technique incorpore en effet à l'objet initialement prévu de nouvelles caractéristiques créées par l'usage (Simonnot, 2009) : les forums et autres technologies du web 2.0 en donnent une bonne illustration. Autour de l'existence et de l'importance des interactions homme-machine attestée, l'approche fonctionnaliste impose l'objet technique comme déterminant quand l'approche phénoménologique lui oppose la situation, les contraintes du « déjà là ». Au delà de l'opposition, il reste la question de penser la signification de « l'être-au-monde » (Leroi_Gourhan, 2013) façonnée par les outils techniques.

Comment apprendre avec les technologies s'interroge Linard soulignant l'importance des interactions avec la machine. Il y a lieu alors de se poser, d'une part la question du rôle des interactions avec la machine et d'autre part de questionner l'intentionnalité du sujet? Dans cette communication, nous souhaitons aborder la dimension symbolique de l'usage des moteurs de recherche par les adolescents, l'activité cognitive. La difficulté d'atteindre l'expérience vécue des collégiens seulement par l'analyse statistique des traces d'activité nous amènera à traiter de l'interprétation des manières de faire par un second corpus, l'analyse statistique des données langagières produits par les récits d'explicitation de ses propres traces.

2. Les traces d'activité : compréhension de l'activité

Enregistrées par les dispositifs techniques, si les traces d'activité constituent le fonds de commerce des moteurs de recherche (Cardon, 2013), elles font aussi la fortune des sciences humaines en pointant de nouvelles possibilités d'observation. Promesse d'une possible description objective de l'activité pour la construction de modèles. En résumé, les traces de l'activité permettent de mieux comprendre l'activité mais achoppent sur l'analyse des intentionnalités. Rappelons en effet, que la trace n'est qu'une réalité partielle, elle doit être interprétée en fonction du contexte pour restituer l'activité (Jeanneret, 2011). En effet, si les traces de l'activité constituent bien un matériel de description, étape première, elles ne sont

[1] OPC : Outils à potentiel cognitif (Depover et al, 2007)
[2] EMI et l'EDMI : Education aux médias et à l'information (LOI n° 2013-595 du 8 juillet 2013)

pas sans poser de difficultés à l'interprétation, les nouvelles façons de faire nous laissent bien souvent désarmés. Comment saisir les rationalités mises en œuvre? A partir de quelles données l'activité cognitive peut-elle devenir intelligible?

Nous proposons l'apport d'une enquête menée auprès d'un public de collégiens, réalisée dans le cadre d'une thèse effectuée en 2012-2013, soutenue en 2015, pour laquelle nous avions choisi une approche centrée sur la théorie de l'activité. Nous avons procédé par étapes successives et cumulatives, en quête d'abord de repérage des rationalités mises en œuvres, des formes d'interaction dégagées de l'analyse statistiques des traces, suivie d'une analyse discursive des récits d'explicitation pour interpréter ces rationalités. Pour se faire nous avons choisi un parti pris épistémologique, celui de nous dégager des usages de la langue en optant pour l'analyse des formes lexicales produites par les récits d'explicitation. En faisant l'hypothèse d'un emploi normatif de la langue, l'analyse statistique des formes lexicales mesure alors les écarts à la norme de l'emploi d'une forme singulière plus qu'une autre et nous offre ainsi de nouvelles possibilités interprétatives.

2.1. Les transactions ou les moyens pour atteindre le but de l'activité

En rupture avec le modèle de l'activité mentaliste, la psychologie soviétique décrite par Léontiev (1976) décline l'idée d'intention du sujet au profit des moyens mis en œuvre pour parvenir au but ou motif aboutissant à une modélisation sous forme de structure hiérarchique, portée par des relations dynamiques entre le sujet et l'objet (sujet, motif et comportement). L'ergonomie du travail s'est largement inspirée de ces travaux, apportant à la structure les concepts de tâche et de détournement (Clot 2002). L'intérêt de la tâche dans la description du processus d'activité est d'introduire un ensemble de conditions pour réaliser le but de l'activité, les moyens que le sujet se donne. Ce n'est plus seulement l'intentionnalité qui fait alors le jeu de l'action, le contexte est aussi de la partie. Dans l'environnement homme-machine, si l'intention est indissociable du contexte de production de l'activité, la machine et ce qu'elle renvoie, contribuent aussi à donner du sens à l'activité. Dans le cas de l'environnement des automates de recherche, le contexte connaît une succession de changements, il est redéfini sans cesse, il se caractérise par la notion de labilité.

La théorie de l'activité permet ainsi de penser que la recherche d'information ne se réduit pas à l'expression du besoin d'information mais qu'elle se pose aussi comme une stratégie, un moyen en vue d'exécuter une tâche. Si la théorie de l'activité énonce comme principe d'action, le but, l'intentionnalité du sujet, l'action située introduit dans l'analyse de l'activité les principes d'organisation de l'action pour que le sujet atteigne son but. Elle examine quels sont les principes d'organisation que poursuit le sujet, ceux qu'il invente ou ceux qui sont déterminés par un plan d'action (Suchman, 1987). La théorie de l'activité permet ainsi de passer d'une analyse cognitive basée sur la satisfaction des besoins d'information au jeu des interactions avec la machine, aux formes matérialisées par l'interaction.

2.2. Le corpus

Notre corpus d'analyse se compose de 466 copies d'écran de traces d'activité conservées, soit 20 % environ des traces collectées d'une population de 50 collégiens de classes de troisième d'un même établissement scolaire et des données langagières produites par 7 heures 50 d'entretien d'explicitation recueillies auprès des collégiens. Les traces de transaction ont été collectées au cours de l'activité par un dispositif technique d'enregistrement simultané. Les copies d'écran ont ensuite servi de support aux entretiens.

L'analyse statistique des transactions décrit les grammaires d'usage des collégiens. Elles donnent à voir l'existence d'une stratégie basée sur le jeu d'interaction homme-machine comme ressource principale, interrogeant alors nos systèmes didactiques. Le temps

d'enregistrement, relativement faible, des entretiens d'explicitation au regard du nombre de collégiens interviewés (7h50) pointe la posture réactive des élèves, plutôt qu'analytique. Les collégiens fournissent peu d'explications (Ladage, 2013), ils préfèrent désigner des objets.

3. Le rôle des interactions

La description des stratégies utilisées par les élèves montre qu'ils n'emploient ni une stratégie analytique, telle qu'elle est analysée par Boubée et al (2010), ni une stratégie de navigation. Ils emploient une stratégie mixte qui combine à la fois l'interrogation par requête et la navigation dans les résultats.

3.1. Les règles d'usage

L'analyse statistique des traces d'activité fait apparaître des règles d'usage appliquées par les collégiens quelque soit le type de tâche demandé : celles de poser d'abord sa question en langage naturelle, de reposer plusieurs fois sa question en utilisant l'indexation machine tout en contrôlant essentiellement la longueur de sa requête. S'appuyer sur les premiers résultats du moteur.

En effet 60% des élèves pose d'abord la question telle qu'elle est donnée, 72% des élèves reformule plus de 3 fois leur requête, 50% plus de 4 fois quand 57% d'entre eux limite leur requête à trois mots clé, 88% des élèves ne consulte que la première page de résultats, 65% se limite aux trois premiers résultats.

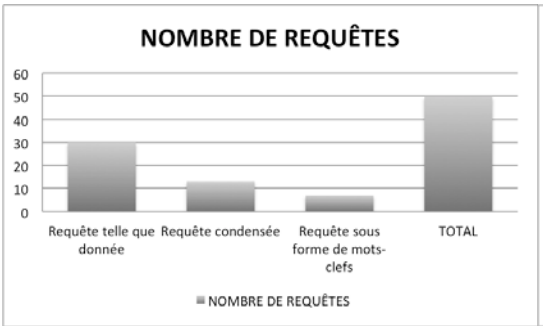


FIG 1 : Répartition des types de requête

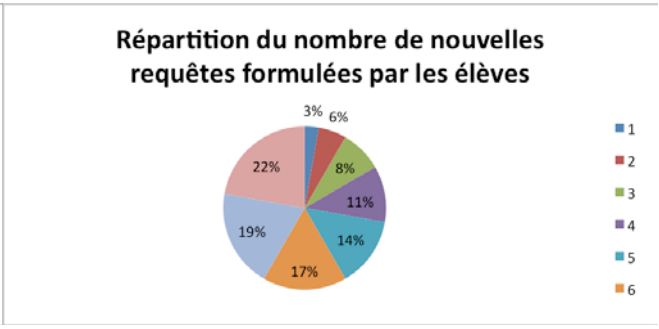


FIG 2 : Répartition du nombre de nouvelles requêtes formulées par les élèves

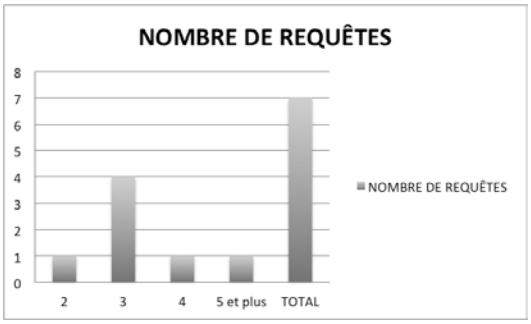


FIG 3 : Répartition nombre de mots clé/requête

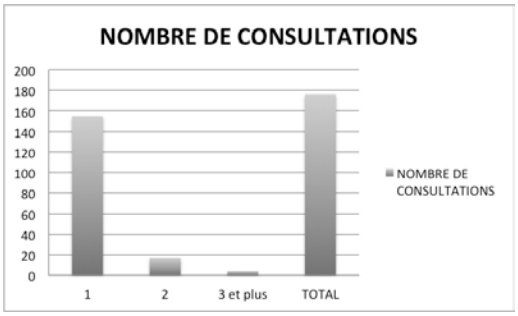


FIG 4 : Profondeur des pages consultées

3.2. La stratégie de requête des collégiens

L'analyse des traces révèle une stratégie de collecte de mots retournés par les résultats du moteur à une requête posée à la volée. Le mot clé prend son sens, émerge du contexte des résultats de la recherche prenant alors une valeur d'indexation du document, de la page web.

Il va permettre, étant donné sa pertinence dans le contexte, de renouveler la requête. La rationalité mise en œuvre par le collégien repose sur le repérage des termes qui ont une propriété indexicale. Dans un premier temps, les résultats de recherche distribuent un index de mots utiles à l'élève, d'ailleurs la première interaction avec l'outil est « directe », l'élève ne fait qu'énoncer la question telle qu'elle lui est posée pour élaborer son lexique. Dans un second temps, il teste l'anthologie constituée au cours d'un exercice de reformulation avant l'examen des résultats classés et renvoyés par le moteur.

3.3. La stratégie des collégiens : un mix entre interrogation et navigation

Les collégiens se constituent, à partir des résultats retournés du moteur, un lexique sur lequel repose leur stratégie de recherche. Le lexique constitué organise alors une ressource pour l'élaboration de la requête. De cette façon, l'importance qu'accordent les collégiens au poids des mots lorsque nous observons la formation des requêtes sur les moteurs de recherche, signale un usage particulier des techniques : non pas « mettre en mots » mais « user des mots ». C'est ce qui explique le contrôle surprenant de la longueur des requêtes et de l'importance de la reformulation dans les stratégies observées, confirmant les résultats déjà apportés (Spink 2004) (Jansen & Spink 2006). Plus qu'un contrôle sémantique de leur requête, les collégiens contrôlent la longueur de leur requête. En guise de traduction du besoin sous forme de mots clé, ils sont en quête de mots pour interagir avec les outils.

3.4. Contourner la navigation par une stratégie de filtrage

Alors qu'il est difficile pour les élèves de naviguer dans l'arborescence des hypermédias (Amadieu & Tricot 2006), ils se servent du classement des moteurs de recherche pour contourner la difficulté. Se limiter aux premiers résultats de recherche établi une règle d'accès direct à la page souhaitée d'un site sans passer par sa page d'accueil. Ainsi échappent-ils à la difficulté de navigation du site hôte. De cette façon, les collégiens exercent une stratégie de filtrage. Ils utilisent les propriétés des moteurs de recherche comme mode d'accès aux sites volumineux, qu'ils connaissent de réputation et qui sont bien référencés (Wikipédia, Google image).

Quelle stratégie ? Si les jeunes ont une préférence notoire pour « la recherche directe », l'expression est tirée de nos entretiens, l'exercice des requêtes prend bien sa part dans la stratégie mise en œuvre par les collégiens sans pourtant être une stratégie analytique. Le jeu des requêtes relève d'une autre visée que celle d'exprimer ses besoins adressés au système, il correspond à une technique d'accès direct aux pages des sites bien référencés. Les adolescents utilisent les propriétés techniques des outils tels que Word Rank pour se constituer un lexique de mots clé, puis ensuite, ils s'appuient sur d'autres algorithmes Page Rank pour filtrer les pages de réponses. De cette manière, ils réinventent les règles de la démarche analytique. C'est sur la base de ces observations que nous qualifions leur stratégie de stratégie mixte cette stratégie ne confère pas aux collégiens l'indice d'une meilleure connaissance des outils de recherche, en revanche elle atteste d'une adaptation de l'outil à leur activité de recherche.

Guidés par la très grande simplicité des interfaces, les jeunes collégiens, dans un contexte technique de labilité des artefacts, apprennent d'abord à interagir avec l'outil, ils butinent à l'intérieur des résultats et c'est ce « butinage assisté par ordinateur » qui, par bouclage, sert de support à leur stratégie de recherche.

4. La production langagière : retour au sujet

Comment interpréter les types de comportement observés? Basée sur les seules transactions, l'analyse de l'activité reste partielle, puisqu'en effet, l'activité de recherche repose essentiellement sur une activité de lecture. Les traces sont le fait d'un contexte et d'une intention. Sur une tâche de recherche seulement 20% de l'activité correspond aux données de transaction, 80% à celle de consultation. L'analyse des stratégies est donc à enrichir avec le l'intention poursuivi par l'élève. Dans cet objectif nous avons analysé un second corpus d'entretiens d'explicitation. Pour cette analyse nous choisissons l'étude statistique des formes langagières produites par les entretiens. Nous considérons qu'une analyse de contenu des entretiens nous contraindrait à rester au niveau de la parole. Avec quelles données langagières les élèves accompagnent les transactions effectuées? Les données langagières sont de nature à mettre en forme les intentions qui sont au préalable des règles que nous avons observées.

4.1. Se débarrasser de la langue

Les liens entre la linguistique et les sciences sociales sont fructueux tant du point de vue théorique que méthodologique, Habermas (1995). Au niveau méthodologique, la linguistique structurale, les théories de l'énonciation ou encore la pragmatique linguistique tendent à se débarrasser de la langue, à la dissoudre et proposent des voies interprétatives nouvelles, des possibilités d'atteindre l'objectivité scientifique. Si l'ambition sémiologique des structuralistes a vécu, nous conservons l'idée d'une déconstruction de la polysémie langagière, proscrire les mots comme langage, isoler la signification, n'en saisir que la matérialité, nous invitant à rompre avec une compréhension intuitive de la langue. C'est une condition nécessaire qui ne résout pas le problème de l'interprétation : quelles sont les possibilités de rupture avec le sens commun, les possibilités d'analyse du langage autrement que par le langage?

A son époque, c'est autour de l'idée de rupture épistémologique que s'est construite l'analyse du discours formulée par Pêcheux (1968). Au-delà de la signification des mots rapportés hâtivement à son locuteur comme signes, l'analyse du discours porte l'exigence de prendre en compte les conditions de production des énoncés explicitant les théories de l'énonciation. Les mots correspondent à des traces énonciatives, des données lexicales. L'analyse de ces données réclame une précision sur les objectifs et les moyens mis en œuvre en veillant à demeurer dans le même registre interprétatif car la difficulté réside précisément dans le passage d'un niveau à un autre, comment interpréter des données linguistiques à partir d'une théorie de la langue sans faire intervenir les usages de la langue ? « Ces deux niveaux viennent de ce que le langage, comme le dit J. Molino (1989) connaît deux formes d'existence, une existence matérielle, que constituent les « réalisations linguistiques » et une existence sociale, que constituent les actes d'énonciation et de réception (Ramognino, 1999 p.44). Quelle interprétation pouvons-nous faire de nos entretiens en laissant provisoirement de côté le sens que donnent les élèves aux phénomènes qu'ils évoquent afin de respecter un niveau neutre de la langue ?

4.1.1. Le pragmatisme linguistique

Nous cherchons dans les théories de l'énonciation les constructions possibles à partir de la matière langagière. Nous choisissons d'observer la matérialité langagière produite par les élèves en postulant comme forces coercitives, la langue. En effet, lorsque l'élève s'exprime il évoque une réalité présente avec les mots chargés de mémoire, contraint par une norme. Notre analyse porte alors sur la description des formes langagières produites par les récits d'entretien. Avec quelles formes langagières sont associées la réalité du collégien ? Quelles en sont les occurrences, les relations entre elles ?

4.1.2. Le fonctionnement langagier

L'activité langagière mêle plusieurs voix, elle n'a pas seulement trait avec le « ici et maintenant », elle est soumise aussi à sa mémoire. C'est sur ce postulat que nous proposons de décrire l'activité langagière. Il est possible d'inventorier des formes langagières qui structurent l'activité langagière, des formes qui montrent un emploi problématique de la langue, de rendre compte des fonctionnements langagiers sans passer par un modèle interprétatif , sans rechercher les intentionnalités des sujets dans les actes de langage ou rapporter les actes de langage à leur représentation, en postulant une description basée sur la seule autonomie des données langagières pour rendre visible « l'hétérogénéité langagière ». Nous analysons la matière langagière comme données de description d'une activité. Pour plus de rigueur nous employons des techniques de description qui permettent d'être précis et de garantir une automatisation méthodologique de l'observation par rapport aux données langagières. Ainsi avons-nous choisi d'assister notre description d'un programme informatique. Nous cherchons à rendre compte des statistiques de la matérialité langagière comme forme descriptive. Quelles sont les formes langagières présentes dans le corpus textuel ? Quel poids représentent-elles ? Comment sont-elles organisées entre elles ? Quelles sont leurs relations de proximité ? Quelles sont les formes répétées autant de possibilités descriptives qui peuvent nous aider à repérer les traces d'activité cognitive.

Les données langagières ont comme statut épistémologique d'être des traces d'opérations cognitives, éthiques et esthétiques. L'analyse statistique lexicale éclaire un fonctionnement méta langagier qui dissout la polysémie langagière. La description à l'aide de la statistique consiste à vider le sens du terme, de leur contenu pour leur affecter une valeur méta langagière ayant alors la capacité de prendre une valeur différentielle normative calculée à partir de la distribution statistique. La langue est une institution qui fabrique des connaissances, la description de l'activité langagière a pour but d'observer l'activité symbolique. Pour rendre compte de l'expérience des collégiens, nous avons décrit leur activité à l'aide du programme IRaMuTeQ[3], un outil d'analyse de discours assisté par ordinateur.

4.2. L'activité langagière saisi par l'analyse statistique

L'analyse statistique lexicale fournit un certain nombre d'outils d'analyse : le nombre d'occurrences du corpus des entretiens, le nombre de formes lexicales (lemmatisées), le type de forme lexicale, le lexique c'est à dire les occurrences de chaque forme lexicale, les cooccurrences de formes calculées sur leur probabilité d'être associées à une autre forme par rapport à une distribution au hasard. Elle calcule le suremploi ou sous-emploi d'une forme lexicale pour une partie, une variable ou une modalité. Elle établit la distribution des formes en les comparant à la distribution théorique, l'analyse factorielle des formes et variables sélectionnées (AFC), le classement des regroupements de segments (CHD), des classes de RST, une classe de RST représente un regroupement de segments de textes (ST) en fonction du degré d'homologie des regroupements (RST) en mots pleins. Le classement s'effectue à partir de la fonction Classification descendante hiérarchique (CDH) de l'outil. Le calcul d'homologie résulte de la plus ou moins grande distance entre deux classes distinctes, calculé par le Khi-deux entre deux groupes, ce qui peut donner lieu à des subdivisions de classes.

[3] Logiciel libre conçu par Ratinaud P. et Dejean S. (Ratinaud, 2012 ou Ratinaud et Dejean, 2009) développé sur la base de logiciels libres : Python : http://www.python.org; R (R Development Core Team, 2009): http://r-project.org; Lexique 3 (New, Pallier & Ferrand, 2005)

4.2.1. Analyse du matériel langagier, les caractéristiques de notre corpus

Après formatage de notre corpus pour pouvoir être lu par le logiciel IRaMuTeQ et déclaration de variables étoilées et de modalités, notre corpus se compose ainsi :

41 textes avec au départ 899 segments, 31 647 occurrences, 1851 formes lexicales et 806 Hapax, le traitement statistique du corpus porte au final sur nos 41 textes composés de 31646 occurrences, 1260 formes lexicales après lemmatisation avec un nombre moyen d'occurrences par formes de 25.12. Le nombre d'Hapax est de 494 (1,56% des occurrences et 39,21 % des formes). 771,88 occurrences par texte. Deux algorithmes sont lancés sur le corpus une Analyse Factorielle de Correspondances (AFC): une Classification Descendante Hiérarchique (CDH). 738 segments ont été classés sur 899 soit 82.09 % du corpus ce qui représente une perte d'un peu plus de 18% du corpus.

4.3. Forte récurrence des formes employées

Les formes lexicales du corpus analysé sont répétées plus de 25 fois (25,12 occurrences par formes) ce qui traduit une très forte récurrence des termes employés par les élèves. Par ailleurs, si le nombre d'hapax représente seulement un peu plus de 1% des occurrences (1,56 %), les 494 hapax représentent 39% des formes lexicales : ce qui dénote une certaine richesse lexicale (supérieur à 1%). L'importance du poids des occurrences est un indice de contrôle de l'activité. Tout se passe comme si les collégiens décrivent l'activité de façon identique indépendamment du type de question avec des formes lexicales bien communes. Nous en faisons l'analyse.

4.4. Distribution du matériel langagier

Les formes actives lemmatisées d'effectif maximum parmi les 30 premières sont : la forme Aller (318 oc), voir (226 oc), trouver (207 oc), regarder (169 oc), chercher (150 oc), mettre (148 oc), taper (107 oc), penser (93 oc), lire (80 oc), écrire (77 oc), connaître (73 oc).

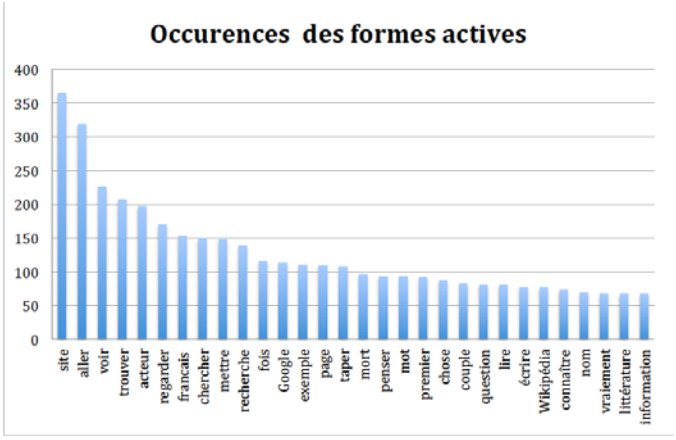
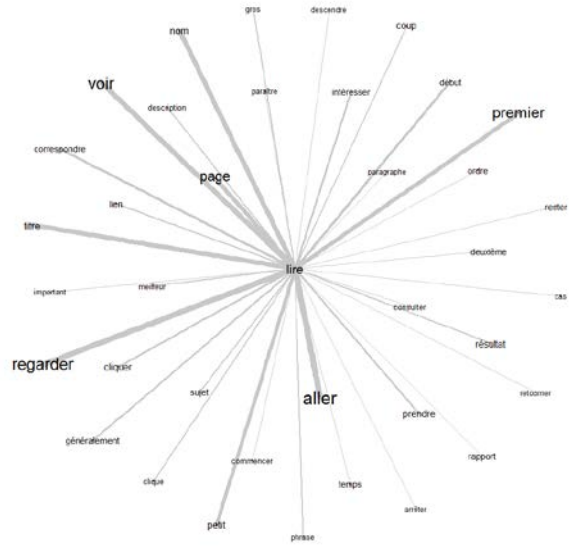


FIG 5 : Distribution des formes actives par nombre d'occurrences.

Cette hiérarchisation fréquentielle des formes actives est caractéristique de l'activité en question, elle s'organise autour de deux espaces sémantiques, d'un côté les synonymes du groupe des verbes : voir, trouver, regarder, chercher, penser, lire, connaître correspondant à une activité de lecture, d'un autre coté les synonymes de : mettre, taper, écrire correspondant à une activité d'écriture. Si le registre de langue est relativement homogène, les collégiens s'emploient à distinguer plusieurs synonymes de ces activités lire et écrire substituant ainsi de

nouvelles formes aux littératies[4] classiques de l'écrit. Ces formes nouvelles expriment des formes de « redisposition » des écrits consultées sur les écrans.

L'AFC, basée sur des calculs d'inertie du nuage de mots que constitue un corpus, fait apparaître les oppositions ou rapprochement, elle détermine pour cela des facteurs sur lesquels les formes se distribuent. Si on observe l'AFC des formes lire, regarder et voir nous observons qu'elles sont fortement corrélées, elles correspondent à l'activité de repérage pour la phase de la collecte. Par ailleurs, la cooccurrence entre aller et voir correspondent à une activité de vérification.



F6 : Graphe similitude des mots associés à « lire »

F 7 : Graphe similitude des mots associés à « regarder »

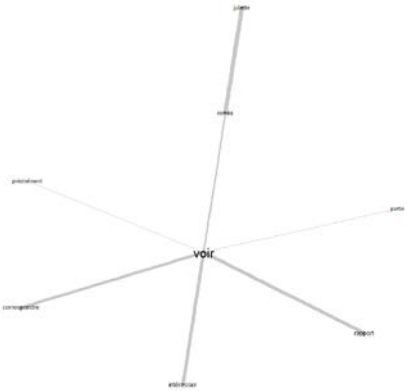


FIG 8 : Graphe similitude des mots associés à « voir »

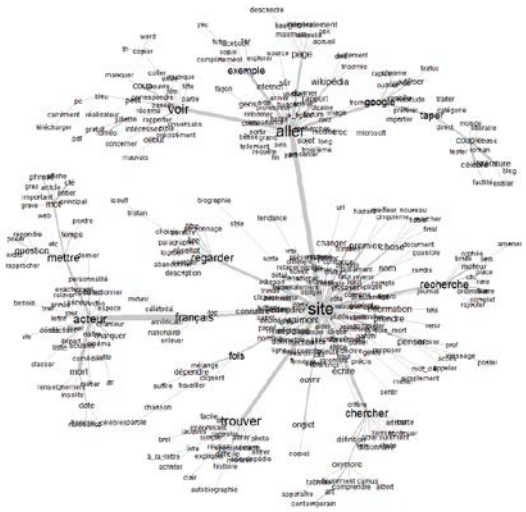


FIG 9 : Graphe des formes lexicales du corpus

[4] Le concept est d'origine anglo-saxonne. Webber, Sheila, Johnston, Bill. Information Literacy : definitions and models. 2006. Il est popularisé en France par celui de « culture de l'information ».

Les techniques de CHD appartenant aux méthodes d'analyse en clusters ou partitionnement des données nous permettent de déterminer l'existence de groupes homogènes de sujets ou de groupes de variables corrélées entre elles. Ces techniques s'avèrent intéressantes pour notre analyse, elles peuvent vérifier si la distribution des formes lexicales des entretiens forme des groupes de discours donc de stratégies distinctes confirmant qu'à chaque forme lexicale correspond une réalité.

4.5. Les formes d'altération des formes lexicales

L'analyse de la classification hiérarchique descendante du corpus analysé signale : 82,09% de segments classés sur 899 segments, 82% de segments construits restent dans la partie stable, 18% de segments ne sont pas analysés. Sur nos données, l'analyse statistique rend compte de 5 classes stables construites par la classification hiérarchique descendante selon la méthode Reinert [5]. Ces classes sont relativement équilibrées en prenant en compte les segments non classés. La première classe est la plus fournie, elle contient 29,4% des segments, la classe 2 :14,36%, la classe 3 : 22,36%, la classe 4 : 14,5%, la classe 5 :19, 38%. Nombre de formes lexicales par classe. Effectifs de chaque classe : non classés 16

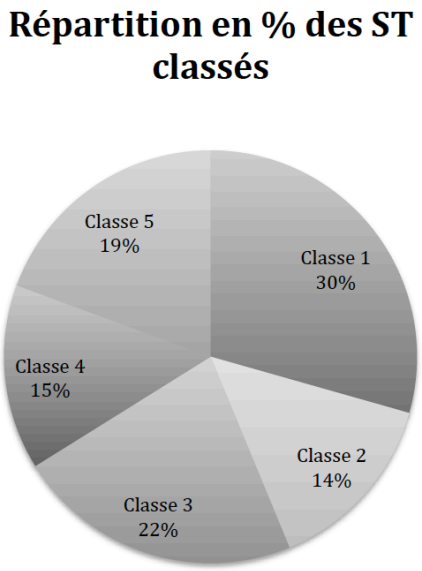


FIG 10 : Répartition en % des ST classés

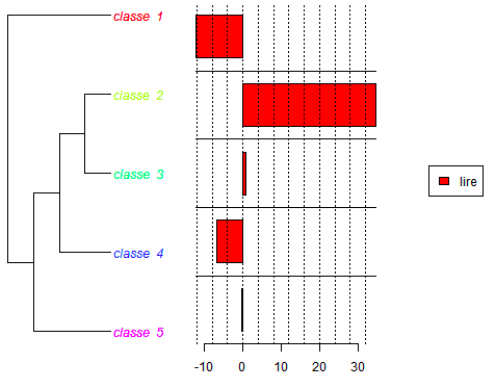


FIG 11 : Chi 2 classé sur la forme « lire »

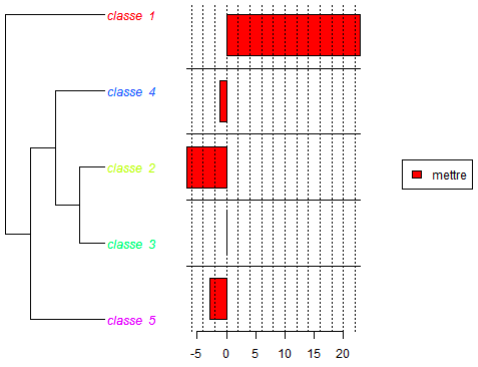


FIG 12 : Chi 2 sur la forme « mettre »

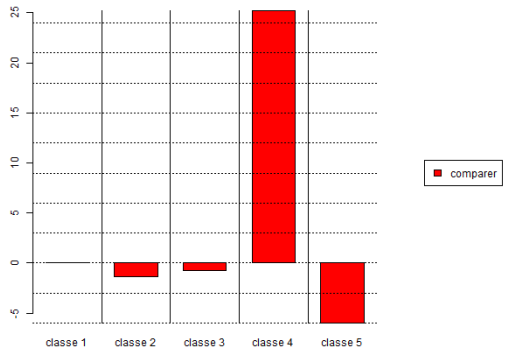


FIG 13 : Chi2 sur la forme « comparer »

[5] Max Reinert a été l'élève de Jean-Paul Benzecri fondateur de l'analyse des données à la française.

Sur nos données, avec 82% de segments classés, 5 classes stables sont construites. Les classes 1 et 3 contiennent 50 % des segments. La classe 1 s'attache à décrire l'importance de la recherche par mots clés autour notamment du verbe « mettre », la classe 2 , la moins fournie, évoque la consultation de ressources considérées comme sûres autour de la forme « lire ». La classe 3 tend à organiser le discours autour du tri des résultats, les termes s'organisent autour des ordres de grandeur. La classe 4 montre une activité de comparaison de sites, la classe 5 est celle des outils autour de la forme « voir ».

Ces 5 classes de discours montrent d'une part que les collégiens ont des stratégies diversifiées qui peuvent correspondre à des profils distincts d'utilisateur ou de recherche. Les stratégies sont centrées sur des phases particulières de la recherche d'information, mots clés ou requêtes, consultations, tris ou comparaisons, elles recouvrent souvent plutôt une question qu'une autre, la classe 1 décrit la question factuelle, la classe 5 la question documentaire, les 3 autres classes plutôt la question 3 alors qui représente peu d'élèves. Elles montrent, par l'étude de distribution des formes langagières des formes cognitives variées.

5. Conclusion

Si l'analyse des traces d'activité permet une description des comportements, que nous désignons comme une stratégie mixte d'interrogation par collecte de mots et de navigation directe, à l'intérieur d'un site volumineux filtré par le moteur de recherche, une stratégie de butinage, la difficulté réside dans la compréhension du jeu d'interaction avec la machine que seules les intentions révèlent. L'analyse des données textuelles s'avère être particulièrement intéressante sur ce point. Elle permet d'amener des hypothèses sur l'activité cognitive à partir de la matérialité langagière et ainsi de rendre compte des intentions poursuivies, restituées par les entretiens en dehors du langage. En effet, l'analyse statistique lexicale éclaire un fonctionnement méta langagier qui dissout la polysémie langagière, elle permet au delà de l'interprétation au niveau de la langue, de montrer une altération des formes cognitives. Les traces énonciatives saisies par les formes statistiques mises en œuvres telles que les CHD ou les classes d'énoncés caractérisées par des concordances de mots, des segments répétés, décrivent des formes qui existent entre le langage et les réalités énoncées. La page analysée statistiquement comme trace d'activité et son rapport aux formes lexicales particulières nous permet de construire les formes cognitives engagées, une forme plurielle de lecture : la page est « lue » alors que le site est « regardé », l'analyse statistique des formes lexicales décrivent deux réalités qui désignent des activités cognitives distinctes. Les collégiens en interaction avec l'environnement technologique montrent qu'ils qualifient leurs activités de manière différenciée, ils inventent en effet de nouvelles formes de littératie que nous nommons « redisposition des écrits ». Bien souvent décriées dans leur manifestation singulière à l'école, ces formes correspondent au phénomène d'autopoïèse décrit par Varela. Une activité cognitive nouvelle accompagne les interactions des collégiens avec les machines interrogeant les littératie de l'école.

6. Bibliographie

Béguin, P., & Clot, Y. (2004). L'action située dans le développement de l'activité. Activités, 1(2), 35–50.
Boubée, N., & Tricot, A. (2011). L'activité informationnelle juvénile. Paris: Lavoisier.
Cardon, D. (2013). Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google. Réseaux, 1(177), 63‑95.
Depover, C., Karsenti, T., & Komis, V. (2007). Enseigner avec les technologies: favoriser les apprentissages, développer des compétences. Québec: Presses de l'Université du Québec.
Ertzscheid, O., Gallezot,G . ; Simonnot, B. A la recherche de la "mémoire" du web : sédiments, traces et temporalités des documents en ligne. Manuel d'analyse du Web, Armand Colin, pp.53-68, 2013.
Habermas, J. (1995). Sociologie et théorie du langage: Christian Gauss lectures, 1970-1971. (R. Rochlitz, Trad.). Paris, France: A. Colin.
Jansen, B. J., & Spink, A. (2006). How are we searching the World Wild Web?: A comparison of nine search ingine transaction logs. Information Processing and Management, 42, 248‑263.
Jeanneret, Y. (2011). Complexité de la notion de trace. De la traque au tracé. In B.
Lebart, L., & Salem, A. (1988). Analyse statistiques des données textuelles. Paris, Dunod.
Leroi-Gourhan. L'homme et la matière : évolution et techniques
Linard, M. (1996). Des machines et des hommes: apprendre avec les nouvelles technologies. Paris; Montréal, Québec: L'Harmattan ; L'Harmattan INC.
Molino, J. (1989), « Interpréter » dans Reichter, C. L'interprétation des textes, Minuit, Paris, P.9-52
Pêcheux, M. (1969). L'analyse automatique du discours, Paris, Dunod.
Popper, K. R. (1982). La connaissance objective. (C. Bastyns, Trad.). Bruxelles, Belgique: Ed. Complexe.
Ramognino, N. (1999). Linguistique et sociologie, un point de vue méthodologique. Sociologie et sociétés, 31(1), 35‑50.
Ratinaud, P., & Dejean, S. (2009). Iramuteq: implémentation de la méthode d'Alceste d'analyse de texte dans un logiciel libre. Modélisation appliquée aux sciences humaines et sociales. (MASHS). Toulouse.
Ravestein, J., Ladage, C., & Johsua, S. (2007). Trouver et utiliser des informations sur Internet à l'école : problèmes techniques et questions éthiques. Revue française de pédagogie, 158(1), 71‑83.
Schutz, A. (1998). Éléments de sociologie phénoménologique. Editions L'Harmattan.
Simonian. (s. d.). Réhabiliter l'homme avec la technologie.
Simonnot, B. (2009). Culture informationnelle, culture numérique : au-delà de l'utilitaire. Les Cahiers du numérique, 5(3), 25‑37.
Suchman, L. A. (1987). Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press.
Varela, F. (1988). Invitation aux sciences cognitives, Paris, Seuil.