Cassandre, un outil pour construire, confronter et expliciter les interprétations

Christophe Lejeune

Chargé de cours temporaire
Centre de recherche en sciences de l'information et de la communication
Université Libre de Bruxelles
e-mail: christophe.lejeune@ulb.ac.be

Résumé

Cassandre est un logiciel libre d'analyse de textes, adossé à la plateforme collaborative Hypertopic. Fruit de la convergence épistémologique entre l'analyse qualitative (en sciences humaines) et le Web socio-sémantique (en informatique), cet outil accompagne la construction semi-automatique du (des) cadre(s) d'analyse du (des) chercheur(s). Une série de comptes-rendus d'utilisation (tantôt collectifs tantôt isolés, plutôt inductifs ou hypothético-déductifs) attestent des apports de l'outil : génération d'un journal de bord, économie cognitive et temporelle, flexibilité, aide au travail en équipe et multiplication des entrées dans le matériau. Ces usages témoignent également des inévitables limites afférentes à l'usage d'un logiciel dans une approche qualitative. En définitive, on plaide donc pour un recours raisonné à Cassandre, comme à tout autre logiciel.

Mots-clés Cassandre, analyse de textes, logiciel, interprétation.

Contexte

Le paysage des logiciels susceptibles d'assister le praticien de l'analyse qualitative est relativement varié. Plusieurs dizaines d'outils existent, de nouveaux apparaissant régulièrement, alors que d'autres ne dépassent pas le stade du prototype. Pour fixer les idées, nous distinguons cinq familles de fonctionnalités (Lejeune, 2010) :

1. les statistiques textuelles (que l'on trouve dans Lexico, Spad-T ou T-Lab),
2. les concordances (que permettent de réaliser Glossanet ou AntConc),
3. l'agrégation automatique (que proposent, à partir de techniques différentes, Alceste, Candide, Leximappe ou Réseau-lu),
4. l'annotation (qu'assitent NVivo ou Weft-QDA)
5. les dictionnaires (que mobilise General Inquirer)

S'agissant de familles de fonctionnalités, elles ne permettent pas de classer les logiciels. En effet, un même logiciel propose souvent des fonctionnalités originaires de familles différentes (Hyperbase combine ainsi des procédures de statistiques textuelles à un concordancier; AtlasTI permet d'établir des statistiques sur base des annotations de l'analyste; la suite Provalis combine, quant à elle, des outils des cinq familles).

Sciences de la communication, de la documentation, de l'éducation, de l'énonciation, de l'information, anthropologie, gestion, histoire, philologie, psychologie, sociologie... Ces ressources informatiques sont convoquées dans un très grand nombre de disciplines. Si, en France, la famille (3) que je qualifie d'agrégation automatique occupe un statut privilégié, les outils d'annotation (4) emportent l'adhésion de la majorité des praticiens de l'analyse qualitative.

Avec leur interface proche du traitement de texte, et leurs fonctionnalités très simples (permettant de souligner et d'annoter des passages du matériau textuel), les logiciels d'annotation s'avèrent à la fois rassurants et flexibles. Pour autant, à en lire la littérature et à s'enquérir auprès de leurs praticiens les plus expérimentés, ces outils ne représentent pas la panacée.

Si les analystes apprécient particulièrement de disposer d'outils reproduisant, sur écran, l'espace de travail effectué sur papier, la paire de ciseau dans une main et le crayon de couleur dans l'autre, ils regrettent également que l'informatique ne leur facilite que très peu les

opérations de codage les plus répétitives. Ils souffrent particulièrement d'être amenés à recommencer la lecture (et l'analyse) intégrale de leur corpus lorsque des hypothèses inattendues émergent en cours de travail (Trivelin, 2003). On peut se féliciter de ce qu'un logiciel encourage les chercheurs à s'imprégner (encore et encore) de leur matériau; on peut également réfléchir à comment l'informatique peut contribuer à ce travail.

Par ailleurs, la plupart des praticiens de l'analyse qualitative témoignent d'une grande ouverture interdisciplinaire ainsi que d'un goût pour le travail en équipe. Même si les intéressés ne le formulent pas explicitement, ces pratiques en appellent également à des innovations technologiques.

Nous avons donc décidé de tenter de relever ces défis adressés aux logiciels d'analyse qualitative :

1. faciliter les opérations intellectuelles d'annotation de texte ainsi que leur révision à tout moment du processus interprétatif,
2. permettre la coexistence et la confrontation d'analyses alternatives,
3. soutenir le travail collectif autour d'une série de textes empiriques communs.

Le logiciel Cassandre

Le logiciel libre d'analyse de textes Cassandre, développé par l'auteur, fait partie de la plate-forme informatique Hypertopic pour l'analyse collaborative de données qualitatives, fruit d'une collaboration internationale entre chercheurs en sciences humaines et en informatique. Conçu pour accompagner la construction d'interprétations d'un matériau empirique textuel, ce logiciel se rapproche de la famille des outils d'annotation.

La première spécificité de Cassandre repose sur une logique d'annotation *semi-automatique* procédant de l'identification, par l'analyste, d'expressions-clés. Pratiquement, l'analyste identifie, pendant sa lecture des textes, les passages qu'il souhaite analyser. Contrairement aux logiciels d'annotation classiques, il ne souligne pas chacun de ces passages dans leur intégralité mais seulement, en leur sein, le marqueur (c'est-à-dire un mot ou une expression-clé) susceptible d'apparaître également dans d'autres passages (Lejeune, 2008). À l'aide de la souris, il glisse ce marqueur dans un espace de travail. Cette opération met automatiquement en parallèle tous les passages qui comportent ce mot-clé ou cette expression.

La plupart du temps, un même phénomène est mentionné dans le matériau sous une variété de formulations. Le chercheur peut donc rassembler différents marqueurs retenus sous une même « étiquette ». Cette « étiquette » permet de désigner le phénomène; elle constitue la définition en intention d'un « registre », dont la liste des marqueurs constitue la définition en extension. Les registres constituent la représentation, au sein de la plateforme informatique, des catégories d'analyse du chercheur.
Une deuxième originalité découle de la nature collaborative de la plate-forme informatique à laquelle il est adossé : cet outil assiste (et encourage) en effet la construction et la confrontation de *plusieurs* analyses d'un même matériau empirique.

Dictionnaire ou registre ?

Les registres peuvent rappeler les premiers usages informatiques de listes de mots-clés en sciences humaines. Dès les années soixante, en effet, les chercheurs ont procédé à des analyses de textes assistées par ordinateur. En collaboration avec IBM, l'équipe de Philip Stone avait alors conçu le logiciel General Inquirer. Celui-ci se propose de coder automatiquement les textes au moyen d'indicateurs définissant (en extension) des catégories d'analyses (appelées dictionnaires). Plus qu'une différence technique, les dictionnaires du General Inquirer se distinguent des registres de Cassandre au niveau méthodologique et épistémologique.

Dépositaire de l'orientation donnée à l'analyse de contenu, le dictionnaire fonctionne comme une grille de codage (« codebook ») appliquée aux textes. Se substituant ainsi aux « codeurs », le dictionnaire assure les tâches répétitives et fastidieuses de repérage des indicateurs définis *ex ante* par le chercheur. Pour leur part, les registres s'élaborent dans les allers et retours entre les sources empiriques et le cadre d'analyse du chercheur. L'élaboration des registres accompagne le processus de « théorisation » et lui est consubstantiel. Les dictionnaires préexistent à la confrontation aux sources, alors que les registres découlent de leur analyse.
En conséquence, une deuxième différence méthodologique caractérise ces deux outils. En analyse de contenu, le dictionnaire est défini une fois pour toute. Le registre, par contre, voit ses marqueurs modifiés au fur et à mesure de l'exploration du corpus, se subdivise lorsque l'analyse s'affine ou change de nom au gré du processus de conceptualisation. La qualité d'un dictionnaire est fonction de son immuabilité; celle d'un registre, de sa plasticité.

En analyse de contenu, les dictionnaires constituent des blocs sémantiques consensuels. Leur « objectivité » est d'ailleurs fonction de leur propension à emporter le « consensus » (Berelson, 1952; Mucchielli, 2006). Les registres, pour leur part, entrent dans la composition d'un cadre d'analyse; plusieurs cadres d'analyse peuvent bien entendu caractériser un corpus. Loin d'être problématique, cette multiplicité des interprétations permet au contraire de rendre compte de la complexité du phénomène.

La rencontre entre analyse qualitative et informatique

Les questions de terminologie ne sont pas triviales : le nom lui-même du projet General Inquirer témoigne, pour les chercheurs en informatique, d'une référence aux recherches en Intelligence Artificielle. Si ce paradigme ne jouit plus aujourd'hui d'une aussi bonne presse qu'à l'époque, les recherches actuelles sur le Web sémantique reprennent une série de thèmes témoignant d'une convergence épistémologique avec l'analyse de contenu. Elles proposent de formaliser les connaissances humaines dans des dictionnaires (appelés « ontologies »). Stables et univoques (donc, consensuelles), les ontologies permettraient aux ordinateurs de « calculer » le sens (unique) contenu dans les documents, sans intervention d'un interprète, ce qui (dans les termes de l'analyse de contenu) en garantit l'objectivité.

À la différence de General Inquirer et des thèses du Web sémantique, Cassandre ne fait pas l'éloge du consensus. Son élaboration découle de la collaboration entre des praticiens d'une analyse radicalement qualitative et des chercheurs en informatique convaincus de l'incalculabilité du sens (c'est le paradigme du Web socio-sémantique). À une vision consensuelle et unilatérale du monde (qu'ils jugent intenable), ces chercheurs préfèrent la co-existence d'une grande variété de schèmes interprétatifs.

Interprétation	Consensuelle (objectivité)	Multiple (inter-subjectivité)
Informatique	Intelligence Artificielle	Web socio-sémantique
Analyse de texte	Analyse de contenu	Analyse qualitative

Figure 1 : Rencontre de l'analyse qualitative et de l'informatique

Quelques projets menés avec Cassandre

Tout comme avec les outils d'annotation, les utilisateurs de Cassandre témoignent d'une grande de variétés d'approches dans la manière dont ils conçoivent leurs registres.

Dans leur recherche sur le vécu des femmes reprenant leur emploi après une interruption de longue durée (les « femmes rentrantes »), Christine Delhaye et Anne-Marie Dieu ont adopté une démarche radicalement empirique. Partant, comme il est classique en analyse qualitative, des entretiens transcrits, elles ont relevés progressivement certains marqueurs qu'elles ont peu à peu structurés en registres (figure 2). Ceux-ci leur ont fourni, au moment de la rédaction de leur rapport, les lignes directrices de leurs conclusions.

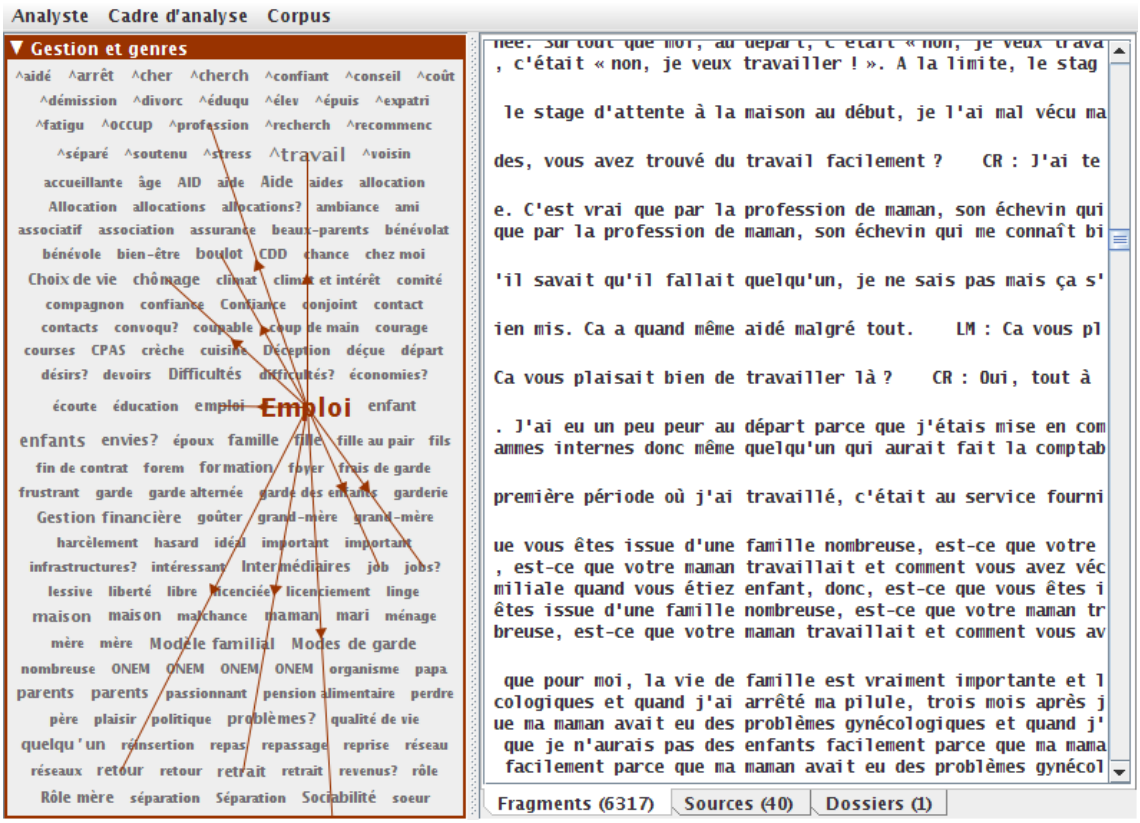

Figure 2 : Femmes rentrantes

Analysant des groupes focalisés (*focus groups*) sur les risques électromagnétiques, une équipe de chercheurs en sciences politiques s'est intéressée au regard porté par des comités de quartier sur la société de consommation (Brunet et al., soumis pour publication). Progressivement, la collaboration au sein de cette équipe a vu émerger deux analyses complémentaires : l'une, plutôt thématique, s'employant à expliciter les ressorts de la perception des risques; la deuxième, plus interprétative, s'est pour sa part concentrée sur l'identification des figures typiques de la veille sur les risques (figure 3). Empruntant les

catégories de la sociologie pragmatique, cette seconde analyse a été notamment le cadre d'élaboration d'un registre des « effets pervers »... qui s'est redéfini, au cours de la recherche, en tant que registre (plus général) des « phénomènes auto-entretenus » (Lejeune, 2008).

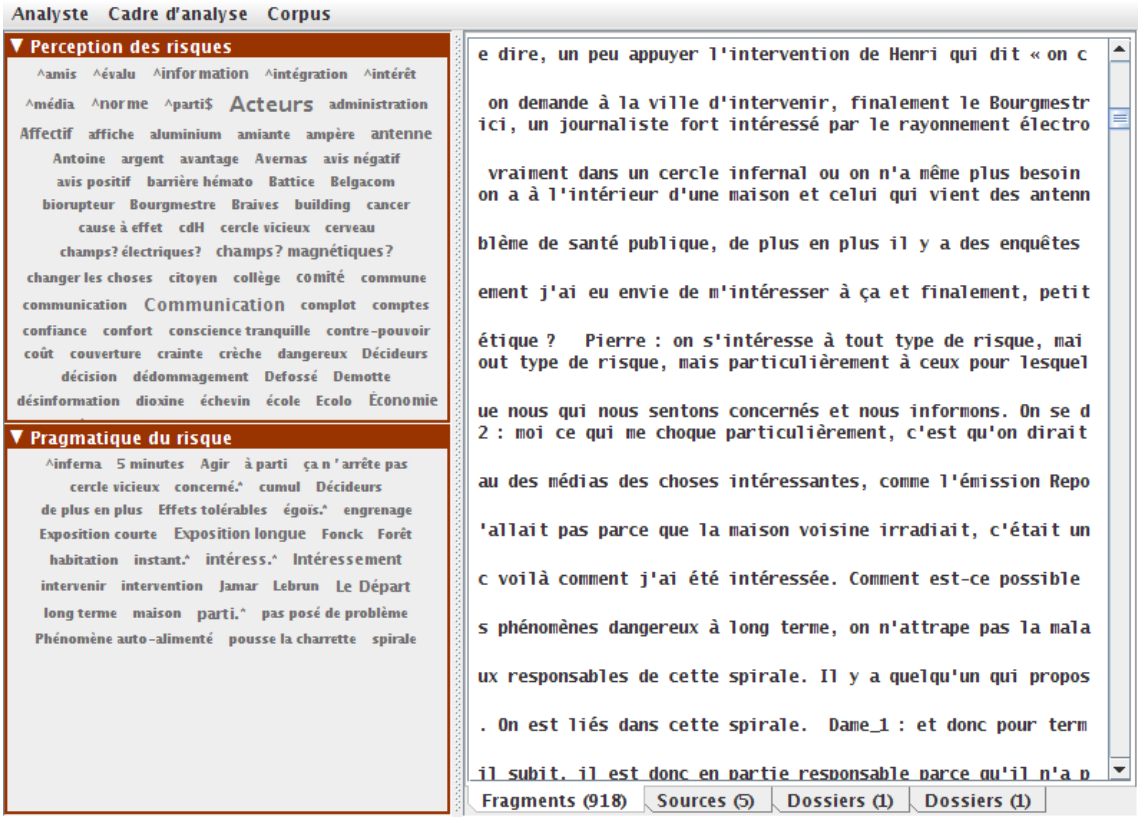

Figure 3 : Deux analyses relatives aux risques électromagnétiques

La plateforme est conçue pour assister des analyses collectives. Un tel travail collectif peut procéder de l'élaboration d'une interprétation commune (comme dans la recherche sur les « femmes rentrantes ») ou de la confrontation d'analyses alternatives (comme dans la recherche sur les risques électromagnétiques). Rien n'empêche pour autant d'y recourir dans un travail plus isolé (comme l'est souvent le travail de thèse).

Pour son analyse d'entretiens auprès de repreneurs de petites et moyennes entreprises familiales, Sarah Santin a ainsi développé des registres lui permettant d'opérer un diagnostic par croisements combinatoires. Les libellés des registres représentent, dans ce cas, les catégories constitutives des hypothèses de recherche, alors que les marqueurs qui les composent découlent du repérage, au sein du matériau empirique, des occurrences de mots-clés ou d'expression qui expriment ces registres. Les registres créés dans cette logique constituent des briques élémentaires soit thématiques soit modales. Les registres thématiques portent sur les sujets abordés par les repreneurs d'entreprise en entretien; les registres modaux portent sur les prises de position (critique, évaluative, favorable) témoignées par les mêmes

acteurs. Ces registres élémentaires permettent ensuite, par combinaison, d'identifier les thématiques associées (par les informateurs) à la pénibilité (ou à la chance) ou bien, de déterminer, pour chacune des thématiques identifiées, ce que l'ensemble des repreneurs en ont dit (Constantinidis et Santin, 2008).

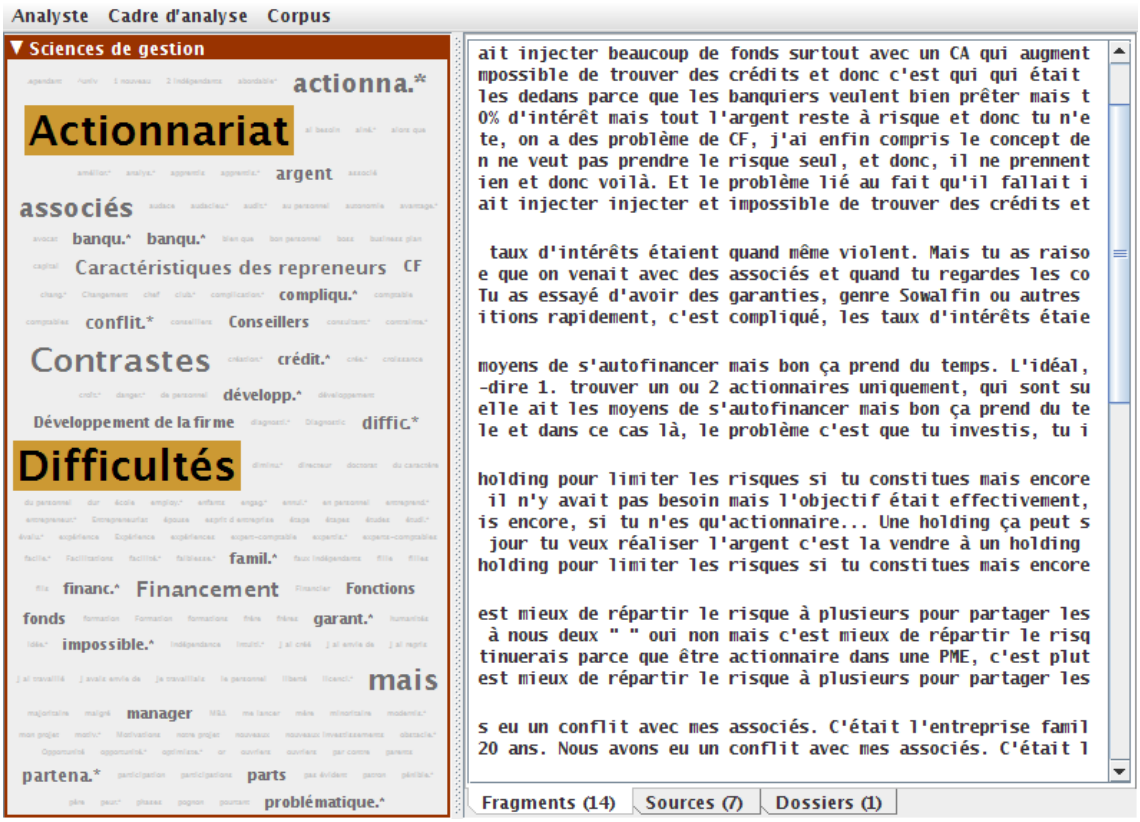

Figure 4 : Combinaison de registres dans la reprise d'entreprises familiales

La logique des registres, très souple, permet de tirer parti de l'informatisation des corpus de textes. Toutefois, il subsiste un grand nombre de phénomènes évoqués ou suggérés en entretiens (ou dans des textes écrits) que des marqueurs échouent à rapatrier. L'exemple suivant, extrait d'une recherche menée par l'auteur sur le témoignage de la confiance dans les collectifs d'utilisateurs de logiciels libres, offre un exemple tout à fait parlant (figure 5). Le premier message posté dans un forum de discussions dédié à des problèmes informatiques, le 1er janvier 2007, rappelle une discussion technique ayant débuté quelques jours auparavant et non encore clôturée. La réponse à cette intervention débute par une ligne souhaitant simplement une « bonne année à toi aussi ». Il s'agit évidemment d'une critique au fait que la première intervention de l'année n'a comporté aucun vœux de nouvel an aux (nombreux) participants au forum. Si le sens (critique) n'échappe à aucun humain, il est bien entendu inaccessible à une machine et, en particulier, impossible à saisir via un marqueur.

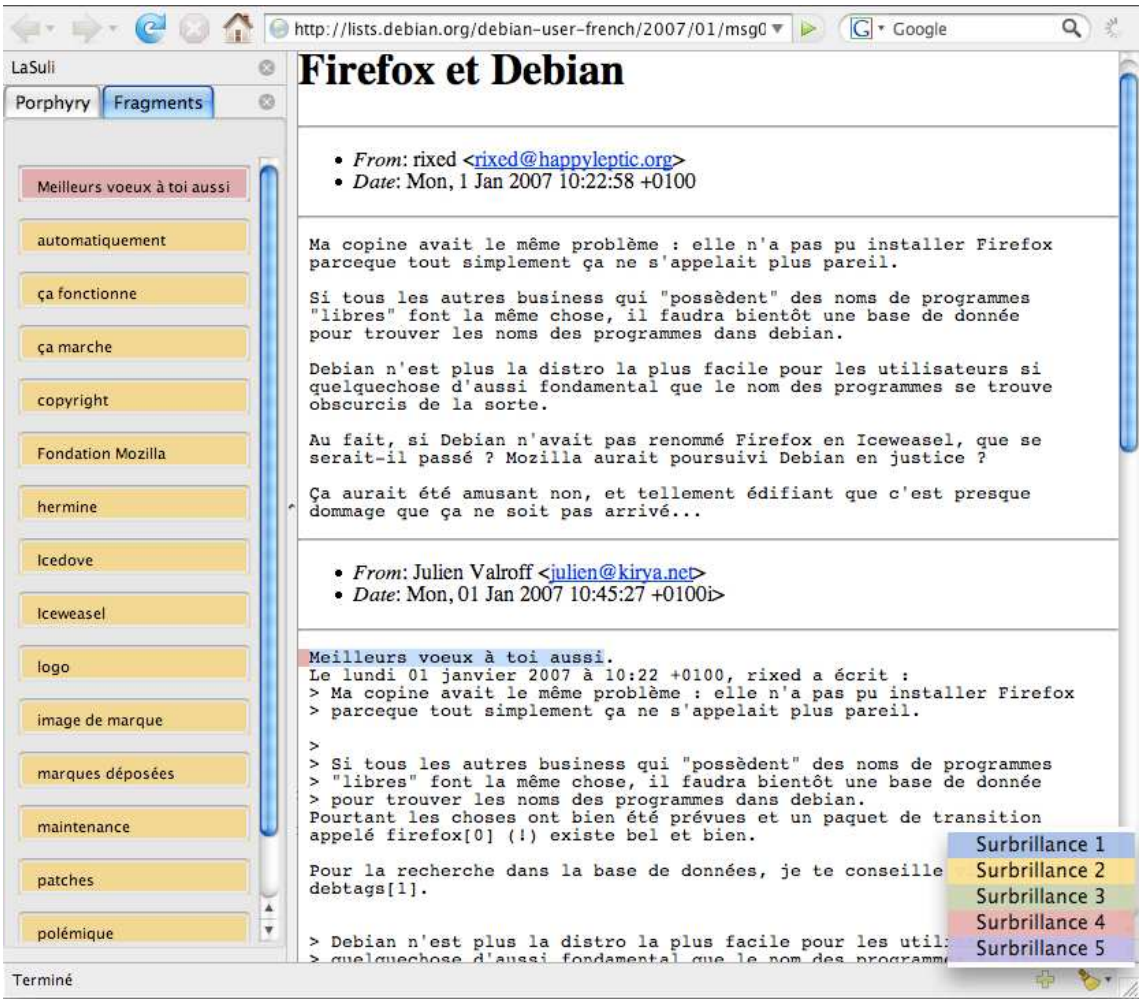

Figure 5 : Allusion échappant aux marqueurs

D'autres doctorants ont tiré profit de l'automatisation qu'autorisent les registres pour adopter une démarche plus hypothético-déductive. Dans son analyse d'entretiens menés auprès des spécialistes de la veille stratégique à propos de leur évaluation de la qualité de l'information disponibles dans le « Web 2.0 », Jérémy Depauw (2009) a procédé de cette manière. Rien n'empêche bien entendu d'opérer de la sorte. Toutefois, pour réussir, une utilisation déductive de tels outils d'analyse qualitative nécessite impérativement une connaissance intime du corpus d'entretiens analysés. En effet, l'analyse par registres n'a de son sens que si les marqueurs sont congruents avec le matériau empirique. Sans une attention aiguë à l'adéquation entre les marqueurs et le texte, l'application d'une grille de codage pré-établie conduit toujours à des contresens. Le gain (temporel et cognitif) de l'annotation par registres est alors définitivement perdu.

Les deux utilisations suivantes illustrent cette limite.

Dans une utilisation pilote de Cassandre, Gautier Pirotte a analysé un corpus composé de rapports d'activités d'organisations non gouvernementales internationales (ONU, UNESCO, FAO). À cette fin, il a développé un cadre d'analyse intégrant les trois axes de sa théorie de la société civile (Pirotte, 2007). Nous lui avons proposé d'appliquer à son corpus un cadre d'analyse alternatif développé ailleurs et intégrant les sept cités des *Économies de la Grandeur* (Boltanski et Thévenot, 1991). Le registre de la cité industrielle (notamment caractérisée par des questions de standardisation) s'est révélé omniprésent dans le corpus. À l'examen, ce diagnostic tenait essentiellement au marqueur « développement ». Dans le contexte original de l'analyse d'écrits relatifs au monde de l'entreprise, l'incorporation du marqueur « développement » au registre de la cité industrielle permet pertinemment de rendre compte de la question du développement industriel. Dans le cadre des rapports annuels des ONG, cependant, le développement caractérisait plutôt le processus suivant lequel les sociétés du Sud se « développent ». Sans un examen vigilant de l'importation du cadre d'analyse, cette polysémie (essentiellement contextuelle) aurait pu conduire à un véritable contresens. Mais on peut également tirer parti de cet artefact; il résulte en effet d'un phénomène éminemment social d'extension du registre industriel à des domaines aussi inattendu que l'écologie (le « développement durable ») et le soutien des minorités (du « développement local » aux « sociétés en développement »).

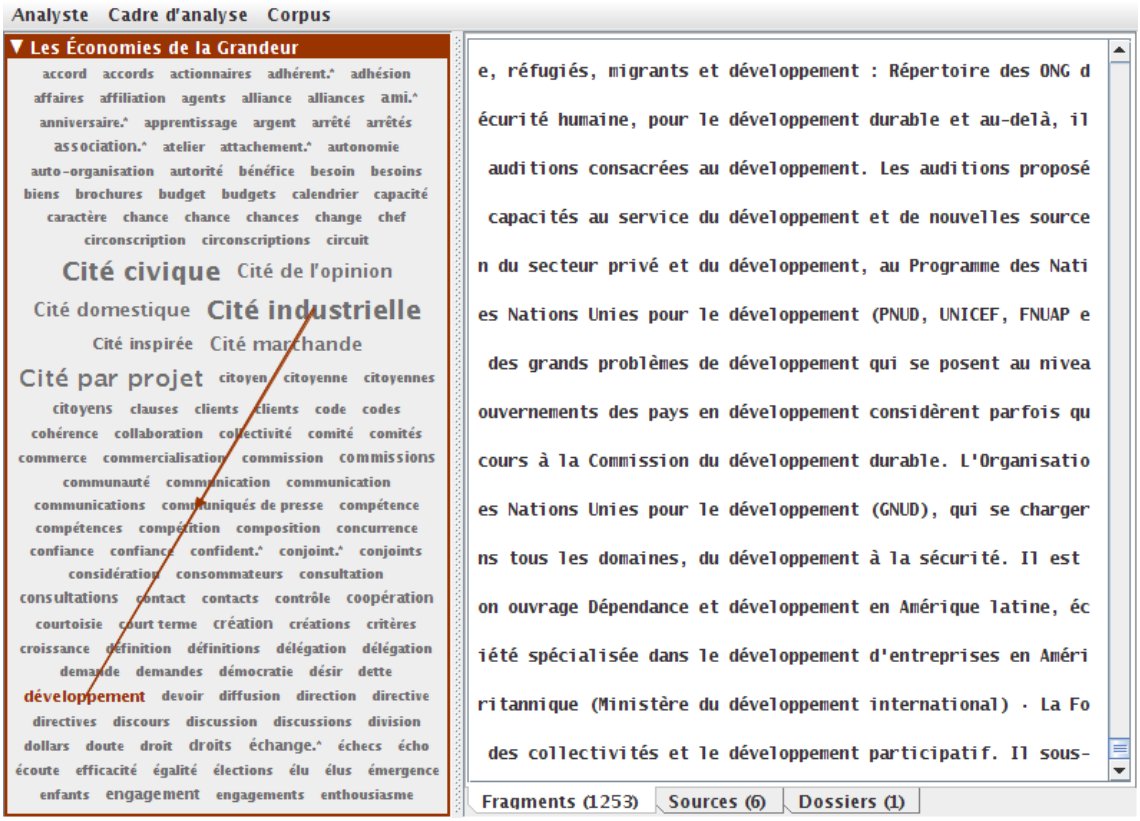

Figure 6 : Développement et société civile

Une autre expérience, menée par une étudiante en sciences de l'éducation, a consisté à importer l'échelle des valeurs de Schwartz (Dupont, 2009). L'expérience fit émerger le même genre de contresens que le cas du « développement » évoqué ci-avant. L'échelle préexistante comportait ainsi la catégorie du pouvoir contenant, notamment, un indicateur relatif à la « richesse ». Tout étant égal par ailleurs, une telle association apparaît relativement fondée. Mais une analyse qualitative peut difficilement faire l'économie de son matériau. Dans le cas présent, le corpus regroupait une série de décrets régulant l'enseignement en Belgique. La question de la « richesse » y renvoie bien évidemment à la pluralité et à l'ouverture des enseignements plutôt qu'à des considérations économiques. L'expérience a donc conclu, ici également, que l'importation d'une grille de lecture antérieure à la recherche devait impérativement s'accompagner de l'examen empirique de la pertinence de chacun des marqueurs qu'elle comporte. En outre, on imagine sans peine que ce même corpus (fut-il homogène) comportent un grand nombre de formulations différentes pour évoquer la question de la richesse (au sens de la variété) des enseignements. Ici également, le recours à un indicateur substantivé témoigne bien plus du raisonnement de l'intellectuel que du

repérage, dans le matériau textuel, de la manière dont le thème, l'idée en question, sont formulés par les informateurs.

Conclusion : apports et limites

Les chercheurs qui ont utilisé Cassandre apprécient particulièrement :

- l'*économie* substantielle, tant temporelle que cognitive, qu'apporte une annotation par registres (et expressions-clés) comparée à l'annotation « au feutre » ou au moyen d'autres logiciels d'annotation;
- la *flexibilité* de l'outil d'analyse par registres (qui accompagne une interprétation toujours réversible);
- la *génération automatisée d'un journal de bord*, qui rencontre leur volonté d'expliciter les étapes de la démarche scientifique;
- la possibilité de *confronter* des analyses alternatives;
- le *travail en équipe* (facilité par l'architecture partagée de la plateforme) en particulier entre chercheurs provenant de traditions ou de disciplines différentes, dès lors envisagées comme complémentaires et non concurrentes.

Cet outil constitue donc l'adjuvant appréciable d'une analyse qualitative de textes. Particulièrement congruent avec une approche émergente, inductive (ou, plus exactement, abductive), il peut également aider dans une approche plus déductive, à condition cependant d'évaluer scrupuleusement la pertinence de l'importation d'une analyse développée dans une autre recherche. Par ailleurs, cette approche apparaît *moins* souple que les manières traditionnelles de procéder, en particulier pour saisir des formulations comme l'allusion, le sous-entendu ou l'évocation.

Au final, on plaide donc pour un recours raisonné à Cassandre, comme à tout autre logiciel. Les limites qui lui sont inhérentes (comme c'est le cas de toute technique) démontrent en effet qu'il n'est pas d'outil « ultime » et que cette technique particulière doit dès lors être envisagée comme complémentaire à d'autres approches. Toutefois, dans la mesure où il apparaît que l'« économie » réalisée est mise à profit pour parcourir *autrement* les corpus, il n'en constitue pas moins adjuvant de l'inventivité du chercheur.

Remerciements

L'auteur tient à remercier Laetitia Godfroid et Orélie Desfriches Doria dont les commentaires ont permis d'améliorer ce texte.
Le développement du logiciel Cassandre n'aurait pas été possible sans Aurélien Bénel, Jean-Pierre Cahier, Manuel Zacklad, Hédi Zaher et Chao Zhou de l'Institut Charles Delaunay à l'Université de Technologie de Troyes. Merci à vous pour cette enrichissante collaboration.
L'auteur tient enfin à exprimer sa gratitude envers les utilisateurs de Cassandre qui, comme Christina Constantinidis, Christine Delhaye, Pierre Delvenne, Jérémy Depauw, Anne-Marie Dieu, Sylvie Dupont, Gautier Pirotte et Sarah Santin permettent, à travers leurs retours d'expérience, de continuellement améliorer cet outil d'analyse qualitative.

Références

Bénel, A. et Lejeune, C. (2009), Partager des corpus et leurs analyses à l'heure du Web 2.0. *Degrés*, Vol. 36-37, n° 136-137, p. m1-20. Disponible via http://hdl.handle.net/2268/12613

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

Boltanski, L. et Thévenot L. (1991). *De la justification. Les économies de la grandeur*. Paris: Gallimard.

Constantinidis, C. et Santin, S. (2008), La reprise d'entreprise familiale par les filles d'entrepreneur : Une lecture en termes de genre, *2èmes Journées Georges Doriot, L'entrepreneuriat familial: États des lieux et perspectives de recherche*, Paris, 15-16 Mai.

Brunet, S. Delvenne, P. Fallon, C. et Lejeune C. (soumis pour publication). Politique et expertise profane en situation de haute incertitude scientifique : le cas des champs électromagnétiques en Belgique francophone. *Politique et Sociétés*.

Depauw, J. (2009). *Qualité de l'information et vigilance collective sur le web. Étude des stratégies d'évaluation des sources en ligne par les professionnels de la gestion de l'information dans les organisations*. Thèse en sciences de l'information et de la communication, Bruxelles: Université Libre de Bruxelles.

Dupont, S. (2009). *Analyse des valeurs des futurs instituteurs primaires et des valeurs déclarées par les différents réseaux d'enseignement*. Mémoire en sciences de l'éducation, Mons: Université de Mons-Hainaut.

Glaser, B. G. and Strauss, A. L. (1970). *Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Lejeune, C. (2004a). L'analisi sociologica di un corpus derivato di fori di discussioni. Arricchimento reciproco delle analisi della conversazione, dei testi e di corpus. In Lancia F. *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*, Milano: FrancoAngeli.

Lejeune, C. (2004b). Représentations des réseaux de mots associés. In Purnelle, G. Fairon, C. et Dister, A. *Le pouvoir des mots. Actes des 7e Journées internationales d'Analyse statistique des Données Textuelles*, tome 2. Louvain: PUL / i6doc. Disponible via http://hdl.handle.net/2268/691

Lejeune, C. (2007). Petite histoire des ressources logicielles au service de la sociologie qualitative. In Claire Brossaud et Bernard Reber (Eds.), *Humanités numériques. Nouvelles technologies cognitives et concepts des sciences sociales*. Paris: Hermès Sciences Publications. Disponible via http://hdl.handle.net/2268/2491

Lejeune, C. (2008). Au fil de l'interprétation. L'apport des registres aux logiciels d'analyse qualitative. *Revue Suisse de Sociologie*, Vol. 34, n° 3, p. 593-603. Disponible via http://hdl.handle.net/2268/6123

Lejeune, C. (2010). Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative, *Recherches Qualitatives*.

Mucchielli, R. (2006). L'analyse de contenu des documents et des communications, Issy-les-Moulineaux : ESF.

Pirotte, G. (2007). *La notion de société civile*. Paris: La Découverte.

Stone, P. J. et al. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press.

Trivelin, B. (2003). Une aide à l'analyse de contenu : le tableau Excel. *Recherches Sociologiques*. Vol. 1, p. 135-147.

Zhou, C. Lejeune, C. and Bénel, A. (2006). Towards a standard protocol for community-driven organizations of knowledge. In Ghodous, P. Dieng-Kuntz, R. and Loureiro, G. *Leading the Web in Concurrent Engineering*. Amsterdam: IOS Press. Disponible via http://hdl.handle.net/2268/4186

Zacklad, M. Cahier, J.-P. Zaher, H. Bénel, A. Lejeune C. et Zhou C. (2007). Hypertopic : une métasémiotique et un protocole pour le Web socio-sémantique. In Trichet, F. *Actes des 18e Journées Francophones d'Ingénierie des Connaissances*, Grenoble: Cepaduès. Disponible via http://hdl.handle.net/2268/4184