Éloge de l'hétérogénéité des structures d'analyse de textes

Aurélien Bénel* — **Christophe Lejeune**** — **Chao Zhou*****

* *Laboratoire ICD/Tech-CICO, Université de technologie de Troyes*
12, rue Marie Curie – B.P. 2060, F-10010 Troyes Cedex
aurelien.benel@utt.fr

** *FAPSE-ISHS-HEC, Université de Liège*
7, Boulevard du Rectorat – Boîte 47, B-4000 Liège
christophe.lejeune@ulg.ac.be

*** *Nostos Technology Ltd.*
13-3-502, Longtengyuan 4, Huilongguan, Changping, C-102208 Beijing
chao.zhou@nostos.com.cn

RÉSUMÉ. *Dans cet article, nous montrons comment, en nous appuyant sur une infrastructure de gestion de « systèmes d'organisation des connaissances », nous avons construit des logiciels pour les chercheurs engagés dans un travail d'interprétation de textes. Fondé sur un modèle « à base de points de vue », ce développement nous a permis de nous interroger sur l'hétérogénéité des structures d'analyse, que celle-ci provienne de la diversité des auteurs de ces analyses ou de leur degré d'automatisation.*

ABSTRACT. *This paper deals with how we used a "knowledge organization system" management infrastructure to build software for researchers involved in text interpretation. Founded on a viewpoints-based model, this implementation highlighted the heterogeneity of analyses structures, whether it comes from authors diversity or from automation level.*

MOTS-CLÉS : *Analyse qualitative, annotation, humanités numériques, sémiotique, points de vue.*

KEYWORDS: *Qualitative analysis, annotation, digital humanities, semiotics, viewpoints.*

RSTI - DN – 13/2010. Applications à base de SOC hétérogènes, pages 41 à 56

1. Introduction

Dans l'analyse qualitative de documents, telle qu'elle est pratiquée par exemple par les sociologues ou les psychologues sur des retranscriptions d'entretiens, une tâche domine : la catégorisation. Il existe donc une certaine analogie entre cette activité et celle des documentalistes[1] (Lejeune, 2004). Partant de ce constat, nous avons essayé de construire des logiciels pour les chercheurs engagés dans un travail d'interprétation de textes, en nous appuyant sur une infrastructure (*Argos*) de gestion de « systèmes d'organisation des connaissances »[2].

Parce que le chercheur, en utilisant ce genre de logiciels, s'oblige à garder une trace de ses parcours d'interprétation, leur usage constitue selon nous un enjeu épistémologique majeur. Faut-il encore que ces logiciels sortent du bureau du chercheur pour se retrouver dans l'espace du débat scientifique. C'est dans ce sens que nous avons privilégié une mise en réseau de ces logiciels. Fondé sur un modèle « à base de points de vue » issu de l'ingénierie des connaissances (Zacklad *et al.*, 2007), ce développement nous a permis de nous interroger sur l'hétérogénéité des structures d'analyse, que celle-ci provienne de la diversité des auteurs de ces analyses ou de leur degré d'automatisation.

2. Diversité des auteurs

2.1. *Structures d'analyse de textes*

Avant d'aborder la question de l'hétérogénéité proprement dite, essayons de préciser ce que nous entendons par « structures d'analyse de textes ».

Dès les prémices de l'informatique, les chercheurs en sciences de l'homme et de la société ont su détecter le potentiel des technologies de l'information pour le traitement de leurs documents de travail (œuvres littéraires, documents d'archive, retranscriptions d'entretiens, coupures de presse...). Les « computers », ainsi détournés du simple calcul sur les nombres, se sont mis au service de l'analyse qualitative de textes. Aujourd'hui, les logiciels disponibles se répartissent en trois familles (Lejeune, 2010b) :

– les descendants des méthodes mécanographiques et informatiques permettant d'analyser le « style » d'un texte par des statistiques sur les mots et leur séquence (Tasman, 1957),

– ceux des *notional families* (Luhn, 1957, p. 314) et de *General Inquirer* (Stone *et al.*, 1966) qui permettent au chercheur d'inventorier des marqueurs à chercher automatiquement dans le texte,

1. Notamment dans les archives audiovisuelles où la frontière entre « annotation » et « indexation » s'estompe (Prié *et al.*, 1998).
2. Cette expression est employée par la Fédération des bibliothèques numériques (Hodge, 2000) pour généraliser toutes les structures permettant de modéliser le sujet d'un document ou de l'une de ses parties.

– ceux de *The Ethnograph* (Seidel *et al.*, 1984) qui permettent au chercheur de « coder » des segments de texte.

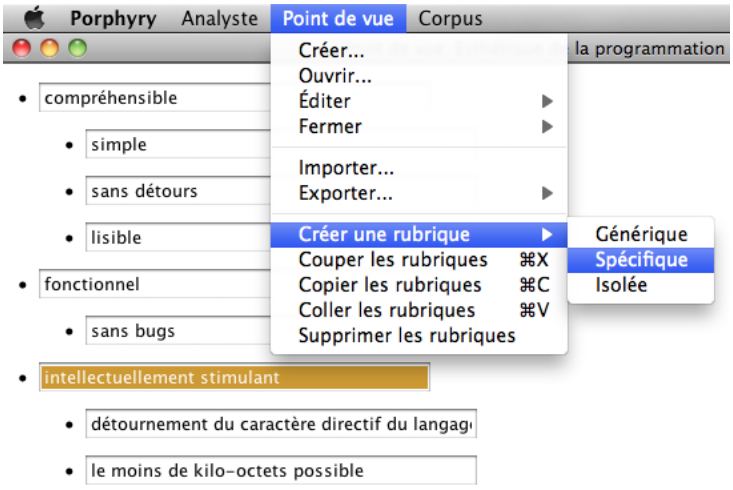

Figure 1. *Création d'une rubrique spécifique (copie d'écran de* Porphyry*)*

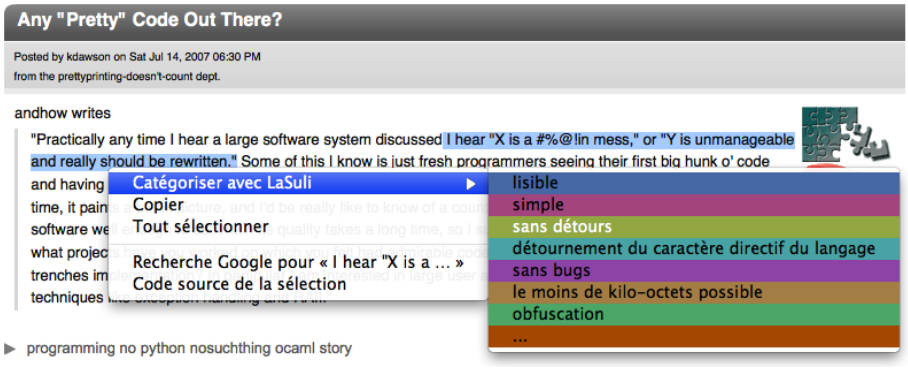

Figure 2. *Catégorisation à l'aide d'une rubrique préexistante (copie d'écran de* LaSuli*)*

Les logiciels de la famille la plus récente (*NVivo*, *Atlas.ti*, *MaxQDA*, *WeftQDA*...) proposent de transposer sur support numérique le geste consistant à « surligner » des fragments de texte en appliquant un code couleur correspondant à une grille d'analyse (Lejeune, 2007). Ce changement de support se justifie principalement par les possibilités qu'il offre de « lecture expérimentale » (Virbel, 1994). En effet, le parcours inverse de la catégorie d'analyse vers les fragments, permet, outre un accès facilité

aux fragments, le croisement (par exemple par comptage dans des tableaux) de plusieurs catégories d'analyse (Bryman, 2008). Ces logiciels permettent en général de regrouper les catégories d'analyse en catégories d'ordre supérieur. La structure qui en résulte est, suivant les logiciels, « arborescente » ou « hiérarchique »[3].

Nos outils *LaSuli* et *Porphyry* s'inscrivent dans cette famille. Pour autant, ils diffèrent sur un certain nombre de points. Les outils existants, bien que se revendiquant de la « Grounded Theory Methodology » (Glaser *et al.*, 2010), suivent souvent curieusement une démarche très « top-down » dans leur interface homme-machine. Malgré les mises en garde de ses fondateurs, le terme même de « codage » laisse penser que la catégorie d'analyse serait pré-existante à l'analyse (cf. figure 2). De même, les fonctions d'organisation des catégories se limitent souvent à la création de catégories de plus en plus spécifiques (cf. figure 1). Sans nécessairement défendre une méthode inductive pour autant, nous avons souhaité permettre également les démarches « bottom-up » (cf. figures 3 à 5). En effet l'interprétation se construit d'ordinaire au travers d'allers et retours entre le matériau empirique et les concepts de la discipline.

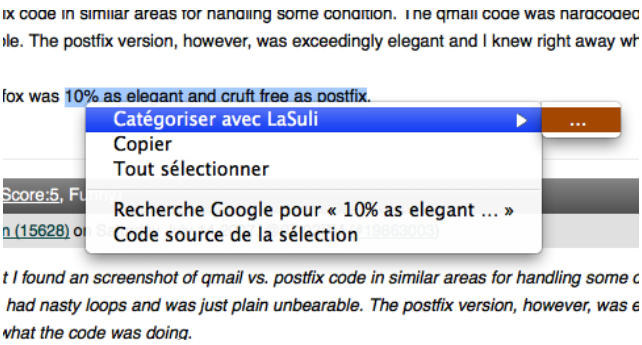

Figure 3. *Catégorisation à l'aide d'une rubrique émergente (copie d'écran de* LaSuli*)*

En analyse de textes, s'il est évident que le chercheur aborde toujours un matériau empirique avec une question de départ, un grand nombre d'éléments de conceptualisation apparaissent néanmoins au cours de la dynamique d'analyse. Dans un mouvement *prospectif*, de nouvelles hypothèses de travail, émergeant de la lecture, peuvent en éclairer la suite. Dans un mouvement *rétrospectif*, ces nouvelles hypothèses peuvent amener à reconsidérer des parties de textes déjà parcourues (Lejeune, 2008b).

Par ailleurs, plutôt que de limiter cette lecture expérimentale à des tableaux de comptage (ce qui est une manière curieuse de faire du qualitatif et qui de plus limite les comparaisons à deux dimensions d'analyse à la fois), nous avons développé dans *Porphyry* un dispositif analogue à la recherche de documents indexés selon des facettes (Allen, 1995), mais permettant une navigation à travers des fragments par ajout de contraintes sur des catégories (Bénel *et al.*, 2000; Zhou *et al.*, 2008).

3. Au sens de (Frécon, 2002).

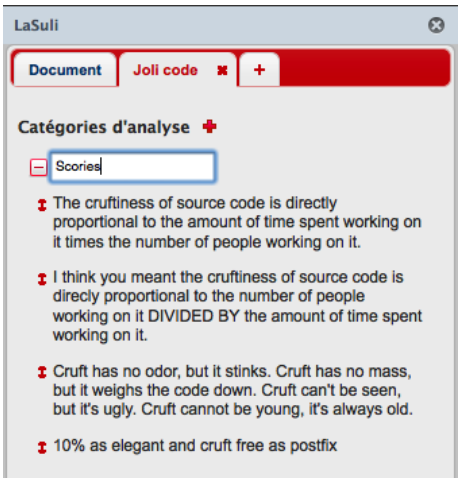

Figure 4. *Attribution d'un nom à une rubrique émergente (copie d'écran de* LaSuli*)*

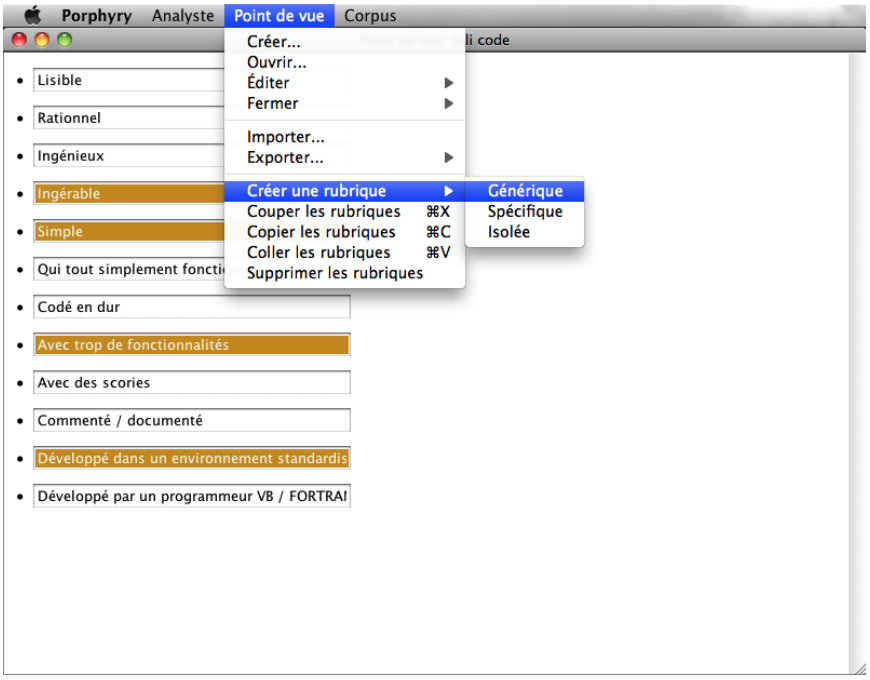

Figure 5. *Création d'une rubrique générique (copie d'écran de* Porphyry*)*

Enfin, à la suite des travaux initiés par Ioannis Kanellos (Tanguy, 1998), on peut rattacher le « codage » des textes au repérage des « isotopies » en sémiotique. Même s'il est avéré qu'il ne s'agissait pas d'un emprunt à la physique (Rastier, 1987), le concept d'isotopie chez Greimas évoque, comme en physique, une similarité permettant de voir une différence. Le carbone 14 est similaire au carbone « normal » par son nombre de protons mais diffère par son nombre de neutrons. De la même manière, au fil d'un texte, la réitération de catégories sémantiques est un indice permettant de rendre visibles les variations.

Par exemple, un sémioticien lisant le chapitre IV du livre de la Genèse sera sensible à la récurrence de la question de l'enfantement : « Adam connut Eve, sa femme ; elle conçut, et enfanta Caïn », « elle enfanta encore son frère Abel ». Cette récurrence introduit une différence radicale : Caïn est « l'homme formé avec l'aide de l'Éternel » tandis qu'Abel est « son frère ». Cette distinction permet alors une lecture renouvelée : l'humanité est du côté de Caïn et de sa descendance, humanité qui invente le meurtre du « frère », mais qui invente également la ville (Hénoc), la musique (Jubal), la technique (Tubal Caïn). Dans la suite du livre, il ne sera plus jamais question de cette descendance, ni donc de l'humanité, mais de la descendance de Seth « l'autre fils donné par Dieu à la place d'Abel, que Caïn a tué » (Calloud, 1998).

Dès lors, l'instrumentation de l'activité d'analyse se heurte à de sérieuses difficultés de modélisation... La différence peut-elle devenir similarité dans un autre contexte ? Est-elle propre à une similarité ? Est-elle une valeur ou le lien entre deux valeurs ? Existe-t-elle indépendamment d'un contexte ? Les théories en la matière sont si riches que les structures qui en résultent doivent, pour être visualisées, transcender parfois les dimensions de l'écran (Perlerin, 2004). Cette difficulté s'explique peut-être par la nature des langages de modélisation et de programmation. L'instanciation d'une classe est de l'ordre de l'*inférence* et non de la *différence*. La singularité est hors d'atteinte d'un dispositif au service de la reproduction du même.

L'importation de la notion piagétienne de « collection figurale » en informatique (Rousseaux *et al.*, 2007) nous amène à reconsidérer le problème de cette modélisation sous un nouveau jour. En effet, l'une des rares voies qu'aurait l'informatique pour aborder la question du sens serait de chercher une analogie avec le cube que l'enfant voit tantôt comme figurant un oiseau ou une plante, tantôt comme le coin d'un carré de cubes. La collection figurale se distingue en effet de la classe par la tension qu'elle instaure entre catégories et singularité. Dans notre cas, la solution consisterait peut-être à cantonner la modélisation à ce qui est de l'ordre de la similarité et à jouer sur la disposition spatiale pour que l'interprète puisse trouver lui-même des singularités, puisse suivre de différence en différence un parcours qui lui soit propre. C'est dans cet esprit que nous avons choisi, dans l'interface de *LaSuli*, de donner à voir les fragments regroupés par catégorie tout en respectant l'ordre de progression du texte (cf. figure 4).

2.2. *Annotation sociale*

La pratique scientifique repose sur la confrontation et le dialogue avec les pairs. Or la plupart des logiciels d'analyse qualitative n'envisagent cette dimension sociale que par des fonctions d'import/export. On pourrait attendre davantage de leur part sachant que la technique d'analyse la plus répandue repose précisément sur une annotation collective et contradictoire des textes. Cette technique, c'est l'*analyse de contenu*, telle qu'elle fût élaborée par Harold Lasswell et ses collaborateurs dans la première moitié du XX^{e} siècle (sous le nom de « sémantique quantitative ») puis formalisée par Bernard Berelson (1952). Cette technique prévoit que les catégories d'analyse soient stabilisées très tôt dans le processus de recherche (en amont de l'analyse proprement dite). Elles sont définies de manière univoque dans un « codebook ». Cette grille de codage est ensuite confiée à une équipe avec la mission d'identifier les passages correspondant aux catégories définies par le chercheur. Des consignes strictes limitent leur tâche au repérage et au codage, afin de proscrire toute interprétation subjective. Des techniques quantitatives (mesure de l'accord inter-juge) permettent d'évacuer la variabilité existant entre les différents codeurs. L'analyse de contenu vise ainsi une objectivité, définie par l'effacement de la subjectivité des codeurs. Si elle est collective, l'annotation qu'opèrent les codeurs de l'analyse de contenu n'est pas pour autant sociale : chacun travaille sans jamais interagir avec les autres.

Notre infrastructure (Zacklad *et al.*, 2007) permet au contraire la construction collective du cadre d'analyse dans le temps, ainsi que la confrontation sur un même texte (ou corpus de textes) des cadres d'analyse de différents interprètes.

L'idée d'une technologie permettant l'annotation sociale n'est pas nouvelle : elle est au cœur du projet originel de l'hypertexte (Nelson, 1980). Absente du web, il ne faudra cependant pas longtemps avant qu'une solution technique ne soit trouvée pour annoter à plusieurs n'importe quelle page web sans avoir à la modifier (Röscheisen *et al.*, 1995) : les annotations seront stockées sur un serveur à part et intégrées à la demande par le navigateur lui-même ou par procuration à un serveur tiers (« proxy »). La modification d'un navigateur n'étant pas à la portée de tous (Kahan *et al.*, 2001), les premiers prototypes vont plutôt s'orienter vers la solution du « proxy » (Röscheisen *et al.*, 1995; Yee, 1998). L'avènement de technologies permettant de développer des « extensions » aux navigateurs va donner naissance à *Annozilla* de Matthew Wilson (Lortal *et al.*, 2005), premier d'une génération de logiciels dont *LaSuli* fait partie. Plus qu'une simple solution technique, cela signifie que l'annotation sociale vient rejoindre l'usager au cœur de son logiciel de lecture numérique (cf. figure 6).

Pour créer des fragments sans modifier le texte initial, le concepteur de logiciel doit choisir une manière de les référencer. On peut en effet référencer un fragment de texte :

– par ses coordonnées (indices du premier et dernier caractère), complétées par son contenu textuel (Yee, 1998),

– par son contenu textuel et par l'identifiant d'un élément structurel (XML) le contenant (Kahan *et al.*, 2001),

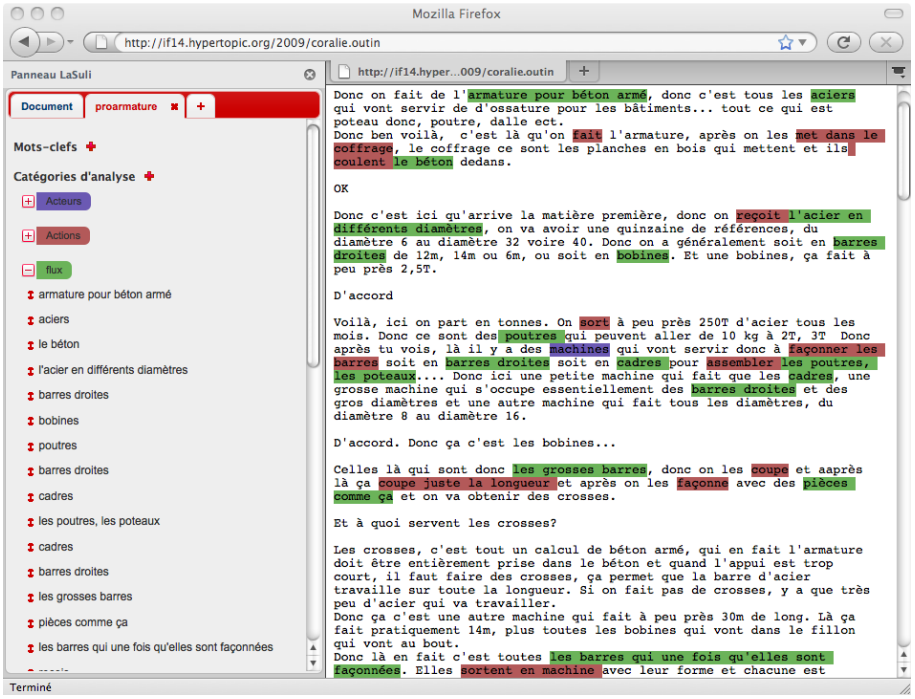

Figure 6. *Une marge pour Firefox (copie d'écran de LaSuli)*

– par son contenu textuel et par le numéro d'ordre de cette occurrence dans le texte (Denoue *et al.*, 2000).

Dans *LaSuli*, nous avons choisi la première solution car elle permet de calculer efficacement le croisement de catégories sur des fragments superposés. La redondance d'information qu'introduit le contenu textuel permet en outre de parer à des modifications légères du texte initial. Cette référence est cependant moins stable en cas de révision que celle de la troisième solution. Si le texte à analyser est susceptible d'être révisé, il est donc préférable de travailler sur une copie.

Un deuxième choix de conception qui distingue *LaSuli* concerne la gestion du « social » dans la visualisation. Donner à voir toutes les analyses sur une même page relève du cauchemar cognitif. Partant de la complémentarité entre la « sagesse des foules » du web 2.0 (O'Reilly, 2005) et notre notion de « point de vue d'expert », nous avons distingué dans notre interface :

– un onglet principal (cf. figure 7) agrégeant tous les points de vue en un nuage de thèmes (catégories de même nom) et signalant dans le texte de manière indifférenciée tous les passages analysés,

– autant d'onglets que de points de vue, pouvant être ouverts à la demande à partir d'une de leurs catégories ou de l'un de leurs auteurs.

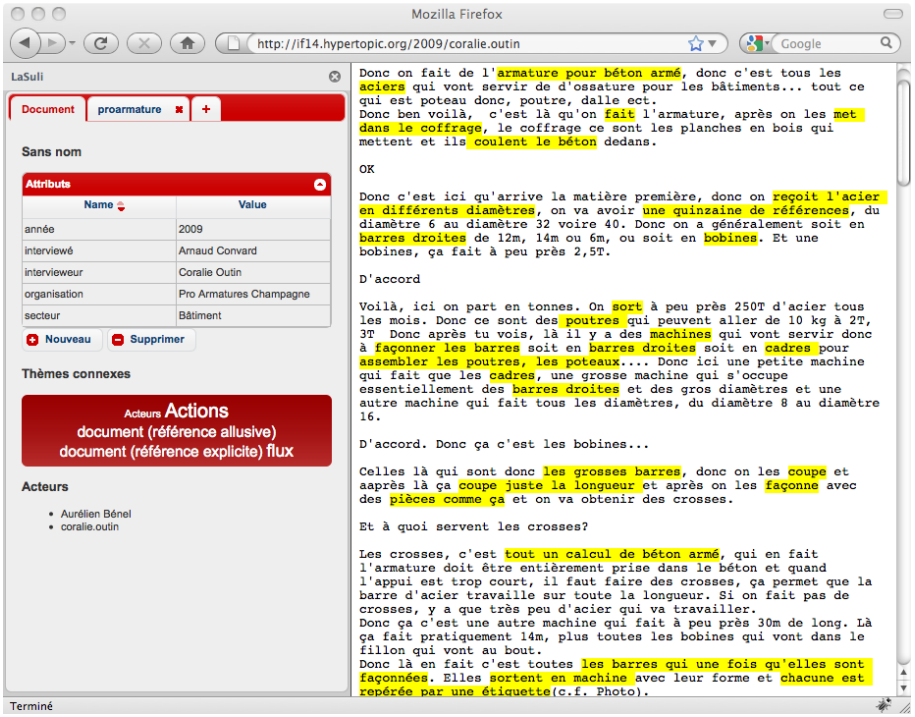

Figure 7. *Différentes analyses et indices de situation (copie d'écran de LaSuli)*

Par ailleurs, l'onglet principal permet de saisir des indices sur la situation de production du texte. Nous sommes conscients que les méthodes comme la sémiotique, issues de l'analyse structurale, interdisent d'ordinaire à l'analyste de prendre en compte le paratexte. Il nous semble cependant nécessaire, dès lors que le texte devient « preuve » dans un débat contradictoire (Bénel *et al.*, 2009), d'assurer sa conservation et son accès à l'aide de ces indices, c'est-à-dire de le constituer comme « document » (Briet, 1951).

3. Degré d'automatisation

3.1. *Registres*

La conception de l'informatique comme changement de support des activités humaines est assez récente. Au début, l'ordinateur est vu comme le moyen d'automatiser des tâches « manuelles ». Le travail de codage de l'analyse de contenu ne fait pas ex-

ception. En effet, lorsqu'elle est dénuée d'interprétation (comme le prône l'analyse de contenu), l'activité de codage se révèle vite fastidieuse voire aliénante.

Dans les années soixante, l'équipe du psychologue Philip Stone met au point, en collaboration avec IBM, le logiciel *General Inquirer*. Il s'agit d'automatiser le codage d'un fragment de texte en se basant sur la présence d'un terme ou d'une expression. Une fois que ces « marqueurs » sont énumérés par catégorie dans des « dictionnaires », réutilisables d'une étude à l'autre, le codage ne requiert plus aucune intervention humaine.

S'il peut s'avérer productif de déléguer à une procédure informatique la recherche de toutes les occurrences d'une expression, une analyse à visée interprétative ne peut se satisfaire de ces « dictionnaires ». Il ne peut pas être question de figer une fois pour toute le cadre d'analyse. La nature univoque et, par conséquent, consensuelle qui garantit l'objectivité des dictionnaires ne correspond pas non plus à la construction des interprétations d'un matériau qualitatif. Ces dernières sont en effet nécessairement situées par le moment, le paradigme, l'orientation théorique et la question de recherche des analystes.

Dans une approche qualitative, la structure d'analyse de textes doit tout d'abord pouvoir évoluer processuellement afin d'accompagner l'analyse du matériau empirique. L'indexation ne peut donc avoir lieu une fois pour toutes : elle doit être modifiée « en temps réel », à chaque modification de la structure par l'analyste. Ensuite, l'outil doit autoriser de multiplier les structures d'analyse de textes. Chaque analyste doit être en mesure de construire son propre cadre d'analyse ; un même analyste peut même produire plusieurs structures alternatives, tout en gardant la possibilité que plusieurs analystes s'accordent entre eux pour partager une seule et même structure d'analyse. À l'idéal d'objectivité, se substitue donc une subjectivité assumée (Paillé, 2006), voire une intersubjectivité raisonnablement discutée (Paillé *et al.*, 2003).

Quand il utilise l'« auto-coding » (*Atlas.ti*) des logiciels de la deuxième famille (Lewins *et al.*, 2007), l'analyste de texte, contrairement à l'analyste de contenu, n'établit pas un « dictionnaire » une fois pour toutes, mais construit des « registres » au fil de l'interprétation (Lejeune, 2008a).

Prenons un exemple réel. Une équipe de l'agence wallonne des télécommunications (AWT) s'intéresse à la façon dont les entreprises de la région wallonne actives dans le secteur des nouvelles technologies de l'information et de la communication se présentent et se positionnent face à la concurrence. L'équipe vise, notamment, à identifier une « grammaire » de la présentation de ces sociétés sur Internet. À cette fin, elle a réuni le corpus des sites web de ces sociétés. Ce corpus se compose de près de 1 000 sites web d'entreprises actives sur un territoire relativement restreint.

Un membre de l'équipe entame la lecture de quelques unes des pages d'accueil des sites en question. Son attention est attirée par des phrases comme « Nous proposons une large gamme de matériel informatique » ou « Notre société offre des services de consultance aux agences et professionnels du web ». Toutefois, contrairement à l'annotation classique, décrite dans les sections précédentes, l'analyste ne surligne pas

l'ensemble de ces énoncés. Au moyen de la souris, il sélectionne plutôt les mots clefs qui lui paraissaient pertinents pour son étude. Intéressé par un catalogue d'activités du secteur, il sélectionne l'expression « matériel informatique » et la copie dans *Porphyry* comme catégorie d'analyse. Transmise au serveur *Cassandre*, l'expression est traitée comme marqueur : tous les passages du corpus qui comportent l'expression « matériel informatique » sont dès lors, automatiquement et instantanément, attachés au marqueur en question. *Porphyry* peut alors afficher le résultat produit par *Cassandre* : l'ensemble de ces passages rassemblés l'un en dessous de l'autre. Il offre ainsi une vue d'ensemble (que les analystes de textes appellent « concordancier ») et ouvre une première piste, celle du positionnement des sociétés étudiées concernant le matériel informatique. L'analyste crée alors, de la même manière, une série d'autres marqueurs, comme « ordinateur », « écran », « imprimante ». Il rassemble ensuite tous ces différents marqueurs sous une même étiquette, qu'il appelle « IT Hardware ». Toutefois, poursuivant son investigation, il découvre des passages mentionnant d'autres produits, mais ne relevant pas de l'informatique. Qu'importe, notre analyste décide qu'il s'agit du même registre. Il sélectionne et glisse dans son premier registre les marqueurs « capteurs », « vannes » et « variateurs ». Afin de mieux rendre compte des marqueurs que rassemble ce registre, il en modifie l'étiquette et la rebaptise, de manière plus large, comme registre des « produits » (cf. figure 8).

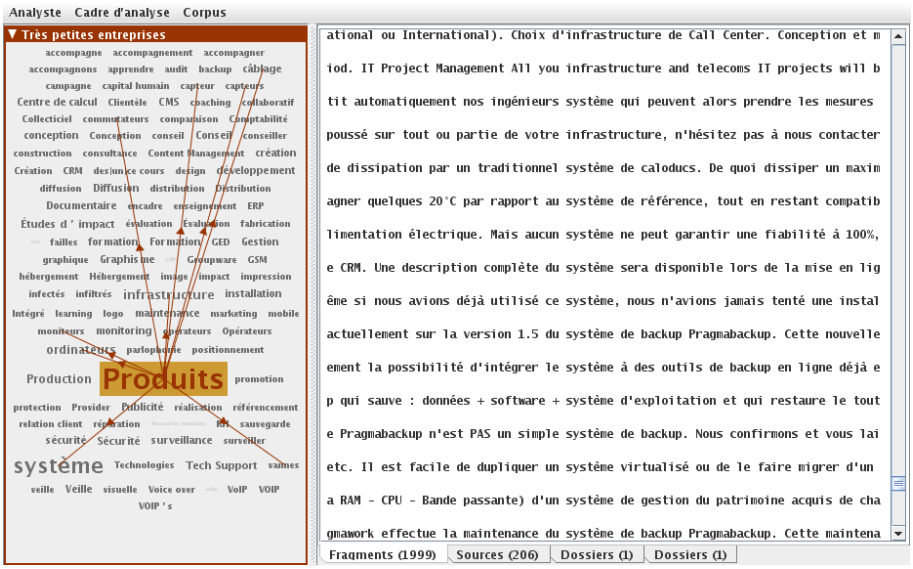

Figure 8. *Visualisation d'un registre de Cassandre (copie d'écran de Porphyry)*

Parallèlement à ce registre des produits, il en crée un deuxième, celui des services. Afin de faciliter son recensement des produits et services, il crée un registre de l'énumération qui comporte les marqueurs « gamme » et « palette ». Cette création, il l'opère de manière réfléchie : grâce au mode de visualisation synthétique du concor-

dancier (*Porphyry*), il vérifie que chaque occurrence de « palette » dans le corpus renvoie bien à une énumération (et non à une palette en bois ou à une palettes graphique). Une telle vérification nécessite une intervention humaine réfléchie. L'interprète humain garde donc toujours la main sur la création et la mise à jour des marqueurs ; elles ne peuvent être automatisées. En conséquence, la pertinence des marqueurs est toujours limitée à un corpus particulier (Lejeune, 2010a). Cela n'empêche pas qu'une structure d'annotation soit mobilisée pour un autre corpus mais cela requiert, nécessairement, que l'usager évalue la pertinence des marqueurs transposés.

Ainsi, à la différence des « dictionnaires », les registres entendent, grâce à leur plasticité, encourager et assister l'élaboration progressive d'interprétations scientifiques congruentes avec les documents analysés. Les registres, à condition qu'ils ne soient pas détournés en tant que « dictionnaires » (ce qui est un usage attesté (Lejeune, 2010a)), permettent donc de s'inscrire dans une perspective qualitative et interprétative.

Les structures d'analyse de texte par annotations et par registres se veulent complémentaires. Reposant sur un protocole informatique commun (Zacklad *et al.*, 2007), nos outils permettent de convoquer ces deux types de structures autour du même texte.

3.2. *Indices lexicométriques*

La famille des logiciels basés sur des statistiques regroupe en fait une grande diversité de modèles : statistiques descriptives, analyse des correspondances, chaînes de Markov, cooccurrences... Des outils comme *Alceste* ou *Candide*, par exemple, fournissent des agrégats de mots susceptibles d'aider le parcours dans de larges corpus textuels. La production de ces informations dites « lexicométriques » est totalement automatisée. Dénuée de la moindre intervention humaine, elle s'inscrit dans la visée « objective » de l'analyse de contenu. Pour autant, ces indices lexicaux ne peuvent se substituer au travail intellectuel du chercheur. C'est d'ailleurs la posture que défend Max Reinert (1990), le concepteur du logiciel de cette famille le plus souvent cité en France (*Alceste*) : seul le chercheur est à même de sélectionner les informations pertinentes dans le flot statistique et de leur donner une signification (c'est-à-dire de les interpréter).

Dédiés aux approches qualitatives, nos outils sont assez éloignés des générateurs automatiques d'indices lexicométriques de la première famille. Il nous est cependant apparu que de telles informations lexicales étaient, dans certains cas, susceptibles de jouer un rôle. Aussi le serveur *Cassandre* calcule-t-il, pour chaque source textuelle, des propriétés formelles ; il s'agit de la liste des dix mots les plus fréquents, de dix mots spécifiques au texte en question (identifiés au moyen du TF*IDF de Salton), des dix mots n'apparaissant qu'une fois dans le corpus (*hapax*) et de dix locutions (bigrammes).

À l'opposé des registres, les indices lexicométriques ne sont ni subjectifs ni interprétatifs. Une fois calculés par *Cassandre*, ils apparaissent donc dans *LaSuli* comme

attributs du document. Pour autant, contrairement aux analystes de contenu, nous ne considérons pas ces informations comme plus objectives (au sens de « plus fiables »). Elles ne constituent en rien une « analyse automatique » mais plutôt une invitation à la lecture attentive. Leur raison d'être est heuristique : une saillance quantitative peut suggérer à l'analyste l'ouverture d'une investigation.

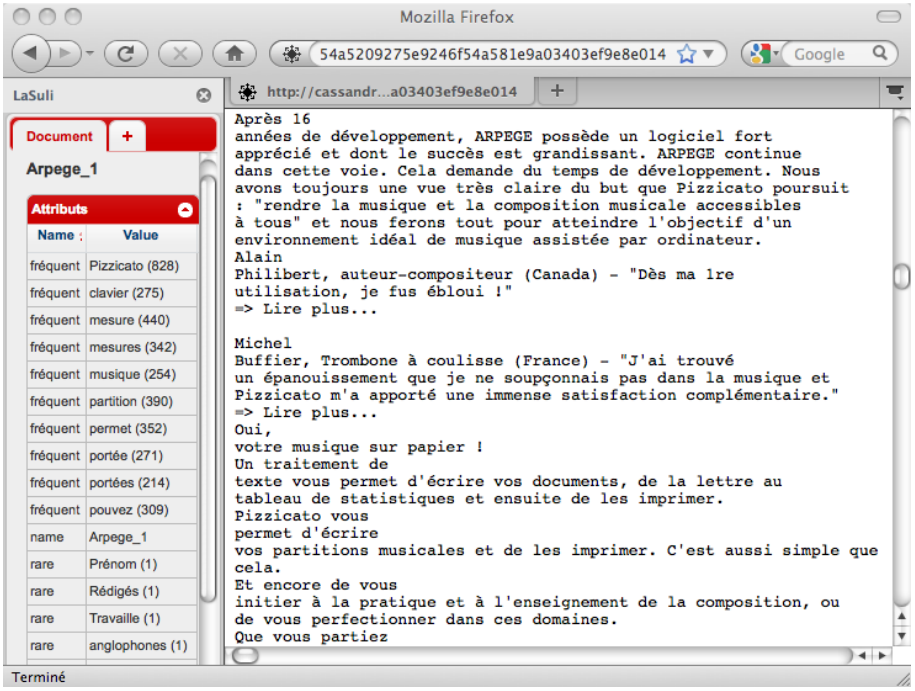

Figure 9. *Quelques propriétés lexicométriques (copie d'écran de LaSuli)*

Dans l'étude de l'agence wallonne des télécommunications, la fréquence affichée de termes comme « management », « information » ou « communication » n'a pas étonné les analystes. Ces informations, jugées triviales, sont constitutives de la composition du corpus. Par contre, l'attention d'un analyste est attirée par la fréquence, dans les sites web de très petites entreprises, de termes comme « permet » ou « pouvez ». Cette singularité statistique suggère à l'analyste une particularité des petites sociétés, qui s'adressent de manière plus directe à leur client que les multinationales. Il formule alors l'hypothèse selon laquelle les sociétés ne vantent pas leurs services de la même manière selon leur taille. Afin d'éprouver cette hypothèse, il recense les termes qu'utilisent les sociétés pour présenter ce qu'elles offrent sans, bien sûr, se limiter au procédé argumentatif mettant en scène le client potentiel projeté dans sa souscription à cette offre. Les termes identifiés sont copiés comme des marqueurs dans l'espace dédié aux annotations heuristiques. Ces marqueurs ont été rassemblés dans le registre dit « de l'offre ». Celui-ci regroupe tous les marqueurs au moyen desquels les sociétés

se mettent en scène comme pourvoyeurs de solutions, de produits ou de services. Ce registre s'alimente au fur et à mesure de l'exploration du corpus de marqueurs comme « proposons » ou « maîtrisons ». La création de ce registre a ceci de particulier qu'elle a été inspirée non par la lecture linéaire du texte, mais pas une propriété formelle, statistique, à savoir la récurrence de certaines formes lexicales dans une partie du corpus (Gobin *et al.*, 1994, p. 69-70).

4. Conclusion

Autour d'*Argos*, un service généraliste de gestion de systèmes d'organisation des connaissances, nous avons construit une architecture en réseau (*LaSuli*, *Porphyry* et *Cassandre*) permettant l'élaboration, la comparaison et la confrontation de structures hétérogènes d'analyse de textes. Dans la fenêtre de lecture du navigateur, *LaSuli* joue le rôle de marge ou d'espace d'annotation. Tout au long de ses lectures, le lecteur-interprète y consigne les annotations soutenant son analyse.

Porphyry joue le rôle de portfolio du chercheur, il lui permet de faire des allers-retours entre le fragment, le document et le corpus. C'est avec ce logiciel que le chercheur structure son cadre d'analyse, le met à l'épreuve du corpus et le compare aux autres cadres d'analyse.

Cassandre permet de remplacer *Argos* lorsque les « systèmes d'organisation des connaissances » sont produits à partir de textes de manière automatique (indices lexicométriques) ou semi-automatique (registres).

Notre infrastructure permet ainsi l'élaboration et la comparaison de structures d'annotation et de marqueurs, éventuellement inspirés d'indices lexicométriques. Nous sommes persuadés que la lecture expérimentale des textes profite grandement de l'hétérogénéité des perspectives subjectives et de celle des modes d'automatisation qui concourent à la construction de ces structures. Pour autant, nous devons reconnaître qu'aujourd'hui nos outils sont majoritairement utilisés par des chercheurs « solitaires » ne construisant généralement qu'une seule grille d'analyse par corpus. On pourrait en conclure que la constitution d'archives scientifiques numériques, le travail en équipe et l'interdisciplinarité ne sont pas des « besoins » en sciences de l'homme et de la société. Mais comment interpréter alors leur omniprésence dans les appels d'offres scientifiques ? N'y a-t-il pas en effet, avec la transition au numérique, une occasion historique de redécouvrir et d'appliquer les principes de l'herméneutique philosophique ?

Remerciements

Ces travaux ont été en partie financés par l'agence nationale de la recherche (ANR) dans le cadre du projet Miipa-Doc n°2008 CORD 014 03 ainsi que par l'agence wallonne des télécommunications (AWT) dans le cadre du projet Vigie n°19104.

5. Bibliographie

Allen R. B., « Retrieval from facet spaces », *Electronic Publishing*, n° 8, p. 247-257, 1995.

Bénel A., Calabretto S., Pinon J.-M., Iacovella A., « Vers un outil documentaire unifié pour les chercheurs en archéologie », *Actes du 18e congrès INFORSID*, Éditions INFORSID, Lyon, p. 133-145, 16-19 mai, 2000.

Bénel A., Lejeune C., « Partager des corpus et leurs analyses à l'heure du Web 2.0 », *Degrés, revue de synthèse à orientation sémiologique*, vol. 136-137, p. m1-20, 2009.

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

Briet S., *Qu'est-ce que la documentation ?*, Editions documentaires, industrielles et techniques, Paris, 1951.

Bryman A., *Social Research Methods*, Oxford University Press, 2008.

Calloud J., « Caïn et Abel : L'homme et son frère », *Sémiotique et Bible*, n° 92, p. 3-34, 1998.

Denoue L., Vignollet L., « An annotation tool for web browsers and its applications to information retrieval », *Content-based multimedia information access (RIAO'2000)*, CID-CASIS, p. 180-195, 2000.

Frécon L., *Éléments de mathématiques discrètes*, PPUR, Lausanne, 2002.

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

Gobin C., Deroubaix J.-C., « Quand la commission se présente devant le parlement », 1994.

Hodge G., *Systems of Knowledge Organization for Digital Libraries : Beyond Traditional Authority Files*, The Digital Library Federation, Council on Library and Information Resources, Washington, D.C., 2000.

Kahan J., Koivunen M.-R., Prud'Hommeaux E., Swick R. R., « Annotea : An open RDF infrastructure for shared web annotations », *Proceedings of the WWW10 International Conference*, Hong Kong, mai, 2001.

Lejeune C., Sociologie d'un annuaire de sites Internet : Les sciences documentaires saisies par l'informatique libre, Thèse de doctorat en sociologie, Université de Liège, 2004.

Lejeune C., « Petite histoire des ressources logicielles au service de la sociologie qualitative », *in* C. Brossaud, B. Reber (eds), *Humanités numériques*, vol. 1, Hermès, p. 197-214, 2007.

Lejeune C., « 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, 2008a.

Lejeune C., « Échecs, blogs et sentiments. La méthode sociologique à l'épreuve de la vie quotidienne », *in* M. Jacquemain, B. Frère (eds), *Épistémologie de la sociologie. Paradigmes pour le XXIe siècle*, De Boeck, p. 157-172, 2008b.

Lejeune C., « Cassandre, un outil pour construire, confronter et expliciter les interprétations », *2e Colloque International Francophone sur les Méthodes Qualitatives*, 2010a.

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

Lewins A., Silver C., *Using Software in Qualitative Research. A Step-by-Step Guide*, Sage, London, 2007.

Lortal G., Lewkowicz M., Todirascu-Courtier A., « Modélisation de l'activité d'annotation discursive pour la conception d'un collecticiel support à l'herméneutique », *in* M.-C. Jaulent

(ed.), *Actes des 16e journées francophones d'ingénierie des connaissances (IC'2005)*, PUG, Grenoble, p. 169-180, 2005.

Luhn H. P., « A statistical approach to mechanized encoding and searching of literary information », *IBM Journal*, vol. 1, n° 4, p. 309-317, 1957.

Nelson T., *Literary Machines*, Mindful Press, Sausalito, 1980.

O'Reilly T., What is Web 2.0 : design patterns and business models for the next generation of software, Journal personnel (blog), 30 septembre, 2005.

Paillé P., « Lumières et flammes autour de ma petite histoire de la recherche qualitative », *Recherches Qualitatives*, vol. 26, n° 1, p. 139-153, 2006.

Paillé P., Mucchielli A., *L'analyse qualitative en sciences humaines et sociales*, Armand Colin, Paris, 2003.

Perlerin V., « Sémantique légère pour le document : Assistance personnalisée pour l'accès au document et l'exploration de son contenu », *Texto !*, décembre, 2004.

Prié Y., Mille A., Pinon J.-M., « Une approche de modélisation de documents audiovisuels en strates interconnectées par les annotations », *Actes des 9e journées francophones d'ingénierie des connaissances (IC'1998)*, p. 143-152, 1998.

Rastier F., *Sémantique interprétative*, PUF, Paris, 1987.

Reinert M., « Alceste, une méthodologie d'analyse des données textuelles et une application, Aurélia de Gérard de Nerval », *Bulletin de Méthodologie Sociologique*, vol. 26, p. 24-54, 1990.

Rousseaux F., Bonardi A., « Parcourir et constituer nos collections numériques », *Actes du 10e colloque international sur le document électronique (CIDE'10)*, Europia, Paris, 2007.

Röscheisen M., Mogensen C., Winograd T., « Beyond browsing : Shared comments, SOAPs, trails, and on-line communities », *Computer Networks and ISDN Systems*, vol. 27, n° 6, p. 739-749, 1995.

Seidel J. V., Clark J. A., « The Ethnograph : A computer program for the analysis of qualitative data », *Qualitative Sociology*, 1984.

Stone P. J., Dunphy D. C., Smith M. S., Ogilvie D. M., *The General Inquirer : A Computer Approach to Content Analysis*, MIT Press, Cambridge, 1966.

Tanguy L., « Traitement automatique de la langue naturelle et Interprétation : Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétative », *Texto !*, mars, 1998.

Tasman P., « Literary data processing », *IBM Journal of Research and Development*, vol. 1, n° 3, p. 249-256, 1957.

Virbel J., « Annotation dynamique et lecture expérimentale : Vers une nouvelle glose ? », *Littérature*, n° 96, p. 91-105, 1994.

Yee K.-P., CritLink : Better Hyperlinks for the WWW, Preprint, University of Waterloo, 1998.

Zacklad M., Bénel A., Cahier J.-P., Zaher L., Lejeune C., Zhou C., « Hypertopic : une métasémiotique et un protocole pour le Web socio-sémantique », *in* F. Trichet (ed.), *Actes des 18e journées francophones d'ingénierie des connaissances (IC'2007)*, Cépaduès, p. 217-228, 2007.

Zhou C., Bénel A., « From the crowd to communities : New interfaces for social tagging », *Proceedings of the 8th international conference on the design of cooperative systems (COOP'08)*, Carry-le-Rouet, p. 242-250, 20-23 mai, 2008.