Socioinformatique des controverses

Outils socio-informatiques pour l'analyse des controverses

Génération automatique de textes

Samuel Szoniecky, Hakim Hachour
Université Paris VIII – Laboratoire Paragraphe
http://www.samszo.univ-paris8.fr/Seminiaire-EHESS-Socioinformatique

La génération automatique est intimement liée à l'usage de l'informatique. Nous l'expérimentons à la moindre frappe sur nos claviers, au moindre clic sur nos écrans. L'informatique nous permet de générer à partir d'un geste très simple : une lettre sur l'écran, une page web qui s'affiche, une existence unique qui se déploie (http://www.samszo.univ-paris8.fr/ChaoticumPapillonae). Le travail des informaticiens consiste à gérer ces multiplicités génératives pour que la touche « A » affiche bien un « A », pour que le lien sur lequel on clique affiche bien la page Web attendue. Dans cette multiplicité de rapports symboliques [8], on peut faire intervenir l'aléa pour générer une virtualité c'est à dire une potentialité de ce qui pourrait arriver [12]. Les générateurs automatiques se situent dans ce « milieu » associant le formalisme symbolique de l'informatique qui requiert unicité et hiérarchie du rapport à l'information, et la virtualité d'un réseau qui multiplie la potentialité des flux évènementiels. Entre cet « arbre-racine » et ce « rhizome-canal » [5, p. 31], ce joue l'expérimentation d'une connaissance « en train de se faire » et qui s'incarne, ici et maintenant, dans une existence informationnelle. Mettre à disposition des dispositifs de lecture et d'écriture de ses existences est sans doute un des enjeux majeurs pour les humanités numériques. A l'heure où l'e-Education envahie les systèmes scolaires et où le « numérique » devient une clef indispensable pour le financement de la recherche, les travaux sur les générateurs automatiques peuvent alimenter les débats sur la nature des contenus, comment leurs éléments traduisent des informations utiles, fiables et vérifiables. Ce rapport entre les contenus et les connaissances associées alimentent alors un débat de fond sur les usages de l'informatique en sciences humaines et de leurs conséquences épistémologiques et pédagogiques.

En 1997, j'expérimentais pour la première fois la génération automatique de texte dans le cadre d'une licence de philosophie à l'université Paris X. Influencé par mes recherches sur John Cage, je profitais de mon service civile à la bibliothèque universitaire de Nanterre pour explorer ce lieu de savoir en utilisant l'aléa comme moteur pour stimuler ma production de connaissances et « trouver une façon d'écrire qui, tout en partant d'idées, ne soit pas un discours sur elles : ou ne soit pas sur des idées mais les produise. » [3, cité par 13, p. 307]

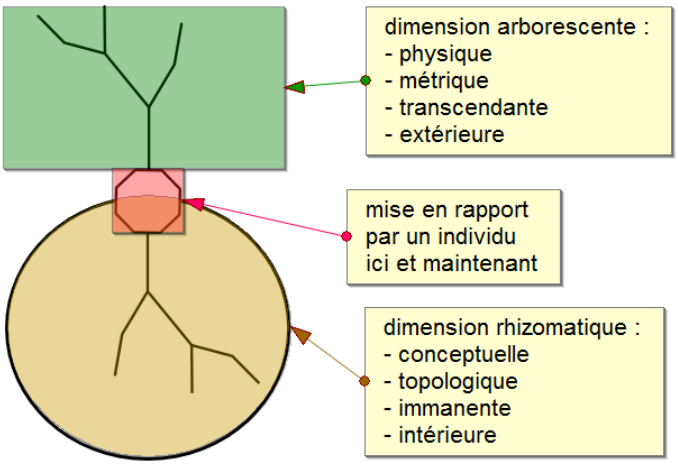
A l'aide de trois dès à dix faces permettant de tirer un nombre aléatoire entre 000 et 999 et d'une application Hypercard (http://fr.wikipedia.org/wiki/HyperCard) que j'ai développée pour l'occasion, j'organisais mes explorations du savoir et la production des devoirs qui étaient demandés par mes professeurs. Les résultats de cette expérience

furent très intéressants. Outre une meilleure connaissance du développement informatique et du catalogage d'une bibliothèque, je découvris que l'Université n'était absolument pas prête pour ce type de pratique qui selon Etienne Balibar relevait d'un « mutant ». Plus précisément, si l'exploration aléatoire d'un corpus à l'aide d'un algorithme et l'archivage dans une mémoire numérique des traces récoltées semblaient des pratiques prometteuses stimulant la sérendipité, en revanche, déléguer à l'aléa la production du discours et au lecteur la responsabilité de son interprétation apparaissaient « hors jeux » et « innotable » (cf. http://www.samszo.univ-paris8.fr/Experience-d-une-mutation).

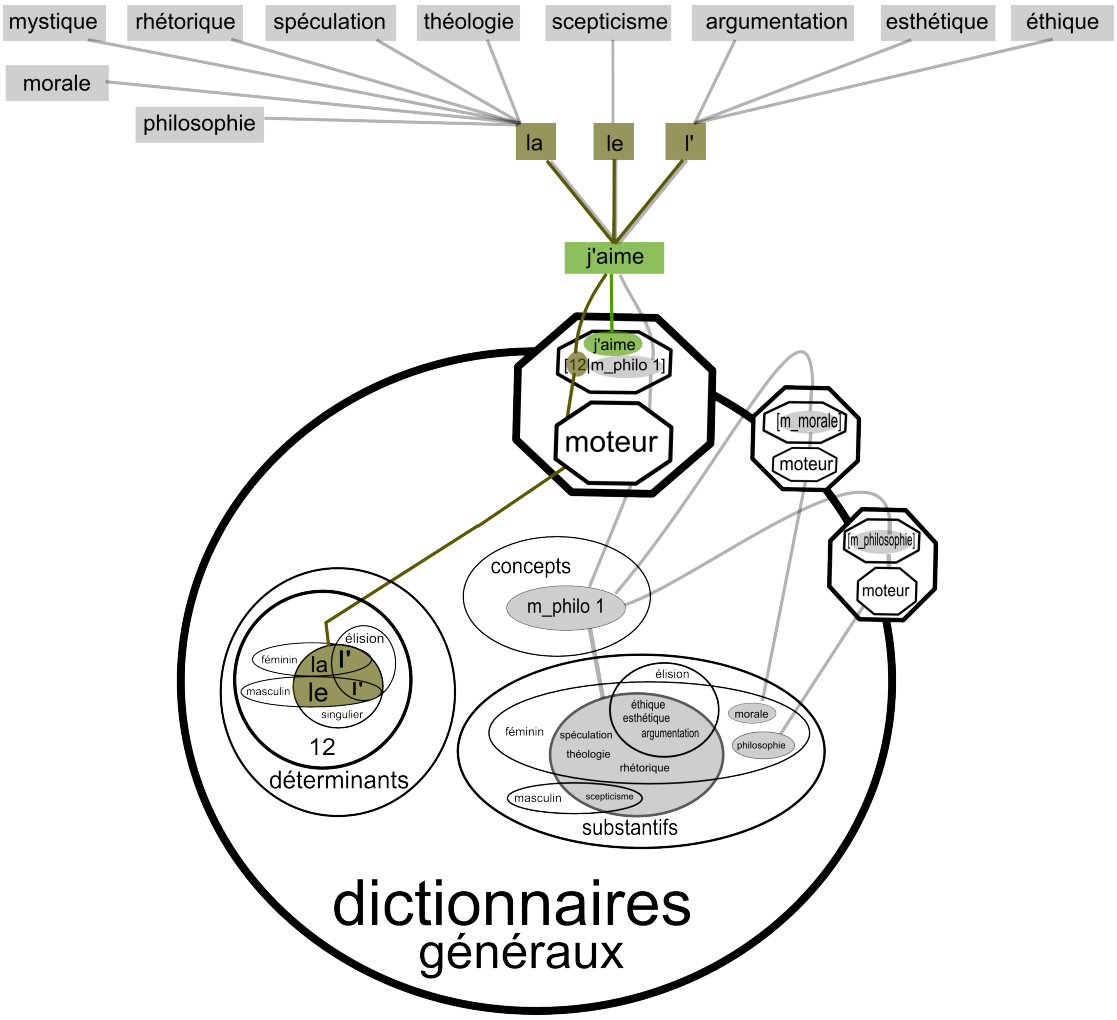
Ma rencontre avec Jean-Pierre Balpe et les travaux que nous avons menés, m'ont confirmé que le terrain de jeux des générateurs automatiques est d'avantage accepté lorsqu'il se situe dans le domaine de l'art plutôt que dans celui des sciences humaines. Notons toutefois, qu'il en est autrement dans le domaine des sciences « dures », particulièrement dans les voies tracées par la théorie du chaos [9], qui utilisent les générateurs ou simulateurs pour constituer et/ou explorer des corpus de recherche. Je ne développerais pas en détail ce point sur lequel il y a beaucoup de chose à dire, précisons seulement que la pertinence de l'aléa dans la production du discours est un débat ancien dont on trouve des traces en psychologie dans les travaux de Gustav Jung sur la synchronicité [11], en histoire dans les recherches de Jean-Pierre Vernant sur l'utilité social de l'aléa pour faire émerger le consensus [16] et bien sûr en art avec notamment la controverse en Pierre Boulez et John Cage [2].

Après avoir expérimenté différents modes de génération automatiques comme produire des hypertextes sur la base de pliages, faire éclore des papillons uniques à partir d'un flux aléatoire, créer des noosphères [4] selon une structure fractale basée sur des fullerènes…, je me suis concentré ces dernières années sur la modélisation de ces existences qui naissent des générateurs automatiques et qui vivent dans les écosystèmes d'information numériques.

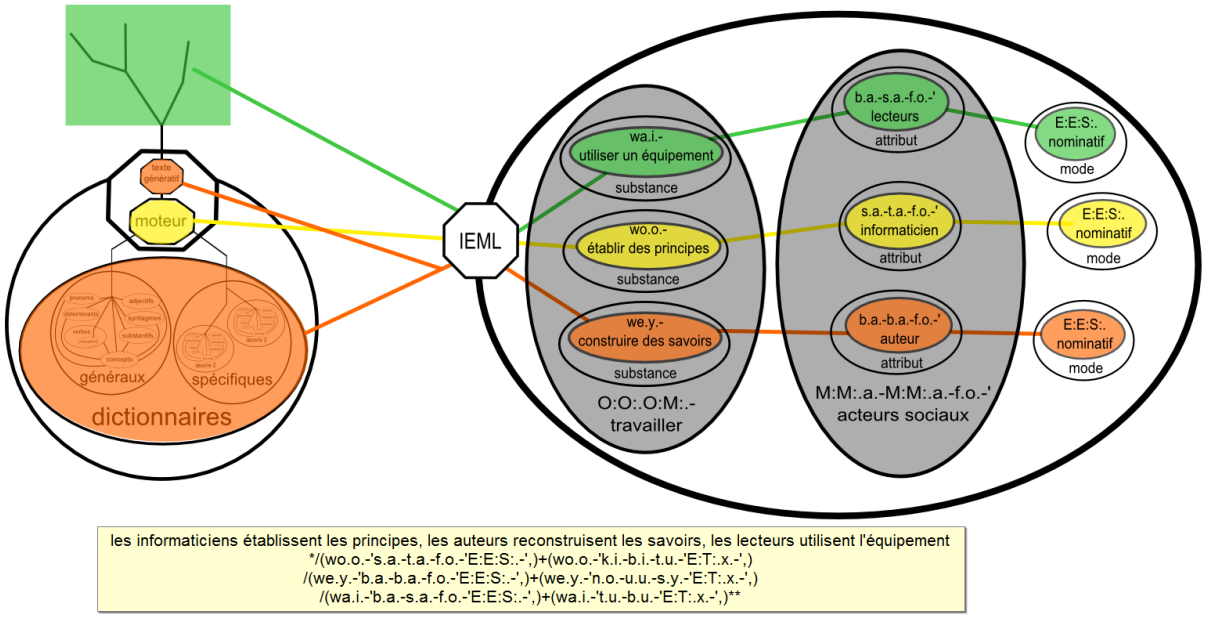
Face aux systèmes d'information dont personne aujourd'hui ne maîtrise l'intégralité des couches informatiques qui les composent, il est important de trouver des modèles génériques pour penser ses complexités, ses usages, ses développements, ses catastrophes… Le modèle que j'utilise est issu de mes recherches pour trouver une nouvelle forme d'interface Homme-Machine qui remplace l'analogie de la bureautique (bureau, dossier, fichier, corbeille…) ou celle de l'architecture [17] par celle du jardin (terroir sémantique, graine, branches, racines… [15]). Ces analogies amènent à définir chaque information comme une existence qui évolue dans un écosystème suivant l'instanciation, par des acteurs, d'une potentialité de rapports entre une dimension arborescente extérieure-physique et une dimension rhizomatique intérieure-conceptuelle. Inspiré à la fois par les propositions de Gilles Deleuze [6] pour « une ontologie corrélat d'une éthique », par les « matrices ontologiques » de Philippe Descola [7] et par la « raison trajective » défendue par Augustin Berque [1], ce modèle d'existence informationnelle se traduit par le diagramme suivant :



Dans le projet de générateur automatique de texte que nous avons mené en partenariat avec Jean-Pierre Balpe (http://www.balpe.name), Digitalarti (http://www.digitalarti.com) et le Labex Arts-H2H (http://www.labex-arts-h2h.fr/), la modélisation des existences informationnelles par un diagramme permet de décrire précisément le fonctionnement du générateur selon un point de vue particulier. A partir d'une phrase générative « J'aime [12|m_philo 1] » et des algorithmes du « moteur », ce générateur de rapports va créer une potentialité de textes en puisant dans les différents dictionnaires composant les dimensions conceptuels de cette existence [18].

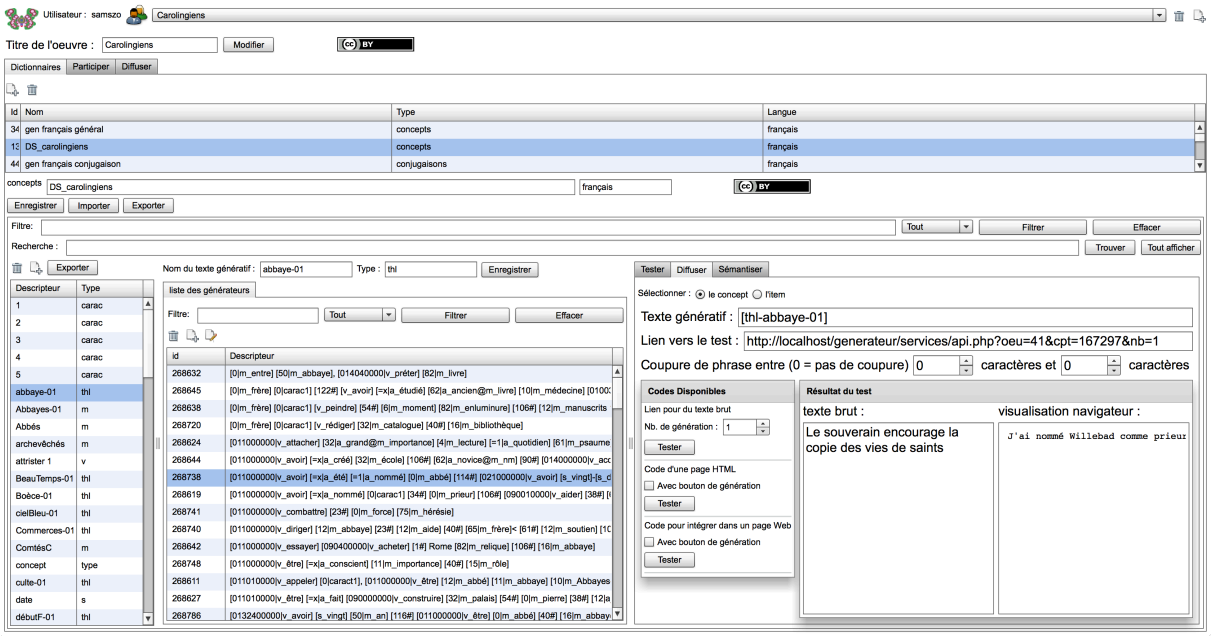


A partir de ce diagramme nous pouvons détaillés encore d'avantage les principes de fonctionnement du générateur en ajoutant une couche sémantique modélisée avec IEML [12] qui décrira, par exemple, le rôle de différents acteurs suivant leurs pouvoir d'agir dans chaque dimension existentielle.



les informaticiens établissent les principes, les auteurs reconstruisent les savoirs, les lecteurs utilisent l'équipement
*/(wo.o.-'s.a.-t.a.-f.o.-'E:E:S:.-',)+(wo.o.-'k.i.-b.i.-t.u.-'E:T:.x.-',)
/(we.y.-'b.a.-b.a.-f.o.-'E:E:S:.-',)+(we.y.-'n.o.-u.u.-s.y.-'E:T:.x.-',)
/(wa.i.-'b.a.-s.a.-f.o.-'E:E:S:.-',)+(wa.i.-'t.u.-b.u.-'E:T:.x.-',)**

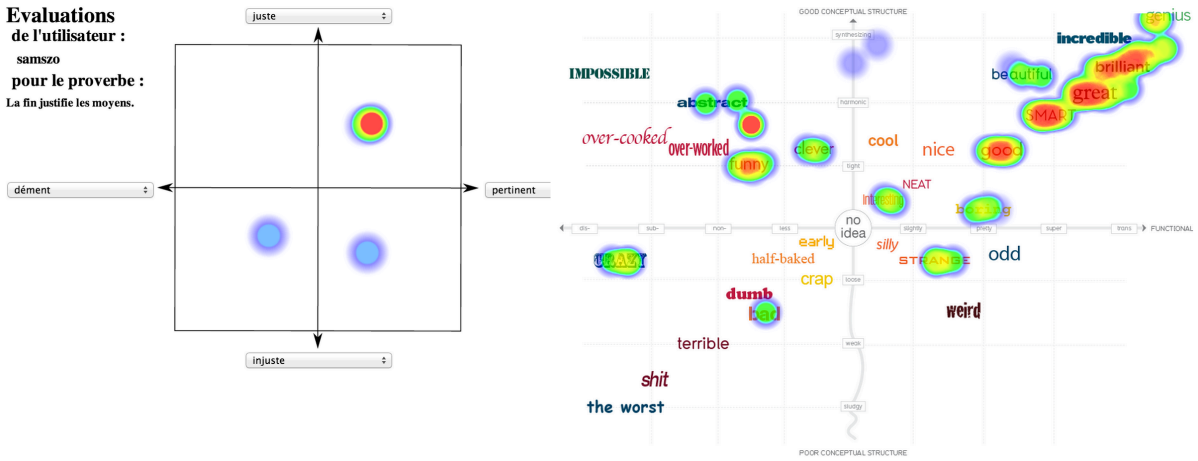
Dans sa version utilisée par Jean-Pierre Balpe, le générateur automatique de texte est dédié aux auteurs qui grâce à une interface d'administration vont créer des dictionnaires spécifiques pour explorer l'univers sémantique constituant leur œuvre littéraire. Ainsi le générateur est utilisé par exemple : pour générer des chansons, des biographies et des critiques dans l'univers du rock (http://chatonsky.net/project/capture), pour générer des descriptions de plantes sur le modèle des herbiers (http://creative.arte.tv/fr/community/herbarius-2059), pour générer des moments carolingiens (http://www.eriver.fr/2014/06/moments-carolingiens.html).



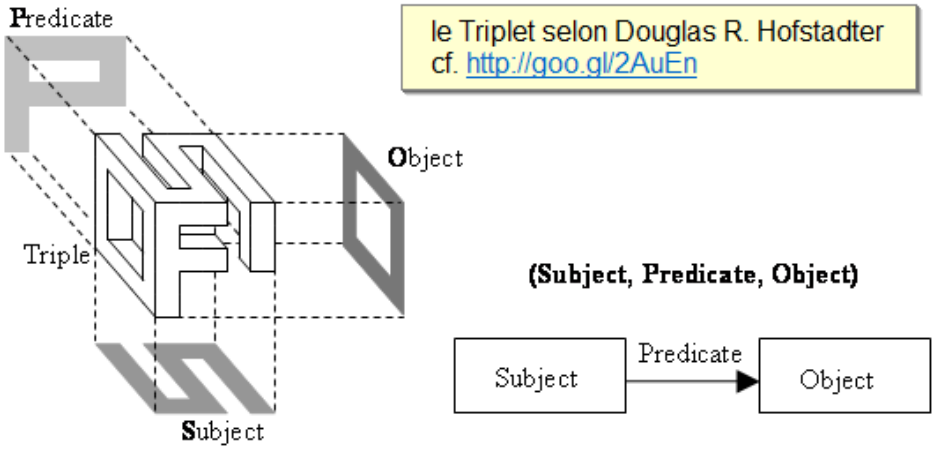
Dans le cadre du projet Gapaii soutenu par le Labex Art-H2H, nous expérimentons des usages du générateur dédiés à la recherche en sciences humaines. Par exemple, en collaboration avec Emmanuel Sander nous mettons en place un atelier laboratoire CreaTIC (http://idefi-creatic.net/ateliers-laboratoires) qui aura pour objectif de récolter des interprétations afin d'analyser dans quelles mesures les interprétants mettent en jeu des analogies spécifiques [10]. Le premier protocole de récolte consiste à faire évaluer par des groupes d'interprètes catégorisés par leur âge et leur langue maternelle, un corpus de proverbes provenant de différentes langues (français, wolof, arabe, tamazirth…). A travers cette expérience nous voulons savoir si la cohérence sémantique des proverbes est conservée lorsqu'on les soumet à des populations pratiquant une autre langue, d'étudier dans quelle mesure la cohérence est négociée dans un travail d'interprétation. Un deuxième protocole consiste à générer des proverbes à partir de structures logiques originales mais en changeant certains des termes. Par exemple dans l'expression « A bon chat, bon rat. », la forme logique peut se traduire par « bon A = bon B » ou plus précisément par « bon animal A = bon animal B », ou encore par « bon prédateur = bonne proie »... En changeant « chat » par « lion » et « rat » par « gazelle » est-ce que le proverbe restera valide pour l'interprétant ? Grâce à ces expériences, peut-on dégager une typologie des analogies suivant la persistance, la profondeur et le frontière de leur « pouvoir d'agir » ?

Ces protocoles utilisent des modes d'évaluation suivant une échelle de difficulté allant du plus matériel au plus abstrait. La première évaluation porte sur le discernement de la forme physique du proverbe : Est-ce les lettres sont bien formées ? Est-ce l'alphabet est connu ? Est-ce que l'orthographe et la grammaire sont respectées ?... Les réponses à ces évaluations sont booléennes (vrai, faux) soit globalement au niveau du proverbe soit au niveau des parties du proverbe : telle phrase, tel mot, telle lettre. La deuxième évaluation consiste à discerner la cohérence sémantique du proverbe. Ces évaluations se font cette fois à partir de cartes conceptuelles [19] qui permettent par un simple clic de catégoriser de façon simple mais très précise le point de vue d'un interprète par rapport à un proverbe ou une partie de ce proverbe. Grâce à une grille sémantique codée avec IEML et superposée à l'image de la carte, le clic crée un rapport entre la forme physique du proverbe et l'espace conceptuel auquel appartient cette forme selon la procédure d'interprétation. Ce dispositif donne pour un interprétant, un positionnement quantifiable (x, y) dans l'espace conceptuel projeté par la grille sémantique sur l'image. On peut alors calculer des distances conceptuelles entre les proverbes, les interprètes et les concepts de la grille. Notons que ces distances conceptuelles non de sens qu'en rapport avec la projection spécifiée par la grille sémantique. Celle-ci est le résultat d'une interprétation particulière de la carte, on peut donc la discuter mais aussi et surtout faire émerger, grâce à ces discussions, un consensus sur les espaces conceptuels que le dispositif d'interprétation met en jeu.

EVALUATIONS DU PROVERBE



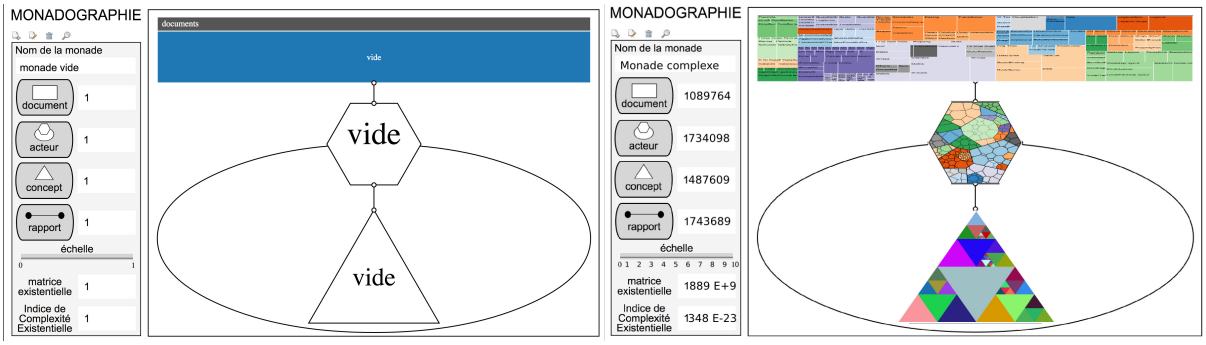
Le troisième mode d'évaluation porte cette fois sur le discernement des structures analogiques contenues dans le proverbe. On demande ici à l'interpréte de décomposer le proverbe en un ensemble de catégories pour lesquelles il doit trouver des éléments analogues. Le travail d'interprétation consiste à définir les réseaux sémantiques qui sont en jeu dans le proverbe sur la base d'un triplet : sujet – prédicat – objet.



Pour reprendre l'exemple « A bon chat, bon rat » vu plus haut, ce proverbe pourra être décomposé ainsi : chat – est – animal, chat – est – bon, chat – est – prédateur, rat – est – animal, rat – est – proie, rat – est – bon… Parallèlement à cette modélisation en catégories, l'interprète doit définir les relations logiques qui existent dans le proverbe afin d'expliciter un raisonnement, par exemple : SI chat-est-bon ALORS rat-est-bon, SI prédateur-est-bon ALORS proie-est-bonne. Pour faciliter ce travail de catégorisation et de définition des structures logiques, un outil graphique sera mis à la disposition de l'interprétant pour lui proposer des listes de catégories et de structures logiques. Pour rendre interopérable les choix de chaque utilisateur, ceux-ci seront écrits en RDF et sémantisés avec le IEML. Un premier exemple de ce type d'outil est en cours de développement dans le cadre du projet ANR Biolographe (http://biolog.hypotheses.org). Une fois les modélisations effectuées, celles-ci sont utilisées pour générer de nouveaux proverbes et les soumettre à d'autres interprétants.

A partir de ces expérimentations, nous pourrons valider la faisabilité d'un dispositif pour la capture des interprétations. L'enjeu principal est de pouvoir calculer à partir des

enregistrements effectués, le profil d'interprétation d'un acteur individuel ou collectif en modélisant son pouvoir de discernement dans les trois dimensions de l'existence que nous avons défini : physique – rapport – concept. Nous travaillons à la représentation de ce profil sous la forme d'une monade [14] qui donne l'étendue existentielle du profil et son niveau de complexité.



Cette capacité de capturer le pouvoir de discernement et son étendue grâce aux générateurs automatiques est déjà à l'œuvre dans des dispositifs comme Facebook qui récolte des multitudes de « j'aime » sur des contenus dont il génère le flux. Nous avons montré ici que la modélisation du discernement et la génération des flux d'information qui en découle peuvent être maîtrisé de façon beaucoup plus précise notamment dans de ressources pour l'e-Education. Ces données récoltées par ces ressources pédagogiques numériques constituent des « Big Data » bien plus puissantes que celles qui existent aujourd'hui car elles sont catégorisées très finement dès l'expression même de l'interprétation. Plus besoin d'inventer des algorithmes très complexes pour prédire une sémantique à partir des données, celle-ci serait déjà codée et à disposition sous la forme de monades qui modélisent ce pouvoir de discernement que Gille Deleuze définissait comme étant l'âme. Que faire alors de toute ces « monades » enregistrées tout au long de la scolarité de nos enfants ? Faut-il laisser « libre » l'économie de ses « âmes numériques » ? Ne forment-elles pas une richesse qu'il faudrait capitaliser dans des banques nationales, européennes, mondiales ?

[1] A. Berque, *Ecoumène : Introduction à l'étude des milieux humains*. Belin, 2009.
[2] P. Boulez, *Par volonte et par hasard*. Seuil, 1975.
[3] J. Cage, *X: Writings, "79-"82*, Reprint. Wesleyan University Press, 1983.
[4] P. T. de Chardin and F. Tardivel, *Hymne de l'univers*. Seuil, 1997.
[5] G. Deleuze and F. Guattari, *Mille plateaux*. Paris: Éditions de minuit, 1980.
[6] G. Deleuze, "La voix de Gilles Deleuze - Spinoza - Des vitesses de la pensée," 12-Feb-1980. [Online]. Available: http://www.univ-paris8.fr/deleuze/article.php3?id_article=91. [Accessed: 22-Apr-2010].
[7] P. Descola, *Par-delà nature et culture*. Paris: NRF : Gallimard, 2005.
[8] J.-G. Ganascia, *L'intelligence artificielle*. Flammarion, 1993.
[9] J. Gleick, *La Théorie du chaos : Vers une nouvelle science*. Flammarion, 1999.
[10] D. Hofstadter and E. Sander, *L'analogie : Coeur de la pensée*. Odile Jacob, 2013.
[11] C. G. Jung, "La synchronicité, principe de relations acausales," in *Synchronicite et Paracelsica*, Paris: A.Michel, 1950.
[12] P. Lévy, *La sphère sémantique : Tome 1, Computation, cognition, économie de l'information*. Hermes Science Publications, 2011.
[13] J. Scott Lee, "Par delà la mimêsis : Mallarmé, Boulez et Cage," *Revue esthétique*, no. numéro 13–14–15, pp. 295–311, Sep. 1998.
[14] S. Szoniecky and H. Hachour, "Monades pour une éthique des écosystèmes d'information numériques," presented at the Digital Intelligence, Nantes, 2014.
[15] S. Szoniecky, "Évaluation et conception d'un langage symbolique pour l'intelligence collective : Vers un langage allégorique pour le Web," Science de l'information et de la communication, Université Paris VIII Vincennes-Saint Denis, 2012.

[16] J.-P. Vernant, *Divination et rationalité*. Seuil, 1974.
[17] C. Lipsyc and M. Ihadjadene, "Architecture de l'information et éditorialisation," *Études de communication. langages, information, médiations*, no. 41, pp. 103–118, Dec. 2013.
[18] S. Szoniecky, H. Hachour, and N. Bouhai, "Générateur hypertextuel pour l'interprétation des médias sociaux dans une topologie sémantique," *Les Cahiers du numérique*, vol. Vol. 7, Empreintes de l'hypertexte sous la direction de Caroline Angé, no. 3, pp. 93–121, Sep. 2012.
[19] S. Szoniecky, "Tweet Palette : cartographie sémantique pour l'interprétation d'un événement," presented at the EUTIC 2012 Enjeux et usages des TIC, Université de Lorraine, 2012, p. 15 p.