

# Dictionnaires de catégories pour la génération automatique de proverbes : vers une économie sémantique de l'interprétation

Samuel Szoniecky

▶ **To cite this version:**

Samuel Szoniecky. Dictionnaires de catégories pour la génération automatique de proverbes : vers une économie sémantique de l'interprétation. La Néologie, les corpus informatisés et les processus d'élaboration des langues de moindre diffusion, Nov 2013, Ghardaïa, Algérie. hal-01098427

**HAL Id: hal-01098427**

**https://univ-paris8.hal.science/hal-01098427**

Submitted on 24 Dec 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC0 - Public Domain Dedication 4.0 International License

# Dictionnaires de catégories pour la génération automatique de proverbes : vers une économie sémantique de l'interprétation.

Samuel Szoniecky, Laboratoire Paragraphe, Université Paris 8

## Introduction

Dans le cadre du projet GAPAÏ financé par le labex Arts-H2H, nous développons une plate forme pour l'évaluation socio-cognitive des réseaux sociaux. Cette plate forme utilise un générateur automatique de texte pour produire des proverbes que l'on soumet aux utilisateurs des réseaux sociaux afin d'observer l'interprétation qu'ils en font.

Pour que ces observations puissent servir à une analyse socio-cognitive précise des réseaux sociaux, il est nécessaire de générer des textes et récolter des interprétations dont la sémantique est finement contrôlable suivant des critères économiques de base : où, quand, qui, combien, viable, juste... Pour ce faire, une des questions cruciales et celle de la catégorisation, tant au niveau de la génération des textes et de l'interprétation de ceux-ci que des analyses qui en découlent. Face à ces problèmes, on peut s'interroger sur l'utilité des dictionnaires. En fournissant des règles précises de catégorisation, les dictionnaires permettent-ils de constituer les bases d'une économie sémantique ? Celle-ci peut-elle servir pour piloter la génération de texte et analyser leurs interprétations et par là même être un outil pédagogique pour l'apprentissage des langues ?

Pour répondre à ces questions, nous nous attarderons, dans un premier temps, sur la construction des dictionnaires utilisés par le générateur de textes. Nous insisterons, dans un deuxième temps, sur les principes de catégorisation que nous avons adoptés pour constituer une économie sémantique. Enfin, nous montrerons comment à partir d'une génération de textes contrôlée sémantiquement, on peut développer des outils pour l'apprentissage des langues. Pour illustrer ces points, nous présenterons nos outils pour créer les dictionnaires, tester la génération des textes, récolter l'interprétation, stimuler l'apprentissage des langues, en bref, pour développer une économie sémantique de l'interprétation.

## 1. La construction des dictionnaires génératifs

La fabrication d'un générateur automatique de texte permet de questionner les structures de cohérence de la littérature et les moyens de les reproduire automatiquement. La formalisation de la littérature par des algorithmes montre qu'à partir de quelques règles simples, il est possible de produire une infinité de textes. Cette formalisation se situe à trois niveaux, elle porte sur le « moteur » du générateur, sur les dictionnaires utilisés par le moteur et sur le langage qui programme le moteur avec un « texte génératif » faisant référence aux dictionnaires (cf. p. 3 : Diagramme du générateur de texte). Nous traiterons ici uniquement des dictionnaires[1] en les détaillant suivant deux types : des dictionnaires formels et des dictionnaires sémantiques. Nous montrerons ensuite la complémentarité de ces dictionnaires pour organiser les langues à la fois suivant des formes hiérarchiques et rhizomathique.

[1] Pour plus de détails cf. [1]

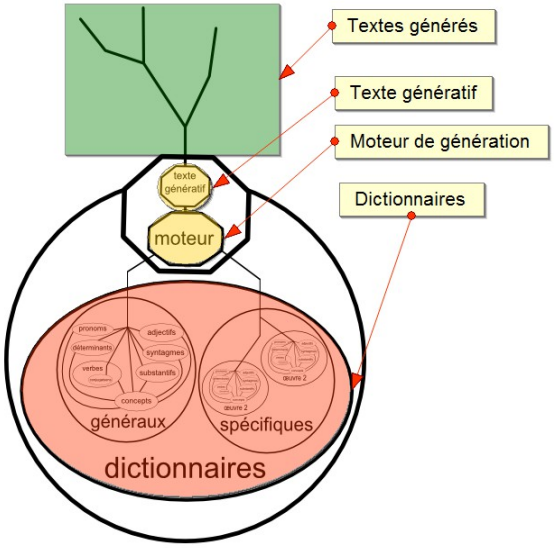



*Figure 1 : Diagramme du générateur de texte*

## *1.1 Les dictionnaires formels*

Les dictionnaires formels fournissent les éléments de structure de la langue dans laquelle les textes sont générés, ils sont organisés dans des collections correspondant à chaque type de structure, à savoir :

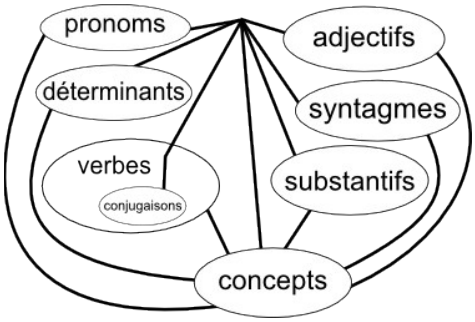



*Figure 2 : Organisation des dictionnaires formels*

- **les syntagmes** = éléments invariables utiliser pour articuler les expression textuelles, par exemple : « aujourd'hui », « alors », « afin »,

« car », « ça »... Dans le dictionnaire français nous disposons aujourd'hui de 215 syntagmes

- **les pronoms** = éléments variables suivant l'élision ou non du libellé. Les pronoms sont organisés en trois catégories :
  - sujet, par exemple « je » ou « j' », « tu » ou « tu »... Il y a 9 pronoms sujets dans le dictionnaire français
  - sujet indéfini, par exemple « ce » ou « c' », « on » ou « on »... Il y a 6 pronoms indéfini dans le dictionnaire français
  - complément, par exemple « se la » ou « se l' », « moi » ou « moi »... Il y a 47 pronoms compléments dans le dictionnaire français
- **les déterminants** = éléments variables suivant le genre, le nombre et l'élision (8 formes) par exemple :
  - masculin singulier : « du », « nul »
  - féminin singulier : « de la », « nulle »
  - masculin singulier avec élision : « de l' », « nul »
  - féminin singulier avec élision : « de l' », « nulle »
  - masculin pluriel : « des », « nuls »
  - féminin pluriel : « des », « nulles »
  - masculin pluriel avec élision : « des », « nuls »
  - féminin pluriel avec élision : « des », « nulles »
- **les verbes et leurs conjugaisons**, éléments variables suivant la personne et le temps, les verbes sont composés d'un préfix, d'un marqueur d'élision et d'un modèle de conjugaison qui renvoie une

terminaison suivant la personne et le temps. Il y a aujourd'hui plus de 1 400 verbes dans le dictionnaire français.

- **les adjectifs**, élément variable suivant le genre et le nombre, ils sont composés d'un préfix, d'un marqueur d'élision et d'une terminaison. Par exemple : abscent → prefix = « absen », élision = oui, masculin singulier = « t », féminin singulier = « te », masculin pluriel = « ts », féminin pluriel : « tes ». Le dictionnaire français possède aujourd'hui plus de 2 500 adjectifs.
- **les substantifs**, éléments variables suivant le nombre, ils sont composés d'un préfix, d'un marqueur d'élision, d'un marqueur de genre, d'une terminaison pour le singulier et d'une terminaison pour le pluriel. Par exemple : cheval → élision = non, genre = masculin, prefix = « chev », singulier = « al », pluriel = « aux ». Plus de 5 000 substantifs sont aujourd'hui dans le dictionnaire français.
- **les concepts** qui permettent d'associer dans des structures autonomes un assemblage cohérent de pronoms, verbes, substantifs, adjectifs, etc... Ce sont ces éléments qui permettent de donner une dimension générative en créant des collections d'assemblages, qui se regroupent dans des collections d'assemblages plus complexes et ainsi de suite dans les limites de calculabilité des machines (boucle sans fin ou trop longue). Par exemple :
    - « [O-ToutPoème] » est composé entre autres de ;
        - « [O-PoèmeStein] » qui est composé entre autres de :
            - « [thl-Steins] » qui est composé entre autres de :
                - « [thl-souvenir-01] » qui est composé entre autres de :
                    - « [0|m_agitation 1] » qui est composé entre autres de :

- désorganisation
- dissension
- émoi
- ...

Le regroupement de ces différents éléments dans des collections permet une première organisation sémantique des dictionnaires suivant un mode hiérarchique simple que nous allons détailler maintenant en montrant qu'il peut être enrichie par d'autre mode d'organisation sémantique.

## *1.2 Les dictionnaires sémantiques*

Parallèlement aux dictionnaires formels qui assurent une cohérence syntaxique des textes générés, les dictionnaires sémantiques vont assurer une cohérence du sens. Pour créer ces dictionnaires, l'auteur va modéliser des structures sémantiques simples qui serviront de base pour la génération de textes dans un univers de signification particulier. Ces structures sont des regroupements d'expression dans une catégorie qui les qualifie. Le travail sémantique de l'auteur consistera à créer ces regroupements par rapport aux univers sémantiques qu'il souhaite explorer.

Prenons par exemple, l'expression « j'aime la philosophie ». Du point de vue de la génération automatique de texte, cette expression est composée formellement d'un verbe et d'un substantif. Pour simplifier l'exposé, nous nous concentrerons uniquement sur le substantif que l'auteur rendra génératif en créant une collection de substantifs qu'il regroupera sous le terme générique de « philo 1 ». Ainsi, il pourra définir un champ sémantique de la philosophie en associant au terme « philosophie » d'autres termes comme : « mystique », « «éthique »,

« théologique »... Le texte génératif pourra dès lors prendre plusieurs formes : « j'aime l'éthique », « j'aime la mystique », « j'aime la philosophie ».

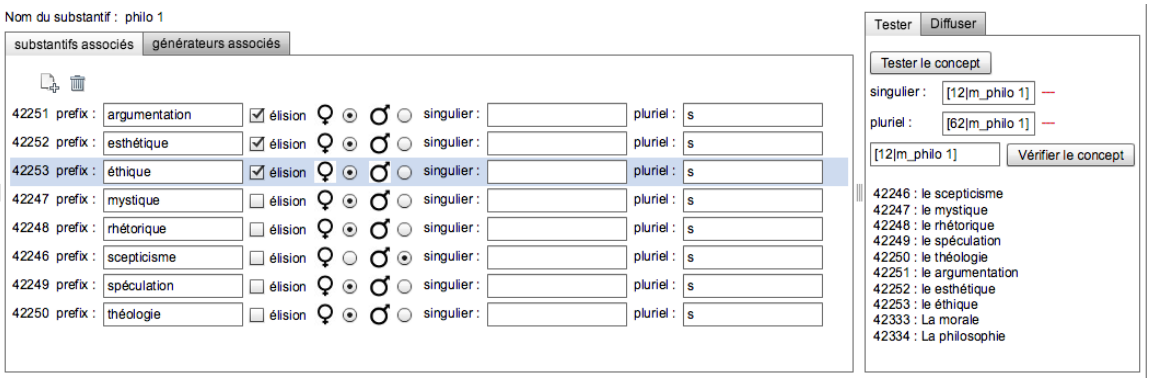



*Figure 3 : Exemple de regroupement sémantique*

Notons que ces regroupements constituent le point de vue particulier d'un auteur sur ce qu'est le champ sémantique de la philosophie. A travers cette collection de substantifs, l'auteur exprime la cohérence sémantique de l'univers qu'il souhaite générer. Pour un autre auteur, celle-ci pourrait être tout autre voir même incohérente, par exemple en ajoutant dans la collection « philo 1 » les substantifs suivants : « bêtise », « mensonge »... Toutefois, on peut imaginer que derrière cet ajout de termes opposés sémantiquement les uns aux autres dans une même collection, l'auteur cherche à exprimer un univers sémantique surprenant où, suivant les générations, le texte exprimera un monde ou son contraire. On le voit, la construction de dictionnaire sémantique relève d'un choix éditorial particulier.

Nous avons montré dans [1], comment le langage d'adressage des concepts IEML permet de modéliser le point de vue sémantique d'un auteur de dictionnaire. Pour ce faire, il suffit de donner pour chaque items d'une collection, l'adresse dans la topologie sémantique IEML [2]. L'illustration suivante montre comment la collection de substantif « philo 1 » se répartie essentiellement dans

les « connaissances organisées » mais aussi dans d'autres espaces sémantiques comme les « accomplissements ».

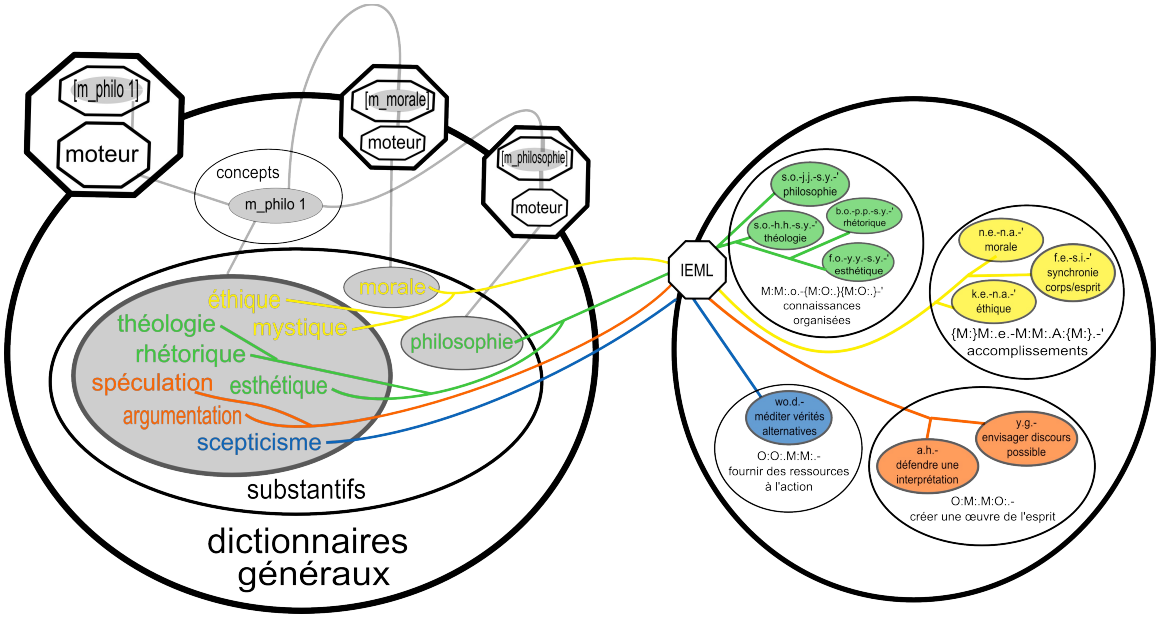



*Figure 4 : Description IEML d'un regroupement sémantique*

La topologie de concepts proposée par IEML permet de concevoir des couches sémantiques organisées suivant une structure ou chaque concept et en relation avec tous les autres suivant un agencement particulier qui correspondra au point de vue d'un auteur. Grâce à ce système rhizomatique, il est possible d'ajouter à n'importe qu'elle expression une description sémantique dont la précision et l'orientation dépendra de l'auteur de cette description. Dès lors, il apparaît que suivant ces principes, la sémantique d'un texte génératif est fonction à la fois des choix d'organisation hiérarchique des dictionnaires et des choix de description rhizomatique de chaque item du dictionnaire. Les outils que nous proposons servent justement à enregistrer ces choix par rapport à une date, un lieu (une adresse IP ou une géolocalisation en temps réel), un auteur et une expression, en ce sens, ils s'accorde avec les principes d'organisation sémantique défendus par Deleuze et Guattari :

> « Ce qui compte, c'est que l'arbre-racine et le rhizome-canal ne s'opposent pas comme deux modèles : l'un agit comme modèle et comme calque transcendants, même s'il engendre ses propres fuites ; l'autre agit comme processus immanent qui renverse le modèle et ébauche une carte, même s'il constitue ses propres hiérarchies, même s'il suscite un canal despotique. » [3] p. 31

Dans cette optique, nous interprétons la proposition d'associer l'arbre et le rhizome comme la mise en rapport par un individu ici et maintenant, d'une dimension physique-métrique-transcendante-extérieure : l'arbre ; avec une dimension conceptuelle-topologique-immanente-intérieure : le rhizome. Ce que nous pouvons représenter par le diagramme suivant qui montre comment ces trois dimensions d'existence s'articulent pour représenter une potentialité de connaissances :

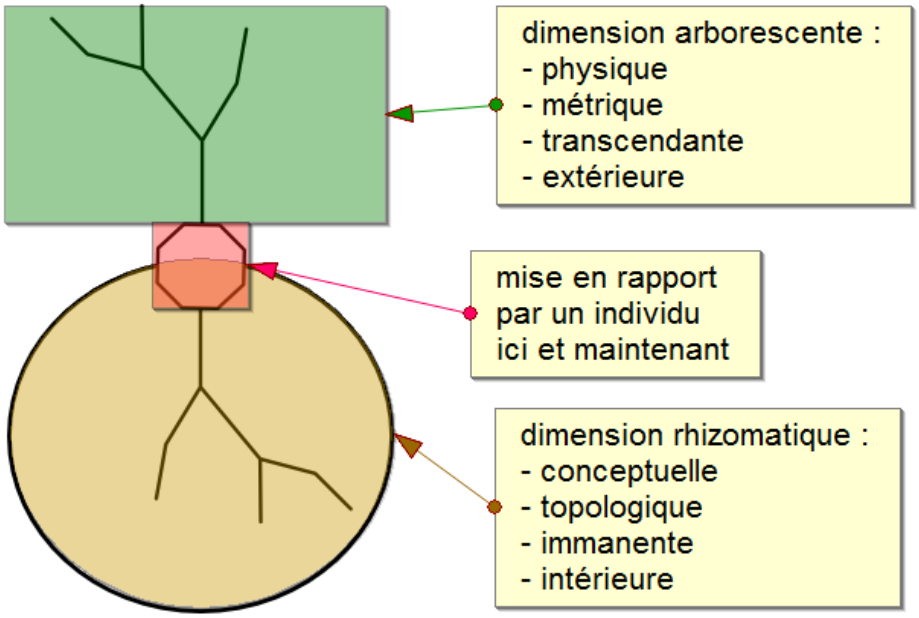
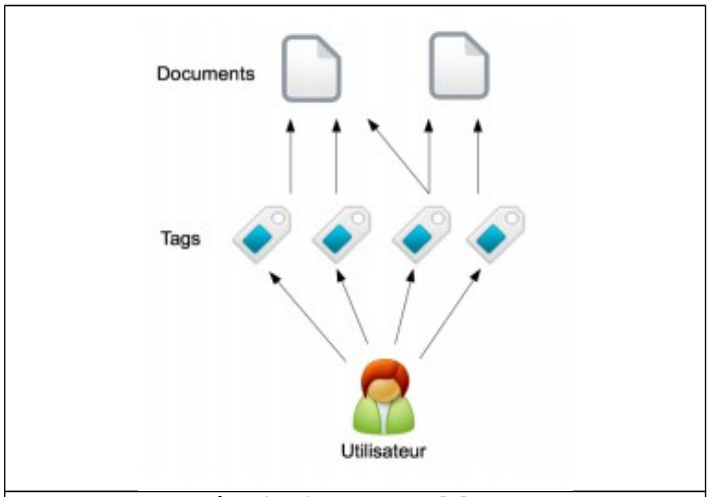



*Figure 5 : Diagramme de relation entre l'arbre et le rhizome*

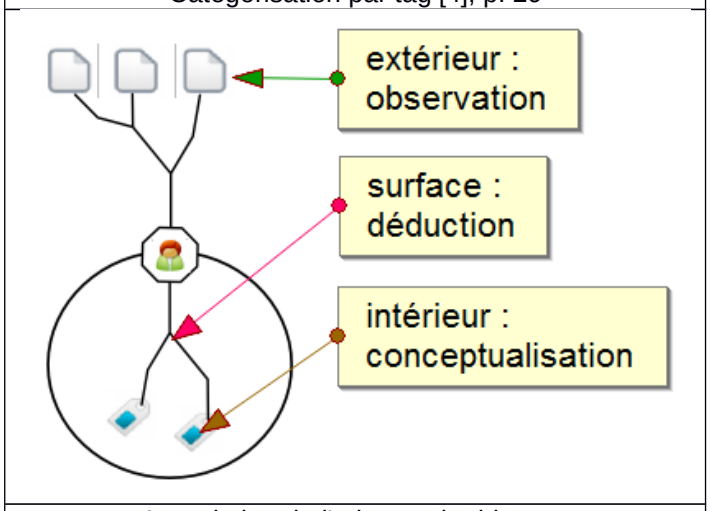
Sur la base de ces principes nous avons construis nos outils de catégorisation afin de récolter la matière sémantique nécessaire pour développer une économie sémantique de l'interprétation.

## 2. Principes de catégorisation

En nous appuyant sur les principes de l'arbre et du rhizome, nous nous détachons des schémas qui décrivent le processus de catégorisation par des mots-clefs en représentant l'ensemble des éléments intervenant dans le processus sans prendre en compte trois dimensions : extérieur, surface et intérieur :




Catégorisation par tag [4], p. 29




Association de l'arbre et du rhizome

*Figure 6: Comparaison entre deux représentations du processus de catégorisation*

Le processus de catégorisation que nous défendons peut être décomposé en trois étapes : l'observation extérieure, la modélisation logique de surface, créer des rapports intérieurs entre documents et concepts.

## *2.1 Etape 1 : observation-extérieure*

*La première étape*, celle de l'observation-extérieure, consiste à décrire les physicalités auxquelles l'utilisateur est confronté dans sa recherche de connaissance. Dans le contexte d'un proverbe, cette étape consiste à évaluer si le texte est formellement juste. Est-ce que l'orthographe et la syntaxe sont juste ? Manque-t-il des mots ou des lettres ? Cette première étape revient en fait à décomposer un document du plus global au plus détaillé, pour créer une représentation de celui-ci sous la forme d'une arborescence. Puis de validité que toute les branches nécessaires sont présentes.

Dans le cadre d'un apprentissage des langues cette étape est fondamental et primordiale, elle consiste pour un apprenant à reconnaître dans une langue les formes adéquates et celles qui ne le sont pas. Pour s'exercer à ce type de pratique, le générateur automatique de texte peut parfaitement simuler des erreurs et proposer aux apprenant des exercices sans cesse renouveler où il conviendra de reconnaître les forme justes et celles qui ne le sont pas.

## *2.2 Etape 2 : modélisation logique de surface*

*La deuxième étape* de ce processus est la modélisation d'une succession de liens logiques sous la forme d'une trame correspondant à l'interface entre l'extérieur et l'intérieur, c'est à dire à une surface. Cette construction utilise les principes de modélisation du « tissu de l'âme » à partir d'une structure fractale hexagonale [5].

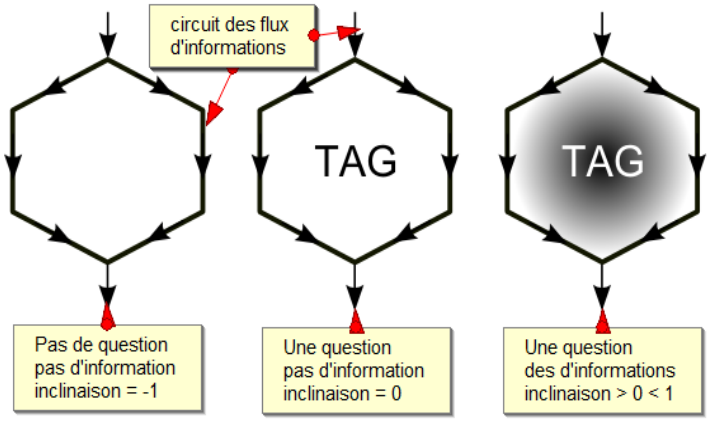



*Figure 7: Utilisation de l'hexagone comme opérateur bolléen*

Pour développer cette organisation, nous constituons cette trame à partir de rhizomes conceptuels. Pour un proverbe, la trame logique correspond par exemple à l'appartenance ou non d'une expression à une succession de catégories et de sous-catégorie. Dans l'exemple ci-dessous le proverbe « A bon chat chat, bon rat » est décrit de façon logique pour montrer que les termes du proverbe renvoie à la logique suivante : animal-prédateur-bon = animal-proie-bon.

Cette trame conceptuelle forme un « crible » ayant une double fonction, à la fois comme représentation des connaissances et comme générateur de l'événement les produisant :

> « L'événement se produit dans un chaos, dans une multiplicité chaotique, à condition qu'une sorte de crible intervienne. » (Deleuze, 1988, p. 103)

La représentation des connaissances associe lecture et écriture pour créer une potentialité de connaissances. En ce sens, nous suivons l'interprétation par Deleuze du « crible » comme étant la forme diagrammatique [6] de la chôra, si chère à Platon [7] et dont on peut donner la définition suivante :

> « Polysémique, le terme chôra évoque l'entrelacement de "l'aspect constitutif" et de l'aspect spatial", "ce en quoi" apparaissent les choses sensibles et "ce de quoi" elles sont constituées. » [8], p. 22

En tant que potentialité générative la chôra doit être associer à une dynamique particulière, à un pouvoir d'agir, pour produire une connaissance. En effet, un champ donnera du blé seulement si on y plante des graines dont la dynamique interne transformera les potentialités offertes par la terre en végétal. Dans notre conception du crible rhizomatique, la dynamique particulière est apportée par l'individu qui, seul ou en réseau, officiera à la manière d'une graine pour générer des connaissances par la mise en relation des rhizomes conceptuels et des arbres documentaires. On peut dès lors observer ce qu'Augustin Berque a qualifié de raison trajective :

> « La raison trajective, elle est en effet dans la pulsation existentielle qui, par la technique, déploie notre corps en monde sur la terre, et qui simultanément, par le symbole, reploie le monde en notre chair. » [9], p. 402

## *2.3 Etape 3 : créer des rapports intérieurs entre documents et concepts*

*La troisième étape* du processus de catégorisation consiste pour un individu à créer des rapports entre des documents-branches et des concepts rhizomes. Par cette démarche, il construira un point de vue particulier qu'il pourra ensuite faire évoluer, partager ou faire disparaître. Placé ainsi face à des choix, l'utilisateur devient acteur de sa connaissance et non plus seulement spectateur, il doit interpréter, se situer et discerner ce qu'il faut prendre en compte ou occulter, ce qui pour lui est important ou non :

> « Or la pratique interprétative s'avère offrir une voie particulièrement propice pour approcher la question de savoir ce qui rend quelque chose important (ou non). Même si cette question peut difficilement être abordée de front, l'effort de jugement réfléchissant et d'explicitation des pertinences qui va de pair avec l'expérience interprétative - en particulier dans les domaines littéraire et philosophique - conduit naturellement à se demander ou non de quoi quelque chose peut apparaître comme important. » [10], p. 82

Le parcours interprétatif que nous venons de décrire consiste à multiplier le potentiel des relations croisées entre une arborescence documentaire et un

réseau de concept, et en même temps à impliquer un tiers qui instanciera des relations particulières par des choix de cohérence. En multipliant ces choix, le tiers modélise un réseau de discernement qui constitue son point de vue, ce qu'on peut aussi appeler son identité informationnelle ou pour reprendre les mots de Leibniz son « tissu de l'âme ». La dimension numérique des documents laisse envisager à la fois un traitement automatique de cette mise en relation mais aussi un traitement manuel dans le sens où le choix d'une relation est effectué par un humain. C'est ce deuxième cas qui nous intéresse particulièrement d'analyser dans le projet GAPAII.

## Conclusion

Notre objectif à travers le projet GAPAII est de fournir à la collectivité des chercheurs les coordonnées des espaces de connaissances explorées et les moyens de conduire cette exploration vers l'interprétation en aménageant un parcours allant de la simple observation à la conceptualisation en passant par la déduction. Nous avons choisi de centrer cette expérience autour des proverbes car ce contexte permet de poser les questions clefs concernant la capacité d'un individu ou d'un groupe d'individu de cerner l'étendue de ces connaissances : ce qu'il sait et ce qu'il lui reste à apprendre.

> « A quel degré de généralité un proverbe s'applique-t-il ? Quelle est l'étendue des situations auxquelles un proverbe peut être attribué sans que cela semble exagéré, voir incongru ? [...] Où se trouvent finalement les frontières de la catégorie implicitement définie par notre proverbe ? » [11], p. 134

A partir des outils que nous développons, notamment ceux liés aux dictionnaires, nous mettrons en place des ateliers-laboratoires auprès des élèves de collège et de lycée afin de valider nos hypothèses scientifiques et l'ergonomie de nos outils. L'objectif est de pouvoir ensuite proposer des outils pour explorer les

connaissances et gérer de façon efficace l'économie sémantique des interprétations.

## Bibliographie

[1] 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.

[2] P. Lévy, *La sphère sémantique : Tome 1, Computation, cognition, économie de l’information*. Hermes Science Publications, 2011.

[3] G. Deleuze and F. Guattari, *Mille plateaux*. Paris: Éditions de minuit, 1980.

[4] M. Crepel, “Tagging et folksonomies : pragmatique de l’orientation sur le Web,” Université Rennes 2, Rennes, 2011.

[5] S. Szoniecky, “Proposition d’une méthode graphique pour le filtrage des flux d’information,” in *Doctorales SFIC 2011, PROBLÉMATISATION ET MÉTHODOLOGIE DE RECHERCHE*, Bordeaux, 2011.

[6] N. Batt, “L’expérience diagrammatique : un nouveau régime de pensée,” in *Penser par le diagramme : de Gilles Deleuze à Gilles Châtelet*, Saint-Denis: Presses universitaires de Vincennes, 2005.

[7] A. Berque, “La chôra chez Platon,” in *Espace et lieu dans la pensée occidentale de Platon à Nietzsche*, Paris: La Découverte, 2012.

[8] J. M. Zamora, “La chôra après platon,” in *Symboliques et dynamiques de l’espace*, Publication Univ Rouen Havre, 2003, pp. 16–32.

[9] A. Berque, *Ecoumène : Introduction à l’étude des milieux humains*. Belin, 2009.

[10] Y. Citton, *L’avenir des humanités : Economie de la connaissance ou cultures de l’interpétation ?* Editions La Découverte, 2010.

[11] D. Hofstadter and E. Sander, *L’analogie : Coeur de la pensée*. Odile Jacob, 2013.