Numexco et l'analyse par attracteurs et par classes des entrées de l'ECHO (Encyclopédie Culturelle Hypermédia de l'Océanie)

Denis Gagnon

Département d'anthropologie, Collège universitaire de Saint-Boniface

degagnon@ustboniface.mb.ca

ABSTRACT. This paper presents two original methods of textual processing analysis: analysis by attractor and by class. The corpus processed is composed of ethnographic data in French found in the Cultural Hypermedia Encyclopedia of Oceania. The goal of this data processing is to compare the pool of attraction and graphs of the semantic network of the attractor "Kava", which was developed by two researchers on the basis of traditional content analysis, with the pool of attraction and graphs developped by Numexco. Numexco, which was developed by LANCI-UQAM, is a software program of the SATIM plateform, by which one can conduct connectionist analyses and extract knowledge from digitized texts.

KEYWORDS: Textual analysis, software textual analysis, connectionism, analysis methodology, cultural anthropology.

RÉSUMÉ : Cet article présente deux méthodes originales d'analyse de données textuelles assistée par ordinateur : l'analyse par attracteurs et par classes. Le corpus traité se compose de données ethnographiques en langue française qui sont à la base des entrées de l'Encyclopédie Culturelle Hypermédia de l'Océanie. Le but de l'expérimentation consiste à comparer le bassin d'attraction et les graphes du réseau sémantique de l'attracteur «Kava» obtenus par deux autres chercheurs suite à une analyse de contenu traditionnelle, à ceux obtenus «à l'aveugle» par Numexco, un logiciel de la plate-forme SATIM développée par le LANCI-UQAM qui permet d'appliquer une méthode connexionniste au problème de l'extraction de connaissances à partir de textes numérisés.

MOTS-CLÉS : Analyse de données textuelles, logiciels d'analyse de textes, connexionnisme, méthode d'analyse, anthropologie culturelle

1. Introduction

Cet article présente deux méthodes originales d'analyse de données textuelles assistée par ordinateur. Numexco[1] est un logiciel de la plate-forme SATIM[2] développée par le LANCI-UQAM[3] qui permet d'appliquer une méthode connexionniste au problème de l'extraction de

[1] Le NUMérique et l'EXtraction des COnnaissances.
[2] Système d'Analyse et de Traitement de l'Information Multidimensionnelle.
[3] Le Laboratoire d'Analyse Cognitive de l'Information de l'Université du Québec à Montréal dirigé par Jean-Guy Meunier.

connaissances à partir de textes numérisés[4]. Les méthodes proposées consistent à effectuer un filtrage numérique sur des corpus afin de classifier et de catégoriser les unités d'information pour approfondir les étapes ultérieures d'interprétation et de construction de réseaux sémantiques. Bien que la plupart des étapes du traitement des corpus soient automatiques, certaines font appel à l'intervention manuelle[5] et l'analyse des résultats relève de l'interprétation et de la créativité du chercheur.

Une expertise dans l'analyse d'ouvrages scientifiques et de transcriptions d'entrevues assisté par ordinateur est déjà acquise avec les logiciels développés par le LANCI[6] et le corpus qui est ici traité se compose de données ethnographiques en langue française (fichiers FileMaker) qui sont à la base des entrées de l'ECHO[7]. Le but de l'expérimentation consiste à comparer le bassin d'attraction et les graphes du réseau sémantique de l'attracteur «Kava[8]» obtenus par Julie Ayotte et Lamont Lindstrom[9] suite à une analyse de contenu traditionnelle, à ceux obtenus «à l'aveugle» par Numexco. Après avoir décrit les étapes de l'analyse de contenu assistée par Numexco, je présente la problématique de cette analyse et les résultats obtenus par Numexco à partir des deux méthodes d'analyse que j'ai développées : l'analyse par classes et segments et l'analyse par attracteur[10]. Enfin, je présente le graphe du réseau sémantique de l'attracteur «Kava» élaboré à partir des données obtenues par Numexco, c'est-à-dire sans avoir lu le corpus.

2. Étapes de l'analyse de contenu assisté par Numexco

Avec Numexco, contrairement à l'analyse de contenu traditionnelle, il n'est pas nécessaire de lire les corpus, souvent très volumineux, mais une problématique solide doit précéder toute analyse. Le choix et la définition des unités de classification sont remplacés par une chaîne de traitement automatique et manuelle qui permet de créer les bases de données qui serviront à l'analyse. Les différentes étapes de cette chaîne de traitement sont la segmentation du corpus, l'extraction, la lemmatisation et le filtrage du lexique, et la création des classes par les réseaux de neurones.

La segmentation du corpus – par nombre de chapitres ou de pages, paragraphes, phrases, lignes ou mots, selon les besoins du chercheur – et l'extraction du lexique se font automatiquement. La lemmatisation du lexique consiste à remplacer chaque mot par son équivalent canonique, une étape qui se justifie par le fait que les déclinaisons n'affectent pas le contenu sémantique des termes, et le filtrage du lexique comporte une étape automatique et une étape manuelle et interprétative. Les termes fonctionnels sont éliminés automatiquement et un filtrage manuel est nécessaire afin d'adapter le lexique à la question de recherche en sélectionnant les termes les plus pertinents et en supprimant ceux qui ne

---

[4] Voir BISKRI, MEUNIER et NAULT 1997, JOUIS *et al.* 1997, et MEUNIER *et al.* 1997.

[5] Certaines des étapes manuelles répétitives et purement statistiques sont en voie d'automatisation.

[6] Voir GAGNON 1999, 2000, 2001a, 2001b, et GAGNON et MEUNIER 2002.

[7] L'Encyclopédie Culturelle Hypermédia de l'Océanie (www.oceanie.org) est un site internet développé sous la direction de Pierre Maranda, professeur émérite au Département d'anthropologie de l'Université Laval.

[8] Le kava (*Piper methysticum*) est une plante rituelle du nord-ouest de l'Océanie qui est appréciée pour les effets narcotiques et anesthésiants du breuvage qui est tiré de ses racines.

[9] Voir annexe 1 et 2.

[10] Un attracteur est un terme autour duquel gravite un nombre donné d'unités d'informations appelées bassin d'attraction.

sont pas porteurs de sens du point de vue sémantique. Enfin, à partir des segments du corpus et du lexique, Numexco construit des classes à partir des termes qui partagent une unité sémantique ou conceptuelle.
Une fois les classes créées, les unifs (unités d'information) sont catégorisés manuellement à l'aide d'un thésaurus[11]. Cette étape interprétative, dont j'ai démontré l'utilité lors des expérimentations précédentes[12], permet d'expliciter les relations que les termes entretiennent entre eux dans une même classe et de passer du niveau de la diversité lexicale à celui des unités sémantiques en regroupant les unifs sous des étiquettes générales[13]

Tableau 1 - Thésaurus utilisé pour l'analyse des entrées de l'ECHO

1 - Êtres humains a - homme b - femme	**8 - Géographie** a - culture b - nature
2 - Anatomie et processus vitaux a - anatomie b - processus vitaux c - altération d - maladie e - guérison f - mortalité	**9 - Règne animal** a - terrestre b - aquatique
3 - Perception	**10 - Règne végétal**
4 - Émotions et sentiments a - neutres b - positifs c - négatifs	**11 - Action** a - alimentation b - horticulture c - autres
5 - Domaine cognitif	**12 - Événement**
6 - Relations sociales a - hiérarchie b - économie c - parenté/alliance d - relations sociales	**13 - Transformation**
7 - Religion a - religion traditionnelle b - christianisme c - général	**14 - Artefact** a - ustensile b - vêtement c - objet

[11] Pour les termes polysémiques ou ambigus, il suffit de consulter la ou les classes dans lesquelles apparaît le terme pour être fixé sur son contenu sémantique. En ce sens, une bonne partie de la catégorisation pourrait être faite automatiquement.
[12] Voir GAGNON 1999, 2000, 2001a, 2001b, 2002a, 2002b.
[13] Voir MARANDA et NZE-NGUEMA 1994.

Pour fin de recherche, nous voyons que chacune des catégories peut être subdivisée. Par exemple, la catégorie 6 (relations sociales) a été subdivisée en relations hiérarchiques (6a), en relations économiques (6b), de parenté ou d'alliance (6c) et de relations sociales générales (6d). Dans le cas d'analyses plus poussées, ces sous-catégories peuvent également être subdivisées (ex.: 6b1 : économie locale / 6b2 : économie globale).
Concernant les différences et les similarités entre l'analyse de contenu traditionnelle et l'analyse de contenu assistée par Numexco, le tableau suivant montre que les principaux apports de Numexco sont la création automatique des segments, du lexique et des classes sous forme de bases de données qui facilitent le traitement statistique pour fins d'analyse et d'interprétation. Au niveau des étapes de l'analyse qualitative/quantitative et de l'interprétation, les étapes demeurent les même pour les deux types d'analyse.

Tableau 2 - Comparaison des deux types d'analyse

ANALYSE TRADITIONNELLE	ANALYSE ASSISTÉE PAR NUMEXCO
1 - Lecture du texte et élaboration du cadre conceptuel	1 - Élaboration de la problématique et choix du type de segmentation
2 - Choix et définition des unités de classification	2 - Segmentation du texte, filtrage du lexique, création de la matrice et des classes
3 - Catégorisation et classification des unités de sens	3 - Élaboration du thésaurus et catégorisation des unifs
4 - Quantification des données et traitement statistique	4 - Exploration des classes et des segments (quantification et traitement statistique)
5 - Analyse qualitative et/ou quantitative	
6 - Interprétation des résultats	

Les bases de données créées par Numexco en format *.mdb sont exportées automatiquement dans le logiciel ACCESS sous la forme des tables et des requêtes présentées dans le tableau 3.

*Tableau 3 - Bases de données en format *.mdb*

TABLES	REQUÊTES
Classes	Fréquence totale
Lexique départ	Fréquence totale par segments
Lexique transféré	Lexique départ
Matrice départ	Lexique transféré
Matrice transférée	**Liste des segments**
	Mots par classes
	Mots par classes départ

L'exploration et l'analyse du corpus sont effectuées à partir des requêtes «Liste des segments» et «Mots par classes». La première, qui renvoie au corpus segmenté, est

convertie en format *.doc, et la seconde, qui présente le contenu des classes, le lexique et la fréquence d'apparition des lemmes par segment et pour l'ensemble du corpus, est convertie en format *.xls. Le type de résultat produit par Numexco permet au chercheur de passer à l'étape de l'analyse et de l'interprétation sans avoir à classifier les unifs, une étape qui se révèle fastidieuse avec des corpus volumineux, et les résultats de l'analyse peuvent être présentés sous forme de graphiques ou de graphes (réseaux sémantiques).

3. Méthodes d'analyse par attracteurs et par classes

Pour cette expérimentation, les bases de données du contenu en langue française du corpus «Kava» traité par Numexco ont été analysées et un graphe du réseau sémantique a été élaboré afin de le comparer avec ceux réalisés par Julie Ayotte et Lamont Lindstrom dans le cadre du projet ECHO[14]. Les étapes de l'analyse de contenu traditionnelle utilisée par Ayotte pour identifier le bassin de l'attracteur consistaient à consulter minutieusement le corpus avec les «explorateurs» du site «Oceanie.org» pour les thèmes généraux et les modes de pensée et à comparer les résultats avec ceux obtenus par Lindstrom.

3.1 Traitement du corpus et des bases de données créées par Numexco

Les données du fichier FileMaker «Kava» ont été exportées en format *.txt et soumises à la chaîne de traitement de Numexco présentée plus haut. Après traitement, le corpus contient 94 segments de 10 phrases et, une fois filtré, le lexique est passé de 4521 unifs à 589 unifs pertinentes réparties en 37 classes.

Tableau 4 - Statistiques du corpus traité par Numexco

Nombre de segments du corpus	94
Nombre de phrases par segments	10
Nombre d'unifs avant filtrage manuel	4521
Nombre d'unifs après filtrage	589
Nombre de classes	37

Après avoir converti la «liste des segments» en format *.doc (fichier seg.doc) et la base de données «Mots par classes» en format *.xls (fichier mcl.xls), trois fichiers *.xls ont été créés pour analyser les résultats : catégories/fréquence/unifs (fichier cfu.xls) ; fréquence/unifs/catégories (fichier fuc.xls) et catégories/fréquences (fichier cf.xls).

[14] Voir annexe 1 et 2.

*Tableau 5 - Fichiers *.doc et *.xls*

Nom	Contenu par colonnes	Utilité
seg.doc	- Base de donnée «Liste des segments»	Fichier de référence. Accès automatique au corpus segmenté et numéroté
mcl.xls	- Base de donnée «Mots par classes»	Fichier de référence. Accès aux classes, aux unifs associées et aux numéros de segments
cfu.xls	- Catégorie (ordre ascendant) - Fréquence (ordre descendant) - Unif (ordre alpha)	Permet de calculer la fréquence des unifs par catégories, l'occurrence des catégories dans le corpus et des unifs à l'intérieur des catégories
fuc.xls	- Fréquence (ordre descendant) - Unif (ordre alpha) - Catégorie (ordre ascendant)	Permet de calculer la fréquence des unifs dans le corpus tout en référant à la catégorie
cf.xls	- Catégorie (ordre ascendant) - Fréquence	Permet de calculer la fréquence des catégories dans le corpus

À partir de ces fichiers, deux types d'analyse sont proposés, l'analyse par attracteur et l'analyse par classes et segments.

3.2 L'analyse par attracteur

L'analyse par attracteur permet d'identifier les unifs qui forment le bassin d'un attracteur donné ; la proximité sémantique entre les thèmes (unifs, catégories) et l'attracteur ; et le réseau sémantique sous-jacent à ce thème. L'exploration par attracteur se fait selon une chaîne de traitement dont chacune des étapes permet d'avoir accès à un niveau d'interprétation particulier.

Dans un premier temps, les classes de termes contenant l'attracteur (le bassin d'attraction) sont sélectionnées[15]. Les unifs apparaissant dans ces classes sont ensuite transférées dans un fichier et classées en ordre décroissant de fréquence. Cette étape offre au chercheur la possibilité d'interpréter le vocabulaire associé à l'objet d'étude. Deuxièmement, les unifs sont catégorisées et sous-catégorisées et la fréquence des unifs dans chaque catégorie est calculée, ce qui permet d'interpréter le contenu sémantique de ce discours à partir des catégories d'unifs. Troisièmement, ces fichiers sont interprétés et présentés sous forme de graphique et le réseau sémantique associé à l'attracteur est présenté sous forme de graphe, et ce, sans avoir à référer au corpus «Kava» du site ECHO. Les trois tableaux et la figure qui suivent présentent quelques-unes des possibilités qu'offrent les étapes de ce type d'analyse[16].

[15] L'unif «kava», qui apparaît 291 fois dans le corpus, a été supprimé pour ne pas alourdir les classes bien que sa présence demeure implicite.

[16] Dans les tableaux suivants, Fréq = fréquence ; et Cat = catégorie.

Tableau 6 - Unifs de fréquence 20 et plus

Fréq	Unif	Cat
95	coupe	14a
82	racine (de kava)	10a
80	cérémonie (lemme)	07a
65	chef	06a
63	roi (lemme)	06a
54	boire	11a
46	homme	01a
38	main	02a
38	eau	11a
37	préparer (lemme)	11c
33	servir	11c
27	tradition (lemme)	05a
27	offrir (lemme)	06b
27	plante	10a
27	boisson	11a
26	île	08b
26	récolter (lemme)	11b
25	rite (lemme)	07a
24	femme	01b
24	mâcher	11a
24	plat	14a
23	breuvage	11a
23	sec (lemme) [kava sec]	10
21	vert [kava vert]	10
21	mission (lemme)	07b
20	*nakamal* (maison des hommes)	08a

Le tableau 6 (fichier fuc.xls) présente les unifs ou lemmes catégorisées — ex.: tradition = tradition(s), traditionnel(les), traditionnellement — qui apparaissent plus de 20 fois dans le corpus. Au niveau interprétatif, nous avons l'essentiel des informations contenues dans le corpus. La coupe cérémonielle remplie du liquide narcotique extrait des racines de kava est traditionnellement passée aux chefs puis aux hommes. Il est préparé par les femmes qui le mâchent et le diluent avec de l'eau et ce breuvage est servi dans la maison des hommes (*nakamal*). Le lemme mission (mission, missionnaire) indique que les données proviennent des rapports de missions océaniennes.

Tableau 7 - Unifs de la catégorie 2c,d,e,f

Cat	Fréq	Unif
02c	13	ivresse

	7	enivrer (lemme)
	5	anesthésiant (lemme), douleur (lemme), narcotique, sudorifique (lemme)
	3	alcoolique, diurétique, drogue, maigre
	2	accoutumance, altération
	1	amaigrir, amincir, ankylose, comateuse, émaciation, embonpoint, nocif
02d	5	maladie (lemme)
	4	blennorragie
	3	cystite
	1	anémie, anorexie, asthme, colique, conjonctivite, dermatose, desquamation, diarrhée, diplopie, éruption, fièvre, hémorroïdes, irritations, lépreux, obésité, prurit, rhumatisme, rhume, suppurants, syphilis, urétrites, vénérien, symptômes
02e	8	médecine (lemme), pharmaceutique (lemme)
	7	guérir (lemme), remède
	3	soigner, soin, tonique
	2	docteur, pilule, thérapeutique
	1	antigonorrhéique, cataplasme, chirurgien, élixir
02f	8	mort (lemme)
	4	enterrer (lemme)
	1	tombe, cadavre, deuil, sépulture

Le tableau 7 présente la fréquence des unifs de la catégorie 2c,d,e,f (anatomie et processus vitaux : altération, maladie, guérison, mortalité). Au niveau interprétatif, nous avons accès aux effets provoqués par la consommation du kava (2c), aux maladies qui découlent d'un excès de l'usage du kava (2d), et au vocabulaire associé à leur traitement.

Tableau 8 - Unifs de la catégorie 13

Cat	Fréq	Unif
13	16	bouillir
	13	feu
	11	broyer (lemme)
	8	extraire (lemme)
	6	délayer, filtrer (lemme), pétrir (lemme)
	4	brasser (lemme), presser
	3	cuire (lemme), mêler, tordre
	2	macérer (lemme)
	1	braise, casser, débiter, dessécher, flamme, fermenter, hacher, infuser, imbiber, laver, malaxer, marteler, mélanger, pilonnage, remuer, rincer, tamiser, tison, tremper, triturer

Le tableau 8 présente la fréquence des unifs de la catégorie 13 (transformation). Cette catégorie n'a pas été subdivisée, car toutes les unifs qu'elle contient décrivent les étapes de la préparation du kava. Cet exemple illustre particulièrement l'utilité de Numexco pour l'extraction des connaissances. Comme ces unifs sont tirées de classes associées à des segments du corpus, le chercheur peut référer rapidement aux parties de textes traitant de cette problématique et, comme nous le verrons plus loin avec un autre exemple,

approfondir la recherche par une analyse par classe qui lui permet de connaître l'ordre de ces étapes et les conditions de cette préparation accomplie par les femmes.

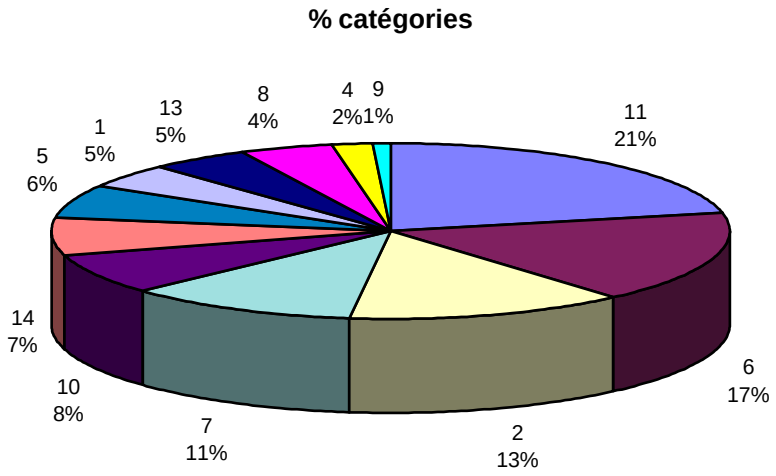


Figure 1 - Ratio des catégories du corpus «Kava»

La figure 1 présente le pourcentage par ordre décroissant des catégories présentes dans le corpus. Ce type de représentation par secteurs, un graphique parmi les ceux offerts par Excel (aire, barre, histogramme, courbe, anneau, surface). Les unifs/lemmes dont la fréquence est la plus élevée à l'intérieur des catégories sont les suivantes :

Catégorie 11 (action) : boire, boisson, breuvage, récolter, servir, produire, assister, participer.
Catégorie 6 (relations sociale) : cher, roi, offrir, donner, social, hôte.
Catégorie 2 (anatomie et processus vitaux) : main, pied[17], peau, sommeil, ivresse, maladie, médecine, mort.
Catégorie 7 (religion) : cérémonie, rite, mission, religion.
Catégorie 10 (règne végétal) : racine (de kava), plante.
Catégorie 14 (artefact) : coupe, plat.
Catégorie 5 (domaine cognitif) : tradition.
Catégorie 1 (êtres humains) : homme, femme.
Catégorie 13 (transformation) : voir tableau 7.
Catégorie 8 (géographie) : nakamal, maison, île.
Catégorie 4 (émotions et sentiments) : les unifs décrivant les effets narcotiques du kava.
Catégorie 9 (règne animal) : cochon.

[17] Les unifs main et pied sont polysémiques et servent d'exemple à l'analyse par classe présentée dans la section suivante.

En résumé, plus de 60% du contenu du corpus présente par ordre d'importance les actions, les relations sociales et les effets biologiques associés à la consommation rituelle et religieuse du kava.

3.3 L'analyse par classes et segments

La fréquence particulièrement élevée d'une unif à l'intérieur d'une catégorie est parfois un indice de polysémie, ce qui en fait un élément potentiel pour devenir un thème du bassin d'attraction. Dans ce cas, l'analyse par classes et segments se relève pertinente, car elle permet de définir la relation entre une unif et ses unifs associées et entre une classe et ses segments. Comme nous l'avons vu plus haut, la fréquence élevée des unifs «main» et «pieds» dans la catégorie 2, respectivement 38 et 17 fois, nous invite à consulter les classes et les segments où elles apparaissent.

Tableau 9 - Exemple de la fenêtre «Mot» de NUMEXCO

MOT
Choisir un mot
"pied"
Segments du mot, classe
11, 22 15, 1 40, 6
Liste des mots des segments
Bananier, horticulture, igname, main, tête
Segment [11] (Extrait) ... Au Centre de Pentecôte, où les ignames kokon et taribal (les jambes et les bras de SALDAM) occupent le coeur du jardin, les ignames vosran (la tête, les pieds et les mains de SALDAM) sont rejetées en périphérie.

Le tableau 9 présente un exemple de la fenêtre «Mot» de Numexco. Nous voyons que l'unif «pied» apparaît dans les segments 11, 15 et 40 qui sont associés respectivement aux classes 22, 1 et 6. Dans le segment 11 (classe 22), nous observons que l'unif «pied» est associée au jardin et non au corps humain.

Tableau 10 - Relations sémantiques des unifs «main», «pied» et «dent»

Cat	**Fréq**	**Unif**	**Relation sémantiques et adéquation classes/segments**
02a	38	**main**	**Analyse des classes**:

			Classe 1 libations (segment 29, unifs : boisson, coupe, coco) horticulture (segment 37, unifs : agriculture, fertilité, femme, homme, igname, jardin). Classe 35 maladie (segment 86, unifs abus, ivresse, maladie, peau). **Analyse des segments** : horticulture (main de l'homme, main de la femme) ; aménagement du jardin (tête, coeur, mains, pieds) ; gestes rituels lors des cérémonies (droite, gauche, haut, bas) ; relations sociales (mains royales, mains de servants) ; maladie (mains écailleuses).
02a	17	**pied**	**Analyse des classes** : Classe 1 horticulture (segment 15, unifs : horticulture, plant, récolte). Classe 6 plante (segment 40, unifs : racine, sec, souche). Classe 22 aménagement des jardins (segment 11, unifs : bananier, horticulture, igname, "main" et "tête" du jardin). **Analyse des segments** : effets du kava (abus, ivresse, maladie, peau se desquame de la "tête aux pieds" et devient douce) ; jardins (cultiver, igname, "jardins pieds de banane", "pied du jardin")
02a	5	**dent**	**Analyse des classes** : Classe 6 porc (segment 35, unifs : don, porc, sacrifice). Classe 16 cochon (segment 21, unifs : don, cochon, sacrifice, rituel). Classe 25 préparation du kava (segment 58, unifs: bouche, broyer, mâcher, préparé, procédé). Classe 32 cochon (segment 6, unifs : fête, sacrifier). Classe 35 maladie (segment 86, unifs : abus, immodéré, maladie, mastication). **Analyse des segments** : cochon (dents courbes = valeur, sacrifice), maladie (abus de kava = dents jaunes), préparation du kava (broyer avec les dents).

Le tableau 10 présente un exemple d'analyse des unifs polysémiques «main», «pied» et «dent» par classes et par segments» dans le but de faire émerger les relations sémantiques qu'elles entretiennent avec leurs unifs associées. Nous voyons que les classes contenant l'unif «main» l'associent au corps humain ou au «corps du jardin», tout comme l'unif

«pied», et que l'unif «dent» est associée au corps humain ou au thème «cochon» en temps qu'unité monétaire.

4. Présentation du graphe «Kava»

Le tableau et le graphe suivants ont été élaborés en tenant compte de la fréquence des unifs, des lemmes et des catégories sélectionnés pour leur adéquation avec le réseau sémantique relié à l'attracteur «Kava» à partir d'une analyse des fichiers mcl.xls ; cfu.xls ; seg.doc. Ces trois niveaux de sens (unif, lemme, catégorie) sont utilisés pour la construction du graphe, car c'est le poids sémantique d'un niveau de sens donné qui est significatif pour la construction des thèmes du graphe. Par exemple, les unifs/lemmes «cérémonie, rite, fête, danse, magie, mana» ont été réunies dans la catégorie «religion traditionnelle». Les unifs «chef, roi, maître» dans la catégorie «hiérarchie sociale». Lorsqu'une unif a un poids sémantique suffisant, c'est l'unif qui est utilisée. Par exemple, les unifs «coupe» et «plat» ont un sens particulier dans la cérémonie du kava qui les distinguent des unifs réunies sous la catégorie «ustensile» (récipient, mortier).

Pour la construction du graphe, les thèmes qui sont en connexion directe à l'attracteur (1er ordre) forment le premier cercle du graphe. Ex.: jardins, récolte, eau, préparation, bol, coupe, plat et autres ustensiles. Ces thèmes sont reliés aux aspects matériels et pragmatiques de l'attracteur, c'est-à-dire qu'ils représentent les conditions de l'existence matérielle de l'attracteur. Les thèmes qui sont en connexion médiatisée par un relais (2^{e} ordre), soit les effets et les contextes, sont regroupés sur le second cercle. Ex.: les effets (euphorie, intoxication, maladie) et les contextes (marchés, cadeau, cérémonie, chef, roi, visiteurs, maison des hommes). Enfin, les thèmes qui sont en connexion médiatisée par deux relais (3^{e} ordre) sont regroupés sur le troisième cercle. Il s'agit des thèmes reliés à la dimension symboliques du kava. Ex.: ancêtre, esprit, possession, mort. Le but est de générer des thèmes qui vont toucher à toutes les facettes du kava et aussi être en lien, s'il y en a, avec les autres graphes du site «Oceanie.org».

Le tableau 11 présente la fréquence des unifs, lemmes et catégories qui ont été sélectionnés pour le graphe. La fréquence permet de calculer automatiquement la distance sémantique du thème face à l'attracteur sans avoir à lire le corpus et à classifier les unités d'information.

Tableau 11 - Unifs, lemmes et catégories sélectionnés pour le graphe

Fréq	Unif	Cat
152	religion traditionnelle (cérémonie, rite, fête, danse, magie, mana...)	07a
139	hiérarchie sociale (chef, roi, maître...)	06a
111	jardin (plante, tige, tubercule, taro, champs...)	10a
102	relations sociales (hôte, famille, visiter, rassembler, alliance...)	06d
101	préparation (bouillir, broyer...)	13a
95	coupe	14a
79	christianisme (mission, monseigneur, ministre...)	07b
67	corps (anatomie)	02a
65	horticulture (récolter, planter, cultiver, défrichage...)	11b
64	intoxication (anesthésiant, narcotique, sudorifique...)	02c
49	médecine (pharmaceutique, guérir, remède...)	02e
48	effets (stupéfier, euphorie, stimuler, engourdissement...)	04
47	femme (fille)	01b
46	don (distribuer, cadeau, économie...)	06b
46	homme	01a
40	ustensiles (récipient, mortier...)	14a
38	main	02a
35	maladie (blennorragie, cystique...)	02d
27	tradition (coutume)	05
24	plat	14a
20	namakal	08a
16	mort (enterrer)	02f
14	maison	08a
14	village	08a
14	cochon (porc)	09a
13	peau	02a
11	sommeil (dormir)	02b
7	ancêtre	07a
6	sacrifice	07c
3	fécondité (fertiliser, sexe)	02b
1	vierge	01b

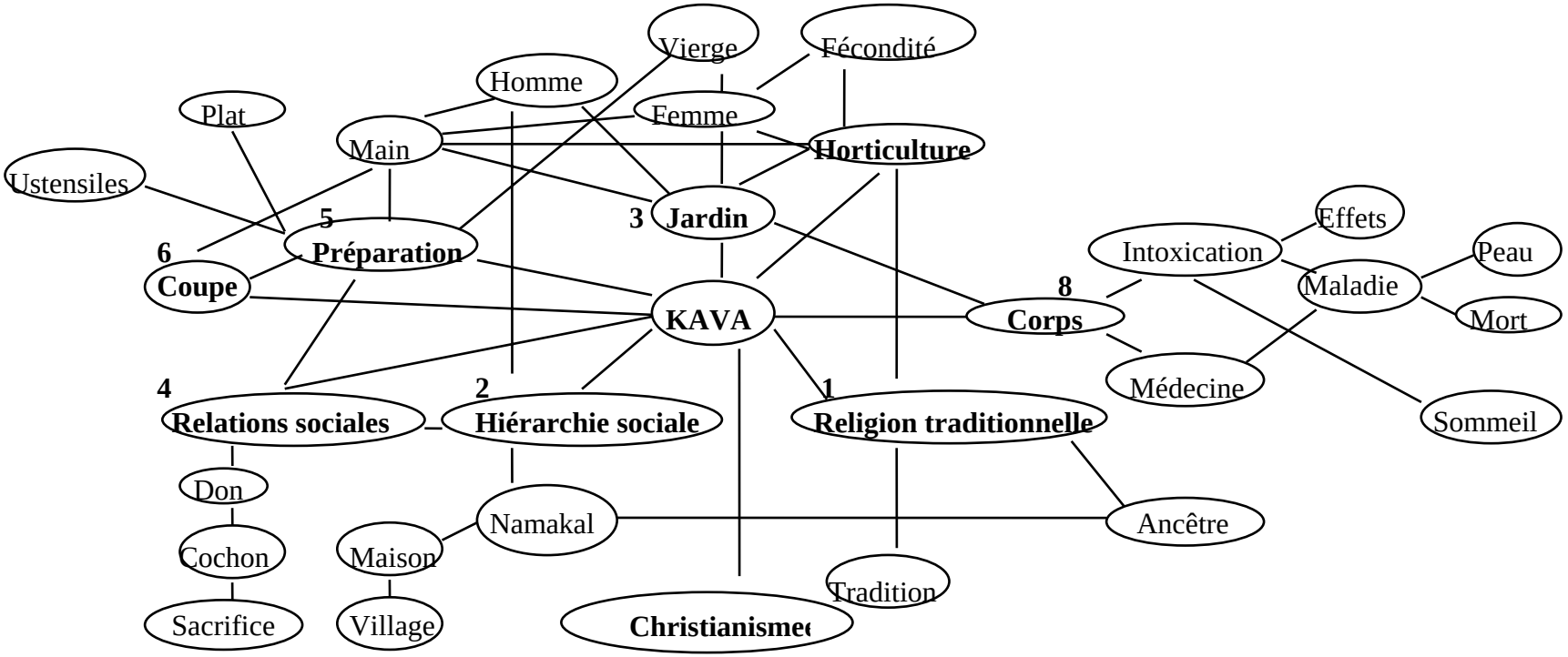


Figure 2 - Graphe «Kava»

Le graphe de la figure 2 présente le réseau et la proximité sémantique des thèmes (caractères gras et numérotés) de l'attracteur «Kava» à partir du calcul de la fréquence des unifs/lemmes/catégories. Bien qu'il s'inspire directement des graphes d'Ayotte et Lindstrom, il a été entièrement réalisé à partir des bases de données de Numexco. C'est-à-dire que le corpus n'a fait l'objet d'aucune lecture à l'exception de quelques références à certains segments dans le but de clarifier le contenu sémantique de certaines unifs, entre autres, les unif «main, pied, dent».

Dans ce graphe, nous voyons toute l'importance du champ religieux associé au kava. Les thèmes «religion traditionnelle» et «christianisme» se classent respectivement aux 1er et 7e rang, tandis que le graphe de Lindstrom ne fait aucune référence au christianisme, un thème qu'on retrouve toutefois dans celui d'Ayotte dont une partie du corpus utilisé contient des descriptions de cérémonies qui ont été faites par des missionnaires. On remarque également l'importance des unifs «coupe» (95 fois) et «plat» (34 fois). Cette fréquence élevée est un indice que ces unifs représentent quelque chose de plus important que de simples couverts. Ce thème est d'ailleurs présent dans le graphe de Lindstrom sous l'intitulé «*bowl*». Le thème «préparation», qui n'a pas été retenu par Ayotte et Lindstrom, est composé des unifs qui décrivent les étapes de l'extraction et de la fabrication du kava.
Dans les graphes de Lindstrom et d'Ayotte, présentés dans les annexes 1 et 2, les cercles rouges représentent les thèmes similaires à leur analyse et à celle réalisée à partir de Numexco. La similarité des thèmes concernant les effets de l'absorption du kava (médecine, maladie, intoxication, poison, mort) et du «rang social» et «jardin» dans le graphe d'Ayotte, est particulièrement frappante. Le corpus français «Kava» que j'ai utilisé ne fait toutefois pas référence à la mythologie, à la sexualité et à l'urbanité, thèmes que l'on retrouve dans le corpus anglais, ce qui explique la présence de quelques thèmes exclusifs aux graphes de Lindstrom et d'Ayotte.

Le graphe de la figure 2 présente le réseau et la proximité sémantique des thèmes (caractères gras et numérotés) de l'attracteur «Kava» à partir du calcul de la fréquence des unifs/lemmes/catégories. Bien qu'il s'inspire directement des graphes d'Ayotte et Lindstrom, il a été entièrement réalisé à partir des bases de données de Numexco. C'est-à-dire que le corpus n'a fait l'objet d'aucune lecture à l'exception de quelques références à certains segments dans le but de clarifier le contenu sémantique de certaines unifs, entre autres, les unif «main, pied, dent».

Dans ce graphe, nous voyons toute l'importance du champ religieux associé au kava. Les thèmes «religion traditionnelle» et «christianisme» se classent respectivement aux 1er et 7e rang, tandis que le graphe de Lindstrom ne fait aucune référence au christianisme, un thème qu'on retrouve toutefois dans celui d'Ayotte dont une partie du corpus utilisé contient des descriptions de cérémonies qui ont été faites par des missionnaires. On remarque également l'importance des unifs «coupe» (95 fois) et «plat» (34 fois). Cette fréquence élevée est un indice que ces unifs représentent quelque chose de plus important que de simples couverts. Ce thème est d'ailleurs présent dans le graphe de Lindstrom sous l'intitulé «*bowl*». Le thème «préparation», qui n'a pas été retenu par Ayotte et Lindstrom, est composé des unifs qui décrivent les étapes de l'extraction et de la fabrication du kava.

Dans les graphes de Lindstrom et d'Ayotte, présentés dans les annexes 1 et 2, les cercles rouges représentent les thèmes similaires à leur analyse et à celle réalisée à partir de Numexco. La similarité des thèmes concernant les effets de l'absorption du kava (médecine, maladie, intoxication, poison, mort) et du «rang social» et «jardin» dans le graphe d'Ayotte, est particulièrement frappante. Le corpus français «Kava» que j'ai utilisé ne fait toutefois pas référence à la mythologie, à la sexualité et à l'urbanité, thèmes que l'on retrouve dans le corpus anglais, ce qui explique la présence de quelques thèmes exclusifs aux graphes de Lindstrom et d'Ayotte.

5. Conclusion

En conclusion, nous voyons que l'analyse par attracteur donne accès à plusieurs niveaux d'analyse et d'interprétation. Si l'élaboration des graphes repose sur une bonne connaissance des données recueillies sur un sujet précis, les bases de données créées par Numexco donnent rapidement accès au contenu du corpus sans qu'il soit nécessaire de le lire en entier. Par le filtrage du lexique et la classification automatique, il favorise l'émergence des unifs les plus importantes (bassins d'attraction) et permet en tout temps de référer aux segments associés aux classes de termes. De plus, il permet d'approfondir l'analyse et l'interprétation des relations sémantiques (graphes) des attracteurs à être traités, en cours de traitement ou déjà traités, tout en facilitant l'interprétation par la schématisation des réseaux sémantiques.

Suite à l'expérimentation avec Numexco sur le corpus «Kava», nous voyons que les deux types d'analyse proposés offrent de multiples avantages à l'analyse des corpus et à l'élaboration des graphes. L'analyse par attracteur — qui consiste à analyser les classes dans lesquelles apparaît l'attracteur — permet d'identifier les thèmes qui forment le bassin d'un attracteur donné et de calculer plus objectivement, par le calcul de la fréquence des unifs/lemme/catégories sélectionnées, la proximité sémantique des unités de sens. Enfin, l'analyse par classe, comme nous l'avons vu avec l'exemple des unifs «main, pied, dent», facilite le repérage des différentes acceptions des termes polysémiques.

Bien que reconnaissant les avantages de Numexco pour le projet ECHO, Julie Ayotte souligne toutefois que les bases de données pour ECHO sont majoritairement en anglais, que Numexco ne traite que les textes en français, et qu'il faut connaître le fonctionnement du logiciel informatique. Le premier problème peut être résolu en utilisant GRAMEXCO, un autre logiciel de la plate-forme SATIM qui peut traiter des corpus multilingues à partir de n-grams[18] plutôt que de mots. Deuxièmement, les logiciels de la plate-forme SATIM sont relativement faciles à employer, un guide d'utilisation est disponible et une formation que quelques heures permet de les maîtriser. Les résultats dépendent toutefois d'une problématique de recherche explicite et de la créativité du chercheur. L'emploi de Numexco, à l'étape actuelle de la recherche pour les graphes PIROGUE, IGNAME et CORPS serait adéquat dans le sens où il permettrait de valider les graphes déjà construits. Enfin, son utilisation en début de recherche autour d'autres graphes ou attracteurs favoriserait un gain de temps important lors de l'élaboration des graphes. Et c'est justement dans ces cas que le traitement informatique d'un corpus numérisé offre de multiples avantages.

[18] Un n-gram est constitué d'un nombre prédéterminé de caractères.

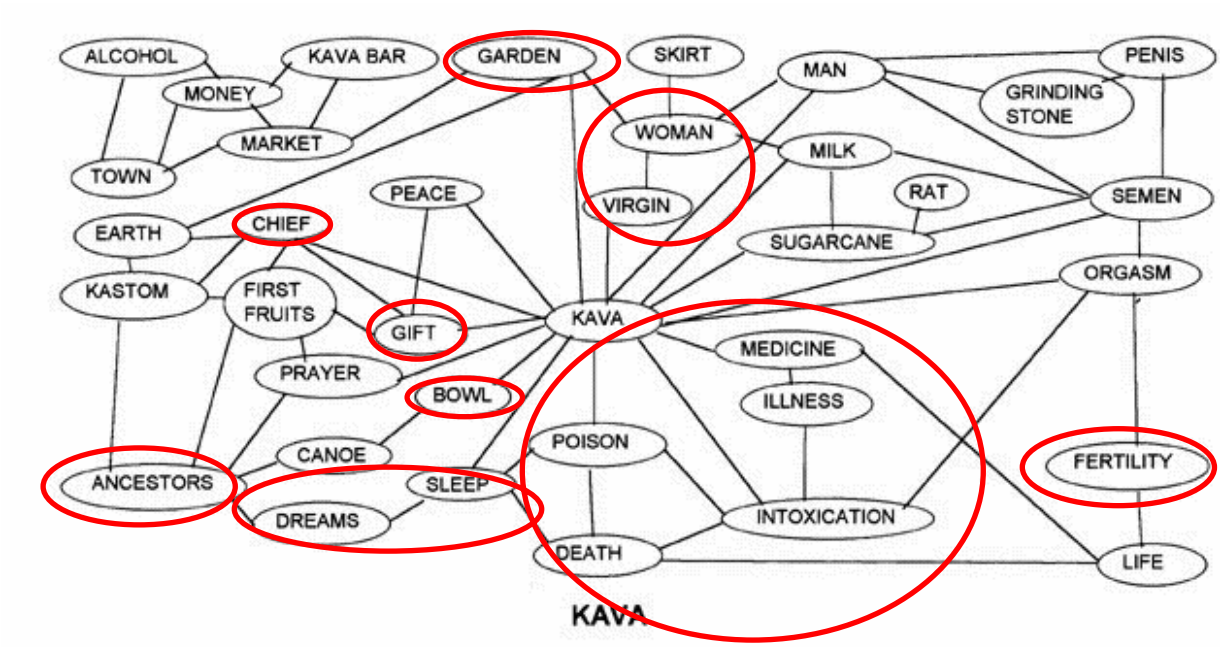
Remerciements
Je remercie Jean-Guy Meunier, Ismaïl Biskri et Dominic Forest du LANCI-UQAM, Pierre Maranda Julie Ayotte et Lamont Lindstrom du projet ECHO, et Sylvia Kasparian du Département d'études française de l'Université de Moncton pour leur précieuse collaboration. Merci également aux évaluateurs anonymes de cet article.

Références

[BISKRI 97] BISKRI Ismaïl, Jean-Guy MEUNIER et Georges NAULT, "Extraction des connaissances terminologiques au moyen des Grammaires Catégorielles : un modèle hybride", dans *Journées Scientifiques et techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF (JST FRANCIL)*, 1997.
[GAGNON 01] GAGNON Denis, 2001a, *Comparaison d'une catégorisation manuelle et d'une catégorisation assistée par ordinateur à l'aide de Numexco sur un corpus anthropologique*, Communication présentée au Séminaire du LANCI, Université du Québec à Montréal, manuscrit, 2001a.
[GAGNON 01] GAGNON Denis, 2001b, *L'analyse de contenu de corpus numérisés et la création de catégories assistée par ordinateur : enjeux, défis et possibilités*. Communication présentée au Congrès annuel de la Société canadienne d'anthropologie, (CASCA), Université McGill, manuscrit, 2001b.
[GAGNON 00] GAGNON Denis, "Application de CONTERM à un corpus ethnographique : relations de genre en Mélanésie", dans *Aladin texte : des stratégies informatiques génériques pour l'analyse et la lecture de texte assistées par ordinateur*, Rapport de recherche remis au FCAR - CEFRIO - CRIM par le LANCI-UQAM, 2000.
[GAGNON 99] GAGNON Denis, *Application de CONTERM à un corpus mélanésien*. Communication présentée à l'Institut des sciences humaines appliquées (ISHA), Université de la Sorbonne (Paris IV), dans le cadre du projet de coopération FRANCIL, manuscrit, 1999.
[GAGNON 02] GAGNON Denis et Jean-Guy MEUNIER, *Classification assistée par ordinateur et exploration par attracteurs sur des corpus ethnographiques*, 70e Congrès de l'ACFAS, Université Laval, 2002.
[JOUIS 97] JOUIS Christophe *et al*., "Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques", dans *Journées Scientifiques et techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF (JST FRANCIL)*, p. 427-432, 1997.
[MARANDA 94] MARANDA Pierre et Fidèle-Pierre NZE-NGUEMA, *L'unité dans la diversité culturelle : Une geste bantu*, Presses de l'Université Laval et Agence de coopération culturelle et technique, 1994.
[MEUNIER 97] MEUNIER Jean-Guy *et al*., "Exploration de modèles classifieurs connexionnistes pour l'analyse de textes assistée par ordinateur", *Actes du colloque LTT97*, Tunis, Tunisie, p. 289-296, 1997.

Annexes

Annexe 1 - Graphe de Lamont Lindstrom



Annexe 2 - Graphe de Julie Ayotte