Random walks in lexical small worlds

Bruno Gaume

Balades aléatoires dans les Petits Mondes Lexicaux

Bruno Gaume

IRIT-UPS, 118 route de Narbonne, F-31062 Toulouse cedex 4, France
gaume@irit.fr *http://dilan.irit.fr*

RÉSUMÉ : Cet article présente une méthode stochastique pour l'étude de la structure des grands graphes de terrain de type petits mondes hiérarchiques. Cette méthode consiste à transformer un graphe en une chaîne de Markov dont les états sont les sommets du graphe en question. Des particules se baladent aléatoirement de sommets en sommets dans le graphe en empruntant les arcs du graphe. Ce sont les dynamiques des trajectoires des particules qui nous donnent les propriétés structurelles des graphes étudiés.
Nous verrons que cette approche, qui est une forme de « connexionnisme structurel », permet de proposer en perspective une modélisation géométrique du sens où les petits mondes hiérarchiques sont non seulement une excellente compression de la forme du sens, mais de plus permettent une navigation et un accès très efficace à l'information, avec une dynamique d'acquisition du général vers le particulier par raffinement (enfant en cours d'acquisition de sa langue maternelle), ainsi qu'une excellente robustesse en cas de déficit (aphasie, apprenant), et un raisonnement à granularité variable ce qui permet de faire chuter la complexité.
Pour illustrer cette approche, des exemples et résultats concrets sont présentés sur des graphes d'origine linguistique, ce qui permet d'envisager un outil de navigation pour le web, dont l'ergonomie cognitive d'accès à l'information est une métaphore de l'acquisition du langage par les jeunes enfants.

MOTS CLEF *: graphe petit monde hiérarchique, chaîne de Markov, lexique, polysémie, synonymie, hyperonymie, métaphore, géométrisation du sens, visualisation, navigation, acquisition lexicale, web, mémoire, information*

Ces travaux (http://dilan.irit.fr/) sont soutenus et financés par : PRESCOT ; PROGRAMME COGNITIQUE : ECOLE & SCIENCES COGNITIVES ; P.I. SOCIETE DE L'INFORMATION ; A.C. SYSTEMES COMPLEXES EN SHS ; P.I. TRAITEMENT DES CONNAISSANCES, APPRENTISSAGE ET NTIC ; ACI JEUNES CHERCHEUSES ET JEUNES CHERCHEURS 2004.

« Dans un état de langue tout repose sur des rapports »

Saussure

1. Introduction

Notre approche dans cet article s'apparente à l'attitude de l'extra-terrestre qui trouvant un dictionnaire d'humain, essayerait d'en décrypter le sens. Or quelle attitude cet e.t. ignorant tout des humains, de leur état de développement cognitif, culturel et scientifique pourrait-il bien adopter ? Il pourrait bien sûr choisir l'approche du logicien. Deux possibilités s'offrent alors à lui :

– La première est de considérer ce dictionnaire en tant que théorie, comme un ensemble d'axiomes, dont il cherchera à savoir s'ils ne sont pas contradictoires, et quelles sont leurs conséquences logiques hormis les vérités universelles qui elles sont indépendantes du dictionnaire en question.

– La deuxième est de considérer le dictionnaire en tant que modèle (qui ne peut être contradictoire, par son existence même en tant qu'objet constitué) et de chercher à en établir une axiomatique adéquate.

Pour la première approche (l'approche théorique) il devrait auparavant définir quel est le langage, vocabulaire et syntaxe dans lesquels cette théorie s'inscrit, mais ceci est sans aucun doute une tache redoutable, même pour un e.t., d'autant plus qu'il ne connaît absolument rien des coutumes langagières d'efficacités Galiléennes « haut/bas » ou Boltzmanniennes [2] « chaud/froid » ou Mendeliennes « parent/enfant » … de la vie sur terre et encore moins celles de l'Académie Française.

Pour la deuxième approche (l'approche empirique) notre e.t. devra définir auparavant quels sont les objets pertinents de ce modèle. On peut penser qu'avec une approche distributionnaliste[3] (grâce aux séparateurs que sont les blancs, les virgules, les points, les retours à la ligne …) et un peu (ou beaucoup ?) de perspicacité il finira par comprendre que les entités de base y sont la définition et le mot. Une fois cela posé il devra ensuite au moins approximer les règles morphosyntaxiques afin d'établir les relations inter-mots du type « chevaux »→« cheval », « mange »→« manger » (voir par exemple (Gaussier 1999). En bref être capable de lemmatiser le dictionnaire qui est l'objet de son étude. Même si la tâche on le sait n'est pas simple, on peut là encore penser qu'une approche distributionnaliste devrait l'aider. Une fois le dictionnaire

[2] « … ''la probabilité d'un état'' si on veut pouvoir lui donner un sens suffisamment empirique, doit être compris comme ''le temps moyen que le système considéré passe dans cet état sur un intervalle de temps infini ou suffisamment grand'' » Boltzmann.

[3] L'analyse distributionnelle de corpus est une description basée sur l'inventaire de la distribution des constituants linguistiques (voir Harris 1951).

lemmatisé notre e.t. se trouve alors face à un objet formel qui est un grand graphe où les sommets sont les entrées du dictionnaire (les mots) avec un arc d'un sommet A vers un sommet B si et seulement si B est dans la définition de A. Que peut-il alors encore distiller de ce grand graphe ? C'est ce à quoi nous allons essayer de répondre dans cet article[4], en adoptant une démarche distributionnaliste, ce que notre e.t. ne manquerait sans doute pas de tenter vu les fruits que cette approche lui aura déjà révélés jusqu'à la construction de ce graphe de lemmes.

Dans ce cadre cet article propose une méthode d'étude de la structure des grands graphes de terrain[5] de type **R**éseaux **P**etits **M**ondes **H**iérarchiques (« Hierarchical Small Worlds » noté par la suite RPMH). Cette méthode consiste à transformer un graphe en une chaîne de Markov dont les états sont les sommets du graphe en question. Des particules se baladent aléatoirement de sommets en sommets dans le graphe en empruntant les arcs du graphe. Ce sont les dynamiques des trajectoires des particules qui nous donnent les propriétés structurelles des graphes étudiés.

Nous verrons que cette approche qui est une forme de 'connexionnisme structurel' propose des applications directes tant en Psychologie Cognitive, Linguistique, Statistique, Visualisation, et permet d'envisager en perspective une modélisation géométrique du sens où les RPMH sont non seulement une excellente compression de la forme du sens, mais de plus permettent une navigation et un accès très efficace à l'information, avec une dynamique d'acquisition du général vers le particulier par raffinement (enfant en cours d'acquisition de sa langue maternelle), ainsi qu'une excellente robustesse en cas de déficit (aphasie, apprenant de langue seconde), et un raisonnement à granularité variable ce qui permettra de faire chuter la complexité.

J'illustrerai mon propos par des exemples sur des graphes de terrain concrets d'origine linguistique du type de ceux qu'aurait peut-être pu construire notre e.t. à partir de son dictionnaire.

Le § 2. présente brièvement les notions de base dont nous aurons besoin pour parler des graphes. On décrit au § 3. une approche stochastique des graphes avec un théorème de convergence qui est une conséquence directe du théorème de Perron Froebenius. Le § 4. est une brève présentation du phénomène « graphe petit monde » ou « small world » et de leurs propriétés, que partagent la plupart des

[4] Notons que l'idée de tirer profit de ce réseau (considéré simplement comme une source textuelle structurée) a été exploitée par Véronis et Ide en 1990 (Veronis & Ide 1990) à travers un réseau de neurones pour la désambiguïsation, mais ces travaux n'ont pas répondu aux attentes de leurs auteurs. (La désambiguïsation consiste à reconnaître le sens d'un mot parmi ceux donnés par exemple dans un dictionnaire, ou bien à distinguer un mot parmi ses différents homographes).

[5] Les graphes de terrains sont les graphes que l'on trouve en pratique, ils sont construits à partir de données de terrains. On les retrouve dans toutes les sciences de terrain. Par exemple le graphe des collaborations scientifiques (les sommets sont les auteurs d'articles scientifiques, et on relie deux auteurs A et B s'ils ont au moins une publication en commun).

graphes de terrain issus des sciences sociales et humaines, des sciences de la vie, ou des sciences et techniques. Au § 5. nous nous focaliserons sur les graphes lexicaux d'origine linguistique qui sont eux aussi des graphes petits mondes, pour au § 6. y appliquer concrètement prox qui est une application directe de l'approche stochastique présentée précédemment au paragraphe 3. Au § 7. nous verrons comment prox nous donne à 'voir la forme' des grands graphes de terrain pour au § 8. brièvement discuter de sa complexité sur les « graphes petits mondes hiérarchiques » dont on peut avantageusement exploiter la structure, ce qui nous permettra d'envisager au § 9. plusieurs pistes de recherches théoriques mais aussi des applications dans différents domaines comme par exemple un outil de navigation pour le web, dont l'ergonomie cognitive est une métaphore de l'acquisition du langage par les jeunes enfants. Au § 10. nous essayerons en guise de conclusion d'amorcer très brièvement une réflexion quand à la démarche distributionnaliste de notre e.t. que l'on pourrait qualifier d'holiste.
Notre démarche étant pluridisciplinaire, (mathématique, informatique, linguistique, psychologie cognitive) nous essayerons dans cet article, tout en restant rigoureux pour chacune de ces sciences, de garder un langage qui nous l'espérons pourra être compris par les lecteurs de ces différentes disciplines, tout en indiquant quelques références qui pourraient sembler superflues pour les spécialistes d'un domaine particulier, mais qui pourraient s'avérer utiles pour les étudiants en sciences cognitives dont la formation de base peut être aussi bien mathématique qu'informatique, linguistique, psychologique, biologique, philosophique, …

Les non spécialistes pourront lire le § 3. sans s'attarder sur les détails techniques, dont les idées générales sont résumées et appliquées à la linguistique au §. 6. Dans ce paragraphe et le suivant, des exemples simples sur le lexique, dont nous avons tous au moins une intuition du fonctionnement par notre pratique quotidienne de la langue, nous permettront d'appréhender comment prox exploite la structure des petits mondes hiérarchiques dans leur généralité.

2. Les Graphes

Dans ce paragraphe nous présentons les notations que nous utiliserons pour parler des graphes. (Pour plus de détails sur les graphes voir par exemple (Diestel 2000, Bollobas 1986, 1998 ou Berge 1983) ou bien encore surfer par exemple à partir de : http://www.math.fau.edu/locke/graphthe.htm

2.1. Définition : graphe, arc entrant, arc sortant, degré

Un *graphe* $G=(V,E)$ est la donnée d'un ensemble non vide fini V de sommets, et d'un ensemble $E \subseteq (V \times V)$ de couples de sommets formant des arcs. On dit qu'un arc $(r,s) \in V$ est un *arc sortant* du sommet r, alors que c'est un *arc entrant* du sommets s.

Le *degré sortant* d'un sommet $r \in V$, est le nombre d'arcs sortants de r, alors que le *degré entrant* d'un sommet $r \in V$, est le nombre d'arcs entrants de r.

2.2. Définition : graphe réflexif

Un graphe G=(V,E) est dit *réflexif* lorsque : $\forall r \in V, (r,r) \in E$.

2.3. Définition : graphe symétrique

Un graphe G=(V,E) est dit *symétrique* lorsque : $\forall r,s \in V$, si $(r,s) \in E$ alors $(s,r) \in E$.

Quand un graphe est symétrique on parle de graphe non orienté et d'arête entre deux sommets r et s plutôt que d'arc. En effet, dans un graphe symétrique deux sommets quelconques $r,s \in V$, soit ne sont pas reliés $(r,s) \notin E$ et $(s,r) \notin E$, soit sont reliés dans les deux sens par les arcs $(r,s) \in E$ et $(s,r) \in E$ et on dit alors qu'ils sont reliés par une arête. Puisque qu'alors pour tout sommet $r \in V$, son degré sortant est égal à son degré entrant, on parle alors de degré d'incidence d'un sommet.

2.4. Définition : chemin de longueur m

Soit G=(V,E) un graphe. Pour tout entier naturel $m \neq 0$, un *chemin de longueur m dans G* est un (m+1)-uplet $c=\langle r_0,\ldots,r_m \rangle$ tel que $\forall i, 0 \leq i < m$: $(r_i,r_{i+1}) \in E$, r_0 en étant l'origine et r_m l'arrivée.
Pour tout chemin $c=\langle r_0,\ldots,r_m \rangle$ de G on posera long(c)=m la longueur du chemin c.

2.5. Définition : graphe fortement connexe

Soit G=(V,E) un graphe. Nous dirons que G est *fortement connexe* si et seulement si : $\forall r,s \in V$, il existe un chemin c de longueur finie dans G dont r est l'origine et s l'arrivée.

2.6. Définition : $C^k_{r,s}$

Soit G=(V,E) un Graphe. Pour tout entier naturel k>0, $\forall r,s \in V$, posons $C^k_{r,s}$ l'ensemble des chemins de G de longueur égale à k ayant r pour origine et s pour arrivée.

2.7. Définition : [G] matrice d'adjacence d'un graphe

Soit un Graphe à n sommets G=(V,E), on notera [G] la Matrice carrée nxn telle que pour tout $r,s \in V$, $[G]_{r,s}=1$ si $(r,s) \in E$ et $[G]_{r,s}=0$ si $(r,s) \notin E$; On appellera [G] la *matrice d'adjacence de G.*

Notation : Quand M une matrice nxm (de n lignes par m colones), et r tel que $1 \leq r \leq n$ et s tel que $1 \leq s \leq m$ on note $[M]_{r,s}$ la valeur située à la $r^{ième}$ ligne et à la $s^{iéme}$

colonne de la matrice M, et on notera dans la suite $[M]_r$ le vecteur ligne ($[M]_{r,1}$, $[M]_{r,2}$, …, $[M]_{r,m-1}$, $[M]_{r,m}$) et $[M]^s$ le vecteur colonne ($[M]_{1,s}$, $[M]_{2,s}$, …, $[M]_{n-1,s}$, $[M]_{n,s}$) . Quand V est un vecteur ligne de dimension n et r tel que $1 \leq r \leq n$ on notera V_r la $r^{ième}$ valeur du vecteur V.

3. Balades aléatoires dans les graphes

Supposons que nous disposions d'un graphe G=(V,E), et que dans ce graphe une particule puisse à chaque instant $i \in \mathbb{N}$ se balader aléatoirement sur ses sommets :

– A l'instant i=1 la particule est sur un sommet $r \in V$.

– Quand la particule est à un instant i sur un sommet $u \in V$, elle ne peut atteindre à l'instant i+1, que les sommets $s \in V$ tels que $(u,s) \in E$. La particule se déplace de sommets en sommets à chaque instant en empruntant les arcs du graphe. On supposera de plus que pour tout sommet $u \in V$, chacun des arcs sortant de u est équiprobable.

Ce sont les trajectoires de la particule dans le graphe qui vont nous permettre d'étudier la structure du graphe. Pour cela nous allons construire une chaîne de Markov homogène[6] à partir d'un graphe, où les états de la chaîne de Markov seront les sommets du graphe.

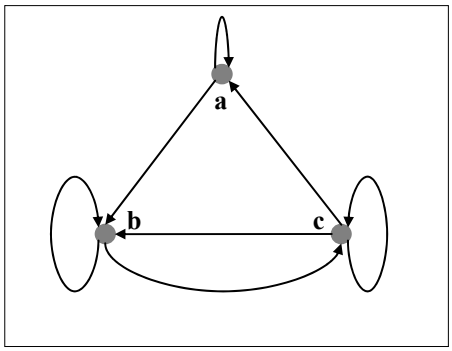
3.1. Définition : $\hat{G}$ la matrice Markovienne d'un graphe G réflexif

Soit G=(V,E) un graphe réflexif à n sommets. Posons $\hat{G}$ la matrice $n_x n$ à coefficients réels positifs ou nuls définie par : $\forall r,s \in V$, $[\hat{G}]_{r,s} = [G]_{r,s} / \sum_{x \in V} \{[G]_{r,x}\}$.
Nous dirons que *$\hat{G}$ est la matrice Markovienne du graphe G*.

Cette définition a bien un sens car $\forall r,s \in V$, $[G]_{r,s} \geq 0$ et de plus si le graphe est réflexif, $\forall r \in V$, $[G]_{r,r} = 1$, donc $\forall r \in V$, $\sum_{x \in V} \{[G]_{r,x}\} > 0$.

[6] Dans une chaîne de Markov la probabilité de passer de l'état i à l'état j au $n^{ème}$ coup ne dépend que de i et de j et de n (le futur ne dépend que du présent et pas du passé). Si on suppose de plus que la probabilité de passer de l'état i à l'état j au $n^{ème}$ coup ne dépend que de i et de j et pas de n (le futur ne dépend que des conditions initiales) la chaîne de Markov est alors dite homogène. Pour une bonne introduction contemporaine aux chaînes de Markov et leurs applications aux sciences sociales : (Berchtold 1998), ou plus spécialisés (Brémaud 2001) ou encore sur les matrices non-négatives (Bermann & Plemons 1994)

Exemple :



[G]	a	b	c
a	1	1	0
b	0	1	1
c	1	1	1

$\hat{G}$	a	b	c
a	½	½	0
b	0	½	½
c	⅓	⅓	⅓

Figure 1 *un graphe G réflexif et $\hat{G}$ sa matrice Markovienne*

3.2. Propriété

Soit G=(V,E) un graphe réflexif à n sommets, alors $\hat{G}$ sa matrice Markovienne est telle que :

1° : $\hat{G}$ est une matrice non négative, c'est-à-dire que $\forall r,s\in V, [\hat{G}]_{r,s}\geq 0$;

2° : $\hat{G}$ est une matrice stochastique , c'est-à-dire que $\forall r\in V, \Sigma_{x\in V}\{[\hat{G}]_{r,x}\}=1$.

□

1° : D'après la déf. 2.7. $\forall r,s\in V, [G]_{r,s}\geq 0$. D'autre part si le graphe G est réflexif alors d'après la déf. 2.2. $\forall r\in V, [G]_{r,r}=1$, d'où $\forall r\in V, \Sigma_{x\in V}\{[G]_{r,x}\}>0$ et donc d'après la déf. 3.1., $\forall r,s\in V, [\hat{G}]_{r,s}=[G]_{r,s}/\Sigma_{x\in V}\{[G]_{r,x}\}\geq 0$, d'où $\hat{G}$ est une matrice non négative.

2° : De plus $\forall r\in V, \Sigma_{x\in V}\{[\hat{G}]_{r,x}\}=$ *(puisque $\forall r\in V, \Sigma_{x\in V}\{[G]_{r,x}\}>0$, et la déf. 3.1.)*
$\Sigma_{x\in V}\{[G]_{r,x}/\Sigma_{y\in V}\{[G]_{r,y}\}\}=$
$\Sigma_{x\in V}\{[G]_{r,x}\}/\Sigma_{y\in V}\{[G]_{r,y}\}=1$
d'où $\hat{G}$ est une matrice stochastique.

■

3.3. Définition : $({}^{F}X_i)_{i\in\mathbb{N}}$ la chaîne de Markov du graphe G débutant équiprobablement sur F.

Soit G=(V,E) un graphe réflexif à n sommets et $\hat{G}$ sa matrice Markovienne.
Pour tout $F\neq\varnothing$, $F\subseteq V$, posons $({}^{F}X_i)_{i\in\mathbb{N}}$ la chaîne de Markov homogène dont l'espace d'état fini est V, définie par la suite des lois des variables $({}^{F}X_i)_{i\in\mathbb{N}}$ que sont les vecteurs lignes $({}^{F}P_i)_{i\in\mathbb{N}}$:

1) $\forall x\in V, [{}^{F}P_0]_x=1_{(x\in F)}/|F|$
2) $\forall x\in V, \forall i\in\mathbb{N}^*, [{}^{F}P_i]_x=P({}^{F}X_i=x)=[{}^{F}P_0\bullet\hat{G}^i]_x$

avec les notations : $1_{(H)}=1$ si H est vrai et $1_{(H)}=0$ si H est faux ; ainsi que : |F| est égal au cardinal de l'ensemble F. D'autre part • est la multiplication matricielle.

Nous dirons que $({}^{F}X_i)_{i\in\mathbb{N}}$ est la chaîne de Markov du graphe G débutant équiprobablement sur F.

Cette définition a bien un sens car :
1) Au premier coup, le vecteur ligne ${}^{F}P_0$ qui est la loi initiale de $({}^{F}X_i)_{i\in\mathbb{N}}$ définit bien une probabilité sur V, (c'est celle d'aller équiprobablement sur l'un quelconque des sommet de F) et
2) d'après prop. 3.2. $\hat{G}$ est une matrice non négative et stochastique.

Pour tout $F\neq\varnothing$, $F\subseteq V$, $\hat{G}$ est donc la matrice de transition de la chaîne de Markov homogène $({}^{F}X_i)_{i\in\mathbb{N}}$ et G sont graphe des transitions possibles[7].

3.4. Propriété

Soit G=(V,E) un graphe à n sommets, fortement connexe et réflexif. Soit $\hat{G}$ sa matrice Markovienne. Alors $\forall r\in V$ $({}^{\{r\}}X_i)_{i\in\mathbb{N}}$ la chaîne de Markov du graphe G débutant par r est telle que : $\forall s\in V$, $\forall i\in\mathbb{N}^*$, $[{}^{\{r\}}P_i]_s=[\hat{G}^i]_{r,s}$.

□
$[{}^{\{r\}}P_i]_s=$ *(par la déf. 3.3.car $({}^{\{r\}}X_t)_{t\in\mathbb{N}}$ est homogène et $\hat{G}$ sa matrice de transition)*
$[{}^{\{r\}}P_0\bullet\hat{G}^i]_s=$*(par la déf. 3.3.car $\forall(x\neq r)\in V$, $[{}^{\{r\}}P_0]_x=0$ et $[{}^{\{r\}}P_0]_r=1$)*
$[\hat{G}^i]_{r,s}$
■

3.5. Définition : Chaînes de Markov ergodiques

Soit $(X_i)_{i\in\mathbb{N}}$ une chaîne de Markov sur S homogène, et π sa matrice de transition.
- La chaîne $(X_i)_{i\in\mathbb{N}}$ est dite ergodique si :

$\exists k\in\mathbb{N}$ tel que $\forall r,s\in S$, $\forall i\geq k$, $[\pi^i]_{r,s}>0$ (1)

Une autre façon de le dire est :
- La chaîne $(X_i)_{i\in\mathbb{N}}$ est dite ergodique si pour son graphe des transitions possibles on a :

$\exists k\in\mathbb{N}^*$ tel que $\forall r,s\in S$, $\forall i\geq k$, $C^i_{r,s}\neq\varnothing$ (2)

On vérifie en effet que (1) si et seulement si (2).

[7] Le graphe des transitions possibles d'une chaine de Markov sur un ensemble d'états finis E dont π est sa matrice de transitions est le graphe G=(V,E) tel que : $\forall r,s\in V$, $(r,s)\in E$ si et seulement si $\pi_{r,s}\neq 0$.

Exemple :
Le graphe de la Figure 1. vérifie (2), en effet $\forall i \geq 2$, $\forall r,s \in S$, $C^i_{r,s} \neq \varnothing$.

Remarque : la formule (2) implique que le graphe des transitions possibles est fortement connexe, mais attention la réciproque est fausse.

Contre Exemple :

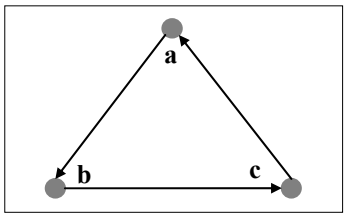


Figure 2 *G=({a,b,c}, {(a,b), (b,c), (c,a)}) est fortement connexe sans vérifier (2)*

Dans le graphe de la Figure 2. on a en effet que $\forall k \in \mathbb{N}^*$, $\exists r,s \in V$, $C^k_{r,s} = \varnothing$.
(par exemple $\forall k \in \mathbb{N}$, $C^{3k+2}_{a,b} = C^{3k+3}_{a,b} = \varnothing$)

3.6. Propriété

Soit G=(V,E) un graphe à n sommets, fortement connexe et réflexif. Soit $\hat{G}$ sa matrice Markovienne. Alors $\forall F$, $F \neq \varnothing$, $F \subseteq V$, $({}^F X_i)_{i \in \mathbb{N}}$ la chaîne de Markov du graphe G débutant équiprobablement sur F est une chaîne de Markov ergodique.
□
Montrons qu'en effet si G est fortement connexe et réflexif alors $\exists k \in \mathbb{N}$ tel que $\forall r,s \in V$, $\forall i \geq k$, $C^i_{r,s} \neq \varnothing$, et donc d'après la déf. 3.5.on aura montré que $({}^F X_i)_{i \in \mathbb{N}}$ est une chaîne de Markov ergodique.

Si G est fortement connexe alors d'après la déf. 2.5., $\forall r,s \in V$, il existe un chemin c dans G dont r est l'origine et s l'arrivée. Donc $\forall r,s \in V$, $\exists k \in \mathbb{N}$ tel que $C^k_{r,s} \neq \varnothing$. Posons $d_{r,s} = \min\{k \in \mathbb{N}^*, C^k_{r,s} \neq \varnothing\}$. Or puisque le graphe G est réflexif il est clair que $\forall k \in \mathbb{N}$ tel que $k \geq d_{r,s}$, $C^k_{r,s} \neq \varnothing$. Posons $d = \mathrm{Max}_{r,s \in V}\{d_{r,s}\}$ qui existe toujours puisque V est fini. Il est alors claire que $\forall r,s \in V$, $\forall i \geq d$, $C^i_{r,s} \neq \varnothing$. (Notons que d est le diamètre du Graphe G et que c'est le plus petit entier $\in \mathbb{N}^*$ vérifiant la propriété)
■

3.7. Théorème

Soit $(X_t)_{t \in \mathbb{N}}$ une chaîne de Markov ergodique sur l'espace d'états S, et π sa matrice de transition.

Alors il existe une unique probabilité stationnaire $[L]_{y\in S}$, c'est-à-dire que $L\bullet\pi=L$.
De plus, on a :

$$\forall r,s\in S,\ \lim_{n\to\infty}[\pi^n]_{r,s}=[L]_s$$

On trouvera une démonstration du thé. 3.7. dans (Senata 1981 ou Bermann & Plemons 1994). Le thé. 3.7. est un corollaire du Théorème de Perron Froebenius (Senata 1981).

3.8. Propriété

Soit G=(V,E) un graphe à n sommets, fortement connexe et réflexif. Soit $\hat{G}$ sa matrice Markovienne. Pour tout F, $F\neq\varnothing$, $F\subseteq V$, $\hat{G}$ étant la matrice de transition de la chaîne $({}^{F}X_i)_{i\in\mathbb{N}}$, il existe alors une unique probabilité stationnaire $[L]_{y\in V}$, c'est-à-dire que $L\bullet\hat{G}=L$.
De plus, on a :

$$\forall r,s\in V,\ \lim_{i\to\infty}[\hat{G}^i]_{r,s}=[L]_s$$

□
C'est une conséquence immédiate du thé. 3.7. et de la prop. 3.6.
■

Mais alors puisque pour tout sommet r de V, et la chaîne de Markov $({}^{\{r\}}X_i)_{i\in\mathbb{N}}$ d'un graphe G on a par la prop. 3.4. que $\forall s\in V,\ \forall i\in\mathbb{N},\ [{}^{\{r\}}P_i]_s=[\hat{G}^i]_{r,s}$, on a donc par la prop. 3.8. que $\forall r,s\in V,\ \lim_{i\to\infty}[{}^{\{r\}}P_i]_s=\lim_{i\to\infty}[\hat{G}^i]_{r,s}=[L]_s$

C'est à dire qu'en partant du sommet r, la probabilité $[{}^{\{r\}}P_i]_s=[{}^{\{r\}}P_0\bullet\hat{G}^i]_s=[\hat{G}^i]_{r,s}=[L]_s$ pour un temps i assez long[8] de se trouver sur un sommet s ne dépend plus du sommet de départ r, mais qu'uniquement du sommet s. ($\forall r,u\in V,\ \forall s\in V,\ \lim_{i\to\infty}[{}^{\{r\}}P_i]_s=\lim_{i\to\infty}[{}^{\{u\}}P_i]_s=[L]_s$)

4. Les propriétés des graphes de terrain

Dans ce paragraphe nous présentons brièvement la notion de Réseaux Petits Mondes Hiérarchiques (RPMH). Pour une présentation des RPMH voir par exemple (Newman 2003a, Watts 1999).

Watts et Strogatz (Watts & Strogatz 1998) ont les premiers montré que la plupart des grands graphes de terrain qui nous intéressent ici ne ressemblent ni aux graphes aléatoires ni aux graphes réguliers. Les grands graphes de terrain, bien que très peu denses, possèdent une connectivité très 'resserrée'. Cela signifie que ces graphes ont une topologie bien particulière, dans laquelle la relation entre structure locale et

[8] En fait ce temps est assez court dans les RPMH comme l'illustrent la Figure 11

structure globale n'a rien à voir avec celle des graphes (aléatoires ou réguliers) classiquement étudiés en théorie des graphes. Ceci explique l'intérêt considérable que ces résultats ont suscité dans les communautés scientifiques concernées. En effet, on peut penser que ces caractéristiques reflètent les propriétés spécifiques des systèmes dont ces grands graphes de terrain rendent compte, et donc que l'étude de leurs structures permettra une meilleure compréhension des phénomènes dont ils sont issus, mais aussi une meilleure exploitation des données ainsi représentées : traitement, modélisation, structuration, indexation, accès à l'information, classification, extraction de sens, visualisation …

Les premières investigations concernant des graphes de grande taille moins réguliers que les graphes 'de laboratoire' sont dues à Erdos & Rényi 1960, qui ont introduit et étudié la notion de graphe aléatoire (les arcs sont aléatoirement déterminés, entre deux sommets r et s du graphe l'arc (r,s) existe ou n'existe pas suivant une probabilité *p*) en tant que modèles pour les graphes dits de 'terrain' : graphes de grande taille (plusieurs milliers de sommets et d'arcs) issus de la biochimie, de la biologie, de la technologie, de l'épidémiologie, de la sociologie, de la linguistique...

Depuis, des recherches récentes en théorie des graphes ont mis au jour un ensemble de caractéristiques statistiques que partagent la plupart des graphes de terrain ; ces caractéristiques définissent la classe des graphes de type RPMH. Ainsi en va-t-il du réseau des interactions protéiques de certaines levures (Jeong & al. 2001), du réseau neuronal du ver *Caenorhabditis elegans* (Watts & Strogatz 1998), du graphe d'internet ($\approx 10^9$ sommets) (Barabasi & al. 2000), du graphe des appels téléphoniques d'une journée aux USA (Abello & al. 1999), de graphes épidémiologiques (Ancel & al. 2001), du graphe des co-auteurs scientifiques (Redner 1998) ou des collaborations cinématographiques (Watts & Strogatz 1998), ou de réseaux lexicaux tirés de WordNet (Sigman & Cecchi 2002) ou des cooccurrences dans un corpus de textes (Ferrer & Solé 2001) …
Ces graphes, comme la plupart des graphes de terrain, sont peu denses, c'est-à-dire qu'ils ont relativement peu d'arcs au regard du nombre de leurs sommets. Dans un graphe à n sommets, le nombre maximum d'arcs possibles est de n^2. En général le nombre d'arcs des grands graphes de terrain est O(nlog(n)) et non $O(n^2)$. Par exemple, le graphe des collaborations cinématographiques[9] possède 13 millions d'arêtes, ce qui peut paraître beaucoup, mais ce qui est très peu par rapport au carré du nombre de ses sommets ($225000^2 \approx 5x10^{10}$).

[9] Les 225 000 acteurs du syndicat du cinéma américain en sont les sommets et il existe un arc (r,s) si et seulement si l' acteur représenté par le sommet r a joué avec l'acteur représenté par le sommet s (Ce graphe est donc réflexif et symétrique).

Watts, (Watts 1999) et Strogatz (Watts & Strogatz 1998) proposent deux indicateurs pour caractériser un grand graphe G non orienté peu dense : son L et son C.

– L= moyenne des plus courts chemins entre deux sommets de G

– **C**= le taux de clustering[10] ou d'agrégation, est défini de la manière suivante : Supposons qu'un sommet S ait K_s voisins, alors il y a $K_s(K_s-1)/2$ arêtes au maximum qui peuvent exister entre ses K_s voisins (ce qui arrive quand chacun des voisins de S est connecté à tous les autres voisins de S). Soit A_s le nombre d'arêtes qu'il y a entre les voisins de S (ce nombre est donc nécessairement plus petit ou égal à $K_s(K_s-1)/2$). Posons $C_s= A_s/(K_s(K_s-1)/2)$

Le C de G est la moyenne des C_s sur les sommets de G. Le C d'un graphe est donc toujours compris entre 0 et 1. Plus le C d'un graphe est proche de 1, plus il forme des agrégats ou clusters (des zones denses en arêtes – « mes amis sont amis entre eux »). En appliquant ces critères à différents types de graphes ils constatent que :

1) les graphes de terrain ont tendance à avoir un L petit (en général il existe au moins un chemin très court entre deux sommets quelconques).

2) Les graphes de terrain on tendance à avoir un grand C, ce qui reflète la tendance qu'ont deux voisins d'un même sommet à être connectés entre eux par une arête. Par exemple dans le world wide web[11], deux pages qui sont liées à une même page ont une probabilité relativement élevée d'inclure des liens l'une vers l'autre. Même si les arcs ne sont pas uniformément distribués (comme dans les graphes réguliers) ils tendent à former des agrégats (zones denses en arcs).

3) Les graphes aléatoires ont un petit L. Lorsque l'on construit de manière aléatoire un graphe ayant une densité en arcs comparable aux grands graphes de terrain, on obtient des graphes dont le L est petit.

4) Les graphes aléatoires ont un C faible : ils ne sont pas formés d'agrégats. Dans un graphe aléatoire il n'y a aucune raison pour que les voisins d'un même sommet aient plus de chance d'être connectés que deux sommets quelconques, d'où leur faible tendance à former des agrégats.

---

[10] Ceci est la définition du C d'un graphe proposée dans (Watts & Strogatz 1998), mais en fait il nous semble que cette définition, bien qu'allant dans la bonne direction, pose encore quelques problèmes. L'idée directrice est de mesurer la tendance d'un graphe à former des cycles courts, or le C tel que défini plus haut mesure seulement la tendance à former des cycles de longueur 3 (Newman 2003a, 2003b). Nous pensons que le C peut être avantageusement remplacé par le R d'un graphe G=(V,E) à n sommets défini de la manière suivante : $\forall s\in V$, posons $T_s=1/(\min\{i\in\mathbb{N}^*, {}^{\{s\}}X_i=s\})$, (où $\min\{i\in\mathbb{N}^*, {}^{\{s\}}X_i=s\}$, est donc le temps qu'une particule met en partant de s pour revenir sur s, on a donc $T_s\in[0,1]$). Posons $R_s=E(T_s)$, et $R=\sum_{s\in V}\{R_s\}/n$. (R est donc la moyenne des moyennes des T_s que sont les R_s).

[11] Les sommets en sont les $\approx 10^9$ de pages disponibles sur internet, et un arc est tracé de r vers s si un lien hypertexte vers la page s apparaît dans la page r.

5) Les graphes réguliers[12] ont un L grand (en général il n'existe que des chemins longs entre deux sommets quelconques).

6) Les graphes réguliers ont un C fort : ils sont formés d'agrégats (du fait de leur régularité).

Les points 1 à 6 peuvent être résumés par le Tableau 1

	C Taux de clustering	**L Moyenne des chemins**
Graphes aléatoires	C petit (pas d'agrégats)	L petit (chemins courts)
Graphes de terrain	**C grand (des agrégats)**	**L petit (chemins courts)**
Graphes réguliers	C grand (des agrégats)	L grand (chemins longs)

Tableau 1 *Graphes de terrain, entre ordre et désordre*

Nous voyons bien là que les graphes de terrain sont entre ordre (par leur C grand), et désordre (par leur L petit).

Watts et Strogatz (Watts & Strogatz 1998) proposent d'appeler **« small worlds »**[13] « petits mondes » les graphes qui ont cette double caractéristique qu'ils identifient dans tous les graphes de terrain qu'ils observent, et dont ils postulent l'universalité. Les graphes d'origine linguistique que nous étudions sont en effet de type *petits mondes* (graphes peu denses, présentant une structuration locale riche – un C fort – et une distance moyenne entre deux sommets très petite sur l'ensemble du graphe – un L faible –).

Des travaux plus récents (Ravasz, & Barabasi 2003) montrent que la plupart des graphes petits mondes, dont les graphes d'origine linguistique, ont de plus une structure hiérarchique. La distribution des degrés d'incidence des sommets suit une loi de puissance « power law », certains nœuds très peu nombreux ayant beaucoup plus de voisins que d'autres, eux-mêmes ayant plus de voisins que d'autres qui eux-mêmes… La probabilité P(k) qu'un sommet du graphe considéré ait k voisins décroît comme une loi de puissance $P(k) = k^{-\lambda}$ (Barabasi & al. 1999, Kleinberg & al. 1999, Adamic 1999, Huberman & Adamic 1999) où λ est une constante

[12] Les graphes réguliers sont des graphes classiquement étudiés en théorie des graphes, tous leurs sommets ont le même degré d'incidence (le même nombre de voisins).

[13] En écho au « small world phenomenon » (Milgram 1967, Kochen 1989, Guare 1990) selon lequel deux personnes A et B sont en relation dans le graphe si A entretient tel ou tel type de relation avec B (A connaît B, A est en contact régulier avec B, A a travaillé dans la même entreprise que B, …). Ces graphes ont été popularisés par le slogan « *six degrees of separation* » (Guare 1990) : pour certains de ces graphes à l'échelle de la planète la longueur moyenne du chemin entre deux humains serait de l'ordre de 6, ce qui est très petit.

caractéristique du graphe, alors que dans le cas des graphes aléatoires, c'est une loi de Poisson qui est à l'œuvre.

Or, des travaux linguistiques et psycholinguistiques (Duvignau 2002, Duvignau 2003, Duvignau & Gaume 2003, Duvignau & al. 2004a, Duvignau & 2004b) qui mettent au jour une organisation du lexique des verbes par cohyponymie[14] intra *vs* interdomaines à partir de l'analyse d'énoncés spontanés corroborent cette propriété supplémentaire :

– aspect petits mondes par le rôle de la cohyponymie intra et interdomaine : {DESHABILLER, DENUDER, PELER, PLUMER, DECORTIQUER, ECAILLER, ECORCHER, DEPIAUTER, EBRANCHER, …} (fort degré d'agrégation/clustering)

– aspect hiérarchique par le rôle de l'hyperonymie : (ÔTER,DESHABILLER), (ÔTER,DENUDER), …, (ÔTER,DEPIAUTER), (ÔTER,EBRANCHER), … (Ôter a un fort degré d'incidence)

A la suite des travaux de Watts et Strogatz (Watts 1999), beaucoup d'articles sont parus où sont analysées les structures de différents graphes de terrain dans les domaines les plus divers (sciences sociales, sciences de la vie, sciences de l'ingénieur), mais les études de graphes d'origine linguistique restent encore trop peu nombreuses.

5. Les graphes lexicaux

Il existe plusieurs types de réseaux lexicaux, suivant la nature de la relation sémantique qui définit les arcs du graphe (les sommets représentant les unités lexicales d'une langue – de quelques dizaines de milliers à quelques centaines de milliers d'éléments, suivant la langue et la couverture du corpus utilisé). Les trois principaux types de relations utilisées sont les suivantes :

– **Relations syntagmatiques**, ou plutôt de cooccurrence ; on construit une arête entre deux mots si on les trouve dans un grand corpus au voisinage l'un de l'autre (typiquement à une distance maximale de deux/trois mots ou plus) cf. (Karov & Edelman 1998, Lebart & Salem 1994).

[14] Cohyponymes : mots dont les sens sont inclus dans le sens d'un autre mot dit hyperonyme. «déshabiller» et «éplucher» sont deux cohyponymes interdomaine de l'hyperonyme «dépouiller»; alors que «éplucher» et «peler» en sont des cohyponymes intradomaine (le végétal)

– **Relations paradigmatiques**, notamment de synonymie ; à partir de bases de données lexicales, comme le célèbre WordNet (Fellbaum 1999), on construit un graphe dans lequel deux sommets sont reliés par une arête si les mots correspondants entretiennent une relation synonymique (Ploux & Victorri 1998) [http://www.crisco. unicaen.fr. ou http://dico.isc.cnrs.fr/dico/fr/chercher]

– **Relations de proximité sémantique** ; il s'agit de relations moins spécifiques qui peuvent prendre en compte à la fois l'axe paradigmatique et l'axe syntagmatique. Nous avons construit un graphe du lexique du Français, en définissant les arêtes de la manière suivante : on construit une arête entre un verbe A et B si l'un est dans la définition de l'autre dans un dictionnaire général (construction du même type que celle décrite dans Veronis & Ide 1990); comme une entrée de dictionnaire général comporte la plupart du temps des définitions, des exemples, des synonymes, et même des antonymes, les arêtes sont alors étiquetées par le type de relation qu'elles représentent : on peut donc, selon les besoins, restreindre le graphe à certaines combinaisons de relations : syntagmatiques et/ou paradigmatiques et/ou même logico sémantiques.

Tous ces graphes sont à l'évidence de type RPMH (graphes peu denses, présentant une structuration locale riche et une distance moyenne très petite sur l'ensemble du graphe, ainsi qu'une structure hiérarchique – par exemple DicoSyn ou le grand Robert). Outre leur intérêt propre dans l'étude du lexique, ils peuvent donc aussi nous permettre de mieux comprendre les grands graphes de terrain dans leur ensemble.

De manière générale, si les définitions d'un dictionnaire sont porteuses de sens, c'est au moins par le réseau qu'elles tissent entre les mots qui en sont les entrées. Notre propos est d'exploiter ce réseau de type petit monde en tirant parti de l'hypothèse que les zones de densité fortes en arcs (les agrégats) identifient des zones de sens proches. Nous illustrons notre approche sur deux types de dictionnaires : un dictionnaire de langue, le Grand Robert[15] et DicoSyn un dictionnaire de synonymes constitué de **sept dictionnaires** classiques (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert) dont ont été extraites les relations synonymiques[16].

[15] Nous avons dû réaliser un important travail de saisie, de lemmatisation et de formatage en XML pour coder le graphe extrait du grand Robert (≈100 000 entrées)

[16] Ce premier travail de fusion, effectué à l'Institut National de la Langue Française (aujourd'hui ATILF: http://www.atilf.fr/) a produit une série de fichiers ; les données de ceux-ci ont été regroupées et homogénéisées au sein du laboratoire CRISCO http://elsap1.unicaen.fr/ par un important travail de correction (par adjonctions ou suppressions de liens synonymiques) sur le fichier final (Ploux & Victorri 1998).

Les dictionnaires sont représentés par des graphes dont les sommets et les arêtes peuvent être définis de multiples façons. L'une d'entre elles consiste à prendre pour sommets du graphe les entrées du dictionnaire et d'admettre l'existence d'un arc d'un sommet A vers un sommet B si et seulement si le mot B apparaît dans la définition du mot A. C'est la position de départ que nous avons adoptée[17]. En effet, cette seule procédure permet d'extraire d'un dictionnaire de langue, ce que dorénavant nous appellerons le **graphe du dictionnaire** en question.

Illustration autour du sommet ÉCORCER :

ÉCORCER [ekóRse] v. tr.; Dépouiller de son écorce (un arbre). Décortiquer, peler (le grain, les fruits)

Figure 3 *Définition de ECORCER – ROBERT –*

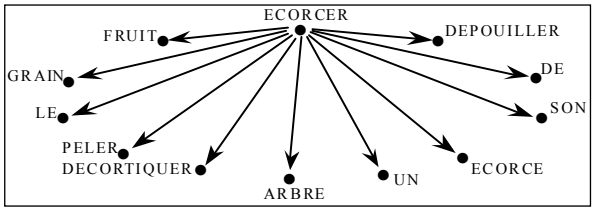


Figure 4 *Extrait du graphe autour de ECORCER – ROBERT –*

En réitérant cette construction pour chacune des entrées du dictionnaire, on obtient le graphe de ce dictionnaire. Si l'on ne s'intéresse qu'aux seuls verbes, voici ce que nous obtenons 'autour' du sommet dénoté par le verbe ÉCORCER :

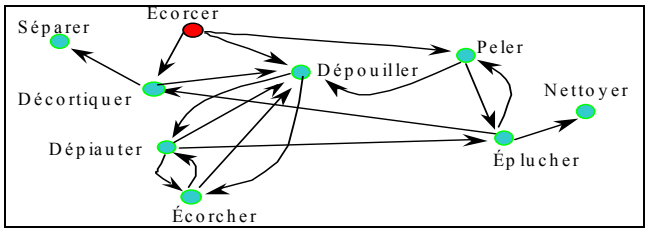


Figure 5 *Extrait du graphe des verbes, autour de ECORCER – ROBERT –*

Les définitions de NETTOYER, SÉPARER... renvoient à d'autres verbes absents de notre schéma pour des raisons de lisibilité (en poursuivant, on rencontrerait très rapidement tous les verbes du dictionnaire). Nous n'avons donc porté sur cette figure qu'une partie des voisins d'ordres 1, 2 et 3 d'ÉCORCER. Une fois ce graphe

[17] En fait, une première phase de lemmatisation est nécessaire.

orienté obtenu, nos algorithmes travaillent à partir de ce que nous appelons un graphe anonyme[18] qui en est la version non orientée.

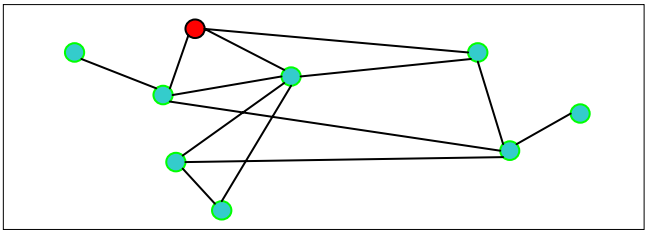


Figure 6 *Extrait du graphe anonyme des verbes, autour du sommet associé à* ECORCER – ROBERT –

Les graphes ainsi obtenus sont des RPMH typiques : par exemple, DicoSyn.Verbe[19] a 9 043 sommets, il a 50 948 arêtes, sur sa plus grande partie connexe (8 835 sommets) son L est égal à 4.1694 et son C est égal à 0.3186, c'est typiquement un petit monde. La courbe représentant la distribution des incidences de ses sommets (Figure 7) est caractéristique des réseaux petits mondes hiérarchiques (Ravasz & Barabási 2003) (en log-log elle forme approximativement une droite).

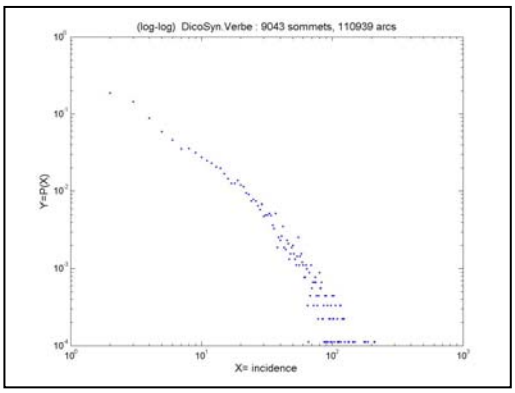


Figure 7 *Courbe log-log de la distribution de l'incidence des sommets (DicoSyn.Verbe : 9043 sommets)*

[18] Nous l'appelons « graphe anonyme » pour insister sur le fait que nos algorithmes tel un e.t. ignorant tout de l'origine de ce graphe ne travaillent qu'à partir de cette seule structure. Par exemple, serait-il possible parmi plusieurs graphes anonymes de distinguer leurs origines (dictionnaire général, dictionnaire de synonymes, world wide web, réseau protéique…) ?

[19] DicoSyn.Verbe est le graphe des verbes extraits de DicoSyn : il existe une arête {A,B} si les verbes représentés par les sommets A et B sont donnés comme synonymes dans DicoSyn.

Dans la section suivante, nous présentons la proximité sémantique entre verbes (c'est une relation de similarité que nous appelons proxémie) à partir de laquelle nous définissons une distance (que nous appelons distance proxémique) entre sommets des graphes de dictionnaire. L'idée importante, contrairement aux méthodes locales telles que par exemple les méthodes à base de cliques (Ploux & Victorri 1998) ou à base d'indices de similarité locale comme ceux de Jaccard, Ochai, … (Hubalek 1982) est de calculer la distance entre deux sommets à partir de la globalité du graphe. Cela signifie que ne sont pas seulement pris en compte les voisins immédiats de deux sommets pour le calcul de leur distance, mais la totalité du graphe par le calcul préalable d'un indice de similarité faisant intervenir tout le graphe, et suivi d'un plongement du résultat dans $\mathbb{R}^n$ où n est le nombre de sommets du graphe. C'est en appliquant cette méthode d'analyse aux dictionnaires que nous mettons au jour la structure de leurs graphes et 'capturons' leurs propriétés topologico-sémantiques parmi lesquelles figurent la proxémie qui organise dans un continuum l'hyperonymie, la cohyponymie intradomaine et la cohyponymie interdomaine (Duvignau 2002, Gaume 2003)

6. La proxémie

Pour simplifier la présentation, dans la suite de cet article nous supposerons que tous nos graphes[20] sont fortement connexes et réflexifs.

A partir d'un graphe G fortement connexe et réflexif donné et d'un entier $\lambda>0$ l'algorithme prox calcule sa matrice $\hat{G}^{\lambda}$, puis une matrice $D_{G,\lambda}$ de distance entre ses sommets comme suit :

– Soit G=(V,E) un graphe à n sommets et [G] sa matrice d'adjacence ;
– Soit $\lambda>0$ un entier[21] naturel ;

[20] Si le graphe G=(V,E) n'est pas déjà réflexif alors pour tout sommet $r\in V$, on ajoute à E les arcs (r,r) il devient ainsi réflexif. Si G n'est pas connexe, on étudie séparément ses parties connexes. Si il est connexe sans être fortement connexe, alors on le symétrise ce qui le rend alors fortement connexe.

[21] Comment Choisir λ : d'une part, un λ trop grand va rapprocher les $[\hat{G}^{\lambda}]_{r,s}$ de leurs limites (voir propriété 3.8.), c'est à dire que $\forall r,s,u\in V$, $[\hat{G}^{\lambda}]_{r,s}\approx[\hat{G}^{\lambda}]_{u,s}$, prox perdant ainsi de son intérêt. D'autre part, un λ trop petit ne va pas se distinguer des méthodes locales. Dans un RPMH, nous proposons de prendre λ entre L et 2L. La particule à partir d'un sommet r quelconque pouvant atteindre en moyenne tout autre sommet s, sans pour cela que la limite $\lim_{\lambda\to\infty}[\hat{G}^{\lambda}]_{r,s}$ soit atteinte, en identifiant ainsi les zones denses en arcs (si il existe une confluence plus forte de chemins entre les sommet r et s qu'entre r et u alors : $[\hat{G}^{\lambda}]_{r,s}>[\hat{G}^{\lambda}]_{r,u}$).

– Soit $\hat{G}$ la matrice Markovienne de G (c'est-à-dire que $\forall r,s \in V$, $[\hat{G}]_{r,s}=[G]_{r,s}/\sum_{x\in V}\{[G]_{r,x}\}$ – voir déf. 3.1)
– Soit $\hat{G}^{\lambda}$ la matrice $\hat{G}$ élevée à la puissance λ ;
– On définit la distance entre sommets en considérant la matrice $\hat{G}^{\lambda}$ comme étant les coordonnées de n vecteurs dans un espace de dimension n, puis en calculant la distance Euclidienne[22] entre chaque paire de vecteurs. La matrice nxn $D_{G,\lambda}$ est donc définie ainsi : $\forall r,s \in V$, $[D_{G,\lambda}]_{r,s}=(\sum_{x\in V}([\hat{G}^{\lambda}]_{r,x}-[\hat{G}^{\lambda}]_{s,x})^2)^{½}$.

Ainsi, l'entrée de prox est le graphe G et un nombre naturel $\lambda>0$ et sa sortie en sont les matrices nxn $\hat{G}^{\lambda}$ et $D_{G,\lambda}$:

$$(G, \lambda) \rightarrow \textbf{[prox]} \rightarrow \hat{G}^{\lambda} \rightarrow D_{G,\lambda}$$

D'après la prop. 3.4., pour tout $r\in V$, $({}^{\{r\}}X_i)_{i\in\mathbb{N}}$, la *chaîne de Markov du graphe G débutant sur {r}* est telle que $\forall s\in V$, $\forall i\in\mathbb{N}$, $[{}^{\{r\}}P_i]_s=[\hat{G}^1]_{r,s}$. C'est-à-dire que $[\hat{G}^1]_{r,s}$ est la probabilité qu'en partant du sommet r, la particule soit en s à l'instant i.

6.1 Prox et la désambiguïsation

Nous avons passé sous silence jusqu'à maintenant un problème pourtant fondamental en traitement automatique des langues : la désambiguïsation (Ide & Véronis J 1998, Victorri & Fuchs 1996). Pour une présentation détaillée de la désambiguïsation par proximité structurelle des entités polysémiques en cotexte calculée avec prox, voir Gaume & al. 2004.
Par exemple pour le français dans le dictionnaire *Le Grand Robert* il y a deux entrées distinctes pour « causer » :

CAUSER_1 « être la cause de. - Amener, apporter, attirer, déclencher, entraîner, faire, motiver, occasionner, produire, provoquer, susciter. *Causer un dommage. Causer du scandale. L'orage a causé de graves dommages aux récoltes…* »

CAUSER_2 « S'entretenir familièrement avec qqn. – Parler, converser, confabuler (vx), deviser, discuter. *Nous causons ensemble. Causer avec qqn…* »

Aussi même si un locuteur du français sait naturellement que dans la définition de « bavarder » :

[22] On peut bien sûr choisir une autre distance (distance angulaire, distance du χ^2, mais nous nous limiterons dans cet article à la distance euclidienne classique)

BAVARDER « Parler beaucoup, longtemps ou parler ensemble de choses superficielles. - Parler; babiller, bavasser (fam.), cailleter, caqueter, causer, discourir, discuter, jaboter, jacasser, jaser, jaspiner (argot), lantiponner (vx), papoter, potiner. *Bavarder avec qqn …* »

le verbe « causer » fait référence à CAUSER_2, notre procédure de construction du graphe (voir § 5.), quant à elle, ne peut désambiguïser. Aussi elle crée un sommet fictif CAUSER (qui n'est pas une entrée du dictionnaire puisqu'on y trouve seulement CAUSER_1 et CAUSER_2) et ajoute ensuite deux arêtes {CAUSER, CAUSER_1} et {CAUSER, CAUSER_2}. Quand « causer » est trouvé dans une définition d'un mot comme « bavarder », alors l'arête {BAVARDER, CAUSER} est ajoutée.

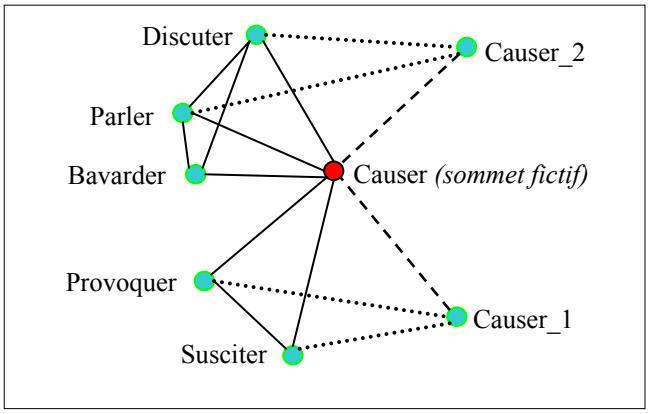


Figure 8 *« Causer » sommet fictif*

Dans la Figure 8 il y a bien sûr beaucoup d'arêtes et de sommets absents de notre schéma par souci de lisibilité. Les arêtes en pointillés {Discuter, Causer_2}, {Parler, Causer_2} sont dues au fait que « Discuter » et « Parler » sont dans la définition de « Causer_2 » ainsi que les arêtes {Provoquer, Causer_1} et {Susciter, Causer_1} qui sont dues au fait que « Provoquer » et « Susciter » sont dans la définition de « Causer_1 ».
On applique ensuite prox à ce graphe pour obtenir une matrice $\hat{G}^{\lambda}$ comme définie plus haut.

$\hat{G}^3$	Bavarder	Parler	Discuter	causer	**Causer_1**	**Causer_2**	Provoquer	Susciter
Bavarder	0.1502	0.1764	0.1764	0.1983	0.0461	0.1342	0.0723	0.0461
parler	0.1411	0.1701	0.1701	0.1956	0.0499	0.1444	0.0789	0.0499
Discuter	0.1411	0.1701	0.1701	0.1956	0.0499	0.1444	0.0789	0.0499
Causer	0.0991	0.1223	0.1223	0.1995	0.1034	0.1237	0.1265	0.1034
Causer_1	0.0461	0.0623	0.0623	0.2067	0.1702	0.0958	0.1864	0.1702
Causer_2	0.1074	0.1444	0.1444	0.1979	0.0766	0.1391	0.1136	0.0766
Provoquer	0.0579	0.0789	0.0789	0.2024	0.1491	0.1136	0.1701	0.1491
Susciter	0.0461	0.0623	0.0623	0.2067	0.1702	0.0958	0.1864	0.1702

Tableau 2 *$\hat{G}^3$ pour le graphe de la Figure 8*

Dans le Tableau 2 nous pouvons observer que :

$[\hat{G}^3]_{Bavarder,Causer\_1}=0.0461<[\hat{G}^3]_{Bavarder,Causer\_2}=0.1342$, ce qui est normal puisque depuis le sommet « Bavarder » il existe une confluence topologique des chemins guidant une particule préférentiellement vers le sommet « causer_2 » plutôt que vers le sommet « causer_1 ».

Dans $({}^{\{bavarder\}}X_i)_{i\in\mathbb{N}}$ la chaîne de Markov du graphe de la Figure 8 débutant par le sommet « Bavarder » on a :

$[\hat{G}^3]_{Bavarder,Causer\_1}=[{}^{\{bavarder\}}P_3]_{Causer\_1}=0.0461<[{}^{\{bavarder\}}P_3]_{Causer\_2}=[\hat{G}^3]_{Bavarder,Causer\_2}=0.1342$
et c'est cela qui va nous permettre de désambiguïser :

Supposons qu'un verbe ait k homonymes[23] (c'est-à-dire que k entrées ont la même graphie dans le dictionnaire étudié), il y aura alors les sommets S, S_1, S_2, … ,S_k dans le graphe où S sera le sommet fictif. S'il y a une arête {A,S}, elle sera alors remplacée par l'arête $\{A,S_i\}$ où S_i est tel que $[\hat{G}^3]_{A,Si}=MAX_{0<z\leq k}\{[\hat{G}^3]_{A,Sz}\}$. On supprime ensuite tous les sommets fictifs du graphe pour obtenir ainsi un graphe désambiguïsé comme dans la Figure 9 :

[23] Deux entrées distinctes sont dites homonymiques dans un dictionnaire si elles ont même graphie. Alors qu'une entrée ayant plusieurs sous sens est dite polysémique. Le choix de décider si une entrée doit être polysémique ou s'il doit exister plusieurs entrées homonymiques n'est pas un problème simple : « *On peut considérer les différents aspects sémantiques (sens) d'un même mot (polysémie) comme des homonymes, lorsque l'origine commune n'est plus sentie* » (Grand Robert).

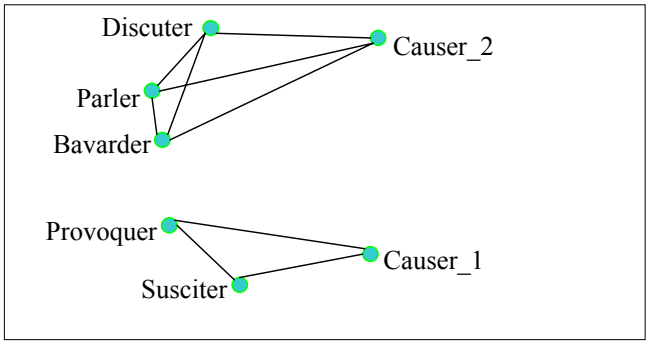


Figure 9 *Graphe désambiguïsé*

On applique alors une nouvelle fois prox, mais à ce graphe désambiguïsé.

Exemple : liste des 100 sommets les plus proches du verbe ÉCORCER (du plus 'prox' au moins 'prox') calculée par prox avec λ=5, sur le graphe construit à partir de DicoSyn.Verbe.

[**1 écorcer**, 2 démascler, 3 peler, 4 gemmer, 5 dépouiller, 6 bourgeonner, 7 décortiquer, 8 baguer, 9 tondre, 10 inciser, 11 éplucher, 12 couper, 13 faufiler, 14 avoir des boutons, 15 marquer, 16 tailler, 17 écaler, 18 démunir, 19 débourrer, 20 ôter, 21 décérébrer, 22 enlever, 23 boutonner, 24 voler, 25 fleurir, 26 gratter, 27 raser, 28 s'époiler, 29 desquamer, 30 plumer, 31 bretauder, 32 écorcher, 33 décerveler, 34 râper, 35 déposséder, 36 piquer, 37 émonder, 38 ouvrir, 39 désosser, 40 blesser, 41 tamiser, 42 monder, 43 écosser, 44 coudre, 45 scruter, 46 raisonner, 47 nettoyer, 48 taillader, 49 examiner, 50 égorger, 51 entailler, 52 empointer, 53 bâtir, 54 entamer, 55 composer, 56 retourner, 57 dégager, 58 débrider, 59 arracher, 60 défaire, 61 fouiller, 62 ébarber, 63 analyser, 64 labourer, 65 scarifier, 66 priver, 67 spolier, 68 disséquer, 69 dégarnir, 70 ciseler, 71 stériliser, 72 séparer, 73 détruire, 74 frustrer, 75 dépecer, 76 retrancher, 77 ruiner, 78 châtrer, 79 déchirer, 80 façonner, 81 faucher, 82 débarrasser, 83 diminuer, 84 dépiauter, 85 sevrer, 86 supprimer, 87 casser, 88 fixer, 89 prendre, 90 retirer, 91 déshabiller, 92 élaguer, 93 faire, 94 manger, 95 tuer, 96 limer, 97 brouter, 98 dénuder, 99 vider, 100 trier, …]

Figure 10 *proxémie de* ECORCER *dans – DicoSyn.Verbe –*

Dans DicoSyn.Verbe le sommet « écorcer » a 8 synonymes : {baguer, décortiquer, démascler, dépouiller, gemmer, inciser, peler, tondre}
Dans la Figure 10, le nombre qui précède chaque verbe est son rang en fonction de sa proxémie à **ÉCORCER**.
La proxémie calculée par l'algorithme prox organise dans un continuum les notions d'hyperonymie, de cohyponymie intradomaine (par les sommets les plus 'prox') et de cohyponymie interdomaines (par les sommets un peu moins 'prox').
L'introduction de la notion de « proxémie » qui recouvre ces trois notions permet de souligner le glissement de sens continu qu'il y a d'un mot en relation synonymique (cohyponyme intradomaine) vers un mot en relation métaphorique (cohyponyme interdomaine) au fur et à mesure que la proxémie au mot de référence diminue.

6.2 prox et la relation d'hyperonymie/hyponymie

L'observation de la dynamique de la particule à partir d'un sommet r vers un sommet s nous indique le rapport sémantique entre les deux sommets r et s.

Par exemple :

• La Figure 11.a est la courbe de $f(t)=[\hat{G}^t]_{écorcer,enlever}$ la probabilité de la particule en partant du sommet « écorcer » d'atteindre le sommet « enlever » à l'instant t ;

• La Figure 11.b est la courbe de $f(t)=[\hat{G}^t]_{écorcer,rire}$ la probabilité de la particule en partant du sommet « écorcer » d'atteindre le sommet « rire » à l'instant t ;

• La Figure 11.c est la courbe de $f(t)=[\hat{G}^t]_{écorcer,écaler}$ la probabilité de la particule en partant du sommet « écorcer » d'atteindre le sommet « écaler » à l'instant t ;

• La Figure 11.d est la courbe de $f(t)=[\hat{G}^t]_{écorcer,pianoter}$ la probabilité de la particule en partant du sommet « écorcer » d'atteindre le sommet « pianoter » à l'instant t ;

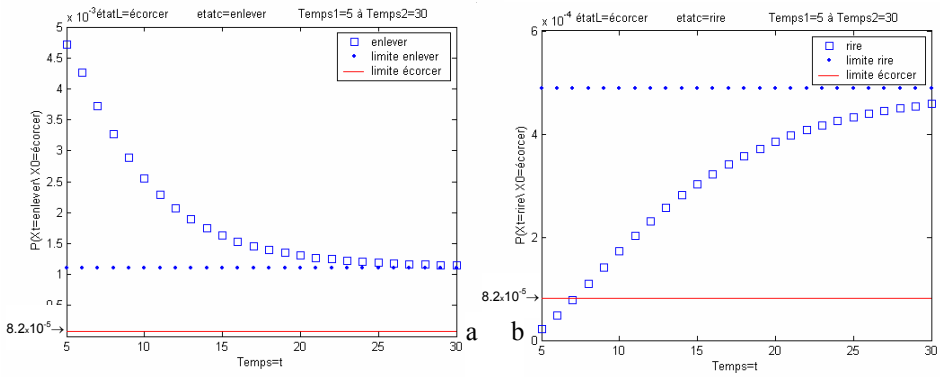
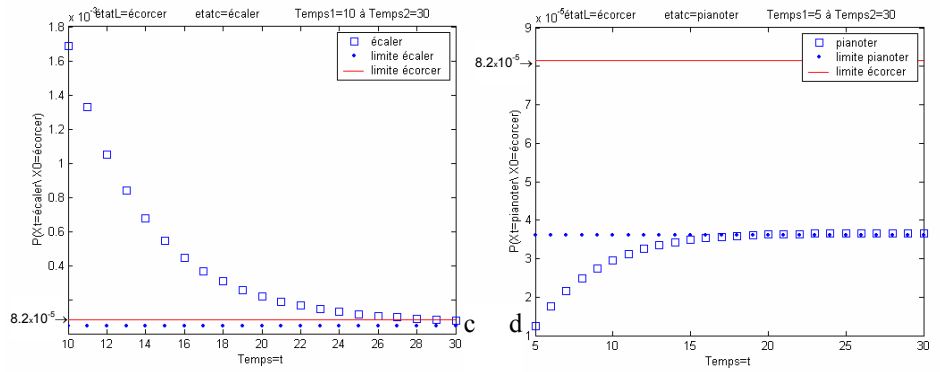


Figure 11 *dynamique de la particule à partir d'un sommet r vers un sommet s*



Bien sûr d'aprés la proposition 3.8 quand le temps t tend vers l'infini chacune de ces courbes tendent vers leur limite :

$\lim_{t\to\infty} [\hat{G}^t]_{écorcer,enlever}=[L]_{enlever}$; $\lim_{t\to\infty} [\hat{G}^t]_{écorcer,rire}=[L]_{rire}$;
$\lim_{t\to\infty} [\hat{G}^t]_{écorcer,écaler}=[L]_{écaler}$; $\lim_{t\to\infty} [\hat{G}^t]_{écorcer,pianoter}=[L]_{pianoter}$;

Définition : Un sommet s est dit hyperonyme du sommet r si :
1) $L_s>L_r$ (dans la Figure 11.a on a $L_{enlever}>L_{écorcer}$ ainsi que dans la Figure 11.b on a $L_{rire}>L_{écorcer}$). Le sommet s pour être un candidat hyperonyme de r doit être plus 'important' que r ;
2) la suite $([\hat{G}^t]_{r,s})_{i>0}$ décroit vers sa limite L_s (dans la Figure 11.a c'est le cas de $f(t)=[\hat{G}^t]_{écorcer,enlever}$) : le sens du sommet s doit être proche du sens du sommet r, au début de sa balade la particule passe par r plus souvent qu'à l'ordinaire ;

• **cas de la Figure 11.a :** « enlever » est un hyperonyme de « écorcer » car : **1)** $L_{enlever}>L_{écorcer}$: « enlever » est donc un sommet plus 'important' que « écorcer » **et 2)** $f(t)=[\hat{G}^t]_{écorcer,enlever}$ **décroit** vers sa limite $L_{enlever}$: « enlever » a donc un sens proche de « écorcer »
• **cas de la Figure 11.b :** « rire » n'est pas un hyperonyme de « écorcer » car : **1)** $L_{rire}>L_{écorcer}$: « rire » est donc un sommet plus 'important' que « écorcer » **mais 2)** $f(t)=[\hat{G}^t]_{écorcer,rire}$ **croit** vers sa limite L_{rire} : « rire » n'a donc pas un sens proche de « écorcer »
• **cas de la Figure 11.c :** « écaler » n'est pas un hyperonyme de « écorcer » car : **1)** $L_{écaler}<L_{écorcer}$: « écaler » n'est donc pas un sommet plus 'important' que « écorcer » **bien que 2)** $f(t)=[\hat{G}^t]_{écorcer,écaler}$ **décroit** vers sa limite $L_{écaler}$: « écaler » a donc un sens proche de « écorcer »
• **cas de la Figure 11.d :** « pianoter » n'est pas un hyperonyme de « écorcer » car : **1)** $L_{pianoter}<L_{écorcer}$: « pianoter » n'est donc pas un sommet plus 'important' que « écorcer » **et de plus 2)** $f(t)=[\hat{G}^t]_{écorcer,pianoter}$ **croit** vers sa limite $L_{pianoter}$: « pianoter » n'a donc pas un sens proche de « écorcer »

Notons qu'aucun des sommets « enlever », « rire », « écaler », « pianoter » n'est un voisin direct de « écorcer », et que les méthodes locales basées sur un indice de similarité calculé sur l'ordre 1 des voisins, ne peuvent donc obtenir de résultats semblables[24]. Prox permet de prendre en compte la totalité du graphe pour définir la proximité entre deux sommets.

[24] La condition **1)** mérite d'être relativisée ou contrainte par un seuil sur les degrés d'incidence de r et s. En effet, les conditions 1) et 2) sont toutes deux vérifiées pour le couple (s=« décortiquer », r=« écorcer ») ce qui classe « décortiquer » comme hyperonyme de « écorcer » alors qu'on préférerait plutôt classer « Décortiquer » qui n'a que 12 synonymes

6.3 prox et les approximations sémantiques par analogie

Les proxémies ainsi calculées entre les mots sont en accord avec les approximations sémantiques par analogie produites par les jeunes enfants. Par exemple, l'énoncé spontané « je **déshabille** l'arbre » pour [j'**écorce** l'arbre], produit à 2 ans et demi, manifeste un rapprochement entre verbes, qui semble valider notre modèle : DÉSHABILLER est prox de ÉCORCER (c'est le 91^{e} mot le plus prox de ÉCORCER – voir Figure 10).
Voici quelques exemples d'approximations sémantiques par analogie (extrait du corpus Duvignau 2002) :

*« je **déshabille** l'orange »* 36 mois (l'enfant pèle une orange) [**Peler/Déshabiller**] (72^{e})

*« maman, tu peux **coller** les boutons ? »* 36 mois (les boutons sont décousus, il faut les coudre) [**Coudre/Coller**] (74^{e})

*« le livre est **cassé** »* 26 mois (le livre est déchiré) [**Déchirer/Casser**] (4^{e})

*« il faut la **soigner** la voiture »* 38 mois (il faut réparer la voiture) [**Réparer/Soigner**] (344^{e})

L'enfant apprendrait d'abord les mots correspondant aux « capitales »[25], et s'en servirait pour désigner une vaste région : l'enfant cherchant à communiquer un événement A [ex : déchirer un livre] pour lequel il ne dispose pas de catégorie verbale constituée (1) ferait une analogie avec un ancien événement B [casser un verre] déjà mémorisé avec une entrée lexicale « *casser »* et (2) utilisant cette analogie, dirait « *le livre est cassé »* pour communiquer l'événement A. Puis l'enfant acquerrait progressivement les mots correspondant aux « villes » de moindre importance, affinant alors sa précision de désignation.

Ce type d'analyse menée à partir d'un corpus de 230 approximations sémantiques par analogie produites par des jeunes enfants (1;8 ans à 4;2 ans) montre que la moyenne du rang du verbe dit par l'enfant (comme *« casser »*) dans la proxémie du

---

dans Dicosyn.Verbe (donc 12-8=4 de plus que « écorcer ») comme un synonyme de « écorcer » c'est-à-dire comme cohyponyme intradomaine de « écorcer ». D'autre part il faudra évaluer l'adéquation de cette modélisation sur un échantillon de couples de mots tirés au hasard.

[25] Souvent l'enfant dit un hyperonyme pour l'hyponyme, mais ce n'est pas toujours le cas par exemple *« il faut la **soigner** la voiture »* [**Réparer/Soigner**] où c'est ici l'inverse. Notons que pour λ=5, « soigner » est le 344^{e} mot le plus prox de « réparer » alors que « réparer » est le 194^{e} mot le plus prox de « soigner » (quand X est un hyperonyme de Y, on a : le rang de X dans la proxémie de Y est plus petit que le rang de Y dans la proxémie de X).

mot 'juste' (comme *« déchirer »*) est de 239, (ce qui est peu sur les 10 860 verbes présents dans le graphe extrait du Grand Robert sur lequel on été comparé les proxémies avec les approximations sémantiques par analogie produites par des jeunes enfants (Duvignau 2002, Duvignau & Gaume 2003, Duvignau & al. 2004a, Duvignau & al. 2004b).

Sur la base de ces premiers résultats, nous postulons qu'élaborer des dictionnaires électroniques [26] en s'appuyant sur une théorie linguistique de l'organisation sémantique du lexique qui s'avère être en adéquation avec des processus d'acquisition précoce du lexique et qui se retrouve chez le locuteur adulte (Gaume & Duvignau 2004), leur conférera un caractère d'ergonomie cognitive tant pour l'apprentissage L1 (langue maternelle) ou L2 (langue seconde), que la traduction automatique, le résumé automatique, l'aide à la rédaction, la fouille de données, la classification automatique, la terminologie, ou encore la visualisation du sens. Cela devrait permettre d'améliorer leur utilisabilité en vue de mener une réflexion positive sur la normalisation des dictionnaires électroniques (Veronis 2002a, 2002b).

7. Organisation 'géosémique' des verbes du français

Par souci de concision nous nous limiterons ici aux graphes fortement connexes et réflexifs, mais il suffit que le graphe étudié soit irréductible et apériodique, pour que tout ce qui est dit dans cet article reste valable[27].

Soit donc G=(V,E) un graphe à n sommets, fortement connexe et réflexif.
On a l'habitude d'illustrer les graphes en représentant les sommets par des points et en rejoignant 2 points r,s par une flèche $r \rightarrow s$ si et seulement si $(r,s) \in E$. On dit qu'on a visualisé le graphe. Si l'on visualise un graphe G c'est pour nous permettre d'en bien comprendre intuitivement la structure, car en effet, face à sa matrice $n_{x}n$ d'adjacence [G], dès que n devient grand, un cerveau humain n'a plus aucune intuition quant à la structure du graphe. Mais un graphe n'est pas un objet géométrique, il n'a pas une forme propre, un graphe est un objet relationnel.
Aussi toute visualisation qui respecte la condition :

Une flèche relie r vers s si et seulement si $(r,s) \in E$

est une visualisation fidèle à la structure relationnelle du graphe en question.

[26] Ceci pouvant s'étendre aux réseaux sémantiques et bases de données de manière générale (représentation des connaissances, accès à l'information, stockage …).

[27] On peut aussi appliquer prox à un graphe pondéré. Dans ce cas les arcs sortants d'un sommet ne sont pas équiprobables.

On voit bien qu'il existe une infinité de visualisations relationnellement fidèles pour un graphe. L'enjeu de la visualisation étant de nous aider à 'penser la structure du graphe' on souhaite en général placer les sommets dans l'espace 2D ou 3D de manière à ce que la structure géométrique reflète des propriétés pertinentes de sa structure relationnelle d'origine. Par exemple pour un graphe de dictionnaire on aimerait placer les sommets dans l'espace de manière à ce que 2 sommets sémantiquement proches soient aussi proches dans l'espace de visualisation. Se pose alors le problème de calculer les coordonnées des sommets dans l'espace 2D ou 3D de visualisation du graphe.

Il est possible de placer les sommets d'un graphe en calculant leurs coordonnées dans un espace propre de son Laplacien (Mohar 1991, Kuntz & al. 2001, Lebart 2001) ou bien par la construction d'une mesure de similarité ou de dissimilarité entre sommets du graphe à partir de sa matrice d'adjacence (en utilisant un indice de type Dice, Jaccard,...). On peut ainsi appliquer à cette matrice une méthode de type Analyse en Composante Principale (ACP) (diagonalisation de la matrice de Torgerson) (Jouve & al. 2001, Ferré & Jouve 2002) ou une méthode de type Multidimensional Scaling pour positionner les sommets du graphe dans des espaces de dimension réduite. Ces méthodes peuvent s'appliquer à des graphes orientés ou non, pondérés ou non.

Cependant quand le nombre de sommets grandit, il est impossible d'afficher sur un écran la totalité des sommets d'un graphe de grande taille, ou alors l'image affichée est illisible du fait de la quantité des informations affichées.

Dans ce paragraphe nous présentons une méthode basée sur la proxémie pour calculer les coordonnées des n sommets d'un graphe dans l'espace $\mathbb{R}^n$ qui nous permet d'extraire la forme locale du graphe autour d'un sommet (une carte locale (Gaume & Ferré 2004), ainsi qu'une méthode de visualisation de la forme globale du graphe, son résumé (une carte globale). Nous verrons ensuite que cette approche nous permet d'envisager un outil de navigation pour le web.

7.1. La Visualisation proxémique locale

Nous prendrons dans la suite pour illustrer notre propos le graphe qui est la plus grande partie connexe de Dicosyn.Verbe, (ce graphe est fortement connexe, réflexif et symétrique, il a 8835 sommets et 110533 arcs).

On peut considérer la Matrice $\hat{G}^{\lambda}$ comme une matrice nxn des coordonnées de n vecteurs dans $\mathbb{R}^n$ (C'est ce point de vue qui nous a permis de calculer la matrice nxn $D_{G,\lambda}$ qui est la matrice des distances euclidiennes entre paires de sommets dans $\mathbb{R}^n$).

Notons que, pour λ=1, des choix adéquats de métrique conduisent à une analyse spectrale du Laplacien ou à l'analyse de contiguïté mentionnées plus haut.

Un autre point de vue est de considérer la Matrice $\hat{G}^{\lambda}$ comme une matrice où $\forall r,s \in V$, $[\hat{G}^{\lambda}]_{r,s}$ nous indique la similarité des sommets r et s. C'est-à-dire que si $[\hat{G}^{\lambda}]_{r,s} > [\hat{G}^{\lambda}]_{r,u}$ alors nous pourrons dire que r est plus similaire ou 'prox' de s que r ne l'est de u. Cela permet de définir la proximité entre deux sommets non pas par la simple existence d'un arc entre ces deux sommets mais par la prise en compte de la structure globale du graphe. Notons que la construction de similarités ou dissimilarités à partir de la matrice d'adjacence et d'indices tels que celui de Dice par exemple, permet de prendre en compte les voisins d'ordre 2, voir, e.g. (Jouve & al. 1998), mais, en général, pas au-delà.

C'est l'alliance de ces deux points de vue ($\hat{G}^{\lambda}$ comme matrice nxn de coordonnées dans $\mathbb{R}^n$ ou bien $\hat{G}^{\lambda}$ comme matrice nxn de similarité sur les n sommets du graphe) qui va nous permettre de visualiser localement un graphe afin de nous aider à mieux comprendre sa structure.

Le tableau $\hat{G}^{\lambda}$ en tant que matrice de coordonnées dans $\mathbb{R}^n$ contient une information calculée sur l'ensemble du graphe qu'il serait possible de représenter dans $\mathbb{R}^3$ au moyen d'une Analyse en Composante Principale[28] (ACP) de $\hat{G}^{\lambda}$. Mais comme nous l'avons déjà remarqué, les 8835 sommets de notre graphe affichés sur un écran donnent une image illisible. Nous n'allons donc en afficher qu'un 'morceau autour' d'un ensemble de sommets F de la manière suivante.

Si l'on veut observer la structure du graphe à n sommets G=(V,E) autour d'un ensemble non vide de sommets $F \subseteq V$ par une carte R-locale 'autour' de F (la région dans G comprise dans un 'rayon R autour de F'), alors **a)** on extrait $A_{\lambda,F,R} \subseteq V$, où $A_{\lambda,F,R}$ est l'ensemble des R sommets de plus fortes coordonnées[29] pour un instant λ donné dans le vecteur ligne $[{}^{F}P_{\lambda}]=[{}^{F}P_0 \bullet \hat{G}^{\lambda}]$ puis **b)** on extrait de $\hat{G}^{\lambda}$ la matrice $M_{\lambda,F,R}$ qui est la matrice carrée RxR formée de l'intersection des R lignes $[\hat{G}^{\lambda}]_x$ telles que $x \in A_{\lambda,F,R}$.avec les R colonnes $[\hat{G}^{\lambda}]^y$ telles que $y \in A_{\lambda,F,R}$ **c)** on normalise[30] ensuite les R lignes de $M_{\lambda,F,R}$ (pour chaque $x \in A_{\lambda,F,R,}$ on remplace la ligne $[M_{\lambda,F,R}]_x$ par

[28] Analyse en Composante Principale, (ACP) est une méthode projective qui permet de projeter une forme F de $\mathbb{R}^n$ dans $\mathbb{R}^d$ (avec d<n), en perdant le moins d'information qu'il est possible sur la forme F.

[29] Par exemple si G est DicoSyn.Verbe, λ=6, F={écorcer} et R=100, alors dans $A_{6,\{écorcer\},100}$ seront présents les sommets de la Figure 10

[30] C'est-à-dire que pour chaque $x \in A_{\lambda,F,R,}$ on remplace le vecteur de coordonnées $[M_{\lambda,F,R}]_x$ par le vecteur dont l'extrémité est l'intersection de $[M_{\lambda,F,R}]_x$ avec l'hypersphère unité.

$[[M_{\lambda,F,R}]_x/||[M_{\lambda,F,R}]_x||)$. On pratique ensuite une ACP sur $M_{\lambda,F,R}$ où l'on ne garde alors que les 3 premières dimensions (qui conservent alors le plus d'informations pertinentes dans la réduction de $\mathbb{R}^n$ à $\mathbb{R}^3$) pour obtenir ainsi $V_{\lambda,F,R}$ qui est la visualisation 3D de la région[31] dans G comprise dans un 'rayon R autour de F' pour un instant λ donné.

Exemple : La Figure 12 nous montre $V_{6,\{jouer\},100}$ la visualisation 3D autour du singleton {jouer}, c'est-à-dire les 3 premières coordonnées sur l'ACP de $M_{\lambda,F,R}$ pour λ=6, F={jouer}, R=100. Nous pouvons voir que la structure géométrique de $V_{6,\{jouer\},100}$ (la forme ainsi obtenue) reflète bien la structure polysémique du verbe « Jouer » avec ses 4 sens principaux.
Pour plus de détail consulter http://dilan.irit.fr où l'ensemble du lexique du Français sera bientôt en ligne ainsi que d'autres langues)

[31] Pour tout ensemble non vide $F \subseteq V$, si R=n, ce seront donc les n sommets du graphe tout entier qui s'afficheront dans $V_{\lambda,F,n}$

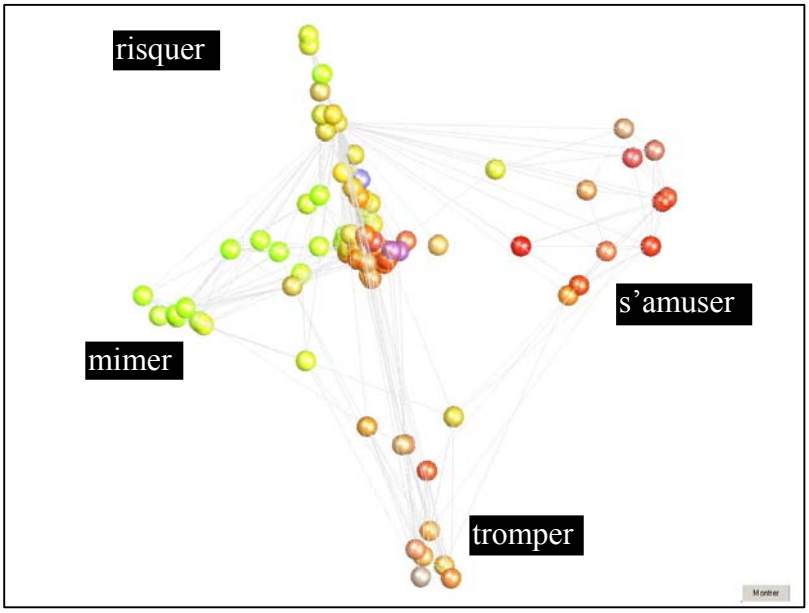
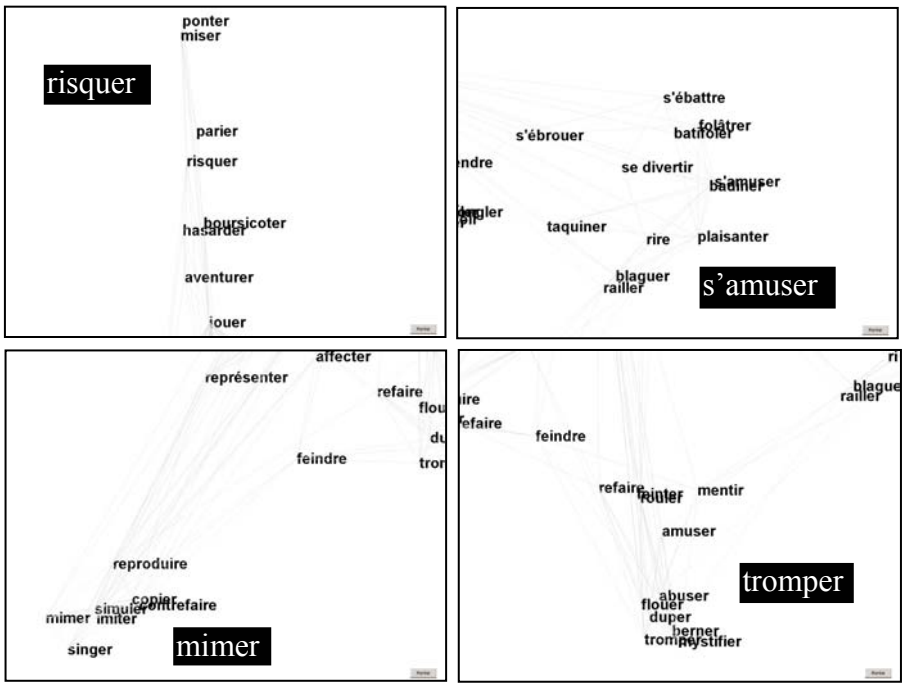


Figure 12 *$V_{6,\{jouer\},100}$, Forme conceptuelle 3D du verbe « Jouer »*

Ci-dessous des gros plans sur les 4 pôles principaux



C'est un peu comme si la particule 'activait/excitait' la zone sémantique de « jouer », son pouvoir évocateur dans le réseau du sens restreint aux verbes. On retrouve ici le même type de zones sémantiques que dans les visualisations pratiquées à partir du modèle basé sur les cliques décrit dans (Ploux & Victorri 1998) http://elsap1.unicaen.fr/dicosyn.html ou http://dico.isc.cnrs.fr/dico_html/). Cependant les méthodes locales comme celles à bases de cliques ou d'indices locaux de similarités comme celui de Dice par exemple, ne permettent pas d'aller au-delà de l'ordre un ou deux des voisins, ne pouvant par exemple pas obtenir les résultats illustrés par la figure 10 ni ceux de la figure 11 (car aucun des sommets illustrés ne sont voisins de « écorcer ») et encore moins obtenir une approche globale telle que présentée au § 7.2. car les méthodes locales sont mal adaptées pour exploiter la structure des RPMH.

Avec prox on peut observer les zones sémantiques d'une paire ou de plusieurs mots en visualisant leurs formes comme par exemple la paire (« commencer », « finir ») voir Figure 13 ou (« monter », « descendre ») voir
Figure ***14*** ou bien encore (« aimer », « détester ») voir Figure 15.

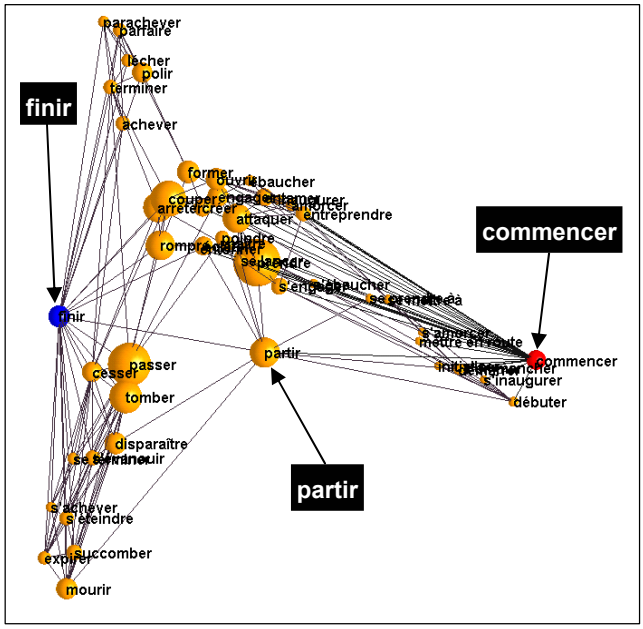


Figure 13 $V_{6,\{commencer,finir\},50}$

Dans la Figure 13 on peut voir qu'à la charnière de l'articulation sémantique de {«commencer », « finir »} il y a « partir ».

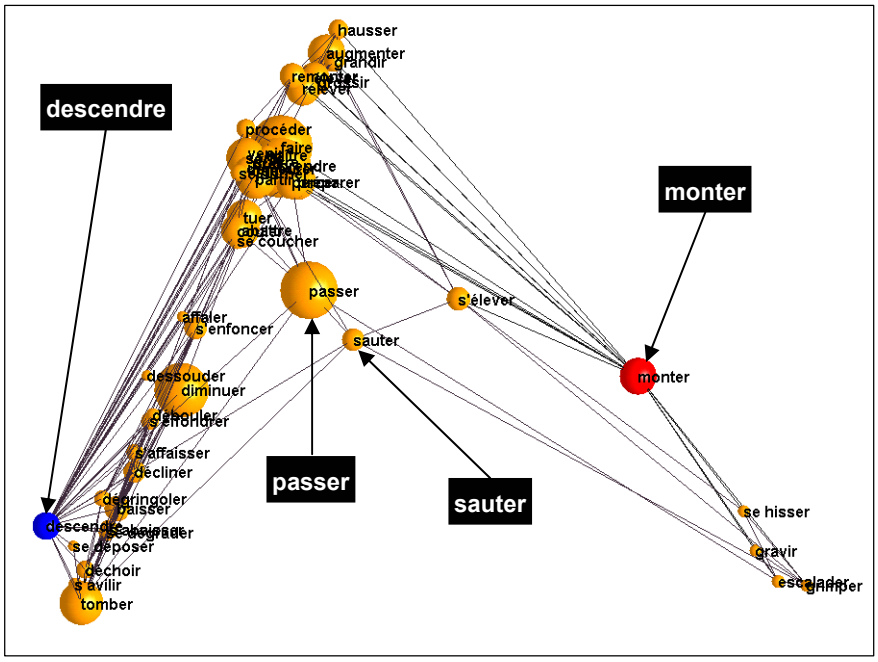


Figure 14 $V_{6,\{\text{« descendre », « monter »}\},50}$

Dans la
Figure ***14*** on peut voir qu'à la charnière de l'articulation sémantique de {« descendre », « monter »} il y a « passer » et « sauter », bien qu'aucun de ces deux mots ne soit directement connecté à « monter ».

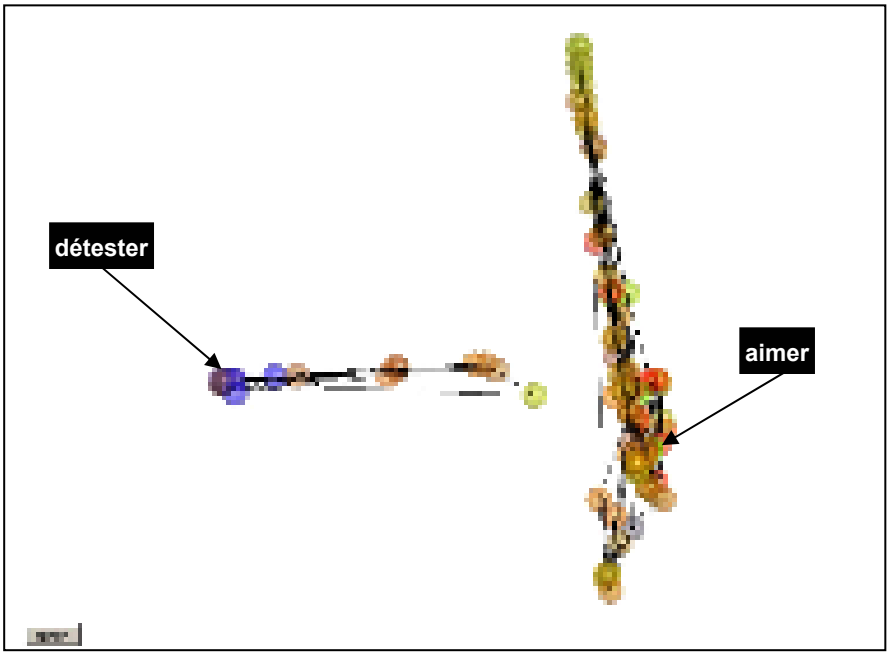
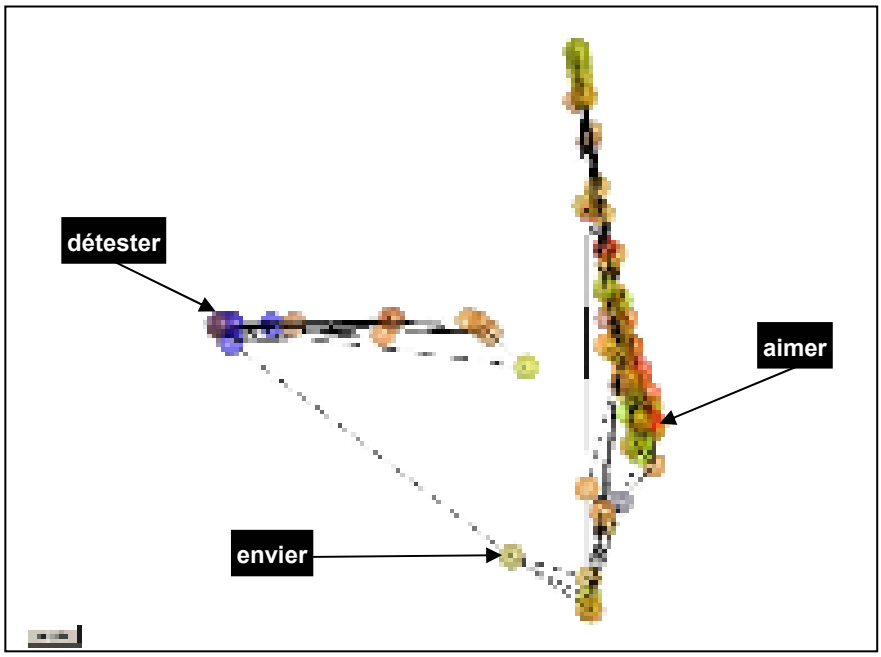


Figure 15 *ci-dessus a :* $V_{6,\{«détester», «aimer»\},78}$, *ci-dessous b :* $V_{6,\{«détester», «aimer»\},79}$



Dans la Figure 15 on peut voir qu'à la charnière de l'articulation sémantique de {« détester », « aimer »} il y a « envier » dont un éclairage (à la main) est décrit dans la Figure 16.

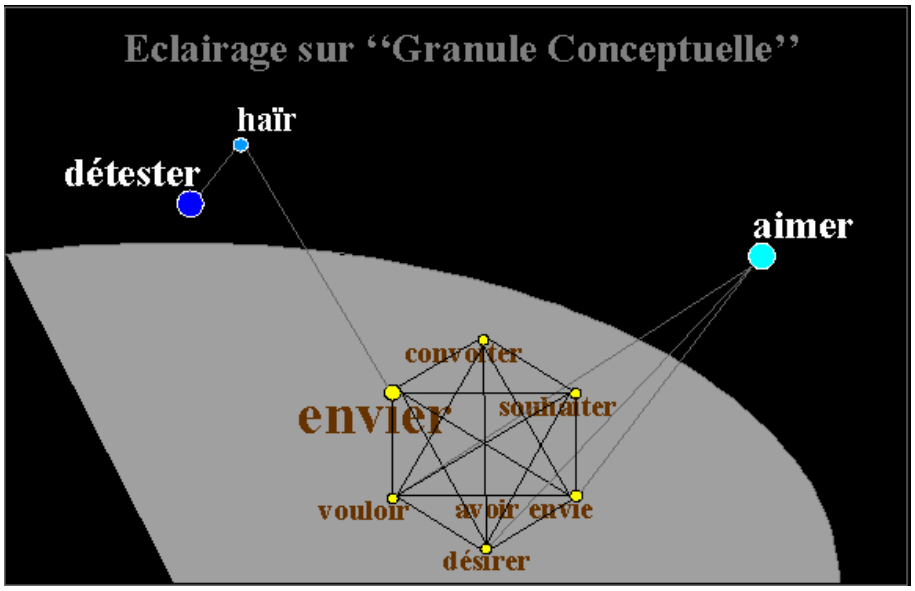


Figure 16 *amour→désir→envie→haine→détestation*

On voit ici émerger dans la Figure 16 les cliques : *« Nous soutenons que les différentes cliques dans lesquelles un mot apparaît représentent différents axes de similarité et aident à représenter les différents sens d'un mot »* (traduit de Habert et al. 1996).

Ce sont ces cliques qui sont au cœur du modèle proposé dans (Ploux & Victorri 1998). Effectivement, l'ensemble de sommets {« avoir envie », « convoiter », « désirer », « envier », « souhaiter », « vouloir »} forme une clique, ce que l'on pourrait appeler 'particule de sens' ou encore 'granule conceptuelle'. C'est d'ailleurs la granule conceptuelle qui crée la première jonction entre « aimer » et « détester » par :

« aimer »→« désirer »→« envier »→« haïr »→« détester »

En effet $V_{6,\{\text{« détester », « aimer »}\},R}$ ne devient connexe que si R>78 (voir Figure 15).

7.2. La Visualisation proxémique globale

La méthode décrite ci-dessus nous permet de faire des zooms autour d'un sommet {s} en extrayant de $\hat{G}^{\lambda}$ les R sommets qui sont les plus prox du sommet {s} à l'instant λ. On pourrait de la même manière visualiser le graphe dans sa globalité en

posant R=n (le nombre de sommets), mais du fait de la quantité d'informations alors affichées, cela devient illisible car les graphes de terrain possèdent en général un très grand nombre de sommets. Nous exposons ici une méthode qui va nous permettre d'apprécier la 'forme' globale d'un graphe même de très grande taille. Je commencerai par une métaphore.
Sur une carte du monde, ne sont cartographiées que les grandes agglomérations et les grands axes qui les relient ; si l'on veut plus de détails concernant une région donnée, on consulte une autre carte de cette région à une échelle plus fine. On pourrait dire que nos visualisations locales (quand F est un singleton du type {jouer}) jouent le rôle de cartes régionales : par exemple dans la région de « jouer », Figure 12 , c'est R qui nous donne l'étendue de la région autour de « jouer ».
Pour avoir une vision globale d'un graphe G=(V,E) sans toutefois être contraint d'en afficher tous les sommets mais uniquement les 'sommets capitales' qui sont au cœur des grandes agglomérations il nous suffit de poser F=V.

Exemple : Les Figure 17 et Figure 18 nous montrent $V_{6,V,200}$ la visualisation 3D (extraction des trois premiers axes de l'ACP sur $M_{6,V,200}$) autour de V (l'ensemble de tous les sommets) sur un rayon R=200 pour λ=6.

A partir de Dicosyn.Verbe, la forme ainsi obtenue est à peu près un tétraèdre.

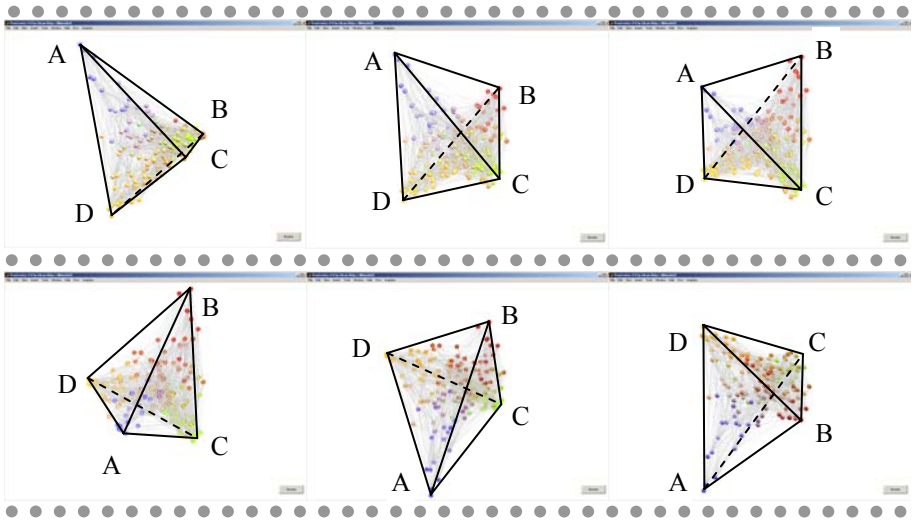


Figure 17 *$V_{6,V,200}$: Le tétraèdre conceptuel des verbes du français (200 verbes)*

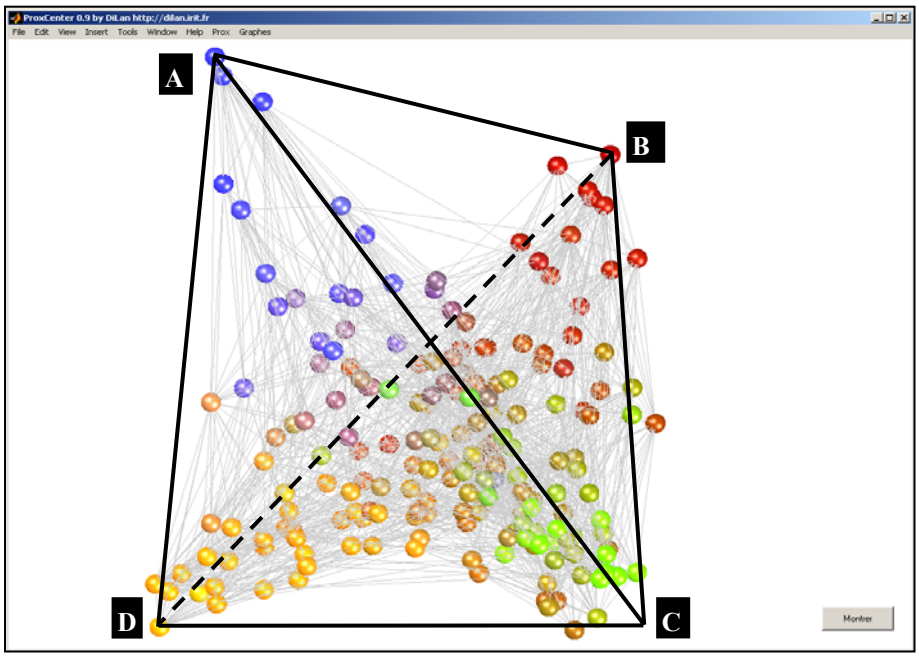
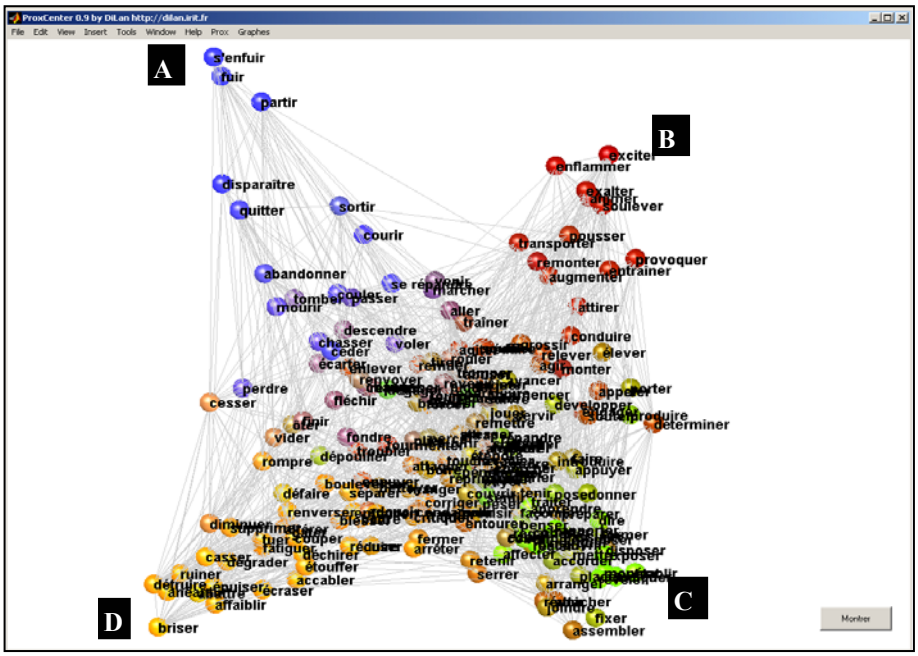


Figure 18 $V_{6,V,200}$ *: Le tétraèdre conceptuel des verbes du français (200 verbes)*



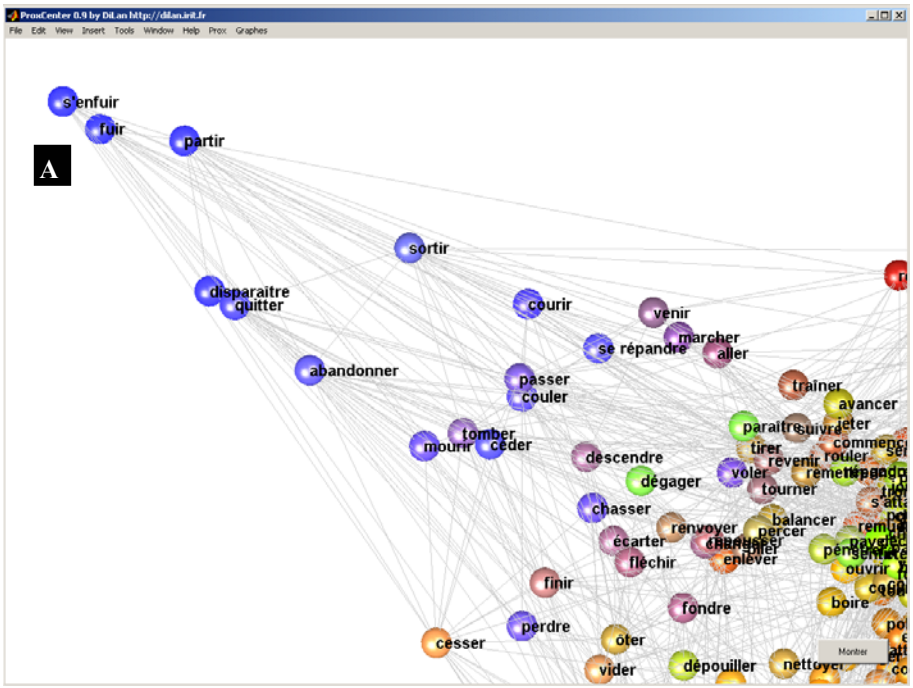


Figure 19 *$V_{6,V,200}$: Zoom sur la zone du sommet A du tétraèdre*

On trouve dans la zone du sommet A du tétraèdre conceptuel des verbes du français les verbes « partir », « fuir », « disparaître », « abandonner », « sortir », …« quitter » qui est entre « disparaître » et « abandonner ».
Si l'on remonte depuis le sommet A vers l'arête [B,C] on rencontre les verbes « s'enfuir », « fuir », « partir », « sortir », « passer », « courir », « venir », « marcher », « aller », « suivre », « avancer », « revenir », « introduire », « faire ».
Si l'on remonte toujours depuis le sommet A mais maintenant vers le sommet D en parcourant l'arête [A,D] on rencontre les verbes « s'enfuir », « fuir », « disparaître », « quitter », « abandonner », « mourir », « cesser », « perdre », « diminuer », « supprimer », « casser », « anéantir », « détruire ».

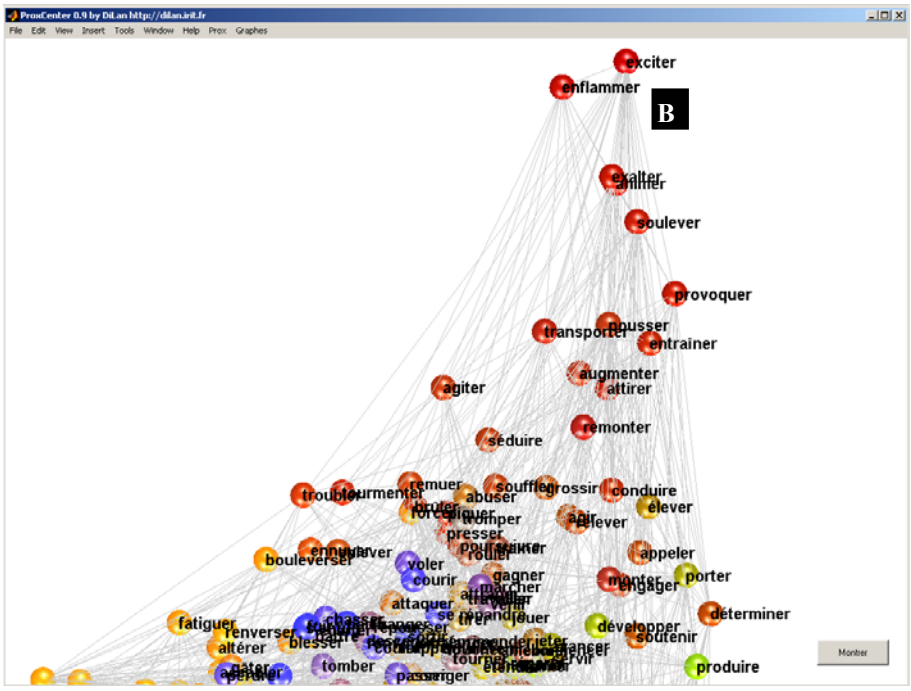


Figure 20 *$V_{6,V,200}$: Zoom sur la zone du sommet B du tétraèdre*

On trouve dans la zone du sommet B du tétraèdre conceptuel des verbes du français les verbes « exciter », « enflammer », « exalter », « animer », « soulever », « transporter », « soulever », « provoquer », « agiter », « augmenter », … « entraîner » qui est entre « attirer » et « provoquer ».

Si l'on remonte depuis le sommet B vers le sommet D en parcourant l'arête [B,D] on rencontre les verbes « exciter », « enflammer », « agiter », « tourmenter », « troubler », « ennuyer », « bouleverser », « fatiguer », « accabler », « ruiner », « détruire », « anéantir », « briser ».

Si l'on remonte toujours depuis le sommet B mais maintenant vers le sommet C en parcourant l'arête [B,C] on rencontre les verbes « exciter », « exalter », « animer », « soulever », « provoquer », « entraîner », « augmenter », « élever », « conduire », « déterminer », « produire », « former », « dire », « établir », « exposer », « indiquer », « montrer », « révéler ».

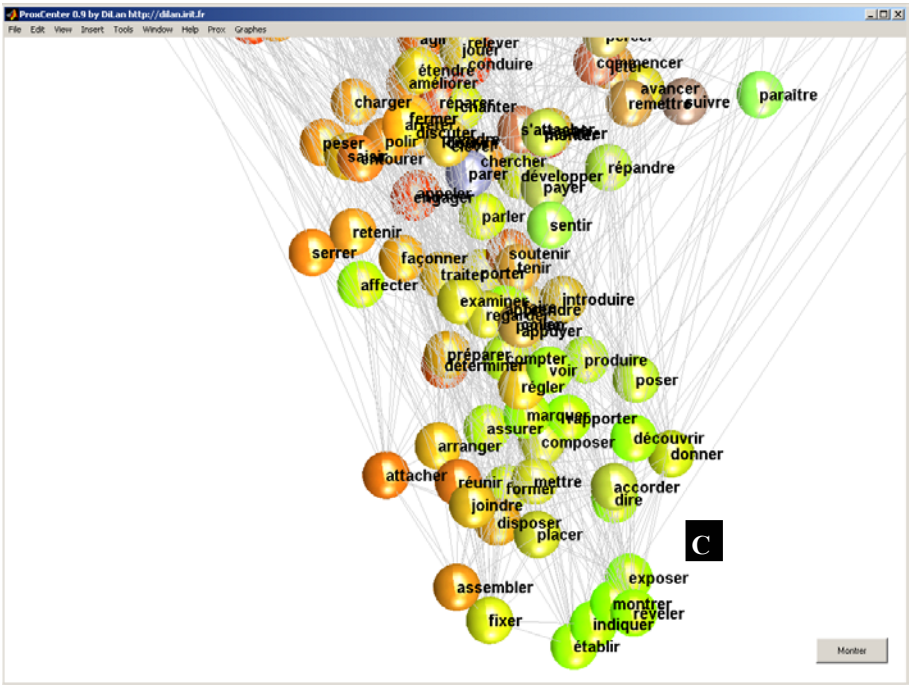


Figure 21 *$V_{6,V,200}$: Zoom sur la zone du sommet C du tétraèdre*

On trouve dans la zone du sommet C du tétraèdre conceptuel des verbes du français les verbes « assembler », « joindre », « accorder », « fixer », « établir », « indiquer », « montrer », « révéler », « exposer », « marquer », « dire », « composer », … « réunir » y est entre « attacher » et « joindre », et « révéler » y est entre « montrer » et « indiquer ».

Si l'on remonte depuis le sommet C vers le sommet D en parcourant l'arête [C,D] on rencontre les verbes « fixer », « assembler », « joindre », « réunir », « arranger », « attacher », « retenir », « serrer », « fermer », « arrêter », « cesser », « rompre », « séparer », « couper », « étouffer », « supprimer », « diminuer », « casser », « affaiblir », « abattre », « anéantir », « briser ».

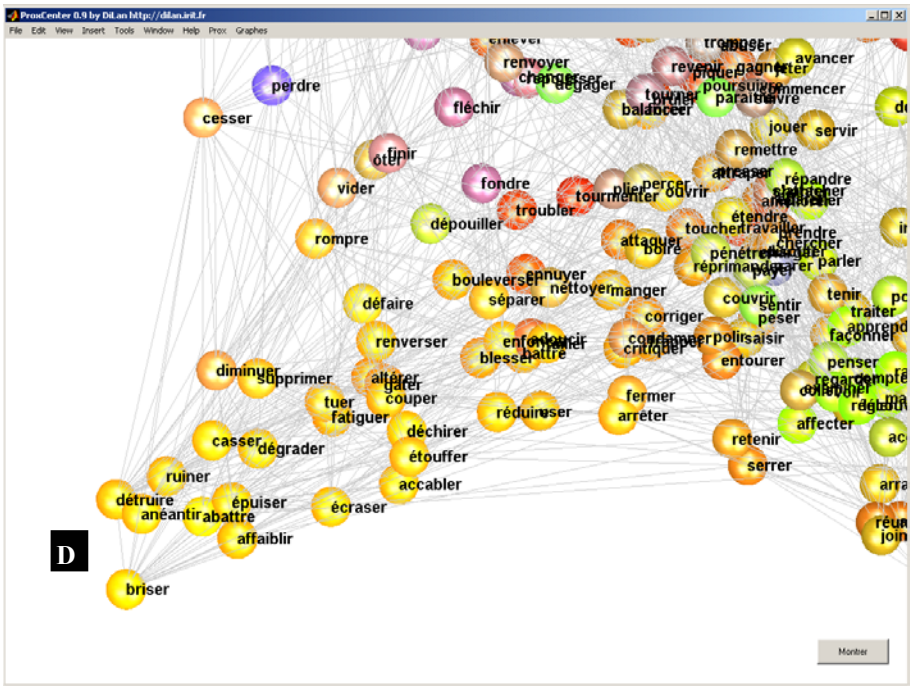


Figure 22 *$V_{6,V,200}$: Zoom sur la zone du sommet D du tétraèdre*

On trouve dans la zone du sommet D du tétraèdre conceptuel des verbes du français les verbes « briser », « détruire », « anéantir », « abattre », « affaiblir », « ruiner », « épuiser », « écraser », « casser », « dégrader », … le verbe « tuer » y est entre « altérer », « dégrader » et « supprimer ». Le verbe « accabler » n'y est pas mais pour un R plus grand, le verbe « accabler » est sélectionné, et il se trouve dans cette région conceptuelle entre « écraser », « fatiguer » et « bouleverser ».

7.3. La Visualisation proxémique sur les graphes de laboratoires

On peut se demander quelle est la pertinence de l'approche proxémique pour la visualisation de graphes artificiels plus réguliers que les graphes de type RPMH. Sans entrer dans une étude détaillée ici, voici deux exemples de visualisation de graphes artificiels simples :

1) La grille de dimension 2 à 100 sommets (un carré de 10 sommets de coté)

2) Un graphe à 8 sommets constitué de quatre sommets d'incidence 1, chacun d'eux étant relié à l'un des sommets d'une clique à 4 sommets.

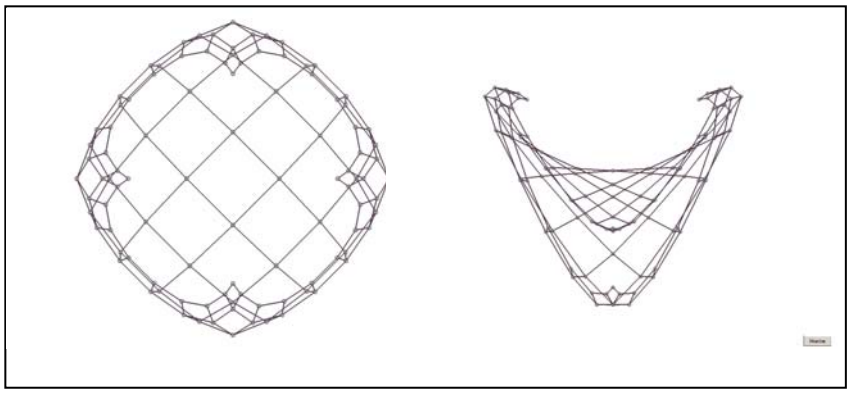


Figure 23 *Visualisation globale de la grille de dimension 2 à 100 sommets*

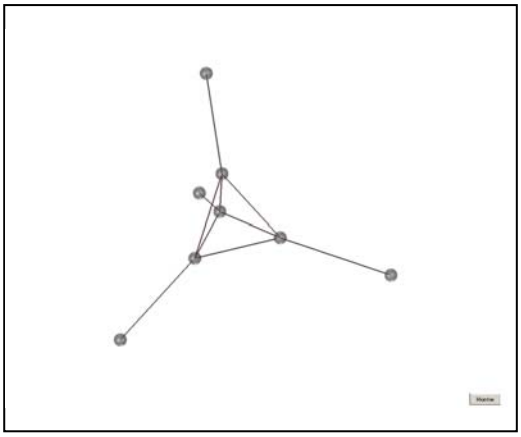


Figure 24 *Visualisation globale d'une clique à 4 sommets avec 4 sommets pendants*

On peut voir dans les Figures 23 et 24 que les visualisations 3D pratiquées reflètent assez bien la structure relationnelle des graphes concernés :

- La grille (Figure 23) est visualisée par un carré dont les angles opposés sont incurvés dans le même sens orthogonallement au plan du carré.

- Dans la Figure 24, la clique à 4 sommets est représentée par un tétraèdre régulier. Les quatre sommets pendants sont eux aussi placés sur les sommets d'un tétraèdre qui est homothétique au tétraèdre formé par les sommets de la clique par rapport à son centre.

8. Structures et complexité dans les RPMH

8.1 Les RPMH sont des systèmes complexes

L'entrée de prox est un graphe à n sommets G=(V,E) avec un nombre naturel $\lambda>0$ et sa sortie est constituée des matrices nxn $\hat{G}^{\lambda}$ et $D_{G,\lambda}$:

$$(G, \lambda) \rightarrow \textbf{[prox]} \rightarrow \hat{G}^{\lambda} \rightarrow D_{G,\lambda}$$

$\forall r,s,u \in V$, si $[D_{G,\lambda}]_{r,s}=d([\hat{G}^{\lambda}]_r, [\hat{G}^{\lambda}]_s)<d([\hat{G}^{\lambda}]_r, [\hat{G}^{\lambda}]_u)=[D_{G,\lambda}]_{r,u}$ dans $\mathbb{R}^n$ (où d est par exemple la distance euclidienne classique) cela veut dire que la dynamique de la particule quand elle débute sur le sommet r, est plus semblable à la dynamique de la particule quand elle débute sur le sommet s, que quand elle débute sur le sommet u. C'est-à-dire que prox rapproche géométriquement deux sommets du graphe dans $\mathbb{R}^n$ d'autant plus que leurs relations à l'ensemble du graphe sont semblables. Ce sont les dynamiques de la particule entièrement déterminées par la structure relationnelle $E \subseteq (V \times V)$ sur l'ensemble des sommets V qui dessinent les structures géométriques ainsi projetées dans $\mathbb{R}^n$.

Les RPMH sont des systèmes complexes dont émerge une forme globale à partir de l'ensemble des relations locales. Nous pouvons par exemple voir dans la Figure 25 la complexité de la structure du champ sémantique de « jouer » qui est 'éparpillé' dans $\mathbb{R}^n$, ce qui est dû à la structure polysémique du verbe « jouer ». On voit par exemple que les sommets « rouler », « tromper », « abuser », « exposer », « enlever », « user », … sont prox de « jouer » qui couvre ainsi un vaste champ sémantique réparti sur plusieurs régions de $\mathbb{R}^n$.

On voit bien là que faire des regroupements par des méthodes de « clustering » sur un tel graphe serait chose délicate. C'est l'un des intérêts de prox de ne jamais simplifier les graphes par clusterisation, ce qui permet d'associer vision globale et vision locale, mais aussi de superposer la structure géométrique avec la complexité polymorphique de la structure relationnelle où tous ces différent points de vue sont toujours calculés sur la globalité de la structure relationnelle du graphe.

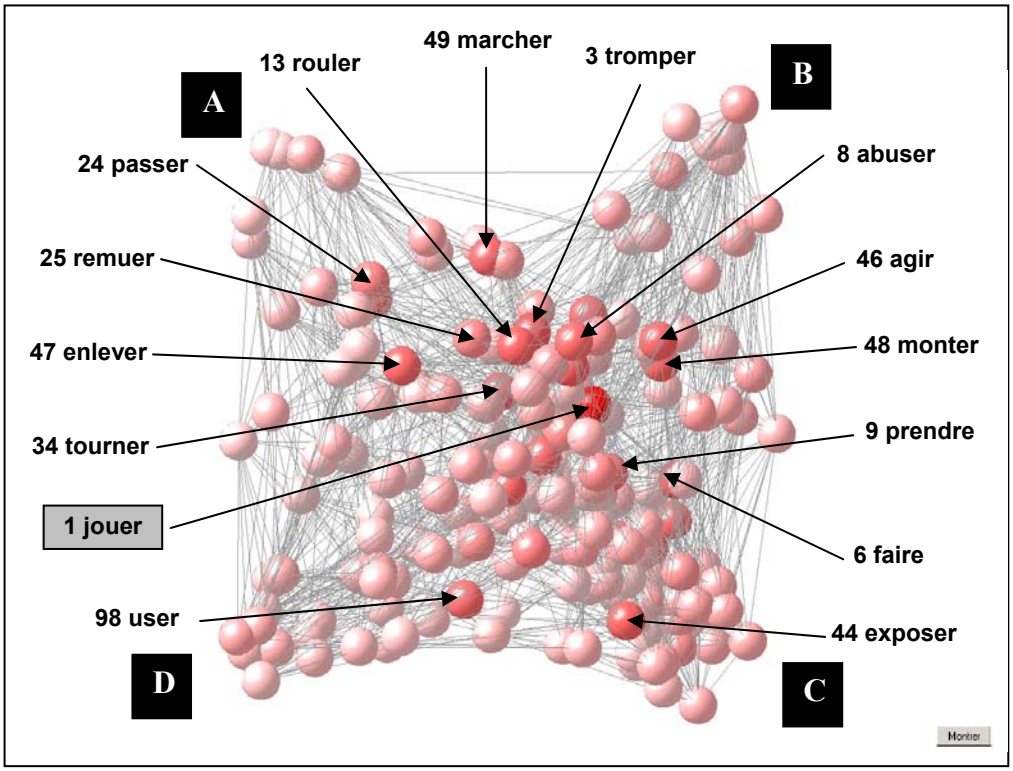


Figure 25 *La proxémie de « jouer » pour λ=6 dans* $V_{6,V,200}$

Dans la Figure 25, la vision globale $V_{6,V,200}$, plus un sommet r est foncé, plus $[\hat{G}^6]_{jouer,r}$ est grand (le nombre qui précède chaque mot est son rang dans la proxémie de « jouer »).

8.2 Complexité algorithmique de prox

L'algorithme de calcul d'un vecteur Vect=$[\hat{G}^{\lambda}]_s$ est :

```
Vect←fonction Ligne(Ĝ,λ,s)
    Vect←[Ĝ]s ;
    pour i de 2 à λ
       Vect←Vect•[Ĝ];
    fin
fin
```

De manière générale pour un graphe G=(V,E) à n sommets, l'espace mémoire nécessaire pour calculer $[\hat{G}^{\lambda}]_s$ est donc de l'ordre de $(2n+n^2)S$ (où S est l'espace nécessaire pour coder un réel), soit $O(n^2)$.
Le temps de calcul de $[\hat{G}^{\lambda}]_s$ est de l'ordre de $\lambda(T^{+}+T^{*})n^2$ (où T^{+} et T^{*} sont respectivement les temps de calcul pour une addition et une multiplication), soit $O(n^2)$.

Mais quand le graphe d'origine à n sommets G=(V,E) est un RPMH peu dense comportant K arcs, alors la matrice $\hat{G}$ est creuse car seulement K de ses n^2 valeurs sont non nulles, et de plus pour la plupart des RPMH on a : K<nlog(n).

Or lorsqu'on exploite efficacement la structure creuse des matrices $n_x n$ comme $\hat{G}$ comportant K valeurs non nulles, l'espace mémoire et le temps de calcul des algorithmes de multiplication Vecteur•Matrice restent O(K) et ne font pas intervenir l'ordre n des matrices.
Aussi quand le graphe d'origine à n sommets G=(V,E) est un RPMH peu dense comportant K arcs, le temps de calcul et l'espace mémoire nécessaire au calcul d'un vecteur ligne $[\hat{G}^{\lambda}]_s$ sont O(K) où K<nlog(n).

Cependant le nombre de sommets des RPMH peut être très grand, par exemple le nombre de pages du web est de l'ordre de 10^9, et même en restant inférieur à O(nlog(n)), le temps calcul et l'espace mémoire effectifs restent prohibitifs quand n atteint voir dépasse 10^9.

Deux voies sont alors envisageables :

1) La parallèlisation : car pour une visualisation donnée $V_{\lambda,F,R}$ on peut facilement paralléliser les calculs des R vecteurs $[\hat{G}^{\lambda}]_x$ pour tous les $x \in A_{\lambda,F,R}$ (le calcul de chacun des R vecteurs $[\hat{G}^{\lambda}]_x \in M_{\lambda,F,R}$ étant indépendant). Notons que la parallèlisation permet de diminuer le temps de calcul, mais pas l'espace mémoire nécessaire à ce calcul.

2) Les méthodes approchées de type Monte Carlo : elles se révèlent très efficaces sur les RPMH, tant pour le temps de calcul (convergence rapide sur les valeurs pertinentes, voir Figure 26) que pour l'espace mémoire. Seules les informations concernant les sommets les plus souvent visités, sont gardées en mémoire vive, sorte de mémoire cache généralisée, la mémoire est structurée en tas, et les sommets les plus souvent visités sont placés en haut du tas, les autres étant sur disque et rarement utilisées.
Pour cela il suffit de mimer une quantité q de balades de la particule dans le graphe par randomisation, sur une profondeur λ donnée. Or d'une part λ n'a pas besoin

d'être grand[32] pour des résultats pertinents grâce au L petit des RPMH qui nous intéressent ici, et d'autre part le q n'a pas besoin non plus d'être grand, car pour un λ fixé les méthodes de Monte Carlo convergent vite sur les valeurs pertinentes dans les RPMH grâce à leur C fort, leur L petit et leur structure hiérarchique[33].

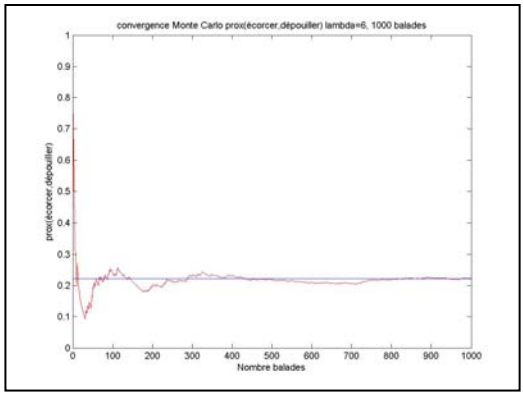


Figure 26 *Convergence vers* $[\hat{G}^6]_{écorcer,dépouiller}$ *Méthode Monte Carlo :* $\lambda=6$, $q=1000$

La Figure 26 nous montre la vitesse de convergence de la suite de coordonnées $(U_{dépouiller})_{0\leq i\leq 1000}$ vers $[\hat{G}^6]_{écorcer,dépouiller}$, pour q=1000 balades aléatoires de profondeur λ=6 depuis le sommet « écorcer » dans Dicosyn.Verbe. Bien que sur l'ensemble des q=1000 balades de profondeur λ=6 la particule aie visité qλ=1000x6=6000 sommets pour calculer la suite $(U)_{0\leq i\leq 1000}$ des 1000 vecteurs convergeant vers le vecteur $[\hat{G}^6]_{écorcer,}$ la particule est passée sur seulement 1165 sommets différents (donc peu de mémoire RAM nécessaire, ceci est dû à la structure des RPMH). De plus quand la mémoire est structurée en tas, en y plaçant les sommets les plus souvent visités en haut, les informations les concernant sont d'autant plus rapidement accessibles. Ceci nous permet par exemple de calculer une forme approchée de la visualisation $V_{6,\{jouer\},100}$ (voir Figure 27 où plus un sommet est prox de jouer, plus sa position dans la visualisation approchée est proche de sa position exacte).

[32] De manière générale, pour des résultats pertinents, prendre λ compris entre L et 2L, où L reste très petit dans les RMPH.

[33] La **''prégnance''** d'un état s depuis un état r, si on veut pouvoir lui donner un sens suffisamment empirique et calculable, doit être comprise comme « en partant d'un état r, le temps moyen que le système considéré passe dans l'état s sur un intervalle de temps relativement court » et c'est la structure du réseau qui détermine ce temps moyen passé sur s.

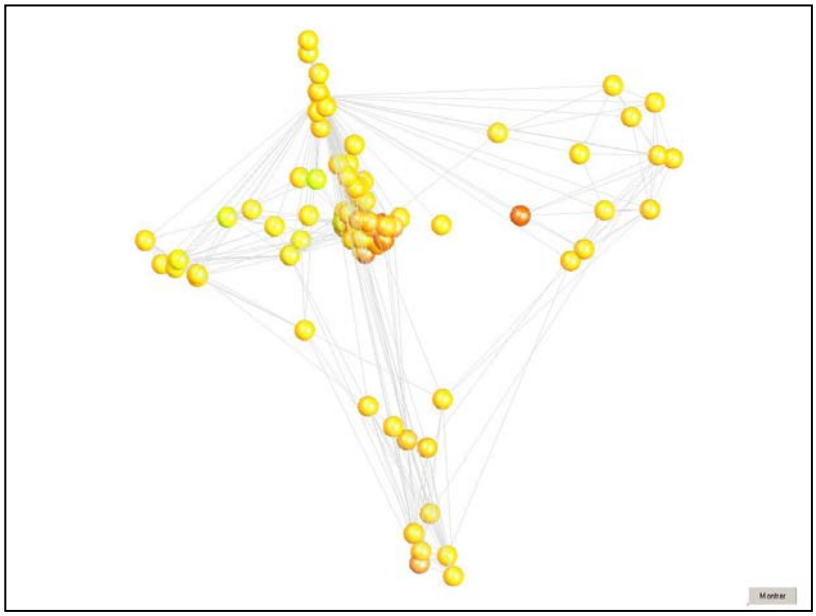


Figure 27 *Forme approchée de $V_{6,\{jouer\},100}$: Méthode Monte Carlo : λ=6, q=1000*

La forme de la figure 27 calculée pour q=1000 balades de profondeur λ=6 est l'approchée de la forme affichée dans $V_{6,\{jouer\},100}$. Ce sont exactement les mêmes sommets qui ont été sélectionnés pour y être affichés, et on peut voir que cette forme est très proche de celle de la Figure 12.

9. Perspective

9.1. Pour une navigation à ergonomie cognitive sur le web

Nous pensons que les graphes d'origine linguistique, outre leur intérêt propre dans l'étude des grands corpus linguistiques, peuvent aussi nous permettre de mieux comprendre les propriétés structurelles des grands graphes de terrain dans leur ensemble comme par exemple le world wild web (Barabasi & al. 2000).
En effet, tout comme les dictionnaires le web est un RPMH de $\approx 10^9$ sommets, c'est-à-dire qu'il ressemble par plusieurs aspects aux graphes de dictionnaires. Si les verbes du français forment un tétraèdre, quelle est alors la forme du web ?
En pratiquant des visualisations $V_{6,\_,100}$ de 100 sommets on atteint tous les mots d'un dictionnaire de verbes de 10^4 sommets comme DicoSyn.Verbe en 3 clics depuis sa visualisation globale à 100 sommets $V_{6,V,100}$. Deux clics ne suffisent pas bien que $100^2=10^4$=nombre de sommets, car la relation d'accessibilité, d'un sommet depuis la

visualisation globale à 100 sommets $V_{6,V,100}$, n'est pas un arbre mais un demi-treillis supérieur, ce qui permet d'atteindre un sommet donné par des chemins d'accès différents par exemple pour atteindre « écorcer » :

«enlever»→ «dépouiller»→«écorcer» ou bien **«couper»→ «inciser»→«écorcer»**

suivant l'état mental de l'utilisateur face au mot «écorcer», voir Figure 28.

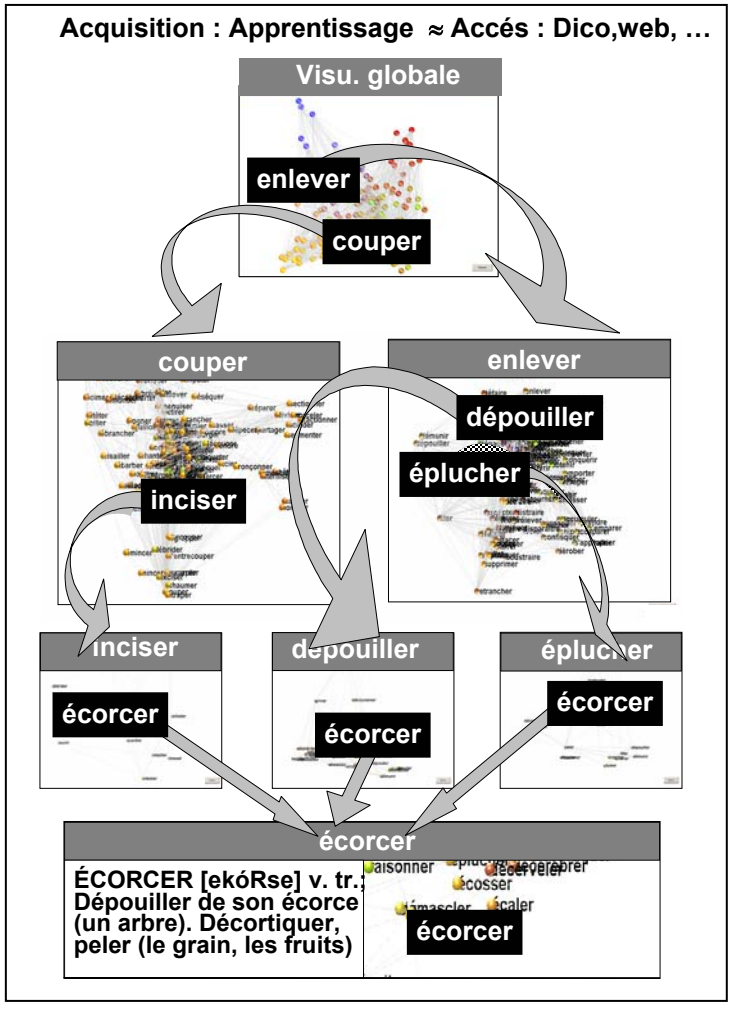


Figure 28 *Dynamique d'accès à l'information*

Le nombre de pages disponibles sur le web est de l'ordre de 10^9, soit $\approx 10^9$ sommets pour son graphe des hyperliens[34]. La visualisation globale à 100 sommets $V_{6,V,100}$ du web en est une sorte de cartographie à une grande échelle, (précisément à une échelle de $10^2/10^9=1/10^7$, seules les 'grandes' pages ou 'capitales' sont affichées). Si un clic sur un sommet S d'une visualisation renvoie $V_{6,S,100}$ (la visualisation de rayon 100 autour du sommet S), chaque clic peu alors selon le sommet, diminuer l'échelle de la visualisation d'un facteur 100, soit 5 à 6 clics pour atteindre l'échelle 1 si nécessaire dans le cas des pages les plus 'périphériques selon prox'. Sur le web, à part quelques 'fibres'[35], la plus grande partie des pages du web devraient donc pouvoir être atteinte en moins de 6 clics depuis la visualisation globale à 100 sommets $V_{6,V,100}$ du web.
La faible complexité de prox nous permet donc d'envisager un outil de navigation/visualisation pour le web, dont l'ergonomie cognitive d'accès à l'information est une métaphore de l'acquisition[36] du langage par les jeunes enfants, ce qui est plutôt de bon augure pour une démocratisation universelle de l'information à l'ère où le web tisse sa toile à travers les métissages linguistiques et conceptuels.

9.2. Théorie des Graphes

Les suites de matrices $(\hat{G}^i)_{i\in\mathbb{N}^*}$ permettent de définir une suite convergente de distances $({}^i d(G,U))_{i\in\mathbb{N}^*}$ entre deux graphes G=(V,E_1) et U=(V,E_2) à n sommets :

$$({}^i d(G,U))=(\Sigma_{x,\in V}\{\text{Dist}([\hat{G}^i]_x, [\hat{U}^i]_x)\})/n$$

où Dist est la distance euclidienne[37] dans $\mathbb{R}^n$, et $[\hat{G}^i]_x$ et $[\hat{U}^i]_x$ sont respectivement les $x^{ième}$ vecteurs lignes des matrices nxn $[\hat{G}^i]$ et $[\hat{U}^i]$. La distance d permet de définir dans un graphe G=(V,E) une notion de valeur informationnelle d'un arc $(x,y)\in E$ en posant :

---

[34] Nous ne parlerons ici que des hyperliens entre pages, mais il serait plus intéressant de les combiner avec des poids sémantiques calculés à partir du contenu des pages.

[35] Ces 'fibres' pouvant être le résultat d'une volonté consciente de la part de l'auteur de la page (ou d'un groupe d'auteurs, afin que la page en question soit peu visible par les moteurs de recherches). En effet la meilleur compréhension de la structure du web commence déjà à influer rétroactivement sur sa structure (webring, … ou bien par exemple : certains auteurs de pages, souhaitant augmenter leur 'page rank' dans Google, cherchent à renvoyer des liens vers leurs propre pages par divers moyens …)

[36] Validations et expérimentations conduites dans le cadre de : PROGRAMME COGNITIQUE : ECOLE & SCIENCES COGNITIVES ; A.C. SYSTEMES COMPLEXES EN SHS ; P.I. TRAITEMENT DES CONNAISSANCES APPRENTISSAGE ET NTIC ; ACI JEUNES CHERCHEUSES ET JEUNES CHERCHEURS 2004 (voir § 6.3).

[37] On pourrait choisir une autre distance que la distance euclidienne.

$$\forall(x,y)\in E,\ {}^{i}I((x,y)) = {}^{i}d(G,G_1) \text{ où } G1=(V, E-\{(x,y)\})$$

Certains arcs (x,y) ont une valeur informationnelle faible (typiquement un arc dans une zone dense, le supprimer, le neutraliser ou le détériorer ne change que très peu la dynamique de la particule dans le graphe. Même si la particule passe moins facilement (ou ne peut plus passer du tout) par l'arc (x,y) il existe d'autres arcs (x,r_i) et (r_k,y) qui permettent à la particule d'emprunter les chemins $\langle x,r_i,.. r_k,y\rangle$ en remplacement de l'arc (x,y). Par contre certains arcs sont plus fondamentaux ($^{i}I((x,y))$ est supérieur à la moyenne), ils maintiennent la cohésion du graphe G tout entier (typiquement un arc entre deux zones denses : « raccourci », le supprimer perturbe fortement la dynamique de la particule dans le graphe). Dans cette approche un arc n'a pas de valeur informationnelle intrinsèque, mais relativement à la globalité du graphe. Notons que l'on peut aisément étendre la définition de la valeur informationnelle à un ensemble d'arcs ou de sommets ou, plus généralement, à un sous graphe du graphe G.

9.3. Linguistique et psycholinguistique

L'introduction de la notion de « proxémie » organise les trois notions linguistiques d'hyperonymie/hyponymie, synonymie et métaphore dans un continuum sémantique, elle permet d'affiner la relation lexicale de synonymie et de reconsidérer la notion de métaphore. De plus, cette notion offre un cadre conceptuel nouveau pour l'établissement d'une catégorisation des verbes : traditionnellement la catégorisation des verbes se construit sur la base des structures syntaxiques et des restrictions de sélections (propriétés sémantiques des arguments du verbe) et ne rend pas compte du rapport d'analogie entre verbes : ainsi « soigner » et « ravaler » ne sont pas regroupés sous une même catégorie dans cette approche du fait de la différence de leurs restrictions de sélection (complément /animé/ vs /inanimé/). Au contraire, notre travail permet de regrouper ces deux verbes dans le même « agrégat » intitulé REMETTRE-EN-ETAT, ces agrégats étant eux-mêmes structurés en domaines :

REMETTRE-EN-ETAT/CORPS→ soigner, …
REMETTRE-EN-ETAT/BÂTIMENT→ ravaler, …
REMETTRE-EN-ETAT/VÊTEMENT → rapiécer, …

« Ce renversement invite à voir la métaphore (« soigner une voiture », « se faire ravaler le visage ») non plus comme un phénomène dérivé de la langue mais comme une manifestation directe de l'analogie comme principe d'ergonomie cognitive qui structure implicitement le lexique » (Duvignau 2002).

Nous proposons par exemple de développer un dictionnaire électronique 'proxémique'. Un tel dictionnaire permettra de trouver un verbe comme « écorcer »

sans le connaître en utilisant un verbe connu et analogue comme « déshabiller » et un mot permettant de cerner le domaine comme « arbre ». En effet, si l'on regarde la définition de « écorcer » il y apparaît les mots « écorce », « arbre », « grain », « fruit » qui se révèlent proches lorsque prox est appliqué sur les substantifs. Ainsi parmi les verbes proches de « déshabiller » qui sont eux-mêmes proches de « arbre » on trouve :

DÉSHABILLER/ARBRE → tailler, baguer, décortiquer, démascler, entailler, écorcer, effeuiller, émonder, gemmer, inciser

Une telle approche permet d'envisager des applications directes dans le domaine de la fouille de données et de la linguistique computationnelle : désambiguïsation, terminologie, indexation, analyse sémantique …
D'autre part la mise au jour d'une similitude entre la proximité sémantique inter-verbes observée chez l'enfant et la distance calculée par prox dans les graphes de dictionnaires permet d'ores et déjà de substituer la notion « d'approximation sémantique par analogie » à la notion « d'erreur » (Duvignau 2002, Duvignau & Gaume 2003, Duvignau & al. 2004a). Afin d'évaluer la pertinence de l'approche proxémique en tant qu'outil de modélisation cognitive des mécanismes de l'analogie chez le jeune enfant, une étude détaillée des productions verbales chez l'enfant L1 et L2 monolingue et bilingue permettra de préciser le rôle et les mécanismes de l'analogie en ce qui concerne la catégorisation des verbes. Ceci fournira des éléments pour contribuer au débat sur la nature de ces mécanismes : de type comparaison (Gentner 1989, Holyoak & Thagard 1995) ou de type projection (Hofstadter 1995, Sander 2000, Sander 2003a, Sander 2003b).

9.4. Raisonnement à granularité variable

Lorsque nous raisonnons nous ne travaillons jamais sur l'ensemble des connaissances stockées dans la base de connaissance qu'est notre mémoire à long terme (la carte détaillée du monde entier). Nous travaillons sur un 'morceau' de mémoire à long terme (la carte détaillée de la région qui nous intéresse), ou bien sur une carte globale à 'gros grains' lorsque l'on veut faire une analyse synthétique. Aussi les structures topologico-sémantiques implicites des graphes de dictionnaires apportent un éclairage nouveau en Intelligence Artificielle et Représentation des Connaissances, notamment la baisse de la complexité par diminution des espaces de recherche sur des graphes contraints par ces structures que sont les petits mondes hiérarchiques.
Par exemple si l'on cherche un chemin entre 2 villes de A vers B comment faisons nous, nous humains ?

1) On cherche une ville C_A qui est une capitale régionale de la région R_A de A et un chemin $A \rightarrow C_A$ sur une carte détaillée de R_A ;

2) On cherche une ville C_B qui est une capitale régionale de la région R_B de B et un chemin $C_B \rightarrow B$ sur une carte détaillée de R_B ;
3) On cherche un chemin $C_A \rightarrow C_B$ sur une carte nationale peu détaillée ;
4) $A \rightarrow C_A \rightarrow C_B \rightarrow B$ est le chemin que l'on cherchait ;

Nous pouvons faire de même pour un graphe G=(V,E) où l'on cherche un chemin entre deux sommets A et B.

1) Chercher un sommet C_A qui est une capitale régionale[38] de la région R_A, c'est-à-dire tel que $C_A \in ([A_{V,R,t}]) \cap ([A_{\{A\},R,t}])$ (voir § 7.1) et un chemin $A \rightarrow C_A$ dans $G_{\{A\},R,t}$ le sous graphe de G qui a pour sommets ceux de $A_{\{A\},R,t}$ qui est un petit graphe (il a R sommets)
2) Chercher un sommet C_B qui est une capitale régionale de la région R_B, c'est-à-dire tel que $C_B \in ([A_{V,R,t}] \cap [A_{\{B\},R,t}])$ et un chemin $C_B \rightarrow B$ dans $G_{\{B\},R,t}$ le sous graphe de G qui a pour sommets ceux de $A_{\{B\},R,t}$ qui est un petit graphe (il a R sommets)
3) Chercher un chemin $C_A \rightarrow C_B$ dans $G_{V,R,t}$ le sous graphe de G qui a pour sommets ceux de $A_{V,R,t}$ qui est un petit graphe (il a R sommets)
4) $A \rightarrow C_A \rightarrow C_B \rightarrow B$ est le chemin que l'on cherchait

Or (complexité de chercher $A \rightarrow C_A$ dans $G_{\{A\},R,t}$) + (complexité de chercher $B \rightarrow C_B$ dans $G_{\{B\},R,t}$) + (complexité de chercher $C_A \rightarrow C_B$ dans $G_{V,R,t}$)<<(complexité de chercher $A \rightarrow B$ dans G).

9.5. Théorie de l'information, complexité et cognition

Il est remarquable que des graphes d'origines aussi diverses que les réseaux des interactions protéiques de certaines levures, le réseau neuronal du ver *Caenorhabditis elegans*, le graphe d'Internet, le graphe des appels téléphoniques, les graphes épidémiologiques, les graphes des co-auteurs scientifiques, le graphe des collaborations cinématographiques, les graphes d'origine linguistique … soient tous de type RPMH. Nous pensons que la théorie de l'information et de la complexité peut apporter des réponses et poser des questions fondamentales relatives aux rapports entre cognition et RPMH. En effet, les matrices $\hat{G}^{\lambda}$ sont des matrices pleines de réels (elles décrivent des formes complexes dans $\mathbb{R}^n$) alors que les matrices [G] sont des matrices creuses de bits. Le passage $\hat{G}^{\lambda} \rightarrow [G]$ est donc une forme de compression de l'information contenue dans $\hat{G}^{\lambda}$ et prox son algorithme de décompression locale qui par les extractions des matrices $M_{\lambda,F,R}$ dessine $V_{\lambda,F,R}$, sorte de mémoire active de travail sur F. Cela ouvre une piste intéressante quant à la

[38] Si $C_A \in A_{V,R,t}$ on peut dire que c'est une R-capitale de G et si $C_A \in A_{\{A\},R,t}$ on peut dire qu'elle est dans la R-région de A.

modélisation du sens. En effet, si le sens a une forme géométrique[39] de type $\hat{G}^{\lambda}$ alors les RPMH de type [G] sont non seulement une excellente compression de la forme du sens, mais de plus ils permettent une navigation et un accès très efficace à l'information recherchée, une dynamique d'acquisition du général vers le particulier (enfant L1), avec une excellente robustesse en cas de déficit (aphasie, apprenant L2) (Jakobson 1963, Nespoulous 1996), ainsi qu'un raisonnement à granularité variable ce qui permet de faire chuter la complexité, mais aussi de faire du raisonnement par défaut[40].

10. Conclusion

On peut qualifier la démarche décrite ci-dessus de holiste. En effet, la forme globale tout comme les formes locales (comme par exemple $V_{6,\{jouer\},100}$ –Fig 12) ainsi calculées ci-dessus qui encodent le sens paradigmatique des mots, ne sont que des morceaux polymorphiques 'découpés' de la grande forme géométrique multidimensionnelle $\hat{G}^{\lambda}$ du graphe dans son entier calculée par prox à partir de sa seule structure relationnelle dans sa globalité[41]. Les mots tirent leurs sens-formes de leur inscription dans le tout qu'est le graphe. **En dehors de son rapport à un tout qu'est la langue le mot ne serait qu'un objet indépendant**. Dans une perspective ensembliste, les mots existent d'une certaine manière avant la langue, alors qu'au contraire dans une perspective holiste la langue existe d'une certaine manière avant les mots. Dire de la langue qu'elle est une représentation *a priori* signifie seulement que le sens des mots n'est que le produit de l'analyse de leurs rapports. C'est la langue qui doit être donnée comme objet d'étude, seul le graphe tout entier a un sens, ou plutôt le rapport des sommets entre eux. Quand on visualise le sens de « jouer » on ne visualise en fait qu'un 'autour proxémique' du sens du graphe tout entier, plus ou moins grand selon R. Si on le pouvait on visualiserait le graphe dans $\mathbb{R}^n$ tout entier, mais comme on ne le peut pas, on le cartographie dans $\mathbb{R}^3$ le plus fidèlement

[39] « Les Gestaltistes ont alors dégagé les lois de ces totalités, telles que les lois de ségrégation entre les figures et les fonds, les lois de frontières, les lois de ''bonnes formes'' ou de ''prégnance'' (les bonnes formes sont prégnantes parce que simples, régulières, symétriques… » J. PIAGET, Épistémologie des sciences de l'homme.

[40] En effet, dans les cartes globales $V_{\lambda,V,R}$, « Tweety » vole par défaut car il est dans la région sémantique des « oiseaux qui volent » alors que dans la carte locale $V_{\lambda,\{pingouin\},R}$ il « ne vole plus » car avec une échelle plus fine, il s'y dessine plus précisément la 'forme polysémique et non prototypique' de Tweety le pingouin (Kleiber 1990, Kayser 1992)

[41] En effet pour calculer les coordonnées d'un sommet u affichées dans une visualisation comme $V_{\lambda,\{s\},R}$ la particule passe par tous les chemins $\langle s, r_1,\ldots,r_{t-1},u\rangle$ du graphe G, et même si un sommet r_i n'est pas affiché dans $V_{\lambda,\{s\},R}$ la particule y est quand même passée pour pouvoir calculer les coordonnées de u dans $V_{\lambda,\{s\},R}$.

possible grâce aux techniques projectives de type *principal co-ordinates analysis* et soit :

a) localement (ex : $V_{6,\{jouer\},100}$) : avec les visualisations autour d'un singleton : par 'morceau proxémique' dont la taille peut s'ajuster selon R (R joue ici le rôle de l'étendue de la carte – plus R est grand, plus l'étendue 'autour' du singleton est vaste ; si R=n (le nombre de sommets) alors c'est la carte du monde entier).

b) globalement (ex : $V_{6,V,200}$) : avec une granularité agrégative qui peut s'ajuster selon R (R joue ici le rôle de la finesse de l'échelle – plus R est grand, plus l'échelle est fine ; si R=n (le nombre de sommets) alors c'est l'échelle 1).

Ce n'est pas le sens des parties qui donne le sens du tout, mais le sens du tout qui donne du sens à ses parties. Ce n'est pas le sens des éléments qui donne son sens au graphe, mais le sens du graphe dans la globalité de sa structure qui donne du sens à ses éléments (sommets, degrés, cliques, zones denses, chemins, cycles …).

Dans notre modèle holiste du sens géométrisé, chaque mot M d'un réseau lexical de n mots donne naissance à trois objets géométriques :

1) $\mathbf{e_M}$ qui est le $M^{ième}$ vecteur de la base orthonormée canonique de l'espace sémantique à n dimension : $e_M=[x_i, \ldots, x_n]$ où $x_i=1_{(i=M)}$.

2) $\mathbf{G_M}$ qui est la M^{ieme} Granule de sens. G_M est un point dans l'espace sémantique à n dimensions dont les coordonnées dans la base $(e_i)_{1\leq i\leq n}$ sont données par le vecteur $[\hat{G}^{\lambda}]_M$ (dans les Figures 18 à 22, ce sont donc des granules de sens qui sont visualisées dans l'espace sémantique).

3) $\mathbf{S_M}$ qui est le sens potentiel hors contexte du mot M : S_M est une fonction de poids sur l'ensemble des n Granules de sens : $S_M(G_i)=[\hat{G}^{\lambda}]_{M,i}$ (c'est donc S_{jouer} qui est visualisé dans la Figure 25 : plus un sommet r est foncé, plus $S_M(G_r)$ est grand).

Je remecie les relect/rices/eurs de la revue *Information, Interaction, Intelligence*, qui, par leurs questions et leurs conseils toujours pertinents, m'ont permis d'améliorer substantiellement cet article.

11. Références

Abello & al. (1999) : Abello J., Pardalos P.M., Resende M.G.C., *On maximum cliques problems in very large graphs*. External memory algorithms, J. Abello and J. Vitter, Eds., DIMACS Series on Discrete Mathematics and Theoretical Computer Science, vol. 50, pp. 119-130, American Mathematical Society, 1999 http://www.research.att.com/~mgcr/doc/vlclq.ps.Z

Adamic (1999) : Adamic L. A., The small world Web. http://www.hpl.hp.com/shl/papers/smallworld/smallworld.pdf

Ancel & al. (2001) : Ancel L. W., Newman M. E. J., Martin M., Schrag S., *Applying Network Theory to Epidemics*, Control Measures for Outbreaks of "Mycoplasma pneumoniae", *SFI Working Paper*, n° 01-12-083, 2001 http://www.santafe.edu/sfi/publications/Working-Papers/01-12-083.ps.gz

Barabazi & al. (1999) : Barabazi A.-L., Albert R., Jeong H., *Scale-free characteristics of random networks : The topology of the World Wide Web*.(http://www.nd.edu/~networks/proceeding.ps

Barabasi & al. (2000) : A.-L. Barabasi, R. Albert, H. Jeong, and G. Bianconi, *Power-Law Distribution of the World Wide Web*, Science 287 2115a (in Technical Comments) 2000 http://www.nd.edu/~networks/Papers/comments.pdf

Berchtold (1998) : Berchtold A., *Chaînes de Markov et modèles de transitions, applications aux sciences sociales*, HERMES

Berge (1983) : Berge C., *Graphes*, Bordas édition. Gauthier-Villars.

Bermann & Plemons (1994) : Bermann A., Plemons R.J., *Nonnegative Matrices in the Mathematical Sciences* Siam : Classics in applied Mathematics, 1994

Bollobas (1986), Bollobas B. : *Combinatorics: Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability*, Bollobas, 1986

Bollobas B. (1998) : Bollobas B., *Modern Graph Theorie*, Graduate text in mathematics, SPRINGER, 1998

Brémaud P. (2001) : Brémaud P.,*Markov Chains, Gibbs Fields, Monte carlo simulation, and Queues,* SPRINGER, 2001

Diestel (2000) : Diestel R., *Graph Theory (second edition)* SPRINGER, 2000

Duvignau (2002) : Duvignau K., *La métaphore berceau et enfant de la langue*. Thèse de l'Université Toulouse – Le Mirail, 2002

Duvignau (2003) : Duvignau, K., *Métaphore verbale et approximation*. In Regards Croisés sur l'Analogie, in Revue d'intelligence Artificielle, n° 5-6, (2003)

Duvignau & Gaume (2003) : Duvignau K., Gaume B., *Linguistic, Psycholinguistic and Computational Approaches to the Lexicon: For Early Verb-Learning*. Cognitive Systems, Janvier, Vol 6 (1) 2003

Duvignau & al. (2004a) : Duvignau K., Gardes-Tamine J., Gaume B., *Pour un enseignement spécifique du lexique verbal chez le jeune enfant : quelques données et propositions*. In Le langage et l'homme 2004 (à paraître)

Duvignau & al. (2004b) : J.-.L Duvignau K. , Gaume B., Nespoulous J.-.L (sous presse) *''lexicalisation, proximité sémantique et stratégies palliative''* In *Revue Parole*, numéro spécial : « Handicaps langagiers, Sciences du Langage et de la Cognition, Sciences et Technologies de l'Information et de la Communication : apports mutuels. De la caractérisation des handicaps à la mise en place de stratégies palliatives, comportementales et/ou technologiques », UMH, Belgique (à paraître en 2004)

Erdös & Renyi (1960) : Erdös P. and Renyi A., ***Publ. Math. Inst. Hung. Acad. Sci*** 5, 17-61, 1960

Fellbaum (1999) : Fellbaum C., *La représentation des verbes dans le réseau sémantique WordNet*. In Langages, Sémantique lexicale et grammaticale, 136

Ferrer & Solé (2001) : Ferrer, R., Solé, R. V. (2001). The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1482):2261—2265, 2001
http://www.santafe.edu/sfi/publications/Working-Papers/01-03-016.pdf

Ferré & Jouve (2002) : Ferré L., Jouve B., Vertex partitioning of a class of digraphs, Mathématiques Informatiques et sciences humaines, 158, 59-77, 2002.

Gaume & Duvignau (2004) : Gaume B., Duvignau K., "Pour une ergonomie cognitive des dictionnaires électroniques, in *Fouille de textes et organisation de document*, numéro spécial de la revue Doccument Numérique, Hermes , 2004.

Gaume & Ferré (2004) : Gaume B., Ferré L., *Représentation de Graphes par ACP Granulair*, in actes d'EGC 2004 : 4èmes journées d'Extraction et de Gestion des Connaissances, Clermont Ferrand, 20-23 Janvier 2004.

Gaume & al. (2004) : Gaume B., Hathout N., Muller P., Word sense disambiguation using a dictionary for sens similarity measure *in acte COLING 2004,* The 20th International Conference on Computational Linguistics, COLING 2004, Geneva

Gaume (2003) : Gaume B., *Analogie et proxémie dans les réseaux petits mondes*, Regards Croisés sur l'Analogie, in Revue d'intelligence Artificielle, n° 5-6,

Gaussier (1999) : Gaussier E. , *Unsupervised Learning of Derivational Morphology from Inflectional Lexicons* in act ACL'99, Workshop Unsupervised Learning in Natural Language Processing

Gentner (1989) : Gentner, D., *The mechanisms of analogical learning*, in S. Vosniadou and A. Ortony (Eds.), Similarity and Analogical Reasoning, (pp. 199-241). Cambridge: Cambridge University Press

Guare (1990) : Guare, J., *Six degrees of separation* : A play, Vintage Books, New York, 1990

Habert et al. (1996) : Habert B., Naulleau E. et Nazarenko A., Symbolic word clustering for medium-size corpora. *In: 16th International Conference on Computational Linguistics*, pp. 490-495. Copenhague, Danemark, 5-6 août 1996

Harris (1951) : Harris Z, *Méthods in structural linguistics*, Chicago University Press, 1951.

Hofstadter (1995) : Hofstadter, D., *Fluid concepts and creative analogies*. New York: Basic Books.

Holyoak & Thagard (1995) : Holyoak, K. J., & Thagard, P., *Mental leaps: Analogy in creative thought*. Cambridge, MA: The MIT press

Hubalek (1982) : Hubalek Z., *Coefficients of association and similarity based on (presence, absence) : an evaluation*, Bilogical Rev., vol 57, 1982, p. 669-689

Huberman & Adamic (1999) : Huberman B. A., et Adamic L.A., Growth dynamics of the world-wide web, *Nature* 401:131. (http://xxx.lanl.gov/abs/cond-mat/9901071)

Véronis & Ide (1990) : Véronis J., Ide N., *Word Sense Disambiguation with very large neural networks extracted from machine readable dictionaries*. Proceedings of the 14th International Conference on Computational Linguistics, Helsinki http://www.up.univ-mrs.fr/~veronis/pdf/1990coling.pdf

Ide & Véronis J (1998) : Ide N., Véronis J., *Introduction to the Special Issue on Word Sense Disambiguation : The State of the Art*. *Computational Linguistics* 24(1), 1: .40, 1998

Jakobson (1963) : Jakobson R., *Essais de linguistique générale*, 1. Chap II. Deux aspects du langage et deux types d'aphasie : 43-67 Traduction : Ruwet, N.Minuit

Jeong & al. (2001) : H. Jeong, S.P. Mason, A.-L. Barabasi and Z.N. Oltvai, *Lethality and centrality in protein networks*, *Nature* 411 41, 2001 http://www.nd.edu/~networks/cell/papers/protein.pdf

Jouve & al. (1998) : Jouve B., Rosenstiehl P., Imbert M., A mathematical approach to the connectivity between the cortical visual areas of the macaque monkey, Cerebral Cortex (8), 8-39, 1998

Jouve & al. (2001) : Jouve B., Kuntz P., Velin E., Extraction de structures macroscopiques dans des graphes par une approche spectrale. In Extraction des connaissances et Apprentissage, vol. 1, n° 4, HERMES (eds), Paris, 2001

Karov & Edelman (1998) : Karov Y., Edelman S., *Similarity-based Word Sense Disambiguation*. Computational Linguistics (1998). 24(1), 41-59

Kayser (1992) : Kayser D.,Profondeur variable et Sciences Cognitives. in Introduction aux sciences cognitives (sous la direction de D.Andler) pp.195-218 Gallimard, Coll.folio/essais, 1992

Kleiber (1990) : Kleiber G., La sémantique du prototype, catégorie et sens lexical, PUF Linguistique nouvelle (1990)

Kleinberg & al. (1999) : Kleinberg J.M., Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, AndrewS. Tomkins, *The web as a graph: measurments, models,*

and methods. In Proceedings of the Fifth International Conference on Computing and Combinatorics, Tokyo July 26-28, 1999 (COCOON'99). Berlin: Springer-Verlag, pages1-17

Kochen (1989) : Kochen M. (ed.), *The small world*, Ablex, Norwood, NJ, 1989

Kuntz & al. (2001) : Kuntz P., Velin F., Briand H., *Iterative geometric representations for multi-way partitioning*, In proc. of the 5th WSES/IEEEworld Multiconf. on circuits, Systems, Communications and Computer in Advances in Scientific Computing, Computational Intelligence and Applications, WSEAS Press, 108-113, 2001.

Lebart & Salem (1994) : Lebart L. Salem A. (1994), *Statistique textuelle*, Paris Dunod 1994

Lebart (2001) : Lebart L., Classification et Analyse de Contiguïté, La revue de modulad, 27, 1-22, 2001

Milgram (1967) : Milgram, S., *The small world problem*. Psychol. Today 2, 60-67, 1967

Mohar (1991) : Mohar B., *The Laplacian spectrum of graphs* In Y. Alavi, G. Chartrand, Ollermann and A Schwenk, Editors, Graph Theory, Combinatronics and Applications, 871-898, New-York, 1991. John Wiley and Son.

Nespoulous (1996) : Nespoulous J-L, *Les stratégies palliatives dans l'aphasie*. Rééducation Orthophonique, 34 (188):423-433, 1996

Newman (2003a) : Newman M.E.J., *The structure and fonction of complex networks*, http://www.santafe.edu/~mark/recentpubs.html

Newman (2003b) : Newman M.E.J., Ego-centerer networks and the riple effect, *Social Networks* 25, 83-95, 2003

Ploux & Victorri (1998) : Ploux S., Victorri B., *Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes*, Traitement automatique des langues, 39(1):161-182

Ravasz & Barabási (2003) : Ravasz E., Barabási A.L. *Hierarchical Organization in Complex Networks*. Phys. Rev. E 67, 026112, 2003
http://arxiv.org/abs/cond-mat?0206130

Redner (1998) : *How Popular is Your Paper? An Empirical Study of the Citation Distribution*, Redner S., cond-mat/9804163, *European Physical Journal B*, **4**, 131-134 (1998). http://cbd.bu.edu/members/sredner.html

Sander (2000) : Sander, E., *L'analogie, du Naïf au Créatif : analogie et catégorisation*. Paris, L'Harmattan

Sander (2003a) : Sander, E., Les *analogies spontanées : analogies ou catégorisations* In C. Tijus (Ed.). Métaphores et Analogies, pp. 83-114. Paris, Hermès

Sander (2003b) : Sander, E., *Analogie et catégorisation*. In Regards Croisés sur l'Analogie, in Revue d'intelligence Artificielle, n° 5-6, Hermes Sciences, Duvignau K., Gaume B., Gasquet O. (éditeurs) 2003 (à paraître)

Saussure (1972) : Saussure F., *Cours de linguistique générale*, édition critique préparée par Tullio De Mauro, Paris 1972

Senata (1981) : Senata E., *Nonnegative Matrices and Markov Chains, (2nd edition)*, SPRINGER, New York, 1981

Sigman & Cecchi (2002) : Sigman M., Cecchi G.A., *Global organization of the Wordnet lexicon*, Proc. Natl. Acad. Sci. 99(3):1742-7

Victorri & Fuchs (1996) : Victorri B., Fuchs C., La polysémie – Construction dynamique du sens, Paris, Hermès, 1996

Veronis, (2002a) : Véronis J., « Les dictionnaires traditionnels sont-ils adaptés au traitement du sens en T.A.L. ? », Journée d'étude de l'ATALA Les dictionnaires électroniques, Paris, (2002)

Veronis (2002b) : Véronis J., « Vers une lexicographie distributionnelle », Colloque Sémantique et corpus, Toulouse, (2002)

Watts (1999) : Watts D.J., *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, 1999

Watts & Strogatz (1998) : Watts D.J., Strogatz S.H., *Collective dynamics of 'small-world' networks*. **Nature** 393 :440-442, 1998
http://tam.cornell.edu/SS_nature_smallworld.pdf