# Cartographier la forme du sens dans les petits mondes lexicaux

Bruno Gaume

IRIT, UPS, 118 route de Narbonne, 31062, Toulouse cedex 4, France, gaume@irit.fr

## Abstract

In this paper we propose Prox which is a stochastic method to map the local and global structures of real-world complex networks called Small Worlds. Prox transforms a graph into a Markov chain, the states of which are the nodes of the graph in question. Particles wander from one node to another within the graph by following the graph's edges. It is the dynamics of the particles' trajectories that map the structural properties of the graphs that are studied. Concrete examples are presented in a graph of synonyms to illustrate this approach.

## Résumé

Dans cet article on présente Prox qui est une méthode stochastique pour la métrologie et la cartographie des structures locales et globales des grands graphes de terrain de type Hierarchical Small Worlds. Prox 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 cartographient les propriétés structurelles des graphes étudiés. Pour illustrer cette approche, des exemples concrets sont présentés sur un graphe de synonymes.

**Mots-clés :** Lexicométrie, graphe petit monde, synonymie, métaphore, hyperonymie, cartographie sémantique

## 1. Introduction

Des recherches récentes en théorie des graphes ont mis au jour un ensemble de propriétés remarquables que partagent la plupart des graphes de terrain ; ces caractéristiques définissent la classe des graphes de type Hierarchical Small Worlds (HSW). Ainsi en va-t-il des réseaux des interactions protéiques, du graphe du web, du graphe des appels téléphoniques, du graphe des co-auteurs scientifiques, des réseaux lexicaux, … Ces graphes ont une topologie bien particulière, dans laquelle la relation entre structure locale et structure globale n'a rien à voir avec celle des graphes (aléatoires ou réguliers) classiquement étudiés. Ceci explique l'intérêt considérable que les HSW ont suscité dans les communautés scientifiques concernées. En effet, on peut penser que ces caractéristiques reflètent les propriétés 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. Dans cet article nous présentons Prox qui est une méthode particulièrement bien adaptée à la métrologie et la cartographie des HSW. Le § 2 résume brièvement les principales propriétés des réseaux HSW et présente rapidement les réseaux lexicaux qui nous servirons d'exemples concrets dans cet article. Au § 3 nous verrons comment la dynamique d'une marche aléatoire sur un HWS est fortement contrainte et orientée par la structure topologique des HSW. Au § 4 nous montrerons que l'analyse de la dynamique des balades aléatoire sur les HSW permet de cartographier la forme du sens dans les réseaux HSW. Au § 5 nous exploiterons ces dynamiques afin de superposer l'information globale contenue dans un réseau à une

extraction topologique pratiquée sur ce réseau et projetée dans une carte locale. Au § 6 nous discuterons de la complexité de Prox pour ensuite conclure au § 7.

## 2. Les propriétés des graphes de terrain

Pour une présentation des HSW voir par exemple (Newman, 2003). Les graphes de terrain, sont peu denses, dans un graphe à n sommets, le nombre maximum d'arcs possibles est de $n^2$ alors qu'en général le nombre d'arcs des grands graphes de terrain est O(nlog(n)) et non $O(n^2)$. Watts et Strogatz (1998) proposent deux indicateurs pour caractériser un grand graphe G symétrique peu dense : son L et son C.

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

**– C=** le taux de clustering : $C \in [0,1]$, et mesure la tendance qu'a un graphe a posséder des zones plus denses en arcs. (C=Nombre_de_triangles/Nombre_de_fourches, plus le graphe a tendance a posséder des agrégats denses en arrêtes, plus le C du graphe est proche de 1).

En appliquant ces critères à des graphes de terrains d'origines divers ils constatent que :

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

**2) Les graphes de** terrain on tendance à avoir un grand C, ce qui reflète une relative tendance qu'ont deux voisins d'un même sommet à être dirtectement connectés.

**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 étudiés, 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 (au sens de Watts et Strogatz : peu de triangles par rapports fourches).

En écho au « small world phenomenon » (Milgram, 1967) Watts et Strogatz proposent d'appeler **« small worlds »** les graphes qui ont cette double caractéristique (L faible et C fort) qu'ils identifient dans tous les graphes de terrain qu'ils observent, et dont ils postulent l'universalité. Des travaux plus récents montrent que la plupart des graphes petits mondes, ont de plus une structure hiérarchique. La distribution des degrés d'incidence des sommets suit une loie de puissance. La probabilité P(k) qu'un sommet du graphe considéré ait k voisins décroît comme une loi de puissance $P(k) = k^{-\lambda}$, où $\lambda$ est une constante caractéristique du graphe (Barabási, Albert, Jeong and Bianconi, 2000), alors que dans le cas des graphes aléatoires, c'est une loi de Poisson qui est à l'œuvre.

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 deux 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. (Manning, Schütze, 2002).

– **Relations paradigmatiques**, notamment de synonymie ; à partir de bases de données lexicales, comme par exemple WordNet (Fellbaum, 1998), on construit un graphe dans lequel deux sommets sont reliés par une arête si les mots correspondants entretiennent une relation synonymique.

Tous ces graphes sont à l'évidence de type HSW (Gaume, 2004), graphes peu denses, présentant une structuration locale riche (un C fort) et une distance moyenne très petite sur l'ensemble du graphe (un L petit), ainsi qu'une structure hiérarchique, (la courbe d'incidence ≈ loi de puissance). Par exemple, DicoSyn.Verbe[1] est un graphe symétrique et réflexif de 9 043 sommets, il a 50 948 arêtes et sur sa plus grande partie connexe (8 835 sommets) son L est égal à 4.17 et son C est égal à 0.32, c'est typiquement un HSW. La courbe représentant la distribution des degrés d'incidence de ses sommets (voir Fig 1.) est caractéristique des réseaux HSW (en log-log elle forme approximativement une droite dont le coefficient directeur est égal à -1.80 avec un coefficient de détermination égal à 0.92).

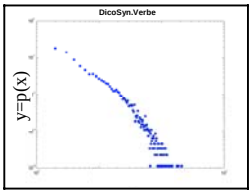



*Fig 1. Courbe log-log de la distribution de l'incidence des sommets DicoSyn.Verbe*

Les graphes lexicaux de substantifs, de verbes, … extraits de dictionnaires généraux, (Robert, Larousse, TLFi) ou de dictionnaires de synonymes (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse , Robert) ou bien encore leur compilation sont tous des HSW typiques.

Aussi nous posons le postulat général suivant :

> *Les propriétés structurelles des réseaux lexicaux qui sont de type petits mondes hiérarchiques révèlent une organisation linguistique sous-jacente et ont une validité cognitive liée à l'efficacité des systèmes d'encodage de la mémoire pour l'acquisition du lexique, l'accès dynamique et contextuel au sens, et la robustesse aux défaillances.*

Les outils de cartographie lexicale présentés dans cet article s'inscrivent dans un programme de recherche qui consiste d'une part à valider et affiner ce postulat général en linguistique et en psycholinguistique en définissant des outils théoriques et appliqués à la métrologie lexicale pour en tester la pertinence du point de vue linguistique et psycholinguistique par l'expérimentation et d'autre part à implémenter nos ressources lexicales et expérimentales avec leurs métrologies associées sur une plateforme pour le TALN (dictionnaires et corpus construits par extractions focalisées sur le web *'focus crawling'*) l'école (apprentissage) la clinique (marqueurs linguistiques de pathologies) et la recherche d'information sur le web (moteur de recherche à ergonomie cognitive).

[1] DicoSyn est 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 à l'ATILF (http://www.atilf.fr/) puis homogénéïsés au CRISCO (http://elsap1.unicaen.fr/). 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.

## 3. Balades aléatoires dans les réseaux lexicaux

**Notation :**

Si U est un vecteur ligne de dimention n nous noterons : $[U]_i$ : la $i^{eme}$ valeur de U ;

Si M est une matrice $n_x m$ alors nous noterons pour tout i,k tels que $1 \leq i \leq n$ et $1 \leq k \leq m$ :
$[M]_{i\,k}$ : la valeur située à l'intersection de la $i^{eme}$ ligne et la $k^{eme}$ colonne de M ;
$[M]_{i\,\bullet}$ : le $i^{eme}$ vecteur ligne de M ;
$[M]_{\bullet\,k}$ : le $k^{eme}$ vecteur colonne de M ;

Supposons que nous disposions d'un graphe connex, symétrique et réflexif, G=(V,E) de n=|V| sommets et m=|E| arcs, et que sur ce graphe une particule puisse à chaque instant t∈ se balader aléatoirement sur ses sommets :
- À l'instant t la particule est sur un sommet r∈V.
- Quand la particule est à un instant t sur un sommet u∈V, elle ne peut atteindre à l'instant t+1, que les sommets s∈V tels que (u,s)∈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∈V, chacun des arcs sortants de u est équiprobable.

Soit $\hat{A}$ la matrice de transition en un pas de la chaîne de Markov correspondant à la balade aléatoire sur le graphe. C'est à dire qu'à chaque étape, la probabilité de transition du sommet r∈V au sommet s∈V est égale à $[\hat{A}]_{r\,s}=[A]_{r\,s}/d(r)$ (où A est la matrice d'adjacence du graphe G et d(r) le degré du sommet r).

Si la loi initiale de la chaîne de Markov est donnée par le vecteur ligne P (c'est-à-dire que $[P]_r$ est la probabilité que la particule soit sur le sommet r à l'instant t=0) alors $[P\hat{A}^t]_s$ est la probabilité que la particule soit sur le sommet s à l'instant t.

Soit F⊆V, un ensemble non vide de k sommets. Notons $P^F$ le vecteur de dimension n tel que : $[P^F]_r=1/|F|$ si r∈F, et $[P^F]_r=0$ si r∉F. Si la loi initiale de la chaîne de Markov est donnée par le vecteur $P^F$, cela correspond donc à une balade aléatoire, débutant sur l'un des sommets de F, tous équiprobables. $[(P^F)\hat{A}^t]_s$ est alors la probabilité que la particule soit sur le sommet s à l'instant t quand la particule commence la balade aléatoire équiprobablement sur l'un des sommets de F à t=0. On remarquera que $[(P^{\{r\}})\hat{A}^t]_s=[\hat{A}]_{r\,s}$ qui est donc la probabilité que la particule soit sur le sommet s à l'instant t quand la particule commence la balade aléatoire sur le sommet r à l'instant t=0.

On démontre[2] que si G=(V,E) est un graphe fortement connexe et réflexif, alors :

$$\forall r,s,u \in V,\ \lim_{t\to\infty}[\hat{A}^t]_{r\,s} = \lim_{t\to\infty}[\hat{A}^t]_{u\,s} = d(s)/\sum_{x\in V}\{d(x)\} \qquad (1)$$

C'est à dire que la probabilité pour un temps t assez long de se trouver sur un sommet s ne dépend plus du sommet de départ r ou u, mais uniquement du sommet s, et est égale à $d(s)/\sum_{x\in V}\{d(x)\}$. Cependant deux types de configurations topologiques peuvent opposer les deux sommets r et u dans leurs rapports avec le sommet s.

**Configuration 1)** les sommets r et s peuvent être reliés par un grand nombre de chemins courts (r et s sont fortement reliés : il existe une confluence forte de chemins de r vers s) ;

[2] C'est une conséquence du théorème de Perron Froebenius (Bermann, Plemons, 1994) car quand le graphe G=(V,E) est réflexif et fortement connexe la matrice de transition $\hat{A}$ de la chaîne de Markov associée à la balade aléatoire sur le graphe G est alors ergodique (Gaume, 2004) (ici la connexité forte et la réflexivité sont nécessaires pour la preuve d'ergodicité).

**Configuration 2)** les sommets u et s peuvent être reliés par seulement quelques chemins courts (u et s sont faiblement reliés : pas de confluence de u vers s) ;

Si la formule (1) exprime que la probabilité pour un temps t assez long de se trouver sur un sommet s ne dépend pas du sommet de départ r ou u, par contre la dynamique vers cette limite dépend fortement du sommet de départ et du type de confluence qu'il entretient vers le sommet s. C'est-à-dire que les suites $([\hat{A}^t]_{r\,s})_{0\leq t}$ et $([\hat{A}^t]_{u\,s})_{0\leq t}$ ne sont pas identiques même si elle convergent vers la même limite $d(s)/\sum_{x\in V}\{d(x)\}$. En effet la dynamique de la trajectoire de la particule dans sa balade est entièrement gouvernée par la structure toplogique du graphe : après t pas, tout sommet s à une distance de t arcs ou moins du sommet de départ peut être atteint. La probabilité d'atteindre un sommet s au $t^{ième}$ pas dépend du nombre de chemins entre le sommet de départ et le sommet s, de leurs longueurs et de la structure du graphe autour des sommets intermédiaires le long des chemins (plus il y a de chemins, plus ces chemins sont courts, et plus le degré des sommets intermédiaires est faible, plus la probabilité d'atteindre s depuis le sommet de départ au $t^{ième}$ pas est grande quand t reste petit). Ainsi si il existe une confluence plus forte de chemins depuis le sommets r vers s que depuis le sommet u vers s, alors pour une balade aléatoire de longueur t pas trop longue, on a $[\hat{A}^t]_{r\,s}>[\hat{A}^t]_{u\,s}$. Au dèbut de sa balade à partir d'un sommet de départ, la particule passe plus probalement sur les sommets vers lesquels le sommet de départ entretient une forte confluence. Par exemple dans DicoSyn.Verbe les sommets « dépiauter » et « rêvasser » ont le même nombre de voisins (d(dépiauter)=d(rêvasser)), et donc d'après (1), $\lim_{t\to\infty}[\hat{A}^t]_{\text{déshabiller dépiauter}} = \lim_{t\to\infty}[\hat{A}^t]_{\text{déshabiller rêvasser}} = 6.3_{\times}10^{-5}$. On peut cependant voir dans la Fig 2. que les deux suites $([\hat{A}^t]_{\text{déshabiller dépiauter}})_{0\leq t}$ et $([\hat{A}^t]_{\text{déshabiller rêvasser}})_{0\leq t}$, sont tres différentes pour t petit, ce qui nous montre que la confluence de « déshabiller » vers « dépiauter » est plus forte que la confluence de « déshabiller » vers « rêvasser ».

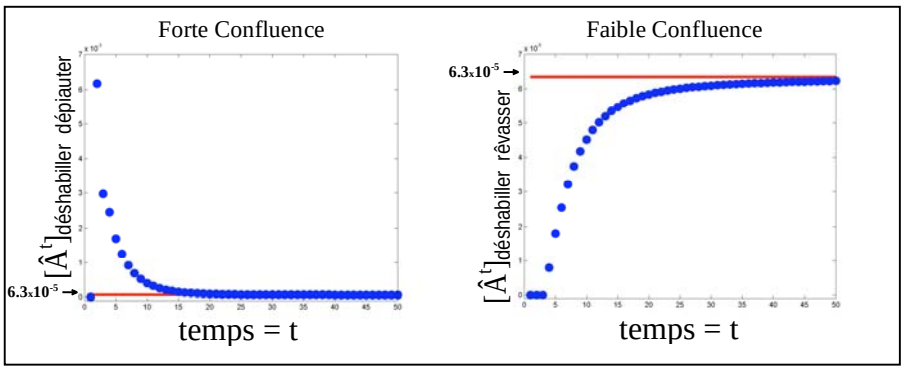



*Fig 2.a : $([\hat{A}^t]_{\text{déshabiller dépiauter}})_{0\leq t}$ et **2.b** : $([\hat{A}^t]_{\text{déshabiller rêvasser}})_{0\leq t}$ dans DicoSyn.Verbe*

Puisque L, la longueur moyenne des plus courts chemins, est petite dans un HSW, nous savons que deux sommets sont en général reliés par au moins un chemin relativement court. Nous choisirons donc t entre L et 2L afin d'atteindre en général presque tous les sommets depuis n'importe quel sommet de départ, sans toutefois atteindre le régime stationnaire de la chaîne de Markov quand t devient trop grand. Dans la suite de cet article nous prendrons pour illustrer notre propos DicoSyn.Verbe dont le L≈ 4, et nous fixerons t=5.

## 4. Formes du sens

Pour un t donné on peut considérer $\hat{A}^t$ comme une matrice où $\forall r,s \in V$, $[\hat{A}^t]_{r\,s}$ nous indique le niveau de confluence du sommets r vers le sommet s. Par exemple la Fig 3. est la liste en ordre décroissant des 100 sommets vers lesquels le sommet « déshabiller » entretient les plus fortes confluences (mesuré à t=5) dans le graphe DicoSyn.Verbe. C'est-à-dire que la Fig 3. est la liste en ordre décroissant de l'ensemble : $\{x \in V, [\hat{A}^t]_{\text{déshabiller}\ x} \geq 0.0026\}$ qui a 100 éléments.

1 →dépouiller, 2 →défaire, 3 →démunir, 4 →déshabiller, 5 révéler, 6 →découvrir, 7 →montrer, 8 →dénuder, 9 →dévêtir, 10 →dégarnir, 11 dégager, 12 dévoiler, 13 →étaler, 14 →ôter, 15 enlever, 16 désaffubler, 17 afficher, 18 →écorcher, 19 →démasquer, 20 arracher, 21 →délacer, 22 voler, 23 dénouer, 24 desserrer, 25 →médire, 26 débarrasser, 27 défubler, 28 exposer, 29 couper, 30 →exhiber, 31 tailler, 32 développer, 33 déchirer, 34 peler, 35 dépourvoir, 36 vider, 37 déchausser, 38 priver, 39 trahir, 40 dire, 41 dépiauter, 42 déposséder, 43 ouvrir, 44 frustrer, 45 déployer, 46 deviner, 47 indiquer, 48 démontrer, 49 tondre, 50 sevrer, 51 détacher, 52 spolier, 53 manifester, 54 effeuiller, 55 présenter, 56 livrer, 57 faire voir, 58 détruire, 59 prouver, 60 marquer, 61 dénantir, 62 signaler, 63 exprimer, 64 retirer, 65 représenter, 66 sortir, 67 porter, 68 lever le masque, 69 abattre, 70 donner, 71 produire, 72 publier, 73 déménager, 74 destituer, 75 élaguer, 76 pénétrer, 77 dessaisir, 78 déclarer, 79 annoncer, 80 déballer, 81 offrir, 82 dénoncer, 83 séparer, 84 expliquer, 85 répandre, 86 déceler, 87 libérer, 88 peindre, 89 tirer, 90 dépecer, 91 égorger, 92 faire montre, 93 témoigner, 94 supprimer, 95 prendre, 96 désassortir, 97 diminuer, 98 voir, 99 faire parade, 100 ruiner

*Fig 3. Les 100 sommets de plus forte confluence depuis «déshabiller» dans DicoSyn.Verbe*

Dans la Fig 3., les mots précédés d'une flèche → sont les 15 voisins du verbe « déshabiller » dans DicoSyn.Verbe. On peut constater, que le verbe « dépouiller » qui est un hyperonyme de « déshabiller » est celui avec lequel le verbe « déshabiller » entretient la plus forte confluence (mesuré à t=5) et que le verbe « effeuiller » (métaphore de « déshabiller ») est lui 54$^{\text{éme}}$ sur 9043. La forte incidence du verbe « enlever » est de 122 et c'est un hyperonyme de haut niveau de déshabiller il est classé 15$^{\text{ème}}$ relativement à « déshabiller ».

Un autre point de vue est de considérer $\hat{A}^t$ comme une matrice $n_x n$ des coordonnées de n vecteurs lignes $([\hat{A}^t]_x .)_{x \in V}$ dans $^n$. Ce point de vue nous permet donc de plonger le graphe G=(V,E) dans $^n$, où un sommet $r \in V$ a pour coordonnées dans $^n$ le vecteur ligne $[\hat{A}^t]_r .$. L'idée est que deux sommets de coordonnées $[\hat{A}^t]_r .$ et $[\hat{A}^t]_s .$ dans $^n$, seront d'autant plus proches dans $^n$ que leurs relations à l'ensemble du graphe seront semblables. Par exemple dans le graphe G1 à 9 sommets de la Fig 4., les sommets 5 et 7 ont exactement les mêmes voisins : {5, 6, 7}, ce qui fait que $\forall t \in {}^*$, $[\hat{A}^t]_5 .=[\hat{A}^t]_7 .$, leurs coordonnées dans $^9$ sont égales, et l'arrête (5,7) a donc une longueur nulle, alors que l'arrête la plus longue est l'arrête (4,6) qui a une longueur de 0.2740 à t=5 (C'est par exemple ce principe décrit dans (Gaume, 2004) qui est directement repris dans les travaux d'application à la classification hiérarchique des sommets d'un HSW dans (Latapy, Pons, 2005)).

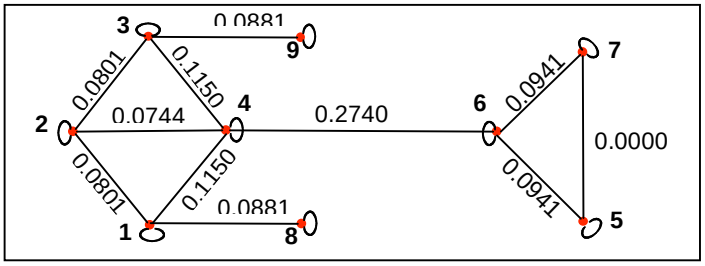


*Fig 4. La longueur géométrique dans $^9$ des arrêtes de G1 à t=5*

C'est l'alliance de ces deux points de vue ($\hat{A}^t$ comme matrice $n_x n$ de coordonnes de n vecteurs lignes dans $^n$ ou bien $\hat{A}^t$ comme matrice $n_x n$ de confluence entre les n sommets du graphe) qui va nous permettre de visualiser globalement ou localement un graphe, en prenant toujours en compte sa structure globale.

La matrice $\hat{A}^t$ en tant que matrice de coordonnées dans $^n$ contient une information calculée sur l'ensemble du graphe qu'il serait possible de représenter dans $^3$ au moyen d'une Analyse en Composante Principale (ACP) de $\hat{A}^t$ en gardant les 3 premiers axes. Par exemple la Fig 5. illustre la forme 3D du graphe G1 de la Fig 4.

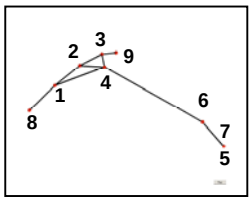



*Fig 5. La visualisation 3D de G1 à t=5 (sur les 3 premiers axes de la PCA de $\hat{A}^5$)*

Cependant les 9043 sommets de notre graphe DicoSyn.Verbe affichés sur un écran donnent une image illisible. Nous n'allons donc en afficher qu'un 'morceau autour' d'un ensemble de sommets F, et ce 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⊆V par une *'carte locale autour de F'*, alors :

**a)** On extrait l'ensemble $T_{t,F,\alpha} \subseteq V$, où $T_{t,F,\alpha}=\{x \in V, [(P^F)\hat{A}^t]_x \geq \alpha\}$. Soit $R_{t,F,\alpha}=|T_{t,F,\alpha}|$.

Pour $\alpha \in [0,1]$, $T_{t,F,\alpha}$ est en quelque sorte une *'extraction topologique autour de F'*, si F={r} est un singleton alors ce sont les $R_{t,\{r\},\alpha}$ sommets $x \in V$ dont les confluences de r vers x sont les plus fortes qui sont sélectionnés :

(la Fig 3. illustre $T_{t,\{déshabiller\},0.0026}$ avec $R_{t,\{déshabiller\},0.0026}=|T_{t,\{déshabiller\},0.0026}|=100$)

**b)** On extrait de $\hat{A}^t$ la matrice $M_{t,F,\alpha}$ qui est la matrice carrée $(R_{t,F,\alpha})_x(R_{t,F,\alpha})$ formée de l'intersection des $R_{t,F,\alpha}$ lignes $[\hat{A}^t]_x$. avec les $R_{t,F,\alpha}$ colonnes $[\hat{A}^t]._x$ tel que $x \in T_{t,F,\alpha}$. $M_{t,F,\alpha}$ est en quelque sorte un *'zoom topologique autour de F'*.

**c)** On normalise ensuite les $R_{t,F,\alpha}$ lignes de $M_{t,F,\alpha}$ (pour chaque $x \in T_{t,F,\alpha}$, on remplace la ligne $[M_{t,F,\alpha}]_x$. par $[M_{t,F,\alpha}]_x ./\|[M_{t,F,\alpha}]_x .\|$. On pratique ensuite une ACP sur $M_{t,F,\alpha}$ où l'on ne garde alors que les 3 premières dimensions (qui conservent alors le plus d'informations pertinentes dans la réduction à $^3$) pour obtenir ainsi $D_{t,F,\alpha}$ qui est la visualisation 3D de la région[3] dans G comprise dans un 'rayon topologique $R_{t,F,\alpha}$ autour de F' pour un instant t donné.

**Exemple 1 :** La Fig. 8.a nous montre $D_{5,\{jouer\},0.0015}$ la carte 3D autour du singleton {jouer}, c'est-à-dire les 3 premières coordonnées sur l'ACP de $M_{t,F,\alpha}$ pour t=5, F={jouer}, α=0.0015 où

[3] Pour tout ensemble non vide F⊆V, si α=0 alors $T_{tF,\alpha}=V$ et $R_{tF,0}=n$, ce seront donc les n sommets du graphe tout entier qui s'afficheront dans $D_{tF,0}$

$R_{t,F,\alpha=100}$. Nous pouvons voir que la structure géométrique de $D_{5,\{jouer\},0.0015}$ (la forme ainsi obtenue) reflète bien la structure polysémique du verbe « Jouer » avec ses 4 sens principaux en français.

**Exemple 2 :** En posant F={monter, descendre} et α=0.001, nous pouvons visualiser $D_{5,\{monter,\ descendre\},0.001}$ (Fig 9.a), où nous pouvons voir qu'à la charnière de l'articulation sémantique de {« descendre », « monter »} il y a « sauter », bien que ce mot ne soit pas directement connecté à « monter ».

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 », Fig 8.a, c'est α qui nous donne l'étendue de la *'région de confluence topologique 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 sémantique*s' qui sont au cœur des grandes confluences il nous suffit de poser F=V.

**Exemple 3 :** La Fig 10. nous montre $D_{5,V,0.0005}$, la visualisation globale 3D autour de V (l'ensemble de tous les sommets) pour α=0.0005 et t=5, où $R_{t,F,\alpha}$=200. À partir de Dicosyn.Verbe, la forme ainsi obtenue est à peu près un tétraèdre qui organise les verbes du français dans un continuum sémantique en faisant émerger 4 axes (les sommets du tétraèdre) : 1) LOCOMOTIF, 2) POSITIF, 3) FIXATIF, 4) NEGATIF.

## 5. Des formes locales aux confluences globales par les couleurs

Dans la figure 8.b (la visualisation globale $D_{5,V,0.0005}$) plus un sommet x est foncé, plus $[\hat{A}^t]_{jouer\ x}$ est grand (la confluence du sommet « jouer » vers le sommet x est forte). Nous voyons bien là que le sommet « jouer » est un verbe très polysémique, et qu'il couvre un grand nombre de sens répartis sur tout l'espace sémantique. Pour construire $D_{5,\{jouer\},0.0015}$, on commence par extraire $R_{5,\{jouer\},0.0015}$=100 sommets dans $T_{t,\{jouer\},0.0015}$, pour être ensuite affichés dans $D_{5,\{jouer\},0.0015}$ (la visiualisation locale autour de « jouer »), et nous ne pouvons plus alors savoir de quelle région sémantique de $D_{5,V,0.0005}$ un sommet affiché a été extrait (ce n'est pas le cas dans nos cartes géographiques locale grâce à la convention Nord/Haut, Sud/Bas qui nous permet de resituer une carte locale dans la géométrie de la globalité). Afin de remédier à ce manque d'information représentée, nous allons voir comment visualiser le rapport qu'entretient une forme locale avec la globalité du graphe relativement à un ensemble de sommets donnés. Choisissons par exemple dans DicoSyn.Verbe les quatre sommets du tétraèdre {fuir, exciter, fixer, briser} (qui remplacerons en quelques sortes les 4 points cardinaux Nord, Sud, Est, Ouest de nos cartes géographiques) choisissons quatre vecteurs couleurs[4] $\{B_{fuir}, B_{exciter}, B_{fixer}, B_{briser}\}$. Pour chaque sommet s de notre graphe attribuons lui une couleur $C_s$ de la manière suivante :

$$C_s = \frac{([\hat{A}^t]_{s\ fuir})B_{fuir} + ([\hat{A}^t]_{s\ exciter})B_{exciter} + ([\hat{A}^t]_{s\ fixer})B_{fixer} + ([\hat{A}^t]_{s\ briser})B_{briser})}{[\hat{A}^t]_{s\ fuir} + [\hat{A}^t]_{s\ exciter} + [\hat{A}^t]_{s\ révéler} + [\hat{A}^t]_{s\ briser}}$$

[4] Les couleurs codées en RVB sont traitées comme des vecteurs de dimension 3.

$C_s$ est le barycentre des quatres vecteurs couleurs, pondérés par les confluences de s vers chacun des quatres sommets respectifs. La couleur $C_s$ d'un sommet s, reflette donc la force des confluences que le sommet s entretient avec les quatres sommets[5] fuir, exciter, fixer, briser. On peut voir dans la Fig 8. que se sont les sommets {voler, coulisser marcher, passer} (verbes de déplacements) qui tirent le plus vers la couleur $B_{fuir}$ (bleu=[0,01]), qui a été associée au sommet « fuir » se trouvant dans l'angle LOCOMOTIF du tétraèdre. Cela nous indique donc que ces 4 sommets entretiennent leurs plus fortes confluences avec des sommets de cette région sémantique, les verbes LOCOMOTIFS.

## 6. Complexité

Soit G=(V,E) un graphe symétrique et réflexif de n=|V| sommets et m=|E| arcs. La Fig 6. résume la procédure de calcul de la carte de $D_{t,F,\alpha}$ où $R_{t,F,\alpha}$ sommets sont affichés.

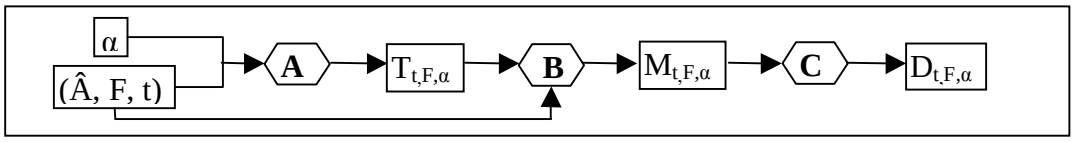



***Fig 6.*** *Organigramme : (Â, t, F, α)* → $D_{t,F,\alpha}$

**– Procédure A :** on calcule d'abord le vecteur $U \leftarrow \text{Ligne}(\hat{A},t,P^F)$ (voir Fig 7.) puis on extrait de U l'ensemble $T_{t,F,\alpha} \subseteq V$, tel que $T_{t,F,\alpha}=\{x \in V, U_x \geq \alpha\}$.

**– Procédure B :** pour chacun des $R_{t,F,\alpha}$ sommets $x \in T_{t,F,\alpha}$, on calcule $H_{x\bullet} \leftarrow \text{Ligne}(\hat{A},t,P^{\{x\}})$ (voir Fig 7.), puis on construit $M_{t,F,\alpha}$, la matrice carrée $(R_{t,F,\alpha})_x(R_{t,F,\alpha})$ formée des $R_{t,F,\alpha}$ colonnes $H_{\bullet x}$ tel que $x \in T_{t,F,\alpha}$.

**– Procédure C :** on normalise chacune des $R_{t,F,\alpha}$ lignes de la matrice $M_{t,F,\alpha}$ puis on pratique une PCA dont on ne conserve que les 3 premiers axes dans $D_{t,F,\alpha}$.

En exploitant efficacement la structure très clairsemée de Â la matrice $n_x n$ qui a m valeurs non nulles, le temps calcul et l'espace mémoire nécessaire pour calculer $U_{out} \leftarrow \text{Ligne}(\hat{A},t,U_{in})$ sont respectivement O(n+tm) et O(2n+m). Le temps calcul cumulé des deux procédures A et B est donc $O((R_{t,F,\alpha})(n+tm))$, et l'espace mémoire est O(2n+m).

Le temps calcul et l'espace mémoire nécessaires pour la procédure C ne font tout deux intervenir que l'ordre de la matrice $M_{t,F,\alpha}$, c'est-à-dire $R_{t,F,\alpha}$ qui n'a jamais besoin d'être grand pour une cartographie pertinente, tant pour les visualisations globales que locales dans les HSW, grâce à leurs structures hiérarchiques et leur C fort. En majorant $R_{t,F,\alpha}$ par K le nombre maximal de sommets affichables dans une carte (K est O(100) et dépend de la charge cognitive maximale qui est acceptée), on peut donc considérer que le temps calcul et l'espace mémoire nécessaire pour la procédure C sont des constants. Le temps et l'espace nécessaire pour calculer $D_{t,F,\alpha}$ sont donc respectivement $O((R_{t,F,\alpha})(n+tm)) \leq O(K(n+tm))$ et O(2n+m). Mais en général dans un HSW, $m<n\log(n)$ , et on a donc que le temps calcul est O(Ktnlog(n)) et l'espace mémoire O(nlog(n)).

[5] En choisissant d'autre sommets on observerait d'autres confluences.

| | |
|---|---|
| $U_{out}$←**function** Ligne(M, t, $U_{in}$)<br>  $U_{out}$←$U_{in}$ ;<br>  **for** i **from** 1 **to** t<br>  $U_{out}$←($U_{out}$)M ;<br>  **end**<br>**end** | **in :**<br>  M : sparse matrice$_{n\times n}$ of real<br>  t : integer<br>  $U_{in}$ : full vector$_n$ of real<br>**out :**<br>  $U_{out}$ : full vector$_n$ of real |

*Fig 7. Algorithme : $U_{out}$←function Ligne(M, t, $U_{in}$)*

Les complexités temps et espace de la fonction Ligne (Fig 7.), qui est au cœur de Prox, sont dans le cas général respectivement : O(tm) et O(m), et pour un HSW : O(tnlog(n)) et O(nlog(n)) (avec t tres petit). Mais quand n devient tres grand ($n>10^9$ pour le web) deux méthodes peuvent être alors envisagées **a)** la parallelisation des $R_{t,F,\alpha}$ calculs de $H_x$.←Ligne($\hat{A}$,t,$P^{\{x\}}$) qui sont indépendants dans la procédure B, ou/et **b)** les méthodes de Monté Carlo qui convergent tres vite sur les HSW et ne nécessitent que tres peu de mémoire vive. Nous avons donc commencé à appliquer Prox aux graphes construit à partir du web (hyperliens, et/ou affinités sémantiques entre pages) afin d'offrir une méthode ergonomique d'accès à l'information où l'utilisateur est guidé dans sa recherche par la forme structurelle du dictionnaire ou du web : http://Prox.irit.fr/ (courrant 2006 avec des graphes de langues différentes --métrologie lexicale comparative)

## 7. Conclusion

Dans cette approche, $[\hat{A}^t]_M$. le sens potentiel d'un mot hors cotexte est calculé à partir de :

$[A]_M$. : l'ensemble des relations que le mot M entretient dans le réseau lexical (local)

$\hat{A}$ : la matrice de transition associée au réseau lexical (global)

Il faut maintenant définir le sens effectif d'un mot en cotexte en adoptant une approche de la construction du sens dans ce réseau par déformations dynamiques locales en cotexte au fil du discours.

Pour cela, afin de prendre en compte l'orientation des relations syntagmatiques il nous faut travailler sur des graphes non nécessairement réflexifs et symétriques. Dans ce sens nous avons développé un algorithme toujours fondé sur les marches aléatoires mais qui reste pertinent sur tous graphes orientés ou/et pondérés sans nécessité d'une quelconque coercition sur sa structure (réflexivisation ou symétrisation comme dans le présent article, ce qui est pertinent pour les relations de type paradigmatiques, mais inacceptable pour les relations de type syntagmatiques où l'orientation de la relation est fondamentale) (Gaume 2006, Gaume et Mathieu, 2006).

## Références

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

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

Duvignau K. (2002). *La métaphore berceau et enfant de la langue*. Phd These, Toulouse II.

Duvignau K. Métaphore verbale et approximation. In *Regards croisés sur l'analogie*. Revue d'Intelligence Artificielle, n° spécial, Vol 5/6. HermèsLavoisier.

Fellbaum C. (1998). *WordNet, an Electronic Lexical Database*, MIT Press, 1998.

Gaume B. (2004). Balades aléatoires dans les petits mondes lexicaux, in *I3* Vol 4, n°2, http://www.revue-i3.org/volume04/numero02/revue_i3_04_02_02.pdf.

Gaume B. *dynamic deformations of words' meaning through cotext*. (en soumission).

Gaume B., Mathieu F. *PageRank Induced Topology for Real-World Networks* (en soumission).

Latapy M., Pons P. (2005). Computing communities in large networks using random walks *proceedings of ISCIS'05*, 2005.

Manning C.D., Schütze H. (2002). *Foundations of Statistical Natural Language Processing*, MIT Press.

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

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

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

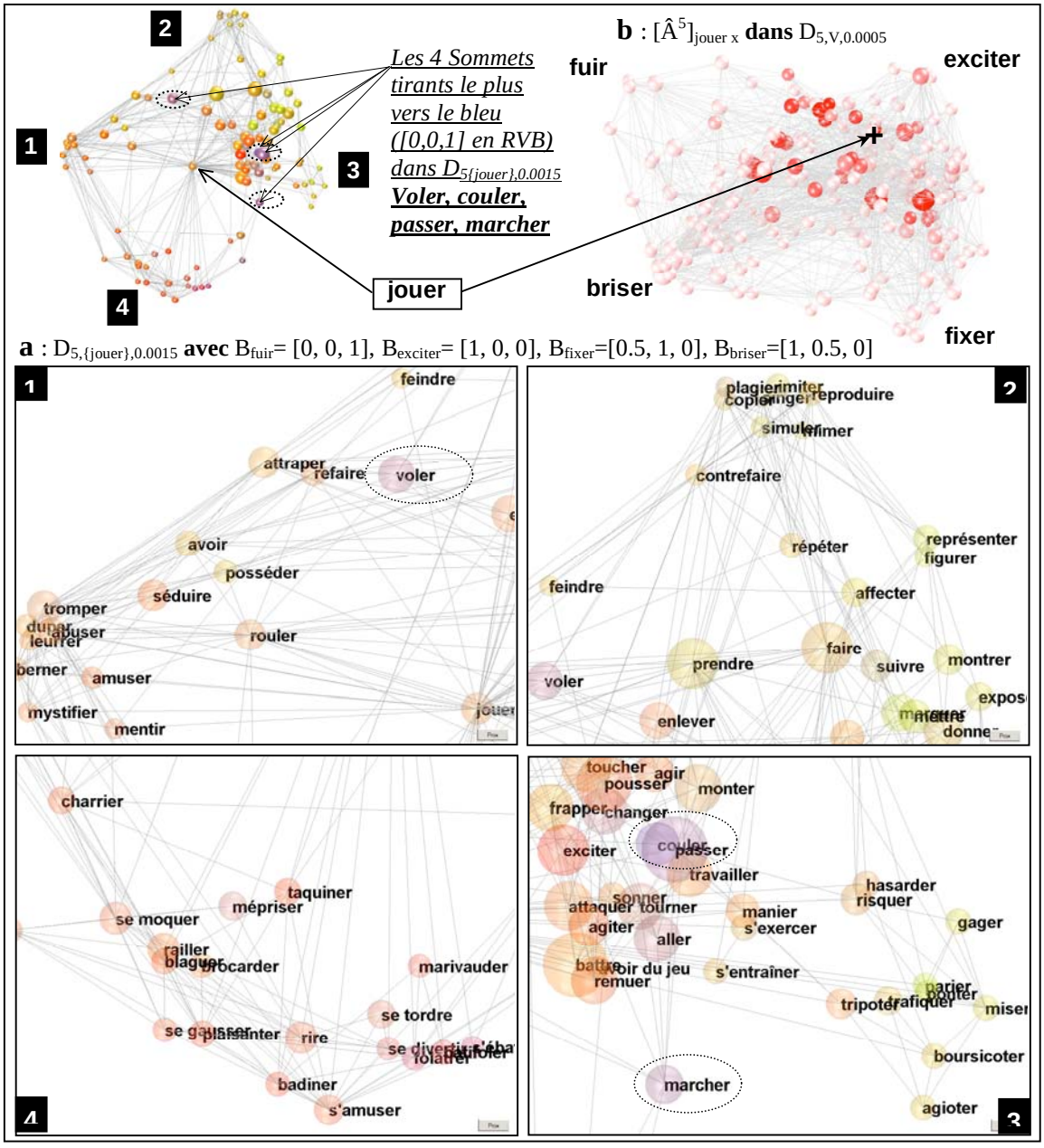



***Fig 8.*** *$D_{5,\{jouer\},0.0015}$ dans DicoSyn.Verbe, la forme conceptuelle de « jouer », $R_{5,\{jouer\},0.0015}$=100*

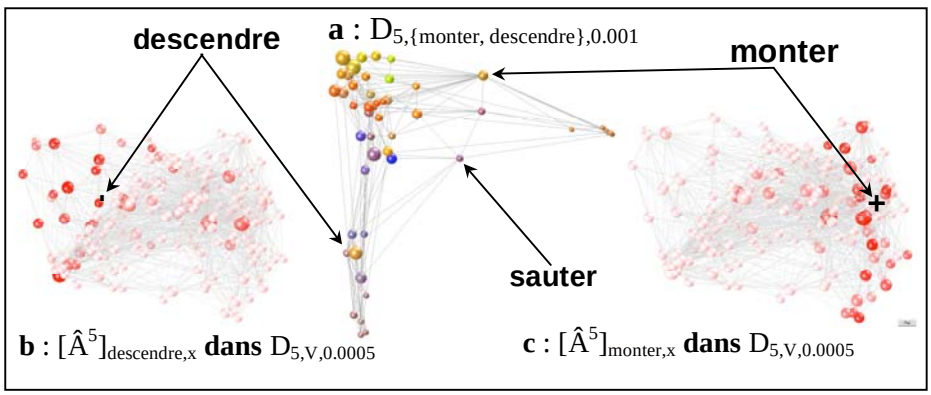
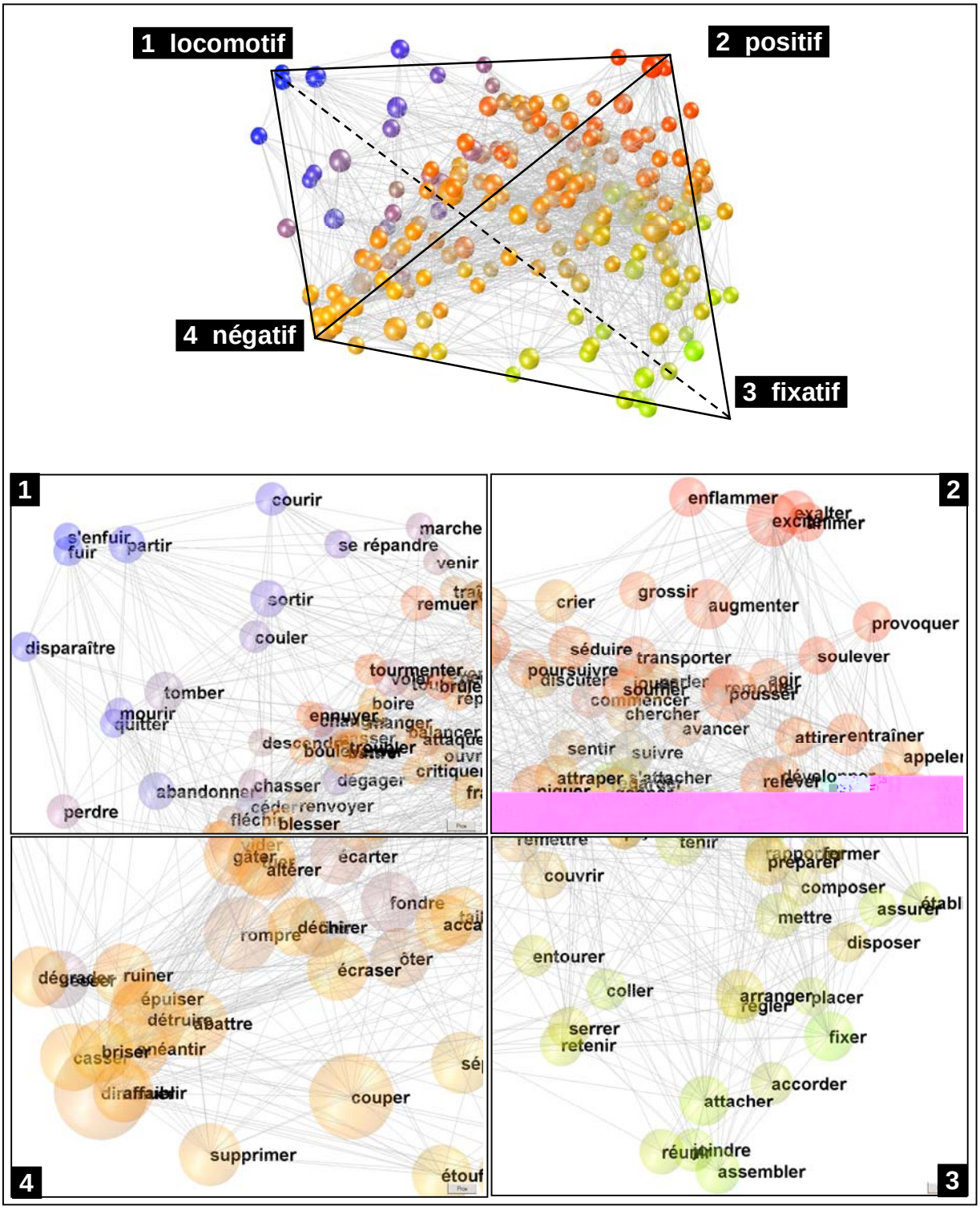



***Fig 9.*** *Zoom topologique sur {« monter », « descendre »}, $R_{5,\{monter, descendre\},0.001}$=50*




*tétraèdre des verbes du français : $D_{5,V,0.0005}$ dans DicoSyn.Verbe, $R_{5,V,0.0005}$=200*