Mesurer la similarité structurelle entre réseaux lexicaux

Bruno Gaume[1] Emmanuel Navarro[2] Yann Desalle[3] Benoît Gaillard[1]
(1) CLLE-ERSS, CNRS, Universié de Toulouse
(2) IRIT, CNRS, Université de Toulouse
(3) ATILF, CNRS, Université de Lorraine
gaume@univ-tlse2.fr, navarro@irit.fr, yann.desalle@gmail.com, benoit.gd@gmail.com,

Résumé. Dans cet article, nous comparons la structure topologique des réseaux lexicaux avec une méthode fondée sur des marches aléatoires. Au lieu de caractériser les paires de sommets selon un critère binaire de connectivité, nous mesurons leur proximité structurelle par la probabilité relative d'atteindre un sommet depuis l'autre par une courte marche aléatoire. Parce que cette proximité rapproche les sommets d'une même zone dense en arêtes, elle permet de comparer la structure topologique des réseaux lexicaux.

Abstract. In this paper, we compare the topological structure of lexical networks with a method based on random walks. Instead of characterising pairs of vertices according only to whether they are connected or not, we measure their structural proximity by evaluating the relative probability of reaching one vertex from the other via a short random walk. This proximity between vertices is the basis on which we can compare the topological structure of lexical networks because it outlines the similar dense zones of the graphs.

Mots-clés : Réseaux lexicaux, réseaux petits mondes, comparaison de graphes, marches aléatoires.

Keywords: Lexical networks, small worlds, comparison graphs, random walks.

1 Contexte

Une ressource lexicale peut être modélisée sous la forme d'un graphe $G = (V, E)$ dans lequel un ensemble de n sommets V représente des entités lexicales (lemmes, contextes syntaxiques ...) et un ensemble de m arêtes $E \subseteq \mathbf{P}_2^V$ représente une relation lexicale entre ces entités. Un des problèmes majeurs concernant ces réseaux lexicaux porte sur leurs désaccords apparents : par exemple, si $G_1 = (V, E_1)$ et $G_2 = (V, E_2)$ sont deux graphes de synonymie standards d'une langue donnée, alors une grande proportion de paires $\{x, y\} \in \mathbf{P}_2^V$ sont liées dans G_1 ($\{x, y\} \in E_1$) mais ne le sont pas dans G_2 ($\{x, y\} \notin E_2$) ; autrement dit, x et y sont synonymes pour G_1 mais ne le sont pas pour $G2$. Un tel désaccord n'est pas compatible avec l'hypothèse d'une synonymie qui reflèterait la structure sémantique du lexique commune aux membres d'une même communauté linguistique.

Pour résoudre cette contradiction apparente, il faut regarder les réseaux lexicaux dans une perspective plus large. La figure 1 est un exemple artificiel de désaccord généralisé entre les arêtes de deux graphes malgré une similarité structurelle. Bien qu'ils n'aient aucune arête en commun ($E_1 \cap E_2 = \varnothing$), ces deux graphes se ressemblent parce que les deux zones denses dessinées par chacun des graphes contiennent les mêmes sommets : $\{1, 2, 3, 4, 11, 12, 13\}$ et $\{4, 5, 6, 7, 8, 9, 10\}$. Cette similarité structurelle est observable en considérant chacun des graphes comme un tout, et non en les comparant arête par arête. Les zones denses de cet exemple artificiel (fig. 1) sont caractéristiques des *graphes de terrain* [1] qui, pour la plupart, sont des réseaux petits mondes hiérarchiques (RPMH) partageant les mêmes propriétés (Newman, 2003; Gaume *et al.*, 2010; Steyvers & Tenenbaum, 2005). Ils présentent une **faible densité en arêtes** (peu d'arêtes par rapport au nombre maximal d'arêtes potentielles), des **chemins courts** (le nombre moyen d'arêtes L sur les plus courts chemins entre deux sommets est faible), un fort **taux d'agrégation** C (des sous-graphes localement denses en arêtes, ou agrégats, peuvent être identifiés alors que le graphe est globalement peu dense en arêtes (Watts & Strogatz, 1998)), et la distribution des degrés d'incidence de leurs sommets approche une **loi de puissance** (Albert & Barabasi, 2002). Nous montrons dans la section 2 que les réseaux lexicaux étudiés dans cet article possèdent les caractéristiques des RPMH. Ainsi, comme le suggère la figure 1, un désaccord apparent entre les réseaux lexicaux n'implique pas nécessairement une incompatibilité structurelle des données modélisées.

Dans cet article, nous étudions les réseaux lexicaux avec une méthode fondée sur des marches aléatoires. Au lieu de caractériser les paires de sommets selon leur seule connectivité binaire (existence ou absence d'une arête entre les som-

1. Les graphes de terrain sont des graphes qui modélisent les données réelles récoltées sur le terrain, par exemple en sociologie, linguistique ou biologie. Ils s'opposent en cela aux graphes artificiels (déterministes ou aléatoires).

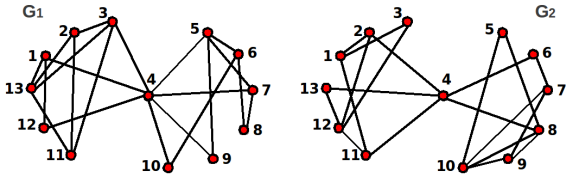


FIGURE 1 – Contradiction entre la variaribilité locale et la similarité globale.

mets), nous mesurons leur proximité structurelle par la probabilité relative d'atteindre un sommet depuis l'autre par une courte marche aléatoire [2]. Parce que cette proximité rapproche les sommets d'une même zone dense en arêtes, elle permet de mesurer la qualité de la divergence de surface entre deux réseaux lexicaux. Notons que ce travail vient à la suite de (Gaillard *et al.*, 2011; Navarro *et al.*, 2012) et de (Navarro, 2013, chap. 3).

Nous montrons dans la section 2 les limites des approches arête-par-arête pour l'analyse et la comparaison des réseaux lexicaux selon lesquelles les réseaux de synonymie d'une même langue seraient significativement différents. Dans la section 3, nous présentons une méthode de comparaison structurelle de graphes basée sur la *confluence*, mesure de proximité entre sommets qui repose sur les marches aléatoires et permet d'analyser structurellement les réseaux lexicaux. En section 4, nous appliquons cette méthode de comparaison de graphes à des ressources construites par des lexicographes et par les foules (crowdsourcing), ressources dont les méthodes d'élaboration diffèrent mais qui tentent de décrire la même relation lexicale : la synonymie. Nous concluons en section 5.

2 Compararer $\mathbf{G_1 = (V, E_1)}$ et $\mathbf{G_2 = (V, E_2)}$ en comparant les ensembles $\mathbf{E_1}$ et $\mathbf{E_2}$ comme des «sacs de liens» sans structures

Nous illustrons notre propos dans cette section sur la comparaison de deux ressources lexicales, toutes deux construites par des lexicographes approximativement à la même époque pour représenter la même relation de synonymie :

– **Rob** $= (\mathbf{V_{Rob}}, \mathbf{E_{Rob}})$: Le dictionnaire Le Robert (Robert & Rey, 1985) a été informatisé au cours d'un parteneriat IBM / ATILF [3]. Cette ressource électronique liste les synonymes des différentes acceptions des vocables du français. Les sommets du graphe lexical *Rob* qui a été construit à partir de cette ressource sont les vocables (les vocables homonymes ne sont pas distingués et sont représentés par un même sommet). La paire $\{x, y\}$ appartient à E_{Rob} si et seulement si une des acceptions de x a été considérée comme synonyme d'une des acceptions de y par l'équipe lexicographique du Robert. Par exemple, le verbe *causer* est à la fois synonyme de *parler* et de *engendrer*.
– **Lar** $= (\mathbf{V_{Lar}}, \mathbf{E_{Lar}})$: Le graphe lexical *Lar* a été construit à partir du dictionnaire Larousse (Guilbert *et al.*, 1971 1978) de la même manière que le graphe *Rob*.

Les caractéristiques (que nous appelons « pédigrés ») des graphes *Rob* et *Lar* sont fournis dans le tableau 1. Ces mesures sont en accord avec la plupart des études sur les réseaux lexicaux (Motter *et al.*, 2002; de Jesus Holanda *et al.*, 2004; Gaume, 2004) qui montrent que les réseaux lexicaux comme la majorité des réseaux de terrains sont des RPMH typiques.

TABLE 1 – Pédigrés des graphes lexicaux *Lar* et *Rob* : n et m sont les nombres de sommets et d'arêtes, $\langle k \rangle$ est la moyenne des degrés d'incidence des sommets, C est le coefficience d'agrégation du graphe, L_{lcc} est la moyenne des plus courts chemins entre tous les nœuds de la plus grande partie connexe (sous-graphe dans lequel il existe au moins un chemin entre deux noeuds quelconques de ce sous-graphe), r^2 est le coefficient de corrélation entre la distribution des degrés d'incidence et la loi de puissance la plus fortement corrélée à cette distribution, λ est la puissance de cette loi.

Réseaux Lexicaux	n	m	⟨k⟩	C	L_{lcc}	λ (r^2)
Lar	22066	73091	6,62	0,19	6,36	-2,43 (0,90)
Rob	38147	99998	5,24	0,12	6,37	-2,43 (0,94)

2. Notons que cette méthode de mesure de proximité entre sommets dans un graphe peut-être utilisée avantageusement pour la modélisation des graphes de terrain (Gaume *et al.*, 2010), sur des tâches de substitution lexicale ou de résolution de métaphore (Desalle *et al.*, 2014b, 2009; Desalle, 2012), pour l'enrichissement de ressources lexicales (Sajous *et al.*, 2011), la navigation dans les réseaux de terrain (Gaume, 2008), la recherche d'informations (Navarro *et al.*, 2011) ou encore pour la détection de pathologies (Desalle *et al.*, 2014a).

3. http://www.atilf.fr

2.1 Distance d'édition

Soit deux graphes $G_1 = (V_1, E_1)$ et $G_2 = (V_2, E_2)$, nous mesurons la similarité des couvertures lexicales de G_1 et G_2 par l'indice de *Jaccard* : $J(G_1, G_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$. Nous avons alors $J(Rob, Lar) = 0,49$. Ces deux graphes ont donc une couverture lexicale commune suffisament large pour que la comparaison entre les jugements de synonymie qu'ils modélisent soit réalisée sur cette couverture lexicale commune : $V_1 \cap V_2$.

Pour mesurer l'accord entre les arêtes de G_1 et de G_2, nous commençons donc par réduire les deux graphes à leurs sommets communs : $G'_1 = \big(V' = (V_1 \cap V_2), E'_1 = E_1 \cap (V' \times V')\big)$ et $G'_2 = \big(V' = (V_1 \cap V_2), E'_2 = E_2 \cap (V' \times V')\big)$. Pour chaque paire de sommets $\{a,b\} \in (V' \times V')$, quatre configurations sont possibles :

- $\{a,b\} \in E'_1 \bigcap E'_2$: accord sur la paire $\{a,b\}$, a et b sont synonymes dans G'_1 et dans G'_2 ;
- $\{a,b\} \in \overline{E'_1} \bigcap \overline{E'_2}$: accord sur la paire $\{a,b\}$, a et b ne sont synonymes ni dans G'_1 ni dans G'_2 ;
- $\{a,b\} \in E'_1 \bigcap \overline{E'_2}$: désaccord sur la paire $\{a,b\}$, a et b sont synonymes dans G'_1 mais pas dans G'_2 ;
- $\{a,b\} \in \overline{E'_1} \bigcap E'_2$: désaccord sur la paire $\{a,b\}$, a et b sont synonymes dans G'_2 mais pas dans G'_1 ;

Une longue tradition dans la recherche sur la comparaison de graphes consiste à déterminer si deux graphes sont isomorphes. Deux graphes $G_1 = (V_1, E_1)$ et $G_2 = (V_2, E_2)$ sont isomorphes s'il existe une fonction bijective $f : V_1 \mapsto V_2$ telle que, pour toute paire de sommets $\{u,v\} \in \mathbf{P}_2^V$, $\{u,v\} \in E_1 \Leftrightarrow \big\{f(u), f(v)\big\} \in E_2$. La comparaison entre graphes consiste alors à rechercher de tels isomorphismes. Dans les graphes étudiés dans cet article, les nœuds sont étiquetés et ne peuvent correspondre que s'ils ont les mêmes étiquettes : la seule bijection possible est donc la fonction identité. Ainsi, pour savoir si deux graphes étiquetés $G_1 = (V, E_1)$ et $G_2 = (V, E_2)$ sont isomorphes, il suffit de vérifier que $E_1 = E_2$.

Une telle similarité est très basique : si aucune arête ne diffère alors les deux graphes sont similaires, sinon ils sont différents (ils ne sont pas isomorphes). Afin d'assouplir cette approche de l'isomorphisme pour fournir une mesure quantitative continue de la différence entre deux graphes, plusieurs alternatives ont été proposées (pour une revue de ces méthodes, voir par exemple (Gao *et al.*, 2010)). Ces méthodes s'inspirent de la distance d'édition entre deux chaînes de caractères (Levenshtein, 1966). La distance d'édition entre entre deux graphes $G_1 = (V_1, E_1)$ et $G_2 = (V_2, E_2)$ est définie par la série d'opérations la plus économique pour transformer G_1 en un isomorphisme de G_2. Habituellement, l'ensemble des opérations possibles ne contient que l'insertion, la suppression et la substitution de sommets et d'arêtes. Cet ensemble peut éventuellement être étendu selon les données que les graphes modélisent. Par exemple, dans le cas d'une segmentation d'image, (Ambauen *et al.*, 2003) introduisent les opérations de cission et de fusion de noeuds.

Dans le cadre de cet article, puisque après la réduction des deux graphes à leurs sommets communs, nous aurons $V_1 = V_2 = V$, les seules opérations possibles vont être la suppression et l'insertion d'arêtes. Si le coût d'édition d'une arête est 1, alors la distance d'édition entre G_1 et G_2 est :

$$ED = |E_1 \cap \overline{E_2}| + |E_2 \cap \overline{E_1}| \tag{1}$$

Remarquons que $ED \in [0, |E_1| + |E_2|]$. Cette mesure de dissimilarité ne prend pas en compte le nombre d'arêtes de G_1 et de G_2. Editer dix arêtes pour rendre deux graphes de quinze arêtes isomorphes n'est pas la même chose qu'éditer dix arêtes pour rendre deux graphes de quinze mille arêtes isomorphes. Cette distance d'édition doit donc être normalisée :

$$GED(G_1, G_2) = \frac{|E_1 \cap \overline{E_2}| + |E_2 \cap \overline{E_1}|}{|E_1| + |E_2|} \tag{2}$$

Maintenant, $GED(G_1, G_2) \in [0, 1]$. Appliquée à Lar'/Rob', $GED(Lar', Rob') = 0,47$. Ce résultat montre que Lar' et Rob' sont dissemblables : les dictionnaires Larousse et Le Robert n'ont qu'un faible accord sur les paires de lexèmes qu'ils jugent synonymes. Ceci peut s'expliquer par le fait que la projection de la notion graduelle de quasi-synonymie sur des jugements binaires de synonymie offre une large place à l'interprétation, même si les juges sont des lexicographes experts comme pour les dictionnaires Larousse et Robert. En fait, on observe souvent un faible accord entre des ressources qui décrivent la même réalité linguistique mais qui sont construites indépendamment, même lorsqu'elles reposent sur des jugements humains qui suivent un même protocole (Murray & Green, 2004).

3 Comparer $\mathbf{G_1 = (V, E_1)}$ et $\mathbf{G_2 = (V, E_2)}$ en comparant la structure engendrée par $\mathbf{E_1}$ sur V à la structure engendrée par $\mathbf{E_2}$ sur V

GED est une mesure quantitative de surface qui analyse les graphes comme des «sacs de liens» sans structure. En comparant les graphes arête par arête, elle ne tient pas compte de la structure globale profonde des graphes bien que celle-ci

soit très spécifique puisqu'il s'agit de RPMH. La présence ou l'absence d'une arête entre deux sommets est un jugement de synonymie qui peut être confirmé ou infirmé par la structure topologique du graphe autour de ces sommets. Dans cette section, nous décrivons une mesure quantitative de la similarité structurelle entre graphes. Cette mesure est basée sur les marches aléatoires, ce qui nous permet d'enrichir l'information sur les paires de sommets par une mesure de proximité structurelle entre sommets : *la confluence*.

3.1 Confluence

Soit $G = (V,E)$ un graphe réflexif[4] et non dirigé, définissons $d_G(u) = |\{v \in V / \{u,v\} \in E\}|$ le degré d'incidence d'un sommet u dans le graphe G et imaginons un marcheur se déplaçant sur le graphe G : au temps $t \in \mathbb{N}$, le marcheur est sur un sommet $u \in V$; au temps $t+1$, le marcheur peut atteindre n'importe quel voisin de u avec un probabilité uniforme.

Ce processus est une simple marche aléatoire (Bollobas, 2002; Kinouchi *et al.*, 2002; Baronchelli *et al.*, 2013). Il peut être défini par une chaîne de Markov sur V à l'aide d'une matrice de transition $[G]$:

$$[G] = (g_{u,v})_{u,v \in V} \text{ avec } g_{u,v} = \begin{cases} \dfrac{1}{d_G(u)} \text{ si } \{u,v\} \in E, \\ 0 \text{ sinon.} \end{cases}$$

Puisque G est réflexif, chaque sommet a au moins un voisin (lui-même) ; G est donc bien définie. De plus, par construction, $[G]$ est une matrice stochastique : $\forall u \in V, \sum_{v \in V} g_{u,v} = 1$. La probabilité $P^t_G(u \rightsquigarrow v)$ qu'un marcheur démarrant sur le sommet u atteigne le sommet v après t pas est :

$$P^t_G(u \rightsquigarrow v) = ([G]^t)_{u,v} \tag{3}$$

On peut alors prouver (Gaume, 2004) à l'aide du théorème de Perron-Frobenius (Stewart, 1994) que si G est connexe, réflexif et non-dirigé, alors $\forall\, u,v \in V$:

$$\lim_{t \to \infty} P^t_G(u \rightsquigarrow v) = \lim_{t \to \infty} ([G]^t)_{u,v} = \frac{d_G(v)}{\sum_{x \in V} d_G(x)} = \pi_G(v) \tag{4}$$

Cela signifie que quand t tend vers l'infini, la probabilité d'être sur un sommet v au temps t ne dépend pas du sommet de départ mais seulement du degré d'incidence de v. Nous noterons cette limite $\pi_G(v)$ dans la suite.

Par contre, la dynamique de convergence vers cette limite (équation (4)) dépend fortement du sommet de départ. En effet, la trajectoire du marcheur est totalement régie par la topologie du graphe autour de ce sommet de départ : après t pas, tout sommet v situé à une distance de t arêtes (ou moins) peut être atteint. La probablité de cet événement dépend du nombre de chemins entre u et v et de la structure du graphe autour des sommets intermédiaires le long de ces chemins. Plus il y a de chemins courts entre les sommets u et v, plus la probibilité d'atteindre v à partir de u est grande. Par exemple, si l'on prend $G_1 = Rob$ et $G_2 = Lar$ et que l'on choisit les trois sommets $u = éplucher$, $r = dépecer$ et $s = sonner$ tels que :

- u et r sont jugés synonymes dans Rob : $\{u,r\} \in E_1$;
- u et r ne sont pas jugés synonymes dans Lar : $\{u,r\} \notin E_2$;
- r et s ont le même nombre de synonymes dans G_1 : $d_{G_1}(r) = d_{G_1}(s) = d_1$;
- r et s ont le même nombre de synonymes dans G_2 : $d_{G_2}(r) = d_{G_2}(s) = d_2$.

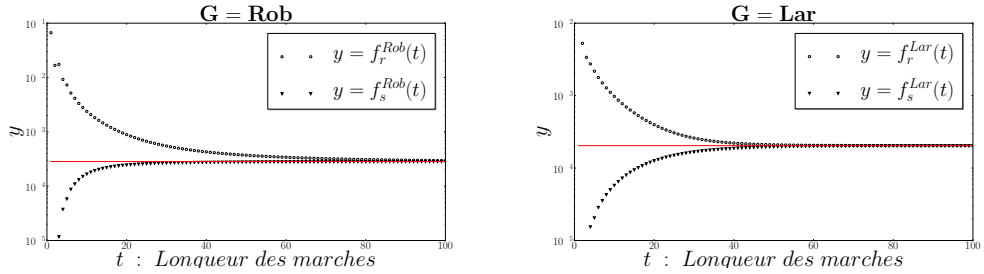
Alors, d'après l'équation (4), les deux séries $(P^t_{G_1}(u \rightsquigarrow r))_{1 \leq t}$ et $(P^t_{G_1}(u \rightsquigarrow s))_{1 \leq t}$ convergent vers la même limite : $\pi_{G_1}(r) = \pi_{G_1}(s) = \frac{d_1}{\sum_{x \in V_1} d_{G_1}(x)}$ tout comme les deux séries $(P^t_{G_2}(u \rightsquigarrow r))_{1 \leq t}$ et $(P^t_{G_2}(u \rightsquigarrow s))_{1 \leq t}$: $\pi_{G_2}(r) = \pi_{G_2}(s) = \frac{d_2}{\sum_{x \in V_2} d_{G_2}(x)}$. Cependant, les deux séries ne convergent pas selon la même dynamique. Au début de la marche, avec t petit, on peut s'attendre à ce que $P^t_{G_1}(u \rightsquigarrow r) > P^t_{G_1}(u \rightsquigarrow s)$ et $P^t_{G_2}(u \rightsquigarrow r) > P^t_{G_2}(u \rightsquigarrow s)$ puisque *éplucher* est sémantiquement plus proche de *dépecer* que de *sonner*. En effet, le nombre de chemins courts entre *éplucher* et *dépecer* est plus grand qu'entre *éplucher* et *sonner*.

La figure 2(a) présente les valeurs de $P^t_{Rob}(u \rightsquigarrow r)$ et de $P^t_{Rob}(u \rightsquigarrow s)$ en fonction de t et les compare à leur limite commune. La figure 2(b) présente ces mêmes valeurs calculées sur *Lar*. Ces figures confirment notre hypothèse : puisque *éplucher* et *dépecer* sont sémantiquement proches, $P^t_{Rob}(u \rightsquigarrow r)$ et $P^t_{Lar}(u \rightsquigarrow r)$ décroissent vers leurs limites même si r et s ne sont pas synonymes (comme c'est le cas dans *Lar*).

En fait, la limite $\pi_G(v)$ ne fournit pas d'information sur la proximité entre u et v dans le graphe ; au contraire, elle la masque par la seule prise en compte de v dans son calcul. Nous définissons donc la t-confluence $CONF^t_G(u,v)$ entre deux sommets u et v sur un graphe G comme suit :

$$CONF^t_G(u,v) = \frac{P^t_G(u \rightsquigarrow v)}{P^t_G(u \rightsquigarrow v) + \pi_G(v)} \tag{5}$$

4. C'est-à-dire que chaque sommet est connecté à lui-même. Si de telles boucles n'existent pas dans les données, elles peuvent généralement être ajoutées sans perte d'information.



(a) $u =$ **éplucher**, $r =$ *dépecer*, $s =$ *sonner*, $f_r^{Rob}(t) = P^t_{Rob}(u \rightsquigarrow r)$, $f_s^{Rob}(t) = P^t_{Rob}(u \rightsquigarrow s)$

(b) $u =$ **éplucher**, $r =$ *dépecer*, $s =$ *sonner*, $f_r^{Lar}(t) = P^t_{Lar}(u \rightsquigarrow r)$, $f_s^{Lar}(t) = P^t_{Lar}(u \rightsquigarrow s)$

FIGURE 2 – Les différentes dynamiques de convergence de la série $(P^t_G(u \rightsquigarrow v))_{1<t}$ vers sa limite pour trois types de relations entre u et v : **(1)** $\mathbf{f_r^{Rob}(t)}$: u et v sont synonymes comme *éplucher* et *dépecer* dans *Rob* ; **(2)** $\mathbf{f_s^{Rob}(t)}$ et $\mathbf{f_s^{Lar}(t)}$: u et v ne sont pas synonymes et sont sémantiquement éloignés comme *éplucher* et *sonner* dans *Rob* et dans *Lar* ; (3) $\mathbf{f_r^{Lar}(t)}$: u et v ne sont pas synonymes mais sont sémantiquement proches comme *dépecer* et *éplucher* dans *Lar*.

$CONF^t_G$ définit une famille de mesures symétriques de proximité entre sommets, une mesure pour chaque longueur de marche t. Par souci de clarté, nous choisissons un t unique pour la suite de l'article. Ce choix est fait en considérant que :
- **Si t est trop grand :** $\forall u_1, v_1, u_2, v_2 \in V,\ CONF^t_G(u_1, v_1) \approx CONF^t_G(u_2, v_2) \approx 0,5$. La mesure $CONF^t_G(u, v)$ n'indique donc pas si les sommets u et v appartiennent ou non à une même zone dense en arêtes de G ;
- **Si t est trop petit :** pour toute paire $\{u, v\}$ telle que la longueur du chemin le plus court entre u et v dans G est plus grande que t, $P^t_G(u \rightsquigarrow v) = 0$ donc $CONF^t_G(u, v) = 0$. Cette mesure n'indique donc pas non plus si les sommets u et v appartiennent ou non à une même zone dense en arêtes G.

C'est pourquoi, dans la suite de cet article, t est fixé [5] à $t = 5$ et $CONF_G = CONF^5_G$.

$CONF_G$ est une mesure de proximité normalisée basée sur les marches aléatoires dans G :
- S'il existe, entre u et v, beaucoup plus de chemins courts qu'entre un sommet quelconque et v (u et v appartiennent à une même zone sur-dense en arêtes) : $P^5_G(u \rightsquigarrow v) > \pi_G(v)$ et donc $CONF_G(u, v) > 0,5$;
- S'il existe, entre u et v, autant de chemins courts qu'entre un sommet quelconque et v : $P^5_G(u \rightsquigarrow v) \approx \pi_G(v)$ et donc $CONF_G(u, v) \approx 0,5$;
- Si il existe, entre u et v, beaucoup moins de chemins courts qu'entre un sommet quelconque et v : $P^5_G(\rightsquigarrow v) < \pi_G(v)$ et donc $CONF_G(u, v) < 0,5$.

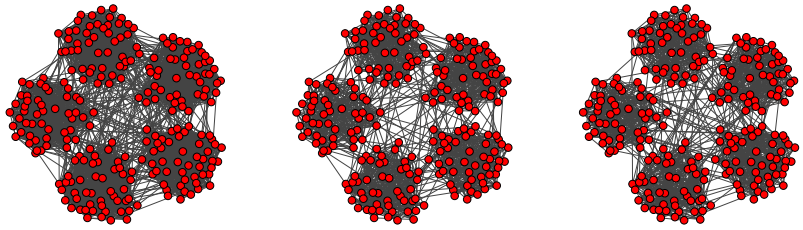
Nous considérons donc comme « proche » toute paire de sommets $\{u, v\}$ telle que la confluence $CONF_G(u, v)$ est plus grande que $0,5$. En d'autres termes, u et v sont proches si la probabilité d'atteindre v à partir de u après une marche aléatoire de cinq pas est plus grande que la probabilité d'être sur v après une marche infinie.

3.2 Une expérimentation contrôlée à l'aide de graphes artificiels

Nous avons construit artificiellement deux types de paire de graphes à comparer :
- **Deux graphes avec 5 zones denses :** nous avons d'abord construit un graphe $G_a = (V, E_a)$ tel que V est l'union de $k = 5$ ensembles $\Delta_1, \ldots \Delta_5$ de $n = 50$ sommets chacun [6] ; les arêtes de E_a ont été placées aléatoirement entre deux sommets u et v à partir de deux probabilités différentes : une probabilité $p_1 = 0,5$ entre deux sommets d'un même ensemble ($u, v \in \Delta_i$), et $p_2 = 0,01$ entre deux sommets appartenant à deux ensembles distincts ($u \in \Delta_i,\ v \in \Delta_j,\ i \neq j$). Nous avons ensuite construit un second graphe $G_b = (V, E_b)$ en choisissant aléatoirement la moitié des arêtes de G_a, et un troisième graphe $G_c = (V, E_c)$ tel que $E_c = E_a \setminus E_b$. Ces trois graphes sont représentés dans la figure 3. Bien que G_b et G_c n'aient aucune arêtes en commun, ($E_b \cap E_c = \varnothing$), G_b et G_c présentent tous deux cinq zones locales denses identiques : $\Delta_1, \ldots \Delta_5$.
- **Deux graphes aléatoires :** nous avons d'abord construit un graphe aléatoire $G^R_a = (V, E^R_a)$ tel que $|E^R_a| = |E_a|$. Nous avons ensuite construit un deuxième graphe $G^R_b = (V, E^R_b)$ en choisissant aléatoirement la moitié des arêtes de G^R_a, et un troisième graphe $G^R_c = (V, E^R_c)$ tel que $E^R_c = E^R_a \setminus E^R_b$. Ni G^R_b ni G^R_c n'ont de zones denses.
- Puisque $E_b \cap E_c = \varnothing$, $E_b \cap \overline{E_c} = E_b$ et $E_c \cap \overline{E_b} = E_c$, donc $GED(G_b, G_c) = \frac{|E_b \cap \overline{E_c}| + |E_c \cap \overline{E_b}|}{|E_b| + |E_c|} = \frac{|E_b| + |E_c|}{|E_b| + |E_c|} = 1$. Ce résultat

5. Avec $t = 5$ nous restons en général proche de la longueur moyenne des plus courts chemins dans les réseaux lexicaux (Motter *et al.*, 2002; Gaume, 2004; de Jesus Holanda *et al.*, 2004). Notons aussi qu'un t petit est favorable à complexité des algorithmes : avec $t = 5$, tous les calculs de confluence (calculs exactes en Python avec un processeur i7) nécessaires pour chacune des paires de graphes analysées dans ce papier ne nécessitent que quelques secondes. Cette approche peut être appliquée à de très grands graphes. Quand les graphes deviennent trop grands, on peut utiliser les méthodes de Monte Carlo qui sont efficaces sur les RPMH pourvu que le degré maximal des sommets soit borné.

6. Si $i \neq j$ alors $\Delta_i \cap \Delta_j = \varnothing$.



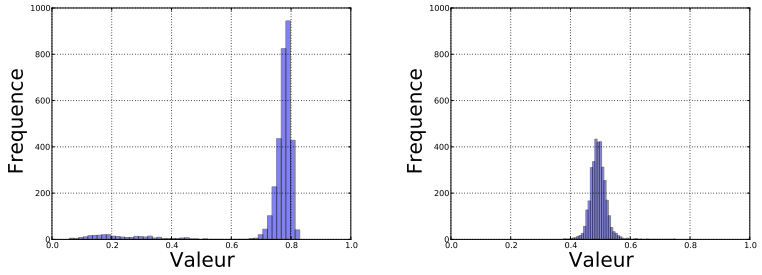
(a) $G_a = (V, E_a)$ (b) $G_b = (V, E_b)$ (c) $G_c = (V, E_c)$

FIGURE 3 – Graphe artificiel avec 5 zones locales denses identiques.

signifierait que ces graphes seraient complètement dissemblables, ce qui est vrai dans le sens où ils n'ont aucune arête en commun mais clairement faux du point de vue de l'« organisation » topologique qu'ils partagent. En effet si deux sommets appartiennent à la même zone relativement dense dans le premier graphe, ils appartiennent également à la même zone relativement dense dans le second.

– Puisque $E_b^R \cap E_c^R = \varnothing$, $E_b^R \cap \overline{E_c^R} = E_b^R$ et $E_c^R \cap \overline{E_b^R} = E_c^R$. Donc, $GED(G_b^R, G_c^R) = \frac{|E_b^R \cap \overline{E_c^R}| + |E_c^R \cap \overline{E_b^R}|}{|E_b^R| + |E_c^R|} = \frac{|E_b^R| + |E_c^R|}{|E_b^R| + |E_c^R|} = 1$.

Toutes les mesures quantitatives de surface comme *GED*, qui ne reposent que sur le décompte du nombre de désaccords, ont le désavantage de ne comparer les graphes que comme des « sacs de liens », étant ainsi insensibles aux contextes topologiques. Mais si nous comparons les distributions de la confluence des arêtes en désaccord dans G_b vs G_c d'un côté (fig. 4(a)), et dans G_b^R vs G_c^R de l'autre côté (fig. 4(b)), la différence est frappante.



(a) G_b/G_c (b) G_b^R/G_c^R

FIGURE 4 – Histogramme de l'ensemble $\{CONF_{G_c}(\{u,v\})$ tel que $\{u,v\} \in (E_b \cap \overline{E_c})\} \bigcup \{CONF_{G_b}(\{u,v\})$ tel que $\{u,v\} \in (E_c \cap \overline{E_b})\}$, en parallèle à l'histogramme de l'ensemble $\{CONF_{G_c^R}(\{u,v\})$ tel que $\{u,v\} \in (E_b^R \cap \overline{E_c^R})\} \bigcup \{CONF_{G_b^R}(\{u,v\})$ tel que $\{u,v\} \in (E_c^R \cap \overline{E_b^R})\}$.

Nous définissons donc $\mu(G_1, G_2)$, une mesure de la similarité structurelle entre les arêtes de G_1 et de G_2 :

$$\mu(G_1, G_2) = \frac{1}{|E_1 \cap \overline{E_2}| + |E_2 \cap \overline{E_1}|} \Big(\sum_{\{u,v\} \in (E_2 \cap \overline{E_1})} CONF_{G_1}(\{u,v\}) + \sum_{\{u,v\} \in (E_1 \cap \overline{E_2})} CONF_{G_2}(\{u,v\}) \Big) \quad (6)$$

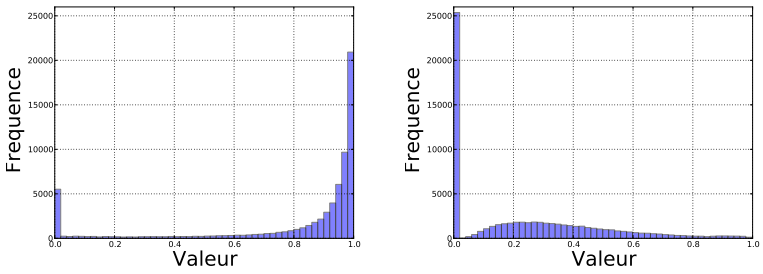
Grâce à la confluence, μ mesure le niveau de proximité structurelle dans G_1 entre les sommets des arêtes directement présentes dans G_2 et absentes de G_1, et le niveau de proximité structurelle dans G_2 entre les sommets des arêtes directement présentes dans G_1 et absentes de G_2.

Bien que $GED(G_b, G_c) = GED(G_b^R, G_c^R) = 1$, avec μ, nous pouvons maintenant voir la différence : sur cinquante réalisations $\mu(G_b, G_c) = 0,74$ (avec un écart type $std < 0,005$) alors que $\mu(G_b^R, G_c^R) = 0,49$ ($std < 0,005$). La différence entre G_b/G_c et G_b^R/G_c^R est identique quantitativement mais différente structurellement.

4 Applications sur les Réseaux lexicaux

Nous commençons par examiner la distribution de la confluence des arêtes contradictoires entre $Lar' = (V', E_{Lar'})$ vs $Rob' = (V', E_{Rob'})$. Nous la comparons à la distribution de la confluence des arêtes contradictoires entre les paires de *graphes aléatoires équivalents* $Lar'^R = (V', E_{Lar'}^R)$ et $Rob'^R = (V', E_{Rob'}^R)$ construits tels que :

$$|E_{Lar'}^R \cap E_{Rob'}^R| = |E_{Lar'} \cap E_{Rob'}|, \quad |E_{Lar'}^R \cap \overline{E_{Rob'}^R}| = |E_{Lar'} \cap \overline{E_{Rob'}}|, \quad |\overline{E_{Lar'}^R} \cap E_{Rob'}^R| = |\overline{E_{Lar'}} \cap E_{Rob'}|$$



(a) Lar'/Rob' (b) Lar'^R/Rob'^R

FIGURE 5 – Histogramme de l'ensemble $\{CONF_{Rob'}(\{u,v\})$ tel que $\{u,v\} \in (E_{Lar'} \cap \overline{E_{Rob'}})\} \bigcup \{CONF_{Lar'}(\{u,v\})$ tel que $\{u,v\} \in (E_{Rob'} \cap \overline{E_{Lar'}})\}$, en parallèle à l'histrogramme de l'ensemble $\{CONF_{Rob'^R}(\{u,v\})$ tel que $\{u,v\} \in (E_{Lar'^R} \cap \overline{E_{Rob'^R}})\} \bigcup \{CONF_{Lar'^R}(\{u,v\})$ tel que $\{u,v\} \in (E_{Rob'^R} \cap \overline{E_{Lar'^R}})\}$

Par construction nous avons $GED(Lar',Rob') = GED(Lar'^R,Rob'^R)$ par contre la différence est clairement visible en comparant la distribution des valeurs de confluence des arêtes contradictoires dans Lar' vs Rob' d'une part (fig. 5(a)) et dans Lar'^R vs Rob'^R de l'autre (fig. 5(b). Quantitativement, la différence entre Lar'/Rob' et Lar'^R/Rob'^R est identique : $GED(Lar',Rob') = GED(Lar'^R,Rob'^R) = 0,47$, mais elle diffère structurellement, $\mu(Lar',Rob') = 0,80$ alors que $\mu(Lar'^R,Rob'^R) = 0,24$ (sur 50 réalisations : $std < 0,005$). Il y'a le même nombre de désaccords, mais ces désaccords sont structurellement faibles entre Lar' et Rob', alors qu'ils sont structurellement forts entre Lar'^R et Rob'^R. C'est ce que nous permet de voir la figure 5 et c'est ce que mesure μ.

Nous comparons maintenant un ensemble de réseaux lexicaux d'origines diverses, ressources construites par des lexicographes et par les foules (crowdsourcing) :

- **Rob** $= (\mathbf{V_{Rob}}, \mathbf{E_{Rob}})$ et **Lar** $= (\mathbf{V_{Lar}}, \mathbf{E_{Lar}})$: voir section 2 ;
- **Wik** $= (\mathbf{V_{Wik}}, \mathbf{E_{Wik}})$: Le wiktionnaire français est construit par les foules sur la base du volontariat. Wiktionary [7] est le compagnon lexical de Wikipedia. Ce dictionnaire multilingue inclus des gloses, des exemples, des relations sémantiques et des liens de traduction que n'importe qui peut modifier. Des instructions sont données aux contributeurs sous la forme de recommandations, mais aucune définition de la relation de synonymie n'est fournie. La construction des graphes de synonymie à partir des « dumps » de Wiktionnary [8] est précisément documentée dans (Sajous *et al.*, 2011). Le graphe $Wik = (V_{Wik}, E_{Wik})$ extrait du wiktionnaire français en janvier 2014 est construit de la même façon que le graph Rob ;
- **Jdm** $= (\mathbf{V_{Jdm}}, \mathbf{E_{Jdm}})$: La ressource *Jeux De Mots* [9] est construite selon une autre forme de crowdsourcing, à partir d'un jeu décrit dans (Lafourcade, 2007). Les joueurs doivent trouver autant de mots que possible qu'ils associent à un terme présenté à l'écran, selon une règle fournie par le jeu. Le but est de trouver le maximum d'associations sémantiques parmi celles que les autres joueurs ont trouvées mais que le joueur concurrent n'a pas trouvées. Plusieurs règles peuvent être proposées, dont la demande d'une liste maximale de synonymes ou quasi-synonymes. A partir des résultats collectés jusqu'en janvier 2014, un graphe de mots liés par des relations sémantiques typées (en fonction des règles) a été construit. Nous travaillons ici sur le sous-graphe des relations de synonymie.

Chacune de ces ressources est découpée en parties du discours (Noms, Verbes, Adjectifs), donnant ainsi trois graphes (ex : $Rob \Rightarrow Rob_N, Rob_V, Rob_A$). Le tableau 2 fournit les pédigrés de ces graphes et montre qu'ils sont tous des RPMH typiques. Dans le tableau 3 nous comparons six paires de graphes par partie du discours.

Entre les graphes *Lar*, *Rob*, et *Jdm* la mesure quantitative de surface *GED* est toujours comprise entre $0,45$ et $0,51$, ce qui indique un accord faible au niveau des liens locaux comparés indépendamment de leurs contextes structurels. Cependant la mesure structurelle μ est toujours supérieure ou égale à $0,70$ ce qui veut dire que malgré la proportion importante de désaccords locaux, ces trois graphes ont une structure globale semblable.

La mesure μ entre *wik* et les autres graphes est toujours inférieure à $0,50$ ce qui veut dire que *wik* diffère au niveau de ses zones denses par rapport à chacun des trois graphes *Lar*, *Rob*, et *Jdm*. Par exemple, la figure 6 montre les sous-graphes sur les voisins de *causer* extraits de Lar_V, Rob_V, Jdm_V et wik_V. On peut y voir un accord entre les trois graphes Lar_V, Rob_V, et Jdm_V au niveau de la *bisémie* du verbe *causer* (PARLER/PROVOQUER) ; le graphe *wik*, quant à lui ne distingue qu'un seul sens : PROVOQUER.

7. http://www.wiktionary.org/
8. Les dumps parsés sont disponibles au format XML à http://redac.univ-tlse2.fr/index_en.html
9. http://www.lirmm.fr/jeuxdemots/jdm-accueil.php

TABLE 2 – Pédigrés des graphes lexicaux (nous renvoyons à la légende de la figure 1 pour la description des colonnes).

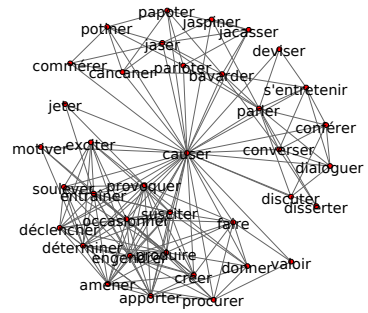
Graphes lexicaux		**n**	**m**	$\langle$**k**$\rangle$	**C**	$\mathbf{L_{lcc}}$	λ ($\mathbf{r^2}$)
Lar	**Adjectifs**	5510	21147	7,68	0,21	4,92	-2,06 (0,88)
	Noms	12159	31601	5,20	0,20	6,10	-2,39 (0,88)
	Verbes	5377	22042	8,20	0,17	4,61	-1,94 (0,88)
Rob	**Adjectifs**	7693	20011	5,20	0,14	5,26	-2,05 (0,94)
	Noms	24570	55418	4,51	0,11	6,08	-2,34 (0,94)
	Verbes	7357	26567	7,22	0,12	4,59	-2,01 (0,93)
Jdm	**Adjectifs**	9859	30087	6,10	0,16	5,44	-2,24 (0,90)
	Noms	29213	56381	3,86	0,14	6,48	-2,66 (0,93)
	Verbes	7658	22260	5,81	0,14	5,06	-2,08 (0,89)
Wik	**Adjectifs**	6960	6594	1,89	0,15	8,48	-2,46 (0,95)
	Noms	43206	37661	1,74	0,13	10,56	-2,51 (0,89)
	Verbes	7203	7497	2,08	0,25	9,22	-2,28 (0,92)

TABLE 3 – Pour comparer deux graphes lexicaux G_1/G_2, on réduit d'abord les deux graphes à leurs sommets communs : $G'_1 = \big(V' = (V_1 \cap V_2), E'_1 = E_1 \cap (V' \times V')\big)$ et $G'_2 = \big(V' = (V_1 \cap V_2), E'_2 = E_2 \cap (V' \times V')\big)$. Ensuite, nous construisons les graphes aléatoires équivalents G'^R_1 et G'^R_2 et calculons : $GED = GED(G'_1, G'2)$, $(\mu) = \mu(G'_1, G'2)$ et $(\mu^R) = \mu(G'^R_1, G'^R_2)$. Chaque valeur (μ^R) sur chacun des graphes aléatoires équivalents, est la moyenne sur 30 réalisations de $\mu(G'^R_1, G'^R_2)$ (tous les écarts type $std < 0,005$).

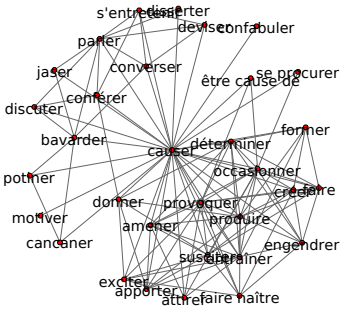
GED (μ) (μ^R) **sur les paires de graphes**			
G_1/G_2	Rob_A	Jdm_A	Wik_A
Lar_A	**0,45** (**0,76**) (0,34)	**0,47** (**0,71**) (0,38)	**0,75** (**0,41**) (0,06)
Rob_A		**0,51** (**0,70**) (0,29)	**0,71** (**0,42**) (0,05)
Jdm_A			**0,54** (**0,43**) (0,03)

G_1/G_2	Rob_N	Jdm_N	Wik_N
Lar_N	**0,48** (**0,70**) (0,20)	**0,48** (**0,70**) (0,19)	**0,72** (**0,31**) (0,03)
Rob_N		**0,47** (**0,70**) (0,13)	**0,71** (**0,29**) (0,02)
Jdm_N			**0,46** (**0,32**) (0,01)

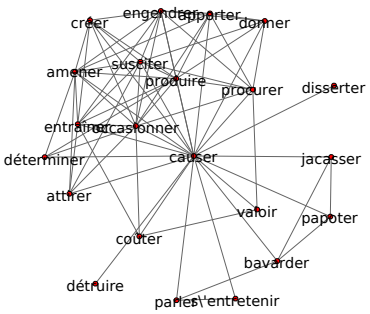
G_1/G_2	Rob_V	Jdm_V	Wik_V
Lar_V	**0,48** (**0,73**) (0,40)	**0,46** (**0,70**) (0,39)	**0,78** (**0,25**) (0,05)
Rob_V		**0,47** (**0,70**) (0,37)	**0,78** (**0,25**) (0,06)
Jdm_V			**0,55** (**0,31**) (0,04)



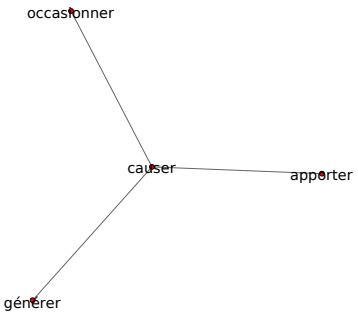
(a) Sous-graphe de Lar_V autour de *causer*



(b) Sous-graphe de Rob_V autour de *causer*



(c) Sous-graphe de Jdm_V autour de *causer*



(d) Sous-graphe de Wik_V autour de *causer*

FIGURE 6 – Accord entre Lar_V, rob_V et Jdm_V sur la polysémie de *causer* (PARLER/PROVOQUER) mais désaccord avec Wik_V (PROVOQUER)

5 Conclusion

« Dans un état de langue tout repose sur des rapports » disait Saussure (1972). Cependant, se limiter à l'analyse de ces rapports au seul niveau local, indépendamment de leurs contextes, n'est pas suffisant. En effet, nous avons montré que si $G_1 = (V, E_1)$ et $G_2 = (V, E_2)$ sont deux graphes standards de synonymie d'une même langue, une grande proportion de paires $\{x, y\} \in \mathbf{P}_2^V$ sont synonymes dans G_1 mais pas dans G_2. Une telle quantité de désaccords n'est pas compatible avec l'hypothèse selon laquelle la synonymie reflèterait une structure sémantique du lexique commune aux membres d'une même communauté linguistique. L'analyse d'une relation lexicale doit être faite au niveau de la structure globale dessinée par la relation. C'est ce que la mesure μ peut faire : avec un niveau de représentation adéquat, elle réconcilie les jugements portés par deux juges différents sur une même relation lexicale.

Ce n'est pas la somme du sens de ses arêtes qui donne le sens d'une relation lexicale, mais le sens de la relation lexicale dans la globalité de sa structure qui donne du sens à ses arêtes : *dans un état de langue tout repose sur la structure des rapports*.

6 Remerciements

Nous remercions les organisateurs de RLTLN2014 pour avoir proposé et organisé ce workshop et les relecteurs qui, par leurs questions et leurs conseils toujours pertinents, nous ont permis d'améliorer cet article. Nous remercions aussi Franck Sajous, Yannick Chudy et Pierre Magistry pour les nombreuses discussions toujours enrichissantes que nous avons eues ensemble.

Références

ALBERT R. & BARABASI A.-L. (2002). Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, **74**, 74–47.

AMBAUEN R., FISCHER S. & BUNKE H. (2003). Graph edit distance with node splitting and merging, and its application to diatom idenfication. In *Graph Based Representations in Pattern Recognition, 4th IAPR International Workshop*, p. 95–106, York, UK.

BARONCHELLI A., I CANCHO R. F., PASTOR-SATORRAS R., CHATER N. & CHRISTIANSEN M. H. (2013). Networks in cognitive science. *CoRR*, **abs/1304.6736**.

BOLLOBAS B. (2002). *Modern Graph Theory*. Springer-Verlag New York Inc.

DE JESUS HOLANDA A., PISA I. T., KINOUCHI O., MARTINEZ A. S. & RUIZ E. E. S. (2004). Thesaurus as a complex network. *Physica A : Statistical Mechanics and its Applications*, **344**(3-4), 530–536.

DESALLE Y. (2012). *Réseaux lexicaux, métaphore, acquisition : une approche interdisciplinaire et inter-linguistique du lexique verbal.* PhD thesis, Université de Toulouse.

DESALLE Y., GAUME B. & DUVIGNAU K. (2009). SLAM : Solutions lexicales automatique pour métaphores. *Traitement Automatique des Langues*, **50**(1), 145–175.

DESALLE Y., GAUME B., DUVIGNAU K., CHEUNG H., HSIEH S.-K., MAGISTRY P. & NESPOULOUS J.-L. (2014a). Skillex, an action labelling efficiency score : the case for french and mandarin. In *Proc. of Cogsci'14, The 36th Annual meeting of the COGnitive SCIence society*, Quebec, Canada. À paraître.

DESALLE Y., NAVARRO E., CHUDY Y., MAGISTRY P. & GAUME B. (2014b). Bacanal : Balades aléatoires courtes pour analyses lexicales, application à la substitution lexicale. In *TALN'14, actes de l'atelier SemDis*, Marseille, France. À paraître.

GAILLARD B., GAUME B. & NAVARRO E. (2011). Invariant and variability of synonymy networks : Self mediated agreement by confluence. In *Proceedings of the The 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, 6th TextGraphs workshop : Graph-based Methods for Natural Language Processing*, Portland, Oregon.

GAO X., XIAO B., TAO D. & LI X. (2010). A survey of graph edit distance. *Pattern Anal. Appl.*, **13**(1), 113–129.

GAUME B. (2004). Balades Aléatoires dans les Petits Mondes Lexicaux. *I3 : Information Interaction Intelligence*, **4**(2).

GAUME B. (2008). Maping the form of meaning in small worlds. *Journal of Intelligent Systems*, **23**(7), 848–862.

GAUME B., MATHIEU F. & NAVARRO E. (2010). Building Real-World Complex Networks by Wandering on Random Graphs. *I3 : Information Interaction Intelligence*, **10**(1).

L. GUILBERT, R. LAGANE & G. NIOBEY, Eds. (1971-1978). *Le Grand Larousse de la langue française (7 vol.) 1971-1978*. Larousse.

KINOUCHI O., MARTINEZ A. S., LIMA G. F., LOURENÇO G. M. & RISAU-GUSMAN S. (2002). Deterministic walks in random networks : An application to thesaurus graphs. *Physica A*, **315**, 665–676. cond-mat/0110217.

LAFOURCADE M. (2007). Making People Play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th Int. Symposium on NLP*, Pattaya, Thailand.

LEVENSHTEIN V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, **10**(8), 707–710.

MOTTER A. E., MOURA A. P. S., LAI Y. C. & DASGUPTA P. (2002). Topology of the conceptual network of language. *Physical Review E*, **65**, 065102.

MURRAY G. C. & GREEN R. (2004). Lexical Knowledge and Human Disagreement on a WSD Task. *Computer Speech & Language*, **18**(3), 209–222.

NAVARRO E. (2013). *Métrologie des graphes de terrain, application à la construction de ressources lexicales et à la recherche d'information.* PhD thesis, Université de Toulouse.

NAVARRO E., CHUDY Y., GAUME B., CABANAC G. & PINEL-SAUVAGNAT K. (2011). Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? In *Proceedings of the Coria 2011 : Conférence en Recherche d'Information et Applications*.

NAVARRO E., GAUME B. & PRADE H. (2012). Comparing and fusing terrain network information. In E. HÜLLERMEIER, S. LINK, T. FOBER & B. SEEGER, Eds., *Scalable Uncertainty Management - 6th International Conference, SUM 2012, Marburg, Germany*, volume 7520 of *LNCS*, p. 459–472 : Springer.

NEWMAN M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, **45**, 167–256.

P. ROBERT & A. REY, Eds. (1985). *Dictionnaire alphabétique et analogique de la langue française 2e éd. (9vol.)*. Le Robert.

SAJOUS F., NAVARRO E., GAUME B., PRÉVOT L. & CHUDY Y. (2011). Wisigoth semi-automatic enrichment of crowdsourced synonymy networks : an application to wiktionary. *LRE Language Resources and Evaluation : Special Issue on Collaboratively Constructed Language Resources*.

SAUSSURE (1972). *Cours de linguistique générale, édition critique préparée par Tullio De Mauro*.

STEWART G. W. (1994). *Perron-Frobenius theory : a new proof of the basics*. Rapport interne, College Park, MD, USA.

STEYVERS M. & TENENBAUM J. B. (2005). The large-scale structure of semantic networks : Statistical analyses and a model of semantic growth. *Cognitive Science*, **29**(1), 41–78.

WATTS D. J. & STROGATZ S. H. (1998). Collective Dynamics of Small-World Networks. *Nature*, **393**, 440–442.