

Cartographie lexicale pour la recherche d'information

Jean Véronis

Equipe DELIC - Université de Provence
29, Av. Robert Schuman - 13621 Aix-en-Provence Cedex 1
Jean.Veronis@up.univ-mrs.fr

Résumé – Abstract

Nous décrivons un algorithme, *HyperLex*, de détermination automatique des différents usages d'un mot dans une base textuelle sans utilisation d'un dictionnaire. Cet algorithme basé sur la détection des composantes de forte densité du graphe des cooccurrences de mots permet, contrairement aux méthodes précédemment proposées (vecteurs de mots), d'isoler des usages très peu fréquents. Il est associé à une technique de représentation graphique permettant à l'utilisateur de naviguer de façon visuelle à travers le lexique et d'explorer les différentes thématiques correspondant aux usages discriminés.

We describe the *HyperLex* algorithm for automatic discrimination of word uses in a textual database. The algorithm does not require a dictionary. It detects high density components in the word-cooccurrence graph, and, contrary to previous methods (word vectors), enables the recognition of very low frequency uses. *HyperLex* is associated with a graphic representation technique that makes it possible to navigate through the lexicon and explore visually the various themes corresponding to the discriminated uses.

Keywords – Mots Clés

Désambiguïsation lexicale, recherche d'information, interfaces graphiques
Lexical disambiguation, information retrieval, graphic interfaces

1 Introduction

La recherche d'information par mot-clés sur le Web, et dans les grandes bases textuelles en général, se heurte au problème de la multiplicité des usages de la plupart des mots. L'homographie et la polysémie omniprésentes dans les langues introduisent un bruit considérable dans les résultats : ainsi, une recherche sur le mot *barrage* verra retourner, au gré des fréquences globales et des heuristiques de classement des moteurs de recherche, des résultats concernant les barrages hydrauliques, les barrages routiers, les matchs de barrage. Extraire les résultats concernant les usages les moins fréquents peut s'avérer particulièrement délicat. Bien sûr l'utilisateur peut généralement compliquer sa requête en croisant des mots-

clés, mais outre que cette technique n'est pas forcément bien maîtrisée du (très) grand public, elle n'est pas toujours très praticable : il ne suffit pas de croiser le mot *barrage* avec le mot *match* pour obtenir les pages concernant les matchs de barrage : de nombreuses pages traitent du thème sans pour autant contenir le mot *match*. Il faudrait alors énumérer les possibilités lexicales et formuler une requête du type *barrage ET (jouer OU jeu OU football OU basket OU basket-ball OU...)*, ce qui est peu économique (et peu sûr). Par ailleurs, sans que le phénomène relève probablement de ce que les lexicographes appelleraient polysémie, un mot peut être impliqué dans des thèmes distincts, qu'il est important de distinguer pour l'utilisateur. Ainsi, bien qu'il s'agisse toujours de barrages sur des routes, il est peu probable que la même requête cherche à retourner pêle-mêle des pages parlant de barrages revendicatifs de camions, ou de barrages d'hommes en armes à des frontières ou des zones en conflit.

Les dictionnaires classiques sont peu adaptés à la tâche. Ils contiennent la plupart du temps des définitions d'une trop grande généralité (« action de barrer », par exemple), et rien ne garantit qu'elles reflètent le contenu exact du corpus textuel interrogé. Nous avons montré de façon expérimentale la difficulté pour des linguistes de faire correspondre correctement les « sens » d'un dictionnaire et les occurrences d'un corpus (Véronis, 1998). De plus, il resterait à catégoriser automatiquement les documents de la base de textes en fonction des « sens » du dictionnaire, tâche d'une difficulté extrême, qui élude les efforts soutenus de la recherche depuis un demi-siècle (cf. [Ide, Véronis, 1998] pour un état de l'art détaillé).

Schütze (1998) a proposé une méthode permettant d'extraire automatiquement la liste des « sens » (nous préférons parler d'« usages ») du corpus lui-même, tout en fournissant une technique robuste de catégorisation. Ces usages correspondent à des groupes (*clusters*) de contextes similaires dans un espace de très grande dimensionnalité formé par des vecteurs de mots ou de cooccurrences proches du mot à désambiguïser, espace rendu utilisable par une technique de décomposition en valeurs singulières classique. Les techniques basées sur les vecteurs de mots se heurtent toutefois à une difficulté majeure et rédhibitoire : la très grande différence de fréquence entre usages d'un même mot (déjà constatée par Zipf, 1945) repousse la plupart des distinctions utiles en-dessous du seuil de bruit du modèle. Ainsi, selon nos estimations, l'usage « match de barrage » concerne moins de 1% des documents contenant le mot *barrage*.

Nous proposons dans cette communication un algorithme radicalement différent, basé sur les propriétés des graphes constitués par les cooccurrences de mots. Cet algorithme, *HyperLex*, permet d'extraire des composantes de forte densité qui reflètent les différents usages des mots. Contrairement aux méthodes vectorielles, l'algorithme *HyperLex* est peu sensible à la fréquence relative et au nombre des différents usages à discriminer. De plus, il est associé à une technique de représentation graphique permettant à l'utilisateur de naviguer de façon visuelle à travers le lexique et d'explorer les différentes thématiques correspondant aux usages discriminés. Nous illustrons *HyperLex* sur des exemples tirés du Web, mais rien n'empêche d'appliquer la même technique à n'importe quelle grande base textuelle.

2 Méthode et hypothèses

Nous partons du même constat que Shütze (1998) et d'autres, selon lequel les cooccurrences constituent de forts indices désambiguïsateurs pour distinguer les différents usages des mots.

Ainsi, en présence du mot *fleuve*, le mot *barrage* renverra presque obligatoirement à l'usage « barrage hydraulique », alors qu'en présence du mot *football*, il renverra très probablement à « match de barrage ».

Partant d'un mot cible donné, nous extrayons de la base textuelle l'ensemble des contextes (nous avons choisi un paragraphe, mais ce pourrait être le document entier ou une fenêtre de taille donnée) contenant le mot-cible. Ces contextes sont lemmatisés, puis filtrés pour éliminer les mots-outils (déterminants, prépositions, etc.), ainsi qu'un certain nombre de mots généraux et tout particulièrement, dans notre application, ceux liés au Web lui-même (*menu, accueil, lien, http, etc.*).

L'ensemble des cooccurrences forme un graphe, dont les nœuds sont les différents mots du corpus. Une arête relie deux nœuds A, B chaque fois que les mots correspondants ont été trouvés en cooccurrence autour du mot-cible (au-delà d'un seuil de fréquence fixé). Nous affectons à chaque arête un poids d'autant plus faible que les mots sont fréquemment associés :

$$w_{A,B} = 1 - \max[p(A | B), p(B | A)]$$

où $p(A | B)$ est la probabilité conditionnelle d'observer A dans un contexte donné sachant que ce contexte contient B , et inversement, $p(B | A)$ celle d'observer B dans un contexte donné sachant que ce contexte contient A . Ces probabilités sont estimées à partir des fréquences :

$$p(A | B) = f_{A,B} / f_B \quad \text{et} \quad p(B | A) = f_{A,B} / f_A$$

Nous prendrons à titre d'illustration les cooccurrences *eau - ouvrage* et *eau - potable*. Le Tableau 1 donne le nombre de contextes dans lesquels ces couples de mots apparaissent ensemble ou l'un sans l'autre dans un corpus de pages Web (cf. section 4). On voit que toutes les occurrences du mot *potable* apparaissent conjointement avec le mot *eau*, alors que c'est le cas seulement d'une partie des occurrences du mot *ouvrage*.

	EAU	~EAU	Total		EAU	~EAU	Total
OUVRAGE	183	296	479	POTABLE	63	0	63
~OUVRAGE	874	5556	6430	~POTABLE	994	5852	6846
Total	1057	5852	6909	Total	1057	5852	6909

Tableau 1. Cooccurrences *eau-ouvrage* et *eau-potable*

On a :

$$p(eau | ouvrage) = 183/479 = 0,38 \quad p(ouvrage | eau) = 183/1057 = 0,17 \quad w = 1 - 0,38 = 0,62$$

$$p(eau | potable) = 63/63 = 1 \quad p(potable | eau) = 63/1057 = 0,06 \quad w = 1 - 1 = 0$$

La mesure reflète donc la plus ou moins grande « distance¹ » sémantique entre mots : lorsqu'elle vaut 0, les mots sont toujours associés (jusqu'à concurrence de la fréquence maximale possible, déterminée par le moins fréquent) ; lorsqu'elle vaut 1, les mots ne sont jamais associés.

¹ Il ne s'agit pas d'une *distance* au sens mathématique du terme, mais d'une *dissimilarité*, l'inégalité triangulaire n'étant pas respectée.

L'hypothèse de base qui sous-tend notre méthode est que les différents usages du mot-cible constituent des sous-ensembles, ou « composantes », de haute densité dans le graphe. En effet, *barrage* (dans l'usage « barrage hydraulique ») doit être en cooccurrence fréquente avec *eau*, *ouvrage*, *fleuve*, *rivière*, *crue*, *irrigation*, etc., et ces mots eux-mêmes ont toutes les chances d'être interconnectés (Figure 1). De même, dans l'usage « match de barrage », *barrage* doit être en cooccurrence fréquente avec *match*, *équipe*, *coupe*, *football*, *victoire*, etc., ces termes eux-mêmes étant fortement interconnectés. Etant donné la complexité du langage (et en particulier le fait que les mots entrant dans les cooccurrences sont eux-mêmes ambigus), il existe aussi des connexions entre les composantes, ce qui interdit l'utilisation d'algorithmes de détection de composantes fortement connexes ou de cliques, mais ces interconnexions entre les composantes doivent être peu nombreuses et leur poids élevé.

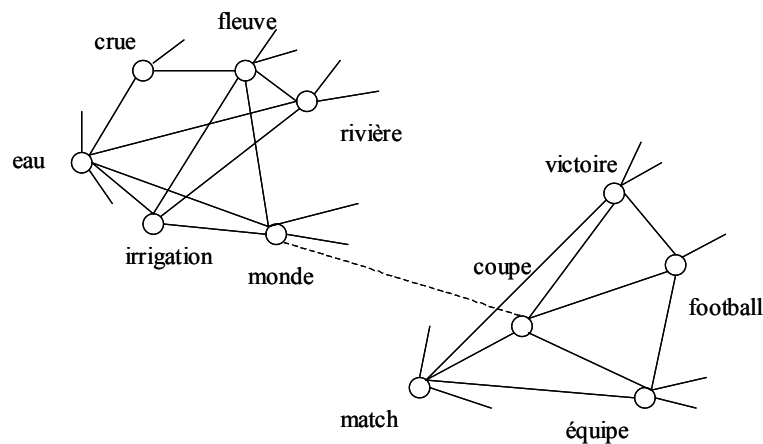


Figure 1. Composantes de forte densité dans le graphe des cooccurrences

Détecter les différents usages d'un mot revient donc à isoler des *composantes de forte densité* du graphe des cooccurrences. La plupart des techniques exactes de partitionnement de graphes sont malheureusement NP-difficiles, et l'on ne peut (étant donné que les graphes obtenus ont plusieurs milliers de nœuds et d'arêtes) qu'utiliser des méthodes approximatives, basées sur des heuristiques. La recherche sur la détection de composantes de forte densité est un domaine particulièrement actif, qui intervient notamment dans les secteurs de la détection de « communautés » ou de « sources autorisées » sur le Web, ou la parallélisation des calculs. Malheureusement, les techniques développées dans ces secteurs ne sont pas directement exploitables, étant donné que l'applicabilité des heuristiques dépend des applications et des propriétés particulières des graphes analysés. La propriété fondamentale que nous exploiterons ici est le caractère zipfien du langage, rappelé plus haut, et qui touche à la fois la fréquence des mots eux-mêmes, la fréquence des usages de chaque mot, et la fréquence des cooccurrences (voir Zipf, 1945, et aussi Baayen, 2000).

3 Détection des composantes de forte densité

Nous créons le graphe des cooccurrences en ne retenant que les mots apparaissant dans au moins 10 contextes, et les cooccurrences correspondant à un poids w en deçà d'un certain seuil (0.9). L'algorithme de détection part du principe que pour chacun des usages du mot considéré (« mot-cible »), un cooccurent (ou « mot-racine ») a une fréquence plus élevée que les autres, en vertu de l'organisation zipfienne des cooccurrences que nous avons mentionnée

plus haut. Ainsi, pour *barrage*, le cooccurrent le plus fréquent dans l'usage « barrage hydraulique » est *eau*, tandis que le cooccurrent le plus fréquent dans l'usage « match de barrage » est *match*.

De plus, nous faisons l'hypothèse que les autres cooccurrents du mot-cible pour un usage donné ont toutes les chances d'apparaître à un moment ou à un autre au contact du mot-racine de cet usage, et donc d'être directement connecté à celui-ci dans le graphe. Par exemple, les voisins directs de *eau* (dans le contexte de *barrage*) sont *ouvrage*, *rivière*, *cours*, *retenue*, *construction*, etc. Ceux de *match* sont *équipe*, *monde*, *coupe*, *football*, *but*, etc. Les mots fréquents qui ne seraient pas des voisins de *eau* ont de bonnes chances de faire partie d'un autre usage.

L'algorithme *HyperLex* utilise cette propriété pour déterminer de façon itérative les « mots-racines » caractérisant chaque usage. Dans un premier temps, *HyperLex* constitue une liste des nœuds du graphe classés par ordre de fréquence décroissante. Pour chaque nœud de la liste, l'algorithme essaie de déterminer si ce nœud est une racine raisonnable pour une composante de forte densité, selon deux critères :

- (1) le nœud doit avoir au moins 6 voisins propres (c'est-à-dire qui ne sont pas voisins d'un mot-racine précédemment déterminé),
- (2) le poids moyen entre le nœud-racine potentiel et ses 6 premiers voisins propres doit être inférieur à un seuil donné (0.8).

Si le nœud satisfait à ces conditions, il est retenu comme mot-racine et il est éliminé de la liste avec tous ses voisins. Le processus est itéré jusqu'à avoir épuisé tous les nœuds du graphe.

Ceci donne les étapes suivantes pour *barrage* dans notre corpus (les mots-racines successifs et leurs voisins propres les plus fréquents apparaissent après le signe \Rightarrow ; les mots qui ne satisfont pas les critères 1 et 2 sont barrés) :

Liste 1: eau construction ouvrage rivière projet retenue crue sécurité hydroélectrique étude travail hauteur mètre zone plan région terre développement cours lac environnement réservoir technique niveau bassin impact fleuve débit population...

\Rightarrow EAU : *construction ouvrage rivière projet retenue crue...*

Liste 2: routier frontière comité match Algérie militaire efficacité armée véhicule Suisse histoire carte auscultation poste Tunisie ALN camion jeu américain connaissance défense document membre liste inspection position financement...

\Rightarrow ROUTIER : *véhicule camion membre conducteur policier groupement...*

Liste 3: frontière comité match Algérie militaire efficacité armée Suisse histoire carte auscultation poste Tunisie ALN jeu américain connaissance défense document liste inspection position financement contenance algérien division Allemagne...

\Rightarrow FRONTIERE : *Algérie militaire efficacité armée suisse poste...*

Liste 4: ~~comité~~ match histoire carte auscultation jeu américain connaissance document liste inspection position financement contenance division Allemagne province CMB INGEMA gabion prix débat congrès renseignement dommage réaction...

\Rightarrow MATCH : *vainqueur victoire rencontre qualification tir football...*

Après cette étape, plus aucun mot ne satisfait les conditions. Au total, *HyperLex* a donc déterminé quatre mots-racines, *eau*, *routier*, *frontière*, *match*, qui reflètent bien les usages de *barrage* dans le corpus.

4 Expérimentation

Nous avons expérimenté *HyperLex* sur 10 mots très polysémiques (Tableau 2), choisis parmi ceux qui ont servi de mots-tests lors de l'action de désambiguïsation *Romanseval*, et qui avaient posé de grandes difficultés à des annotateurs humains (Véronis, 1998). Un sous-corpus de pages Web a été constitué pour chacun de ces mots, à l'aide du méta-moteur Copernic Agent², en interrogeant tout d'abord la forme singulier, puis la forme pluriel. Les pages obtenues ont été filtrées de façon à éliminer les pages qui ne contenaient pas le mot cherché (erreurs du type « Page not found », par exemple), ainsi que les doublons.

Mot	Pages		Contextes		Graphe	
	Brutes	Utiles	Bruts	Utiles	Nœuds	Arêtes
<i>BARRAGE</i>	1702	1372	7256	6924	1203	6138
<i>DETENTION</i>	2112	1270	8902	8728	1418	19007
<i>FORMATION</i>	5974	1590	5248	4885	542	1531
<i>LANCEMENT</i>	2828	1231	3307	3174	617	2521
<i>ORGANE</i>	2786	994	2953	2849	531	1997
<i>PASSAGE</i>	3512	1046	4210	3894	797	2916
<i>RESTAURATION</i>	5327	1227	3522	3287	512	1398
<i>SOLUTION</i>	6287	896	2085	1915	253	1704
<i>STATION</i>	7916	1093	3837	3671	487	971
<i>VOL</i>	5237	818	3001	2579	259	719

Tableau 2. Mots-cibles et caractéristiques quantitatives des sous-corpus

Les paragraphes contenant chaque mot-cible ont été extraits et étiquetés à l'aide du logiciel Cordial Analyseur³. Seuls ont été retenus les noms et les adjectifs. Dans un premier temps nous avons retenu aussi les verbes, mais il s'est finalement avéré que ceux-ci dégradent notablement les performances. Les mots généraux appartenant à un anti-dictionnaire (stoplist) ont été supprimés. Finalement, les contextes contenant moins de 4 mots après filtrage ont été éliminés. Le Tableau 2 donne les caractéristiques quantitatives du sous-corpus recueilli pour chaque mot, ainsi que du graphe de cooccurrences qu'il a permis de construire.

L'algorithme *HyperLex* a été appliqué à chacun des mots, et donne les résultats résumés dans le Tableau 3. La liste des usages détectés pour chaque mot reflète évidemment la composition du corpus, et les biais introduits par les moteurs de recherches et leurs algorithmes de pondération et de classement des résultats (tels que *PageRank* de Google). Par exemple, le mot *formation*, polysémique dans la langue générale, n'apparaît dans notre corpus que dans le contexte d'apprentissage et d'enseignement. On note aussi l'influence d'énumérations, comme pour le mot *station*, où certaines pages contiennent des annuaires de stations-services délivrant du gaz GPL. Ces énumérations résultent dans des usages un peu artificiels pour les mots concernés. L'algorithme n'étant pas ici en cause, nous n'avons pas exploré plus avant la question, mais il conviendrait sans doute de se pencher plus en détail sur le nettoyage des pages dans une utilisation opérationnelle.

Les usages répertoriés couvrent la quasi-totalité des thématiques développées dans les différents sous-corpus. En procédant par sondage (1 page sur 100), nous n'avons constaté l'absence d'aucune thématique majeure. Ceci n'exclut pas que quelques usages très peu

² <http://www.copernic.com>

³ Développé par Synapse Développement : <http://www.synapse-fr.com>

fréquents aient été omis par l'algorithme, mais seule une évaluation beaucoup plus détaillée pourrait les révéler.

Mot-Cible	Racine	Fréq.	Voisins les plus fréquents
BARRAGE	EAU	1057	construction ouvrage rivière projet retenue crue
	ROUTIER	125	véhicule camion membre conducteur policier groupement
	FRONTIERE	106	Algérie militaire efficacité armée Suisse poste
	MATCH	65	vainqueur victoire rencontre qualification tir football
DETENTION	PROVISOIRE	2189	juge liberté loi procédure prison instruction
	DETENU	1049	police centre autorité arrestation torture arbitraire
	ARME	306	autorisation acquisition feu munition vente commerce
	ANIMAL	212	transport compagnie sauvage certificat annexe directive
FORMATION	PROFESSIONNEL	573	centre entreprise organisme stage service programme
LANCEMENT	SATELLITE	301	Ariane programme spatial lanceur orbite fusée
	PRODUIT	124	public entreprise événement convention presse affaire
ORGANE	DON	187	transplantation greffe donneur prélèvement tissu vie
	DELIBERANT	126	public établissement président demande attribution communauté
	REGLEMENT	91	pays appel différend OMC réunion autorité
	TECHNIQUE	62	scientifique convention économique conférence subsidiaire programme
	CONSULTATIF	52	matière civil tête supervision memorandum PAB
	MALADIE	48	cœur traitement spécimen preuve sang intervention
	REPRESENTANT	48	délégué suprême concertation département personnel agent
	PARTI	47	presse chef journal Genève Allemagne rédacteur
PASSAGE	EURO	621	public travail entreprise système national monnaie
	AN_2000	264	programme autorité installation réseau solution matériel
	NIVEAU	172	porte chemin ouverture salle route entrée
	LIBRE	92	cour prestation police assurance caisse prévoyance
	CHEVAL	71	main énergie équilibre trot dos foulée
	PARAMETRE	61	mode appel variable argument langage expression
	GALERIE	53	ville boutique bois panorama époque verrière
	TERRE	53	durée mouvement soleil Vénus Mercure nœud
	MORT	46	rite Dieu naissance Christ vivant Jésus
RESTAURATION	HOTELLERIE	188	formation durée centre professionnel entreprise alternance
	CONSERVATION	157	sauvegarde atelier monument technique historique oeuvre
	HEBERGEMENT	136	activité hôtel région loisir culture contact
	RAPIDE	114	restaurant vente établissement repas marche traiteur
	FICHER	103	système information donnée client espace bande
	PIERRE	70	bâtiment chantier terre polychromie taille sec
	MEUBLE	69	bois table mobilier décoration fabrication antiquité
SOLUTION	GESTION	147	entreprise service logiciel client information système
	JEU	80	monde gratuit astuce joueur gain francophone
	INJECTABLE	65	perfusion glucose HOP commercialisation arrêt Fandre
STATION	SKI	261	hiver piste montagne sport village location
	METEO	132	température Oregon scientifique WS professionnel capteur
	SPATIAL	117	international MIR système programme ISS projet
	TRAVAIL	106	réseau traitement donnée carte Sun environnement
	RADIO	94	navire région réception installation antenne communication
	PRIMAGAZ	86	Paris aire Esso province Marseille Dyneff
	EAU	79	épuration source mer plage Yves rivière
	LIGNE	65	métro quai terminus voyageur correspondance atelier
VOL	AVION	393	billet pilote club sec départ voyage
	LIBRE	260	école parapente loisir montagne formation Paris
	VOILE	219	centre photo vent pilotage forum compétition
	VOLE	44	service recherche numéro base donnée véhicule

Tableau 3. Résultats de l'algorithme sur les 10 mots-cibles
(la colonne *Freq.* donne la fréquence des mots-racines)

Par contre, on peut discuter certaines distinctions ou certains regroupements opérés par l'algorithme. Ainsi, pour *détention*, l'algorithme distingue les aspects juridiques liés à la détention provisoire (sous cette racine), et ceux liés au prisonnier et aux conditions de détention (sous la racine *détenu*). En examinant plus en détail les pages concernées, on constate effectivement des champs lexicaux différents, l'un de type juridique et judiciaire, l'autre lié aux aspects humains de la détention, mais les deux pourraient sans doute être regroupés. A l'inverse, l'algorithme fusionne les stations de radio à usage public (FM, etc.) et les stations radio de navires, les champs lexicaux étant assez proches (*radio, communication,*

bande, MHz, etc.). On voudrait pourtant très certainement différencier ces usages, qui correspondront vraisemblablement à des requêtes différentes. L'algorithme pourrait être amélioré sur plusieurs points, notamment par la prise en compte des distances entre cooccurrents. Ainsi, l'expression « station de radio » n'est utilisée que pour les stations de radio à usage public, alors que pour l'usage maritime, on trouve l'expression « station de navire », et le mot *radio* lui-même est généralement plus éloigné dans le contexte, entrant dans d'autres expressions (*opérateur radio, équipement radio, etc.*).

Enfin, la sélection du mot-racine est dans quelques cas discutable, même si lui-même et ses voisins représentent bien la thématique de l'usage considéré. Par exemple, *métro* serait plus intuitif que *ligne* pour l'usage « station de métro ». L'usage des distances entre le mot-cible et ses cooccurrents pourrait probablement améliorer les choses, comme pour le cas précédent.

5 Visualisation et navigation « hyperlexicale »

La visualisation et la navigation à l'intérieur de grands graphes est un domaine de recherche en pleine expansion, lié notamment à la nécessité de représenter de façon satisfaisante les liens entre serveurs ou entre hyperdocuments à l'intention des administrateurs et webmasters. L'algorithme H3 récemment développé par Tamara Munzner dans le cadre de sa thèse (Munzner, 2000) nous paraît l'un des plus convaincants, et, contrairement à d'autres, directement adaptable à notre problématique. Un outil qui implémente cet algorithme, *H3Viewer*, est de plus disponible⁴, et permet une expérimentation immédiate.

Le problème de la représentation de grands graphes dans un espace euclidien est tout d'abord le « manque de place » de l'espace euclidien, puisque le nombre de nœuds d'un arbre, par exemple, croît de façon exponentielle avec la profondeur, alors que l'espace disponible pour les représenter croît seulement de façon polynomiale par rapport au rayon d'un cercle ou d'une sphère. Il en résulte des distorsions inévitables, qui peuvent perturber considérablement la lecture, et ont souvent pour effet, lorsqu'on se déplace en profondeur dans l'arbre, de faire perdre totalement de vue le contexte environnant. De plus, des stratégies doivent être développées pour gérer la surabondance d'information (multiples liens et nœuds) qui rend très rapidement les représentations totalement illisibles.

L'algorithme H3 résout ces problèmes de façon très élégante à l'aide de la géométrie hyperbolique : dans un espace hyperbolique, l'espace disponible croît de façon exponentielle avec le rayon, contrairement à l'espace euclidien. L'espace hyperbolique (infini) est ensuite projeté dans une portion finie de l'espace euclidien (un disque), comme filmé par une caméra, dont l'utilisateur peut déplacer la position et l'angle. Le résultat apparaît à l'utilisateur en perspective comme une projection dans une sphère (effet « fish-eye » : Figure 2).

H3 prend en entrée un arbre construit à partir du graphe à visualiser. L'arbre que nous construisons essaie de représenter au mieux les composantes de forte densité qui représentent les différents usages du mot-cible. Nous construisons le sommet de l'arbre en ajoutant un nœud qui correspond au mot-cible (par exemple *barrage*), relié par des arêtes de poids 0 à chacun des mots-racines déterminés à l'étape précédente. Après quelques heuristiques

⁴ <http://graphics.stanford.edu/~munzner/h3/>

d'élagage de certaines arêtes (par exemple celles qui relient des mots de fréquences très différentes), nous calculons un *arbre de couverture minimale* ayant pour racine le mot-cible. Cet arbre organise les différentes composantes denses d'une façon qui reflète la force d'association entre mots. Il est ensuite visualisé à l'aide d'*H3Viewer*. L'utilisateur peut naviguer de thème en thème à l'intérieur de la représentation hyperbolique à l'aide de la souris : un clic sur un nœud avec le bouton gauche permet de centrer la représentation autour de ce nœud, un glissement avec le bouton gauche appuyé permet de déplacer un nœud et de changer le contexte, un glissement avec le bouton droit permet une rotation de l'arbre. La Figure 2 montre en (a) la vue initiale offerte à l'utilisateur pour le mot *barrage* et en (b) la vue obtenue en cliquant sur le nœud *match*.

6 Conclusion

Nous avons présenté un algorithme, *HyperLex*, qui permet d'extraire automatiquement d'une grande base textuelle les usages d'un mot donné, sans recours à un dictionnaire. Contrairement aux méthodes précédentes basées sur des vecteurs de mots, notre technique permet d'isoler des usages de fréquence très faible. L'algorithme est couplé avec une technique de représentation visuelle basée sur un espace hyperbolique (Munzner, 2000) qui permet à l'utilisateur de naviguer à travers le lexique et les thématiques dégagées.

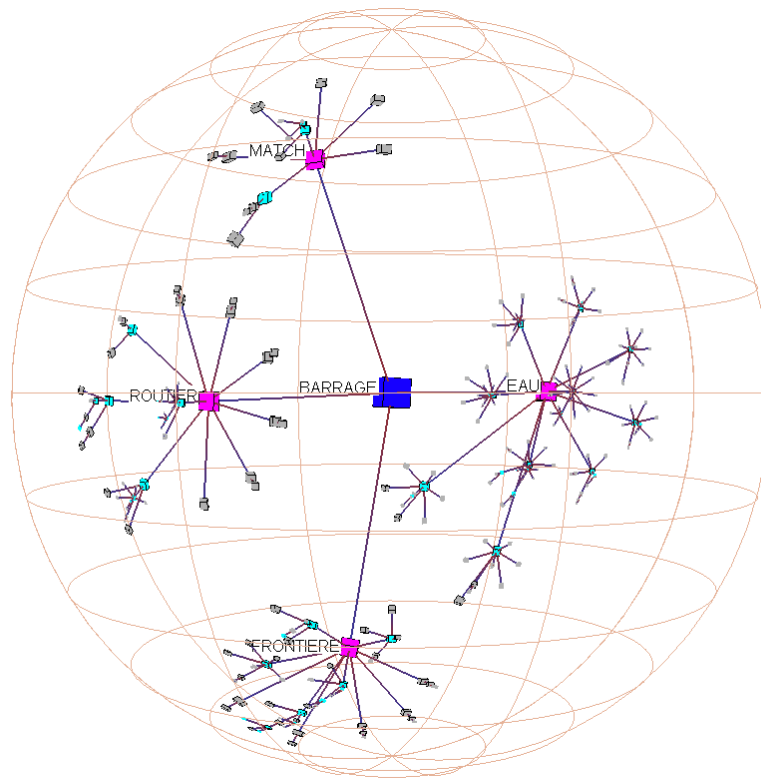
Remerciements

Nous sommes très reconnaissant à Tamara Munzner pour la mise à disposition de l'outil *H3Viewer*, utilisé dans ce travail.

Références

- Baayen, R. H. (2000). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.
- Ide, N. M., Véronis, J. (1998). Introduction to the special issue on word sense disambiguation : the state of the art. *Computational Linguistics*, (24)1, 1-40.
- Munzner, T. (2000). *Interactive visualization of large graphs and networks*. Ph. D. Dissertation, Stanford University.
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, (24)1, 97-124.
- Véronis, J. (1998). A study of polysemy judgements and inter-annotator agreement, *Programme and advanced papers of the Senseval workshop* (pp. 2-4). Herstmonceux Castle (England) [<http://www.up.univ-mrs.fr/veronis/pdf/1998senseval.pdf>].
- Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, 33, 251-266.

a)



b)

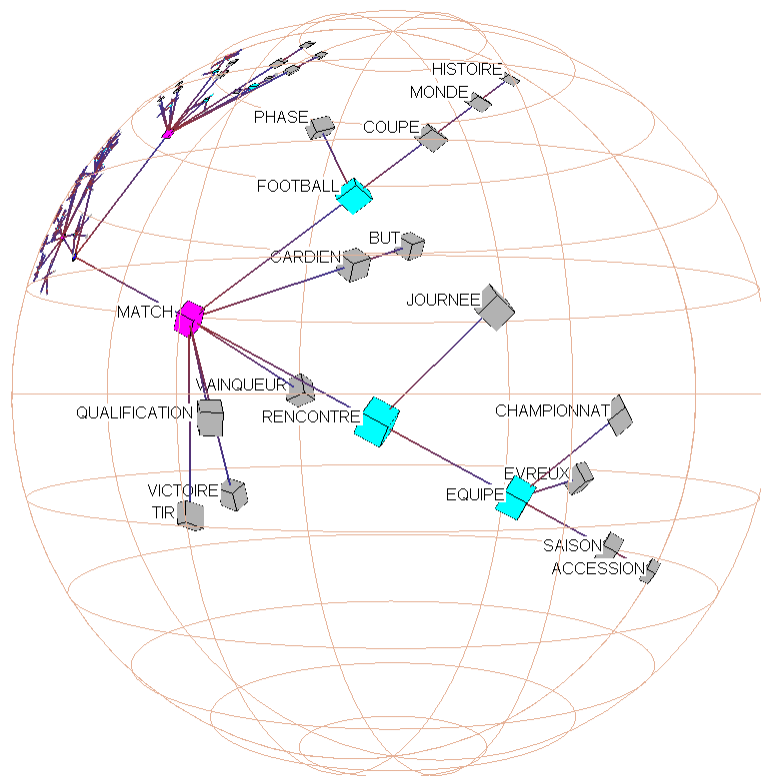


Figure 2. Deux vues du champ lexical de *barrage*