Le dictionnaire électronique des synonymes du CRISCO

Jean-Luc Manguin

Colloque "Sciences humaines et nouvelles technologies"
Tunis, 30 mai - 1er juin 2002
Titre : "Le dictionnaire des synonymes du CRISCO"

Introduction

Le laboratoire CRISCO (Centre de Recherche Inter-langues sur la Signification en COntexte) est une unité mixte de recherche de l'Université de Caen et du CNRS ; les recherches qui y sont menées relèvent de la linguistique générale, et de la linguistique informatique. Dans ce dernier domaine, des travaux sur la modélisation et la représentation du sens des mots ont conduit les chercheurs à construire un dictionnaire des synonymes du français sous forme d'une base de données exploitable par des programmes informatiques.

En 1998, la création du site Internet du laboratoire s'est accompagné de l'ouverture de ce dictionnaire à la libre consultation par les internautes ; mais ce dictionnaire qui n'était au départ qu'un support pour la recherche en sémantique s'est avéré, au contact du public, être également un outil répondant à une formidable demande. Certes son apport à la recherche a continué à se développer, mais son originalité et la demande croissante dont il fait l'objet nous ont amenés à le considérer autrement : du statut de simple ressource, il est devenu lui-même objet de la recherche, qu'elle soit mathématique ou linguistique.

1. Genèse et présentation actuelle du dictionnaire

Le dictionnaire des synonymes du CRISCO est unique en son genre ; c'est en effet le seul dictionnaire des synonymes du français en accès libre sur Internet (sur http://www.crisco.unicaen.fr/) ; il contient 49000 entrées, chacune suivie de ses synonymes, comme par exemple :

démarcation:distinction,délimitation,frontière,ligne,limitation,limite,lisière,marque,séparation

Chaque synonyme est lui-même une entrée, de sorte que l'on peut considérer notre dictionnaire comme un ensemble de 49000 mots-vedettes (entrées) reliés les uns aux autres par un réseau de relations. Il y a relation lorsque les deux vedettes sont synonymes, et l'on dénombre un peu plus de 198 000 relations pour tout le dictionnaire.

Ces relations de synonymie proviennent de sept dictionnaires classiques : deux dictionnaires analogiques (le Grand Larousse et le Grand Robert), deux dictionnaires des synonymes du 19è siècle (Lafaye et Guizot), et trois dictionnaires des synonymes du milieu et de la fin du 20è siècle (Bailly, Bénac et Du Chazaud). Toutes ces informations contenues dans ces dictionnaires nous ont été fournies sous forme de fichiers informatiques par l'Institut National de la Langue Française ; le travail effectué au laboratoire a consisté à harmoniser les mots-vedettes de ces sept fichiers, puis à fusionner toutes ces données. Le tableau ci-dessous permet de mesurer l'ampleur de ce travail :

	Données de départ	**Conservées**	**Ajoutées**	**Données actuelles**
Entrées	63000	48500	600	49100
Relations	219000	192000	6000	198000

Tableau 1 : corrections et ajouts effectués sur les données.

Par ailleurs, la fusion des dictionnaires a fait disparaître (quand ils existaient) les commentaires provenant des ouvrages d'origine ; cet appauvrissement apparent, nécessaire en raison des différences entre les différents dictionnaires de départ, constitue en fait un tremplin pour des méthodes automatiques qui, comme nous le verrons, permettent d'analyser de manière fine et objective le champ sémantique des mots.

Outre cela, la consultation du dictionnaire exploite les possibilités offertes par Internet, notamment l'hypernavigation qui permet aux utilisateurs de trouver rapidement le terme qui leur manque.

L'interface de consultation se présente sous la forme d'un formulaire contenant un champ unique, dans lequel l'utilisateur va taper le mot dont il recherche les synonymes (figure 1).

Figure 1 : écran d'accueil du dictionnaire des synonymes.

En réponse à sa requête, il reçoit une page contenant la liste des synonymes, et un classement fondé sur le nombre de connexions que ces synonymes forment entre eux (figure 2) ; la justification de ce classement automatique est la suivante :

a) Nous considérons le graphe ayant pour sommets le mot demandé et ses synonymes, et pour arêtes les relations directes existant entre tous ces sommets.

b) Le mot demandé est en relation avec tous les autres sommets (puisque ce sont ses synonymes).

c) Il est légitime de penser que le synonyme qui a presque autant de relations que le mot demandé est celui qui lui ressemble le plus.

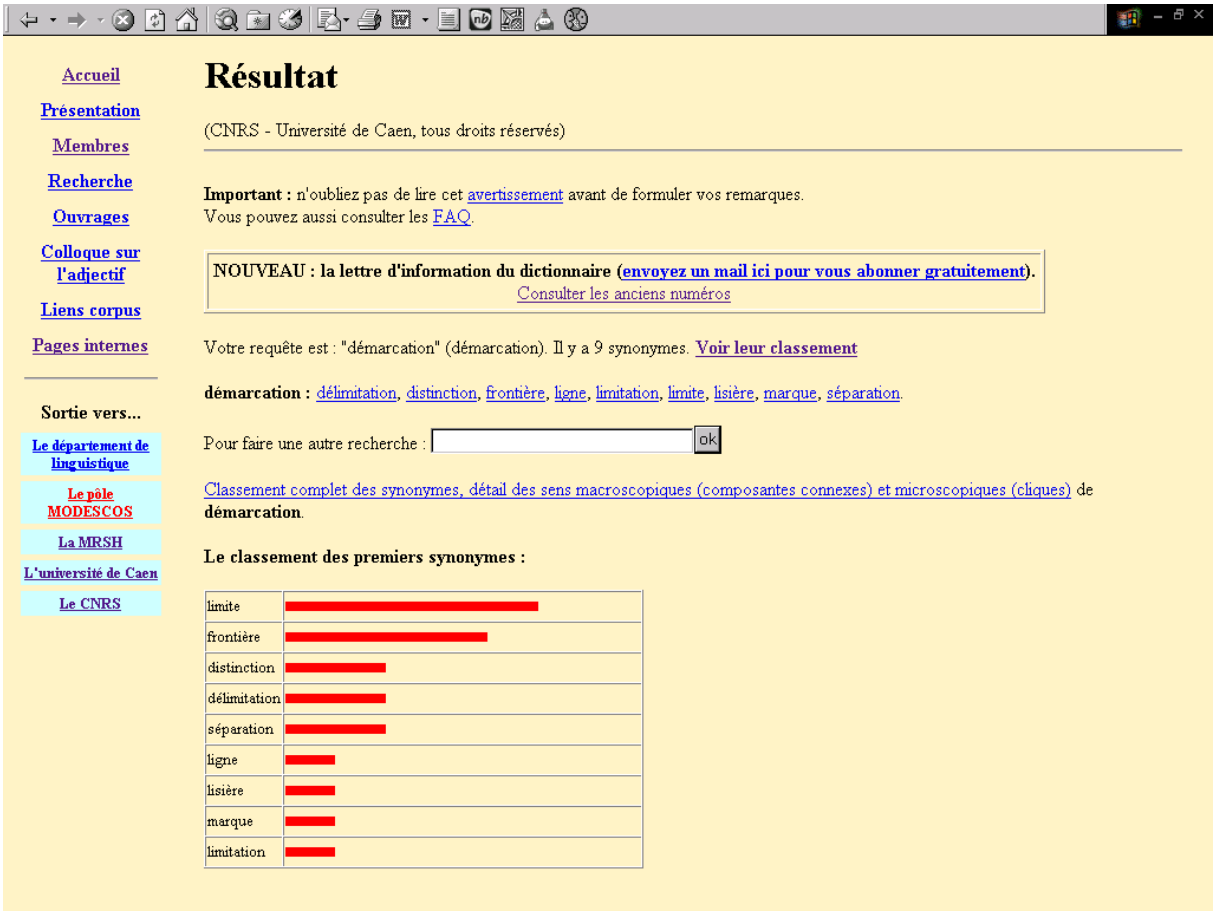


Figure 2 : exemple de réponse du dictionnaire des synonymes.

Comme nous l'avons dit, l'interface de notre dictionnaire adopte l'hypernavigation comme moyen de déplacement dans le dictionnaire ; à la différence d'un ouvrage papier où il faut tourner les pages pour passer de synonyme en synonyme, il suffit ici de cliquer sur un terme de la liste pour obtenir ses synonymes. On peut ainsi effectuer une "traversée" du dictionnaire en seulement sept clics de souris (ce qui ne veut pas dire que l'on a passé en revue les 49000 termes durant ce parcours).

Notre dictionnaire se présente donc comme un outil aux réponses relativement abruptes (puisque tous les commentaires ont disparu), mais cette sécheresse apparente est compensée par le nombre élevé d'entrées, et par la grande rapidité d'obtention des résultats. Il faut toutefois ajouter que la gratuité de sa consultation est certainement une des raisons majeures de son succès.

2. Impact du dictionnaire

Depuis l'ouverture du site en octobre 1998, le dictionnaire a reçu plus de 10 millions de requêtes, dont plus de 6 millions dans les 12 derniers mois ; la fréquentation est toujours en progression, et la croissance du trafic se maintient aux environs de 100 % par an. La figure 3 illustre l'augmentation du nombre de requêtes durant les deux dernières années.

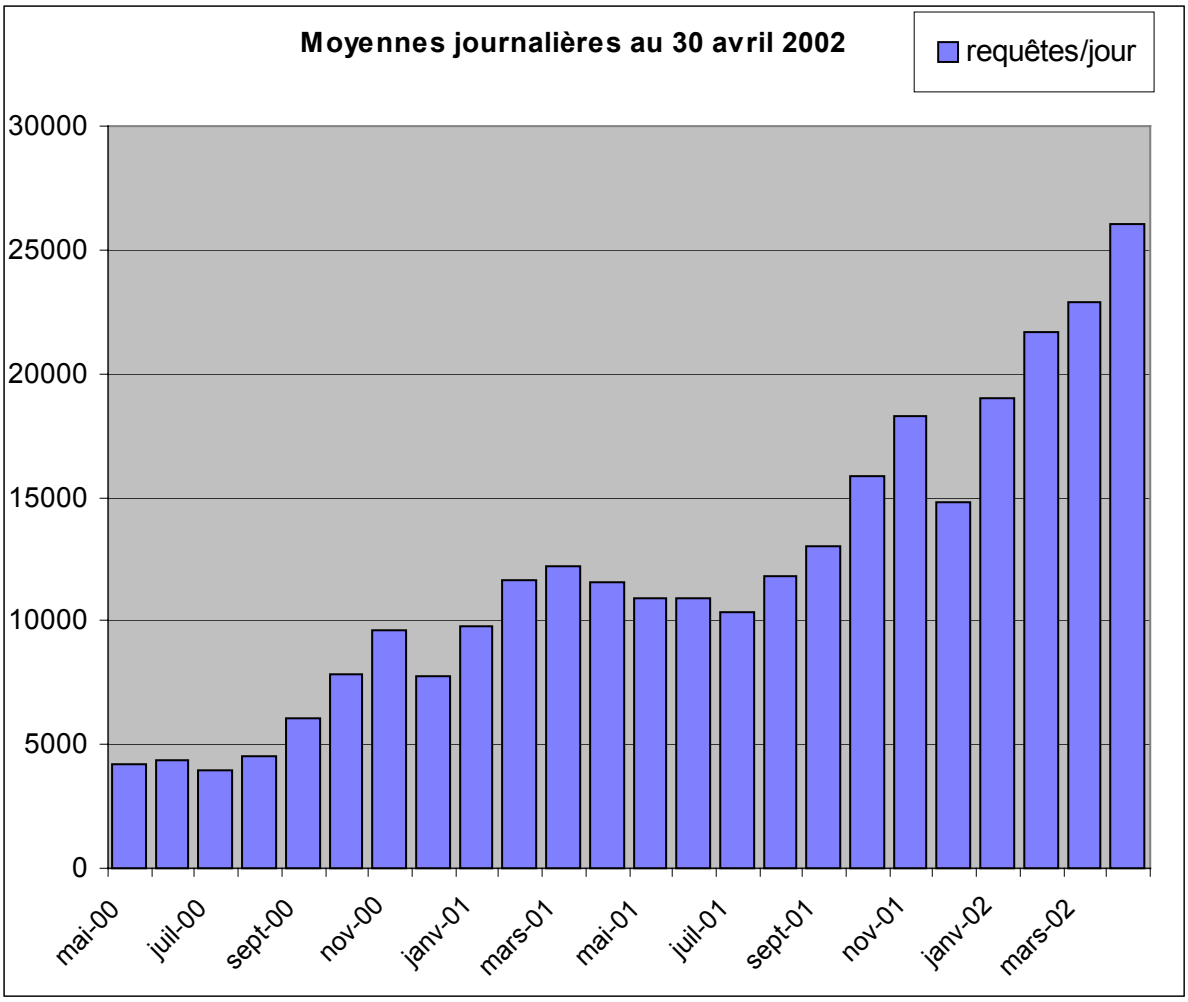


Figure 3 : trafic du dictionnaire dans les 24 derniers mois.

Cela dit, s'il est important de constater que le dictionnaire répond à une forte demande, il est encore plus intéressant d'essayer de connaître la motivation de cette fréquentation ; or, si l'on s'attarde sur la répartition du trafic vers notre ressource en fonction du jour de la semaine, on constate une très nette diminution durant le week-end (figure 4).

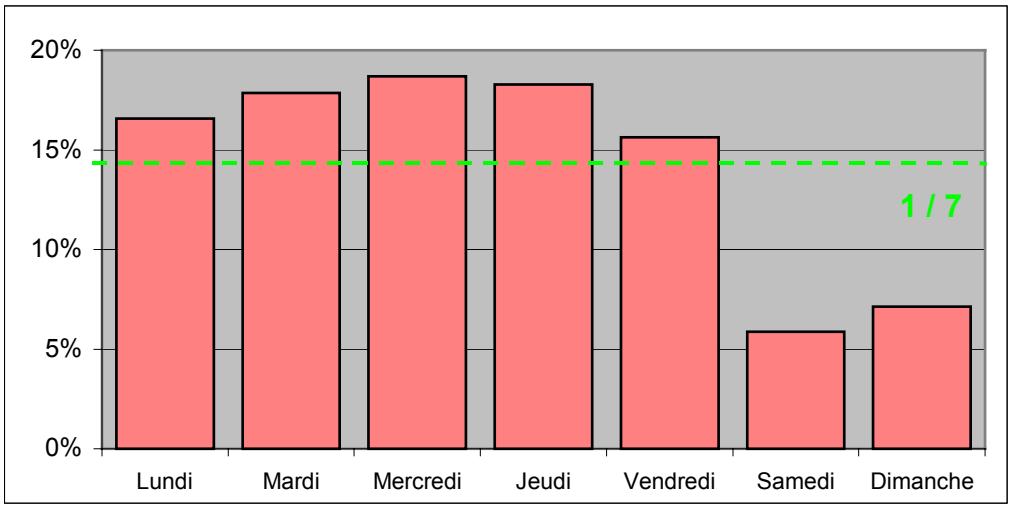


Figure 4 : trafic du dictionnaire suivant le jour de la semaine.

En outre l'histogramme de la fréquentation du dictionnaire suivant locale chez le "client" fait apparaître une courbe quasi-identique à celle de la présence au bureau, avec toutefois un léger prolongement en soirée (figure 5). La comparaison de cette courbe avec celle d'autres sites, soit informatifs, soit ludiques, confirme que notre dictionnaire est essentiellement utilisé à des fins professionnelles.

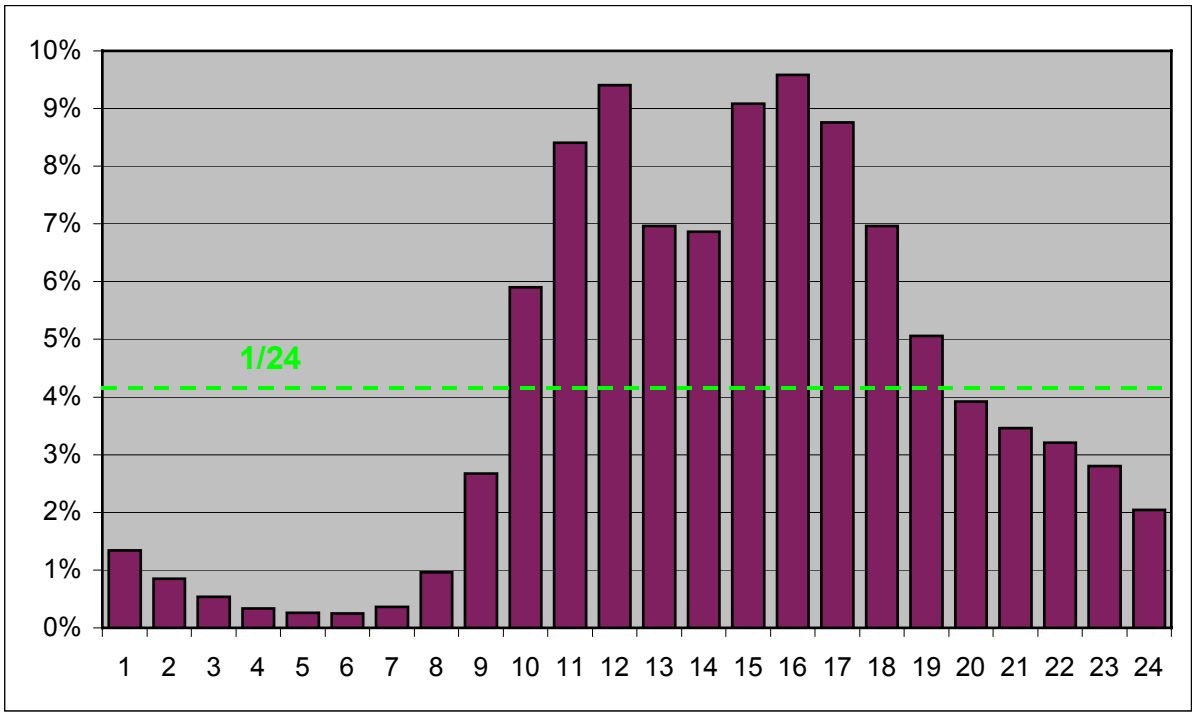


Figure 5 : trafic du dictionnaire suivant l'heure chez le client.

La répartition des appels selon leur origine géographique révèle une très forte majorité francophone (tableau 2), même si 99 % du trafic se répartit suivant 30 pays (ou super-domaines) différents, et si par ailleurs il n'est pas facile de localiser géographiquement les domaines .COM et .NET.

Super-domaine	**Pays ou activité**	**Part de trafic (en %)**
FR	France	30
CA	Canada	25
COM	(Commerce)	17
NET	(Internet)	13
BE	Belgique	4
CH	Suisse	3
EDU	Universités américaines	0,9
DE	Allemagne	0,8
ORG	(Organisations)	0,7
UK	Royaume-Uni	0,7
	Autres	4,9

Tableau 2 : les 10 premiers super-domaines par importance du trafic.

Une analyse plus détaillée des 26 000 domaines qui interrogent notre dictionnaire confirme l'importance de l'usage professionnel de cet outil ; ainsi, le gouvernement québécois, au 7è rang par l'importance de son trafic, représente à lui seul 8 % des appels canadiens ; de même, 10 % du trafic helvétique provient de l'administration fédérale, et ce qui place le super-domaine ORG au neuvième rang, ce sont toutes les requêtes effectuées par l'ONU et l'OECD ; à titre de comparaison, l'ONU totalise autant de requêtes que tout le domaine "Italie".

Notre dictionnaire des synonymes s'avère ainsi constituer un outil de travail pour beaucoup de professionnels de l'écrit ; et si la richesse de son contenu, la rapidité de ses réponses et la gratuité de son accès sont des atouts conséquents, sa présence sur Internet nous permet aussi de connaître les demandes des utilisateurs, et de le faire évoluer vers une plus grande satisfaction des internautes, tout en préservant ses principes de base.

3. Evolution du dictionnaire.

L'intérêt d'un dictionnaire en ligne est double :

- pour l'utilisateur, la consultation d'une telle ressource peut faire appel à différentes techniques qui facilitent la recherche (hypernavigation, corrections, etc.)
- pour l'éditeur de la ressource, les traces laissées par les internautes sont une information du plus haut intérêt ; comme nous allons le voir, ce sont elles qui permettent de déduire le degré de satisfaction des utilisateurs, autrement dit de savoir dans quel(s) cas le dictionnaire répond aux requêtes formulées ; cette connaissance est ensuite le point de départ des enrichissements apportés aux données.

3.1 Les variantes orthographiques.

Dès la mise en ligne du dictionnaire, il est apparu nécessaire de résoudre le problème des variantes orthographiques ; en effet, certains mots ont plusieurs graphies possibles (comme "clé" ou "clef"). Nous avons opté pour la solution du fichier annexe juxtaposé au fichier principal. Certes, nous aurions pu "doubler" les lignes des vedettes présentant cette particularité, mais la solution choisie était plus économique, et s'est avérée par la suite plus souple.

En outre, certains mots n'ont pas seulement deux graphies possibles, mais trois voire quatre (comme "schlinguer", "chlinguer", "chelinguer", etc.), sans parler des mots composés qui peuvent s'écrire avec ou sans trait d'union. Le "fichier des variantes" a donc permis de pallier à cette particularité de toute langue naturelle. En voici un extrait :

chausse-trappe:chausse-trape
chelinguer:schlinguer
chenastre:chenâtre

Lorsque l'internaute tape une variante orthographique dans sa requête (par ex. "chelinguer"), le programme de consultation utilise le fichier des variantes pour rediriger l'interrogation vers le mot associé ("schlinguer"), et obtenir ainsi une réponse.

3.2 Les signes diacritiques.

Il faut jouter à cela que bien des langues naturelles possèdent des signes dits "diacritiques" (accents, cédilles, etc.) qui lui sont souvent spécifiques ; ces signes sont rarement exportés sur

les claviers informatiques d'un autre pays, et par conséquent difficilement accessibles pour beaucoup d'utilisateurs.

De plus, un francophone qui écrit sa langue maternelle a souvent tendance à omettre certains accents (en particulier l'accent circonflexe) par distraction ou par paresse ; outre cela, la langue française contient certains mots dont l'orthographe officielle contredit la prononciation (comme "interpeller", si on le compare à "appeler" ou "rappeler").

Pour éviter que notre dictionnaire ne pénalise sans raison les non-francophones privés de clavier adéquat, ou les francophones hésitant sur l'emploi de certains accents, nous avons élargi le concept du "fichier des variantes" en y incluant les entrées du dictionnaire qui contiennent des signes diacritiques, mais en les débarrassant de ceux-ci :

chaumiere:chaumière
chausse:chausse,chaussé
chausse-trappe:chausse-trape
chaussee:chaussée

A noter que dans le cas de "chausse", l'entrée du fichier des variantes aboutit à deux entrées du dictionnaire ; dans ce cas, la réponse affichée correspond aux synonymes du premier, mais propose à l'utilisateur une autre interprétation possible de sa requête ; il lui suffit alors de cliquer sur un lien pour obtenir les synonymes du second terme. Ci-dessous un autre exemple de ce cas de figure :

tache:tâche,tache,taché
tacher:tacher,tâcher

Cette amélioration de l'interface de consultation permet non seulement de corriger les accents oubliés, mais également ceux qui sont ajoutés ou transformés par erreur (par exemple "à priori" au lieu de "a priori"). Nous avons ainsi une ébauche de correction orthographique des requêtes.

3.3 Les formes fléchies.

Les deux améliorations que nous venons de décrire ne nécessitent pas d'analyser en détail les requêtes formulées par les utilisateurs ; en revanche, la poursuite des améliorations visant à satisfaire au mieux les besoins des internautes oblige à regarder quelles ont été les demandes ; l'analyse des 10 millions de requêtes par un programme automatique fait apparaître un "lexique" de plus de 500 000 mots. Cependant, seulement 68 000 ont été demandés plus de 10 fois (ce qui représente 91 % des requêtes) et parmi ceux-ci, 17 000 l'ont été plus de 100 fois (74 % des requêtes).

A l'intérieur de ces lexiques, il est facile de savoir quels sont les mots absents du dictionnaire ; par exemple, pour les 17 000 mots demandés plus de 100 fois, environ 600 ne font pas partie de notre ressource ; cette absence s'explique par une des trois raisons suivantes :

- le mot est une forme fléchie d'un mot du dictionnaire.
- le mot contient une faute d'orthographe.
- le mot est réellement absent du dictionnaire.

Dans les 600 mots inconnus évoqués précédemment, chacune des trois possibilités représente environ un tiers de l'effectif.

Les formes fléchies rencontrées sont généralement des adjectifs au féminin, quelques noms au pluriel, et parfois des formes conjuguées de verbes. Pour permettre à notre dictionnaire de les prendre en compte, nous avons choisi de faire à nouveau appel au "fichiers des variantes", la forme fléchie à ajouter devenant une nouvelle entrée de ce fichier :

intelligente:intelligent

La redirection de la requête s'effectue donc vers l'adjectif masculin, le nom au singulier ou encore l'infinitif du verbe.

3.4 La correction orthographique.

Le deuxième groupe des mots inconnus est constitué, comme nous l'avons dit, des requêtes qui contiennent une faute d'orthographe. Il faut cependant nuancer cet avis, et les "fautes" en question peuvent se ranger en trois catégories :

- Les "vraies" fautes, comme "mysogine", ou "sinonyme" ou encore "notament", qui sont faciles à interpréter, et à corriger.
- Les graphies désuètes, qui ne sont plus admises, mais qui ont été utilisées autrefois ; ainsi le mot "remords", parfois écrit sans le "s" final, comme dans la correspondance de George Sand (source : base Frantext) :

  "Et pourtant il y a des gens qui me traitent de fanatique, de communiste et de romanesque, parce que je laisse quelquefois percer un ***remord*** *ou un regret de mon peu de vertu. J'ai ici dans ce moment mon ami Delacroix qui est un charmant et excellent homme,"*

  (**SAND George** / CORRESPONDANCE 1844 / 1844 page 491 / *2871 à CHARLES-ARISTIDE PERROTIN*)

- Les mots douteux, pour lesquels il est difficile de savoir ce qu'a voulu demander l'utilisateur ; par exemple, "soutient" offre deux interprétations possibles. La première suppose qu'il n'y a pas de faute, et que l'internaute a tapé une forme conjuguée de "soutenir" ; la deuxième, en revanche, suggère qu'il a commis une faute en demandant le mot "soutien".

Quoi qu'il en soit, le traitement de toutes ces demandes passe une fois encore par des entrées supplémentaires au fichier des variantes, mais cette fois avec une indication qui précise au programme gérant les requêtes qu'une correction orthographique a été effectuée. Par exemple, si le fichier des variantes contient la ligne suivante :

tranquilite:tranquillité!CORR

le résultat d'une requête avec le mot "tranquilité" contiendra la ligne ci-dessous :

Votre requête est : "tranquilité" (CORRECTION : tranquillité).

L'utilisateur est ainsi informé de son erreur et de la correction qui en a été faite. Le système corrige ainsi les erreurs les plus couramment commises, sans passer par un programme heuristique qui chercherait le mot le plus probable parmi les entrées du dictionnaire et qui risquerait de ralentir la consultation. La rapidité de réponse est en effet primordiale sur Internet.

3.5 Les mots manquants.

Enfin, la dernière catégorie de mots inconnus correspond à des mots existant dans la langue française, mais absents du dictionnaire ; cette absence s'explique de deux manières :

- Le mot demandé fait sans doute partie du vocabulaire terminologique ; autrement dit, il s'emploie dans un contexte technique très particulier, comme "oxymore", qui est une figure de rhétorique et rien d'autre ; de tels mots n'ont pas de synonymes, puisqu'ils désignent sans ambiguïté quelque chose de très spécifique. Cette idée peut s'étendre à des mots courants ; par exemple "foie" désigne un organe que tout le monde connaît, mais qui réfère toujours au même objet, sans aucun autre sens figuré, sauf dans l'expression "avoir les foies" (avoir peur, manquer de courage) ; mais cela n'implique pas que "foie" et "peur" soient substituables. Il est impossible de faire entrer ce type de mots dans un dictionnaire des synonymes, puisque par essence, ils n'ont pas de synonymes.
- Ou bien le mot demandé correspond à une lacune réelle de notre dictionnaire ; pour la combler, il faut cette fois enrichir notre base de données, et réaliser un travail lexicographique. De nouvelles entrées sont ainsi périodiquement rajoutées, choisies en prenant dans l'ordre les absences les plus flagrantes ; par exemple, nous avons récemment inscrit "éradiquer", "extrapoler", "évolutif", "look", "sur mesure", "informel", "ubuesque", etc.

En conclusion de cette partie, nous voyons bien qu'Internet nous a permis d'évaluer notre outil en terme de satisfaction des utilisateurs, et ceci relativement rapidement et de manière commode, si l'on songe à ce qu'il faudrait mettre en œuvre pour faire la même évaluation d'un dictionnaire publié sous forme de livre. Cette satisfaction, si on la mesure en considérant qu'une requête "satisfaite" est celle qui reçoit une réponse de la part du dictionnaire, donne le résultat suivant : les mots demandés plus de 100 fois représentent au total 7,2 millions de requêtes, soit 74 % du trafic ; sur ce total, environ 83 000 requêtes n'ont pas reçu de réponse, donc l'indice de satisfaction est, pour cette partie du trafic, de 98,8 % .

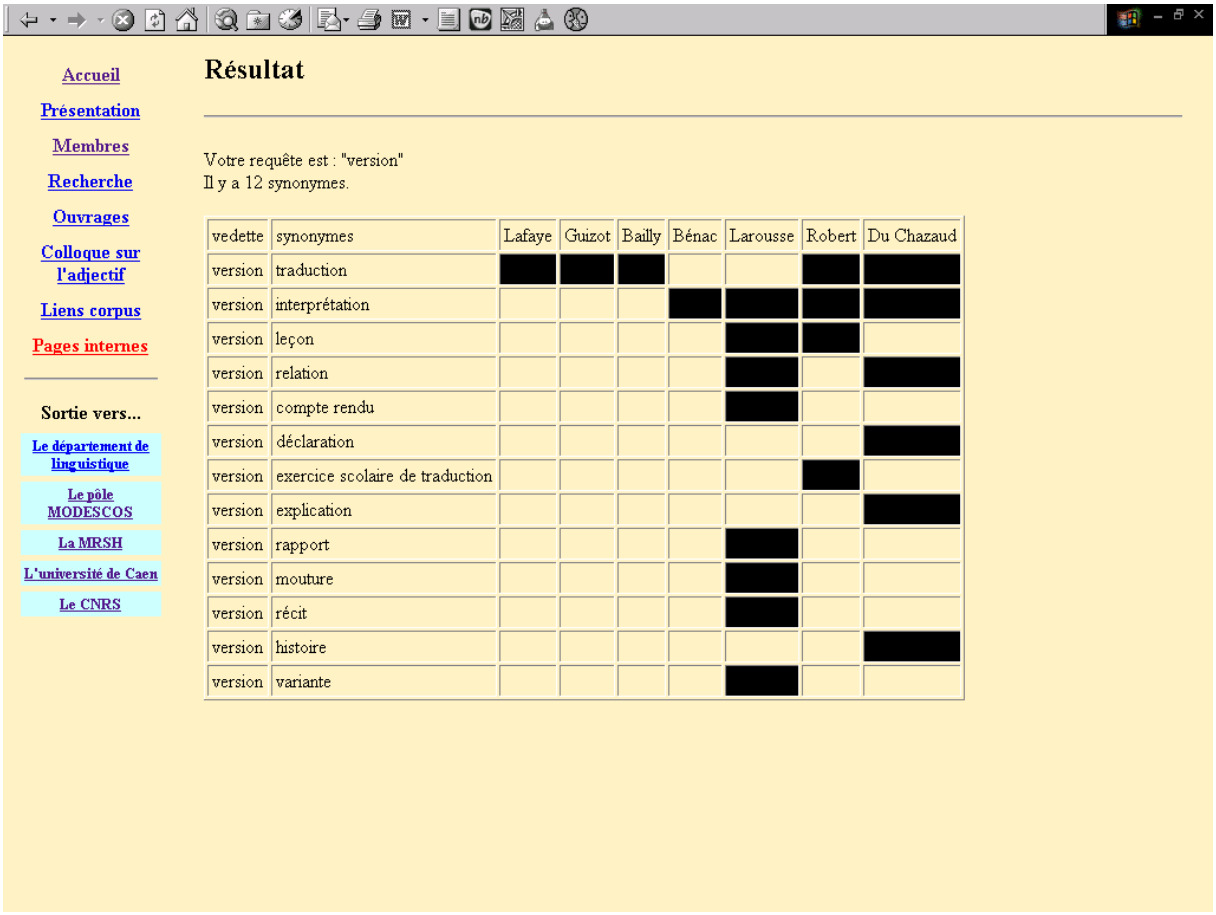
Pour les utilisateurs, l'avantage d'une ressource en ligne devient évident : si l'évaluation peut se faire rapidement, l'évolution qui en découle profite aux internautes, pourvu qu'elle se réalise dans un délai assez bref. La ressource en ligne obtient ainsi un avantage sur la ressource papier, puisque l'utilisateur n'a plus à attendre ni à se procurer une nouvelle édition.

4. Interaction avec la recherche.

En dehors de son utilité pour rechercher des synonymes, notre ressource lexicale recèle un grand nombre d'informations qui peuvent s'examiner sous différents angles. Le dictionnaire devient ainsi un outil pour les chercheurs, en linguistique générale, en traitement automatique des langues, et même en mathématiques appliquées.

4.1 Les comparaisons de dictionnaires.

Tout d'abord, la fusion des données des dictionnaires d'origine s'est faite en conservant l'origine des relations mises en commun ; au moyen d'un programme élaboré au laboratoire, mais réservé à l'usage interne, il est ainsi possible de faire des comparaisons entre dictionnaires pour un mot donné. Le programme délivre un tableau où les synonymes du mot demandé sont classés soit par ordre alphabétique, soit par le nombre de dictionnaires qui les mettent en relation avec la vedette (figure 6).



vedette	synonymes	Lafaye	Guizot	Bailly	Bénac	Larousse	Robert	Du Chazaud
version	traduction	■	■	■			■	■
version	interprétation				■	■	■	■
version	leçon					■	■	
version	relation					■		■
version	compte rendu					■		
version	déclaration							■
version	exercice scolaire de traduction						■	
version	explication							■
version	rapport					■		
version	mouture					■		
version	récit					■		
version	histoire							■
version	variante					■		

Figure 6 : comparaison des dictionnaires pour le mot "version".

4.2 La modélisation en sémantique.

Le travail de modélisation repose tout d'abord sur le fait de considérer le dictionnaire des synonymes comme un graphe ; ce graphe possède 49000 sommets (les entrées du dictionnaire) et environ 200 000 arêtes (les relations de synonymie). Ce point de vue permet d'effectuer des analyses à l'aide d'objets issus de la théorie des graphes (voir Kahlmann, 1975).

Le premier objet que nous utilisons est le sous-graphe, le plus souvent celui dont les sommets sont un mot-vedette et ses synonymes, et dont les arêtes sont les relations qui existent entre ces sommets sélectionnés. La restriction ainsi effectuée nous permet d'étudier un mot en particulier, et d'aboutir à une représentation de son champ sémasiologique, autrement dit de déplier l'éventail de ses significations (voir Fuchs et Victorri, 1996).

Dans le cadre de ce sous-graphe, nous allons en premier lieu chercher des structures de cohésion, et parmi celles-ci, deux sortes qui représentent les extrêmes :

- Celles où il existe un chemin de longueur quelconque entre toute paire de sommets (cohésion lâche).
- Celles où il existe une arête (c'est-à-dire un chemin de longueur 1) entre chaque paire de sommets (cohésion forte).

Pour illustrer notre propos, voici une représentation simplifiée du graphe des relations synonymiques qui accompagnent le mot "baie" (figure 7).

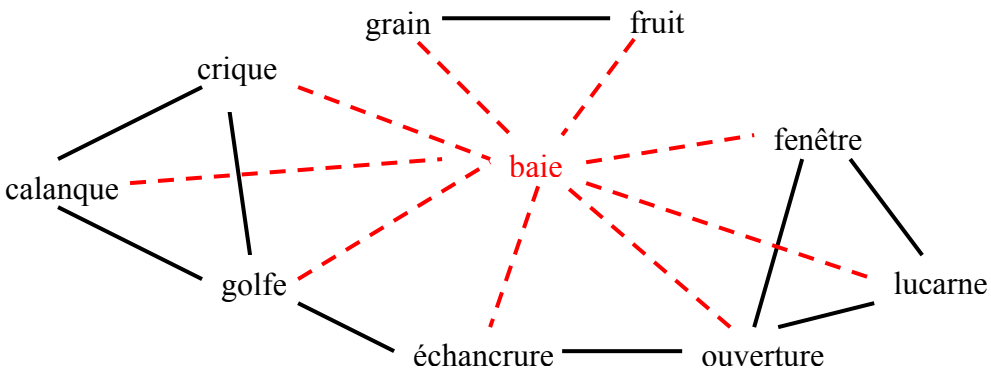


Figure 7 : exemple de graphe de synonymie (simplifié)

4.2.1 La cohésion lâche.

Dans la théorie des graphes, les groupes qui possèdent cette propriété sont appelés "composantes connexes" (voir Berge, 1967). Dans l'exemple de la figure 7, il est facile de constater qu'il n'y a qu'une seule composante connexe ; cela s'explique par le fait que la vedette est en relation avec tous ses synonymes. Aussi, lorsque nous recherchons les composantes connexes dans le graphe lié à un mot, nous ne conservons que les relations qui n'impliquent pas le mot-vedette (dans notre exemple, celles qui sont en traits pleins).

On voit alors apparaître deux composantes connexes : (grain, fruit) et (calanque, crique, golfe, échancrure, ouverture, fenêtre, lucarne). Elles dissocient "baie" en deux homonymes, l'un qui concerne la baie végétale, l'autre relatif à l'idée d'une "échancrure" ou d'une "ouverture", que celle-ci soit dans une côte ("golfe") ou dans un objet architectural ("fenêtre") ; bien entendu, il n'y a pas de relation entre "golfe" et "fenêtre", et si ces deux termes sont regroupés, c'est parce que le mot "échancrure" joue le rôle d'un pivot. Dans la réalité, les composantes connexes obtenues pour "baie" sont les suivantes :

1. *abri, anse, calanque, conche, crique, croisée, enfoncement, fenêtre, golfe, havre, lucarne, ouverture, rade, vide, vue, échancrure*
2. *akène, fruit, grain, graine*

Un autre exemple, avec le mot "pompe" :

1. *accompagnement, affectation, apparat, appareil, bouffissure, cérémonie, emphase, fanfare, faste, fumée, grandeur, grandiloquence, gravité, luxe, magnificence, majesté, phraséologie, rhétorique, richesse, solennité, somptuosité, splendeur, vanité, éclat*
2. *chaussure, godillot, péniche, soulier*
3. *poste d'essence, station-service*
4. *traction*

On le voit, l'analyse du graphe selon ses composantes connexes permet de dégager les homonymes d'un mot vedette ; il faut toutefois signaler que cette méthode se heurte parfois à l'homonymie de certains synonymes ; par exemple, la composante connexe principale, extraite à partir du graphe de synonymie du mot "tour", contient des mots qui vont de "fortification" à "méchanceté", mêlant ainsi deux homonymes de "tour" totalement différents,

LE (vilain) "tour" et LA "tour". La cause de ce phénomène réside dans le mot "vacherie", qui signifie non seulement "méchanceté", mais aussi une "tour" pour abriter les vaches !

Cette imperfection est l'une des raisons qui nous ont conduits à rechercher des groupes à la cohésion plus forte, à savoir ceux dans lesquels tous les sommets sont en relation directe les uns avec les autres.

4.2.1 La cohésion forte.

Les groupes de forte cohésion, pour lesquels chaque sommet est en relation directe avec tous les autres, se nomment "cliques" dans la théorie des graphes ; dans notre exemple simplifié de la figure 7, on trouve 5 cliques : (grain, fruit), (échancrure, golfe), (échancrure, ouverture), (golfe, crique, calanque) et (ouverture, fenêtre, lucarne) ; il n'y a pas de clique (échancrure, golfe, ouverture) car il n'y a pas de liaison entre "golfe" et "ouverture". On voit ainsi que les cliques permettent une meilleure dissociation des sens ; d'autre part, nous pouvons ici considérer les liaisons avec le mot-vedette, car cela ne change absolument pas le nombre de cliques, mais rajoute seulement la vedette dans chacune d'entre elles. Dans la réalité, les cliques qui correspondent aux deux composantes connexes de "baie" sont les suivantes :

1. *abri, anse, baie, golfe*
2. *abri, baie, enfoncement*
3. *abri, baie, golfe, havre*
4. *abri, baie, rade*
5. *akène, baie, fruit*
6. *anse, baie, calanque, crique, golfe*
7. *baie, conche, golfe*
8. *baie, croisée, fenêtre, ouverture*
9. *baie, fenêtre, lucarne, ouverture*
10. *baie, fenêtre, ouverture, vue*
11. *baie, fruit, grain, graine*
12. *baie, golfe, échancrure*
13. *baie, ouverture, vide*
14. *baie, ouverture, échancrure*

Une étude détaillée des cliques, faite pour un ensemble de mots donnés, a permis de montrer que ces ensembles à forte cohésion peuvent être assimilés aux valeurs les plus élémentaires du sens des mots, autrement dit aux nuances les plus fines que nous pouvons appréhender via notre dictionnaire.

Mais l'analyse ne s'arrête pas là, et cet ensemble de cliques va subir une transformation qui nous permet de représenter de manière géométrique le champ sémasiologique du mot-vedette ; pour ce faire, nous considérons que les cliques sont des points d'un espace multidimensionnel, et que les vecteurs unitaires de cet espace sont les synonymes. Le sens du mot-vedette se trouve ainsi matérialisé sous la forme d'un nuage de points qui se déploie dans cet espace imaginaire. Pour arriver à une représentation visible (c'est-à-dire plane) de ce nuage de points, nous appliquons tout d'abord une pondération aux coordonnées de chaque point, puis nous effectuons sur ces coordonnées pondérées une transformation qui s'appelle "analyse en composantes principales", et dont le but est de fournir les plans selon lesquels il est plus intéressant d'observer notre nuage de points. En général, les deux ou trois premiers plans sont ceux qui permettent de refléter au mieux les inter-distances entre les points (pour plus de détails, voir Ploux & Victorri, 1998).

Pour réaliser ces opérations, Bernard Victorri a développé un logiciel nommé "Visusyn" qui fonctionne sous la plate-forme Matlab, que nous utilisons dans nos recherches communes ; la figure 8 donne un aperçu des visualisations obtenues, toujours avec le mot "baie".

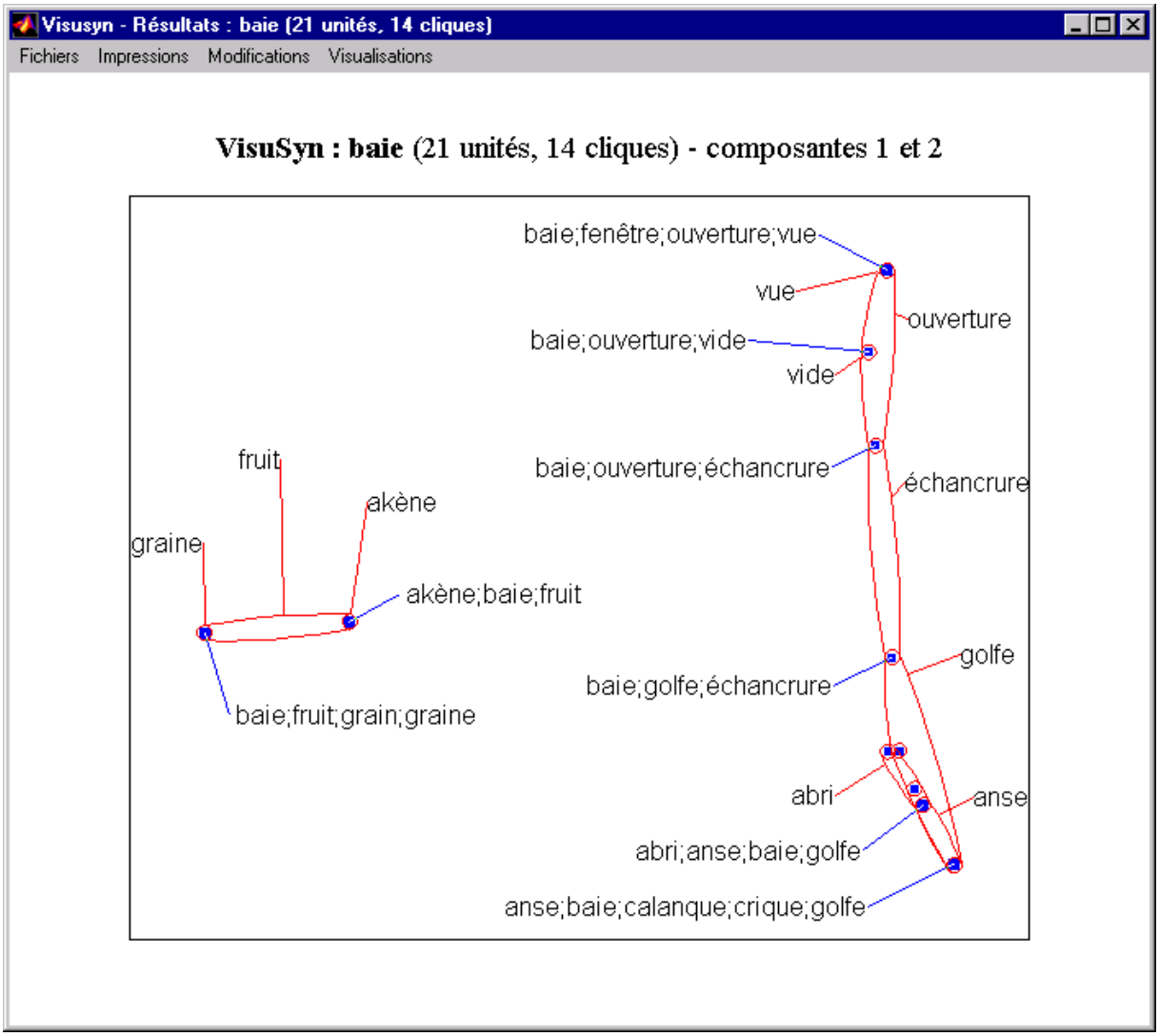


Figure 8 : visualisation du sens par l'analyse des cliques en composantes principales.

Sur la figure, les cliques sont donc des points, et chaque synonyme est représenté par un contour qui englobe les cliques où l'on trouve ce synonyme ; on remarque sur cette représentation la dissociation entre les deux homonymes ("fruit" et "ouverture, échancrure"), ainsi que l'étalement du sens de la composante qui va de "anse" à "fenêtre" dans un continuum où les liaisons sont assurées par les mots "échancrure" et "vide".

4.3 L'analyse de la complexité.

Ce nouveau projet de recherche est très lié au précédent, dans la mesure où le dictionnaire y est également considéré comme un graphe ; de plus, les résultats des études menées permettent aussi de faire de la modélisation en sémantique. La différence réside dans l'objet de la recherche : dans le projet précédent, le but poursuivi était purement sémantique ; ici, par contre, il s'agit de déterminer, voire d'inventer des moyens de mesurer les caractéristiques globales et locales d'un graphe complexe tel que celui du dictionnaire des synonymes.

Il est probable que les caractéristiques locales permettront d'aboutir à d'autres représentations ; un exemple de ces caractéristiques se nomme "indice de similitude", et permet de mesurer la différence des "rôles" joués par deux sommets d'un graphe donné (voir Degenne et Forsé, 1994, et Legendre et Legendre, 1998). En fait, cet indice est égal au nombre de "partenaires" communs aux deux sommets, divisé par le nombre de "partenaires" qu'ils ont à eux deux. Si l'on applique cet indice aux sous-graphes des relations d'un mot-vedette, il est possible d'obtenir une autre représentation du sens, cette fois sous forme arborescente (figure 9).

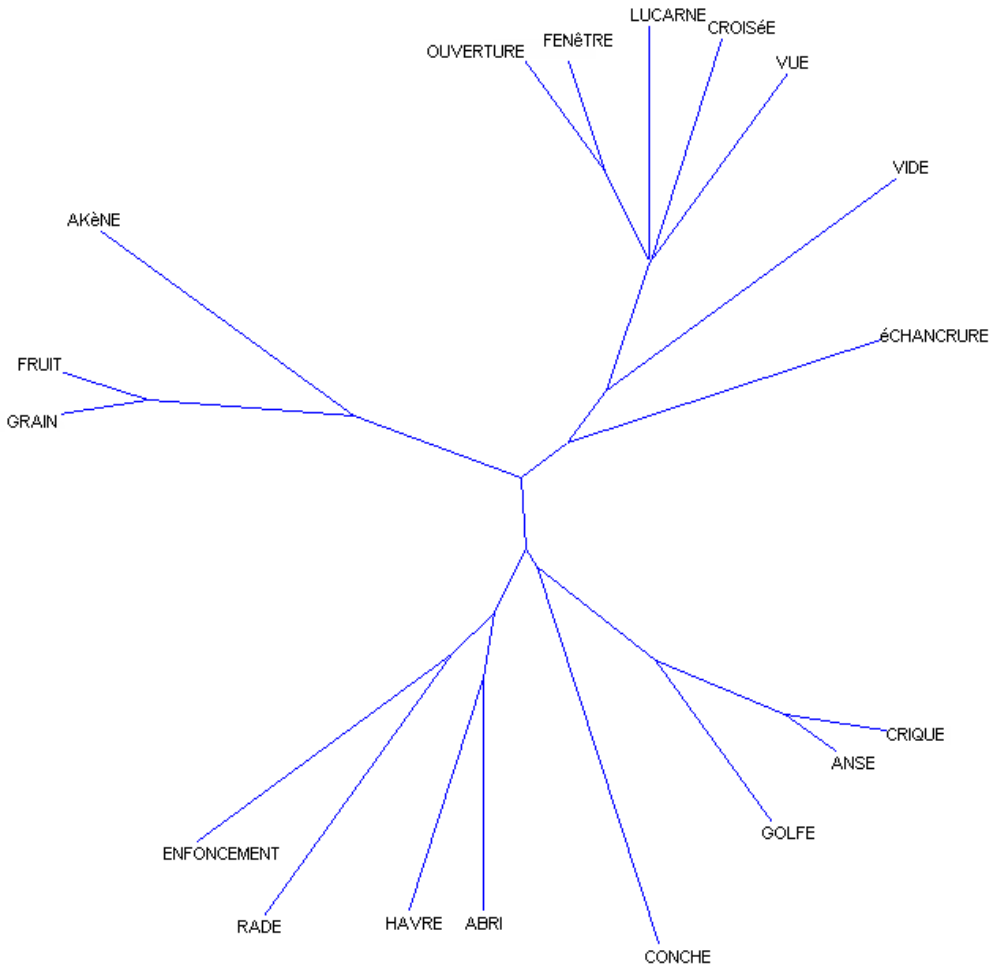


Figure 9 : arborescence du mot "baie", d'après les similitudes entre synonymes.

On peut noter sur cette dernière représentation, que "vide" et "échancrure" occupent une position qui suggère que ces termes lient les deux aspects (maritime et architectural) d'une des composantes du mot-vedette.

Conclusion

Le dictionnaire des synonymes du CRISCO est, comme nous venons de le voir, une ressource particulière à plusieurs égards. En effet, il est plus qu'un dictionnaire électronique ordinaire car il est librement consultable sur Internet, et ce contact externe engendre son évolution et son enrichissement. De plus, son modèle original en fait également un support d'étude pour le domaine dont il est issu (la linguistique), ainsi qu'un objet appartenant à des concepts assez nouveaux des mathématiques appliquées. En outre, toutes les traces laissées

par les internautes lors de leur passage dans notre dictionnaire sont une mine dont nous n'avons commencé à exploiter que certains filons ; il est en effet très probable que les informations dont nous disposons intéresseront des chercheurs en sciences de la communication ou en sociologie.

Enfin, la qualité de son contenu et la renommée qu'il a acquise rend ce dictionnaire susceptible d'engendrer des actions de valorisation ; autrement dit, le modèle de données et son exploitation peuvent aussi intéresser des entreprises qui construisent des outils sémantiques visant à mettre à profit les gigantesques ressources documentaires accessibles en ligne ou à l'intérieur des entreprises.

On voit ainsi qu'un dictionnaire électronique comme le nôtre s'avère jouer un rôle de pivot dans les mutations accomplies ces dernières années par les nouvelles technologies de l'information et de la communication.

Bibliographie

Berge C. (1967), *Théorie des graphes et ses applications*, Paris, Dunod.

Degenne A. & Forsé M.(1994), *Les réseaux sociaux*, Paris, Armand Colin.

François J., Victorri B. & Manguin J.L. (à paraître en 2002), Polysémie adjectivale et synonymie : l'éventail des sens de curieux, actes du colloque *La polysémie*, Paris, novembre 2000, Presses Universitaires de la Sorbonne.

Kahlmann A. (1975), *Traitement automatique d'un dictionnaire de synonymes*, Stockholm, Université de Stockholm.

Legendre P. & Legendre L. (1998), *Numerical Ecology*, Amsterdam, Elsevier.

Manguin J.L. & Victorri B. (1999), Représentation géométrique d'un paradigme lexical, actes de la conférence *TALN 1999*, pp. 363-368.

Ploux S. & Victorri B.(1998), Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *TAL*, Vol 39/1, pp. 161-182.

Victorri B., François J., Manguin J.L.(à paraître), Dynamical construction of meaning in polysemic units, in Willems D. (ed.), *Points of comparison in linguistic theory: from morphology to discourse*.

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