Regroupements de synonymes par indices de similitude : exemple avec l'adjectif "ancien"

Jean-Luc Manguin

▶ **To cite this version:**

Jean-Luc Manguin. Regroupements de synonymes par indices de similitude : exemple avec l'adjectif "ancien". Colloque : Les adjectifs non prédicatifs, Nov 2002, La Plaine St-Denis, France. pp.239-254, 10.15122/isbn.978-2-8124-4338-1.p.0243 . hal-00012327

HAL Id: hal-00012327
https://hal.science/hal-00012327v1

Submitted on 20 Oct 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Regroupements de synonymes par indices de similitude : exemple avec l'adjectif *ancien*.

Jean-Luc MANGUIN

Centre de Recherches Inter-langues sur la Signification en Contexte
CNRS UMR 6170 – Université de Caen
Esplanade de la Paix
14032 Caen Cedex
manguin@crisco.unicaen.fr

0. Introduction

Le but de cet article est, à travers l'exemple de l'adjectif *ancien*, d'exposer une méthode de quantification de la relation synonymique, en vue de regrouper les différents synonymes d'un mot par proximité sémantique. Plusieurs procédés ont déjà été utilisés pour quantifier cette relation avant l'utilisation massive de la micro-informatique, comme par exemple B. BRODDA et H. KARLGREN (1969), ou depuis comme B. GAUME (2002). La technique que nous proposons se base sur l'exploitation des relations synonymiques présentes dans un dictionnaire électronique des synonymes, formalisées en un graphe non-orienté. Elle nous permet de révéler d'une manière quantitative la polysémie d'un élément du lexique, et également de construire des représentations synthétiques du champ sémasiologique de l'item étudié.

L'adjectif *ancien* a été choisi en raison de son caractère « double », c'est-à-dire que son emploi en fonction épithète n'autorise pas toujours la substitution avec une proposition relative dans laquelle il serait attribut, comme dans :

Jeanne a revu hier son ancien mari.
* *Jeanne a revu hier son mari qui est ancien.*

On remarque là un emploi non-prédicatif d'*ancien*, bien distinct de celui que l'on rencontre dans des exemples comme :

Pierre s'exprime souvent avec d'anciennes tournures de phrase.
Pierre s'exprime souvent avec des tournures de phrase qui sont anciennes.

Il est souvent possible de repérer les deux emplois par la position de l'adjectif dans le syntagme nominal, mais ce n'est pas toujours le cas ; en voici un autre exemple, fourni par la base *Frantext* :

> « Ce n'est pas une famille illustre, mais c'est une bonne et très ancienne famille de province » (Marcel PROUST, *Sodome et Gomorrhe*, chapitre II).

Cela dit, notre but n'est pas de réaliser une étude exhaustive de l'adjectif *ancien* au point de vue syntagmatique, mais de montrer comment les données paradigmatiques exploitées mathématiquement peuvent séparer ses différents synonymes en deux groupes ; la substituabilité d'un ou plusieurs termes d'un groupe avec *ancien* pourra nous éclairer sur le type d'emploi rencontré. Parmi les études récentes de ces questions, on pourra se reporter notamment à J. GOES (1999), à M. NOAILLY (1999), au numéro 136 de la revue *Langue Française*, dirigé par Catherine SCHNEDECKER, ainsi qu'à V. LENEPVEU (2001) et M SALLES (2001).

1. La liste proposée par le dictionnaire électronique des synonymes.

Comme nous l'avons dit, le point de départ de notre méthode est la liste des synonymes du mot étudié, fournie par le Dictionnaire Électronique des Synonymes de notre laboratoire[1]. Rappelons que ce dictionnaire a été élaboré à partir de la compilation des relations synonymiques contenues dans sept dictionnaires (GUIZOT, LAFAYE, BAILLY, BÉNAC, DU CHAZAUD, Grand Larousse et Grand Robert) ; pour les détails de sa construction, on pourra se reporter à S. PLOUX et B.VICTORRI (1998) . Ce dictionnaire nous fournit la liste suivante :

> ***ancien*** *: âgé, aïeul, aîné, ancestral, ancêtre, antédiluvien, antérieur, antique, archaïque, ascendant, authentique, briscard, chevronné, croulant, démodé, d'époque, désuet, devancier, doyen, éloigné, ex-, fané, flétri, gothique, haut, immémorial, long, passé, patriarcal, père, périmé, poussiéreux, précédent, précurseur, prédécesseur, préhistorique, premier, primitif, reculé, révolu, rococo, séculaire, suranné, usagé, usé, vénérable, vétéran, vétuste, vieillard, vieillot, vieux.*

Puisque notre étude porte sur l'adjectif ancien, nous pouvons d'ores et déjà supprimer de cette liste les synonymes nominaux, savoir *aïeul*, *ancêtre*, *ascendant*, *briscard*, *devancier*, *doyen*, *père*, *précurseur*, *prédécesseur*, *vétéran* et *vieillard*. Nous supprimerons également *aîné*, qui n'est pas substituable à ancien en tant qu'adjectif, mais en tant que nom. Enfin, nous éliminons trois adjectifs qui ne sont liés qu'à *ancien* et pas à d'autres de ses synonymes, en l'occurrence *authentique*, *chevronné* et *d'époque*. Leur absence de liaison avec d'autres synonymes les empêche en effet de participer à tout regroupement. Il nous reste ainsi la liste ci-dessous :

> ***ancien*** *: âgé, ancestral, antédiluvien, antérieur, antique, archaïque, croulant, démodé, désuet, éloigné, ex-, fané, flétri, gothique, haut, immémorial, long, passé, patriarcal, périmé, poussiéreux, précédent, préhistorique, premier, primitif, reculé, révolu, rococo, séculaire, suranné, usagé, usé, vénérable, vétuste, vieillot, vieux.*

Parmi cette liste, on peut remarquer deux adjectifs assez polysémiques dont la présence demande une explication : *haut* et *long*. Pour *haut*, mentionné par le Grand Robert, la substituabilité avec *ancien* s'applique quand *haut* prend dans le domaine temporel le sens de « près de l'origine », comme dans la citation :

> « les musées américains si bien dotés, ont fourni, pour l'art des hautes époques, un effort qui surpasse tout ce que nous avons pu faire ». Paul MORAND, *New York*, p. 226.

Pour *long*, répertorié à la fois par le Grand Larousse et le Grand Robert, il s'agit ici selon ce dernier d'un sens métaphorique « qui remonte loin dans le temps, dans le passé », comme dans les exemples *une longue histoire* ou *une longue habitude*.

2. Formalisation en graphe de synonymie.

L'étape suivante de notre méthode consiste à formaliser la structure du dictionnaire des synonymes en un graphe dont l'ensemble des sommets est constitué par la vedette et ses

[1] Ce dictionnaire est en libre consultation sur le site http://www.crisco.unicaen.fr .

synonymes ; si deux sommets sont synonymes, ils sont alors reliés par une arête[2]. Nous considérons également que cette relation est réflexive, autrement dit que chaque sommet est en relation avec lui-même. Si cette réflexivité n'apparaît pas explicitement dans le dictionnaire des synonymes (on ne trouve pas la vedette dans la liste des synonymes), elle n'est toutefois pas dépourvue de réalité linguistique[3]. La figure 1 donne un extrait du graphe de synonymie d'*ancien*.

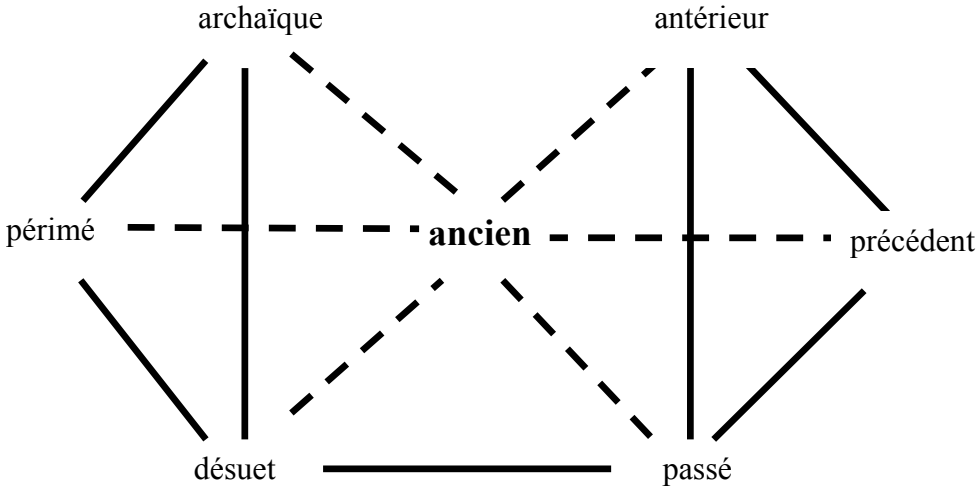

Figure 1 : extrait du graphe de synonymie de l'adjectif *ancien*.

Par la suite, nous étudierons le sous-graphe constitué par la vedette *ancien*, les synonymes que nous avons retenus, et les relations qui existent entre tous ces items. Il peut être intéressant de rechercher dès à présent s'il existe des composantes distinctes à l'intérieur de cet ensemble ; pour cela, il suffit de ne considérer que les relations qui ne passent pas par la vedette, et de chercher les composantes connexes[4] de ce nouveau sous-graphe (par un logiciel d'analyse de graphes) ; dans le cas présent, il n'y a qu'une seule composante connexe, ce qui laisse penser que la dualité de sens donnée dans les exemples reflète plutôt une polysémie d'*ancien* qu'une véritable coupure qui conduirait à un dégroupement homonymique[5].

Il apparaît donc nécessaire que pour parvenir à regrouper les synonymes de manière pertinente, nous devions quantifier précisément la relation synonymique.

3. Les indices de similitude – Similitude relative et absolue.

3.1. *Définition des indices de similitude.*

Il existe plusieurs indices de similitude différents pour les données binaires (ou si l'on préfère, à deux valeurs possibles), mais ils possèdent généralement deux caractéristiques :

---

[2] Pour la terminologie relative aux graphes, on peut consulter C. BERGE (1958).

[3] La symétrie découle de la construction du dictionnaire, mais on pourra se reporter à la thèse d'A. KAHLMANN (1975) où cette question est étudiée de manière détaillée.

[4] Dans un graphe, une composante connexe est un ensemble de sommets où il existe au moins une chaîne (c'est-à-dire une suite d'arêtes) qui relie toute paire de sommets. Autrement dit, il n'y a aucun sommet isolé dans cet ensemble.

[5] Le Trésor de la Langue Française, le Lexis, le Petit Robert et le Grand Robert n'offrent d'ailleurs qu'un seul article pour *ancien*.

- leur valeur est comprise entre 0 et 1.
- pour deux éléments, ils expriment la proportion entre ce qui est commun à ces éléments et ce qui peut appartenir à l'un ou à l'autre.

Dans notre cas, nous utiliserons l'indice de communauté, aussi appelé « indice de Jaccard », défini de la manière suivante : si A et B sont deux sommets du graphe, les relations que les autres sommets (y compris eux-mêmes) entretiennent avec A ou B peuvent se résumer dans le tableau 1, appelé « tableau de contingence ».

		Relation avec A	
		oui	non
Relation avec B	oui	*a*	*b*
	non	*c*	*d*

Tableau 1 : Tableau de contingence de données qualitatives.

L'indice de Jaccard est alors défini selon la formule donnée par LEGENDRE et LEGENDRE (1998) :

$$S = \frac{a}{a+b+c}$$

autrement dit, c'est le nombre de sommets communs à A et B, divisé par le nombre de sommets en relation avec A ou B. Dans l'exemple de la figure 2, l'indice S vaut 4/12 soit 0,333.

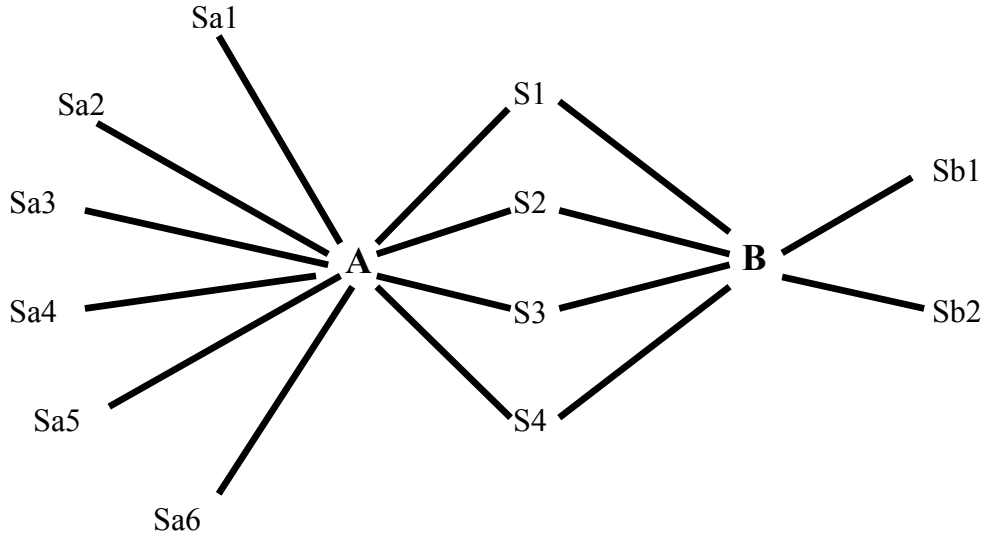

Figure 2 : exemple de graphe pour le calcul de la similitude entre A et B.

Cet indice est symétrique, puisque A et B jouent le même rôle dans la formule de son calcul, et sa valeur est comprise entre 0 et 1 ; elle vaut 0 si A et B n'ont aucun synonyme commun, et 1 si tous les synonymes de l'un des sommets sont aussi synonymes de l'autre.

La figure 2 appelle cependant une remarque si l'on s'aperçoit que A entretient plus de relations « extérieures » à B que communes avec B, tandis que B se comporte de manière inverse. La symétrie de l'indice de similitude ne peut refléter cette discordance, et il devient ainsi opportun de compléter l'indice « standard » symétrique par des mesures qui seront à même de faire apparaître les dissymétries.

3.2. *Similitude relative et absolue*[6].

En nous référant au tableau 1, nous définissons les similitudes S_A (similitude relative à A) et S_B (similitude relative à B) de la manière suivante :

$$S_A = \frac{a}{a+c} \qquad S_B = \frac{a}{a+b}$$

S_A est ainsi égale au nombre de sommets communs, divisé par le nombre de sommets en relation avec A. Dans notre exemple de la figure 2, S_A vaut ainsi 4/10 et S_B 4/6, ce qui montre que la plupart des sens de B sont synonymes de A, et que la réciproque n'est pas vraie.

Nous allons maintenant examiner l'interprétation de ces indices lorsque A est la vedette qui sert de référence à une étude, et B est un de ses synonymes.

3.3. *Application à l'adjectif* ancien.

Le tableau suivant donne les similitudes absolues et relatives les plus élevées pour la vedette *ancien* et ses synonymes (S_A étant la similitude par rapport à la vedette) :

synonyme	S_A	S_B	S
vieux	0,65	0,48	0,38
antique	0,48	0,76	0,42
suranné	0,35	0,58	0,28
démodé	0,35	0,50	0,26
vieillot	0,29	0,71	0,26
passé	0,29	0,25	0,16
vétuste	0,25	0,57	0,21
désuet	0,25	0,46	0,19
ancêtre	0,23	0,63	0,20
…	…	…	…
haut	0,12	0,06	0,04

Tableau 2 : Similitudes entre *ancien* et ses synonymes.

On peut ici remarquer deux types de synonymes totalement différents ; le premier, représenté par *antique*, *vieillot* ou *ancêtre* regroupe les synonymes désambiguïsants de la

[6] Cette construction d'indices non-symétriques se trouve aussi chez M. LAFOURCADE et V. PRINCE (2001), et en traitement de corpus chez D. BOURIGAULT (2002).

vedette, qui en outre n'appartiennent qu'à une seule catégorie grammaticale. Pour ces synonymes, la similitude relative S_B est nettement plus élevée que celle relative à la vedette, ce qui signifie qu'ils sont beaucoup moins polysémiques que celle-ci.

Le second groupe, illustré typiquement par *passé* ou *haut*, contient des synonymes qui ne sont pas désambiguïsants de la vedette ; ils sont polycatégoriels comme elle, et en outre leur champ sémasiologique se développe dans d'autres domaines que celui d'*ancien*.

Il est d'autre part évident que la frontière entre les deux ensembles de synonymes n'est pas délimitée par une valeur fixe : même si 0,5 constitue un seuil pour S_B, il est nécessaire d'apprécier sa proportion avec S_A, et celle-ci dépend de la polysémie de la vedette. Ainsi *désuet* est un synonyme très peu ambigu et monocatégoriel, tandis que *vieux* ne désambiguïse pas l'item *ancien*, même si sa valeur de S_A le classe comme le plus « similaire » à ce mot ; cette forte similitude (0,65) s'explique par une appartenance aux mêmes catégories grammaticales, et à une grande communautés de signifiés[7].

4. Mise en évidence de la polysémie.

Les indices de similitude peuvent aussi révéler de manière indirecte la polysémie d'un élément du lexique ; celle-ci apparaît au vu des valeurs des indices de similitudes entre les synonymes de l'item considéré, et plus précisément à leur répartition dans l'intervalle [0,1]. Si nous étudions une vedette qui possède N synonymes, nous construisons le tableau à (N+1) lignes et (N+1) colonnes contenant les valeurs des S_A (nous nous plaçons en effet dans le domaine de la vedette). Sur l'ensemble des valeurs du tableau, nous calculons ensuite les quartiles[8] q_0 à q_3 (q_4 vaut toujours 1, puisqu'il représente la valeur maximale, et dans notre cas, tout élément est complètement similaire à lui-même). La figure suivante permet de comparer les valeurs de ces quartiles pour les items *ancien*, *haut* et *antique*.

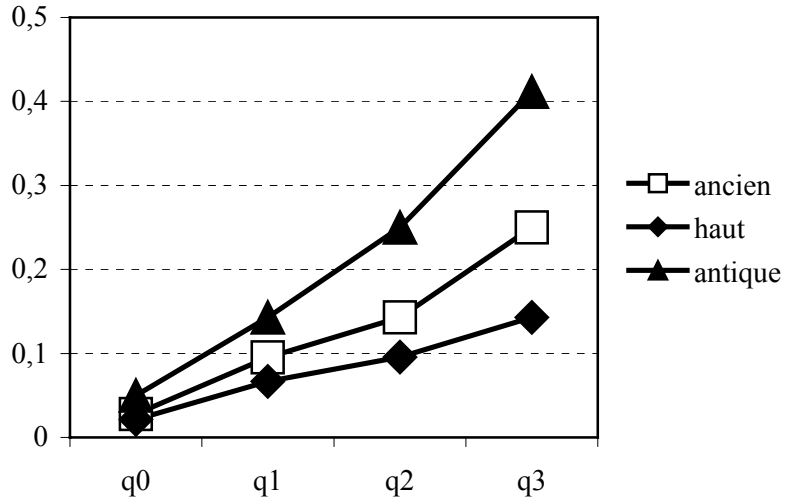

Figure 3 : répartition des indices de similitudes pour *ancien*, *haut* et *antique*.

L'interprétation de cette figure est la suivante : considérons par exemple la médiane q_2, à laquelle 50 % des indices sont inférieurs, et qui vaut 0,1 pour *haut*, 0,15 pour *ancien* et 0,25 pour *antique*. Les différences de valeur de cette médiane signifient que pour *haut*, de

[7] Il faut aussi remarquer que c'est le seul synonyme d'*ancien* mentionné par nos sept dictionnaires sources.
[8] Les quartiles sont des paramètres connus en statistique ; pour leur définition, on pourra se reporter par exemple à A. VESSEREAU (1947).

nombreux synonymes sont faiblement similaires entre eux, à l'inverse de ce qui se passe pour *antique*. Cette faible similarité traduit le fait que les synonymes de *haut* sont rarement synonymes entre eux ; autrement dit, la qualité[9] des relations synonymiques est moins bonne pour *haut* et ses synonymes, que pour *antique* et les siens ; en termes linguistiques, cela signifie que la polysémie de *haut* recouvre des signifiés qui se différencient de manière assez forte (plus forte que dans le cas d'*ancien*, qui est pourtant polycatégoriel lui aussi).

On pourrait toutefois objecter que les valeurs des quartiles présentées soient d'une certaine manière corrélées au nombre de synonymes des items étudiés, puisqu'en effet *haut* possède 94 synonymes, *ancien* 52 et *antique* 33 seulement. Mais il n'en est rien : nous avons par la suite calculé les quartiles pour l'item *ancien* débarrassé des synonymes mentionnés au paragraphe 1, et pour le même item auquel nous avons soustrait de manière aléatoire le même nombre de synonymes. Ainsi, dans la figure qui suit, les indices de similitude dont nous donnons les quartiles ont été calculés sur 37 synonymes ; le calcul effectué pour *ancien* et ses 52 synonymes est donné à titre comparatif.

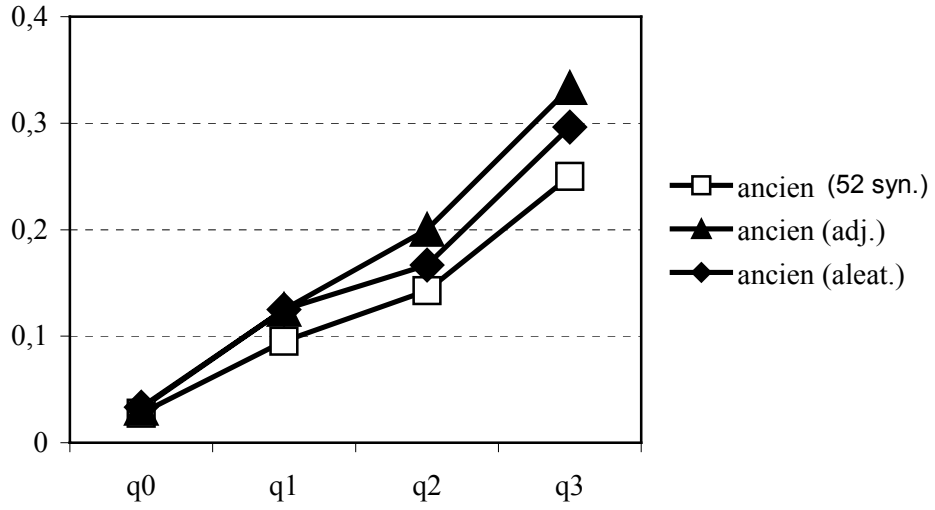

Figure 4 : répartition des indices pour *ancien* avec différentes liste de synonymes.

Il apparaît clairement sur la figure 4 que la suppression aléatoire d'un certain nombre de synonymes ne permet pas d'atteindre la même qualité de relations synonymiques que celle obtenue en effectuant un choix qualitatif des synonymes ôtés ; ainsi, la médiane de la répartition pour la composante adjectivale d'*ancien* est de 20 % supérieure à celle qui contient le même nombre de synonymes sélectionnés au hasard.

La répartition dans l'intervalle [0,1] des indices de similitude entre une unité lexicale et ses synonymes est donc un bon critère d'appréciation de la polysémie de cette unité, pourvu que l'on se réfère à des données paradigmatiques de référence.

5. Les regroupements de sens.

Nous avons annoncé au début de cet article que l'un de nos buts était le regroupement des synonymes dans des ensembles de bonne pertinence sémantique ; nous pouvons désormais l'atteindre, puisque nous avons pu établir une relation quantifiée entre les différents éléments que nous désirons classer. En effet, les indices de similitude sont suffisants pour effectuer une

[9] Qualité, et non densité, car celle-ci, égale au nombre de relations divisé par le nombre de sommets ne révèle rien dans notre cas ; par exemple, la densité pour *ancien* vaut 9,35, pour *antique* 11,55, mais pour *haut* 9,68.

classification hiérarchique ascendante, en considérant que ceux-ci reflètent une similarité et non une distance (autrement dit, la relation d'un élément avec lui-même donne la valeur maximale, tandis que dans le cas d'une distance elle est minimale, et généralement nulle).

Les algorithmes de classification hiérarchique ont fait l'objet d'une littérature assez abondante, c'est pourquoi nous nous bornerons simplement à en rappeler quelques principes élémentaires. L'opération commence toujours par regrouper ensemble les éléments qui sont les plus proches, autrement dit dans notre cas ceux qui présentent la plus forte similitude ; quand un groupe vient d'être créé, les similitudes avec les éléments restants sont calculées par rapport à ce groupe, suivant différents principes. Le principe du lien simple consiste à être le plus tolérant possible, c'est pourquoi la similitude entre un élément restant et le groupe sera égale à la valeur maximale des similitudes entre cet élément et les membres du groupe ; ce principe a pour effet de créer des groupes les plus larges possibles. Au contraire, le principe du lien complet utilise la valeur minimale des similitudes, et provoque de ce fait l'apparition de groupes plus nombreux et disjoints. Enfin, le principe du lien moyen consiste, comme son nom l'indique, à prendre la moyenne des similitudes. On pourra se reporter à l'ouvrage de J.M. BOUROCHE et G. SAPORTA (1980) pour plus de détails.

Nous avons utilisé le logiciel UCINET IV pour effectuer nos classifications hiérarchiques ascendantes, et nous donnons en annexe deux résultats obtenus l'un par lien moyen, et l'autre par lien complet. Nous avons conservé la représentation en histogramme proposé par cet outil, car elle montre clairement le niveau quantitatif des regroupements ou des ruptures. Ainsi, dans le cas de la classification par lien moyen (figure 5), nous remarquons tout de suite une dissociation importante entre le groupe (*précédent*, *antérieur*, *ex-*, *primitif*, *premier*) et le reste des synonymes, qui correspond à la distinction des emplois d'*ancien* dont nous avons parlé dans l'introduction. La seconde rupture importante est celle qui dissocie le groupe principal en une partie dont les termes reflètent un jugement négatif, et une autre où le jugement est neutre voire positif ; la partie « positive » contient les termes *éloigné*, *haut*, *reculé*, *long*, *vénérable*, *patriarcal*, *séculaire*, *immémorial*, *ancestral*. La composante restante recèle des termes plus ou moins péjoratifs ; cette appréciation négative est sans doute moins perceptible dans des mots comme *âgé* ou *passé*, mais devient très nette dans d'autres comme *démodé*, *poussiéreux* ou *croulant*.

Sur la figure 6, nous pouvons observer la dissociation provoquée par l'utilisation de la méthode par lien complet ; en effet, nous retrouvons là pas moins de sept groupes de synonymes, qui affichent des distinctions parfois assez fines ; ces sept groupes sont :

- *éloigné*, *haut*, *reculé*, *long*.
- *vénérable*, *patriarcal*, *séculaire*, *immémorial*, *ancestral*.
- *âgé, vieux, antique, usé, usagé, passé, flétri, fané.*
- *vétuste, périmé, poussiéreux, archaïque, vieillot, suranné, démodé, rococo, désuet, préhistorique, antédiluvien.*
- *révolu, gothique, croulant.*
- *précédent, antérieur, ex-.*
- *primitif, premier.*

Il importe de garder ici à l'esprit que ces groupes sont issus de données paradigmatiques présentes dans les dictionnaires de synonymes, et qu'il serait sans doute vain de chercher à les interpréter comme qualifiant chacun une classe d'entités ; leur forte proximité sémantique n'est généralement que le reflet des considérations de lexicographes, même si dans le cas

présent les plus fortes similitudes (par exemple *suranné* et *démodé*, ou bien *ancestral* et *immémorial*) se rencontrent pour des termes qui ont des emplois syntagmatiques semblables[10].

Pour en terminer avec les regroupements, il faut signaler la grande stabilité des petits groupes à très forte similitude, qui se retrouvent quelle que soit la méthode employée (y compris la méthode par lien simple dont nous n'avons pas fait figurer les résultats) ; ces huit groupes ont des valeurs de similitudes supérieures au seuil « critique » de 0,5 :

- *éloigné, haut.*
- *vieux, antique.*
- *usé, usagé.*
- *périmé, poussiéreux, archaïque.*
- *suranné, démodé, vieillot.*
- *rococo, désuet.*
- *immémorial, ancestral.*
- *précédent, antérieur.*

Par leur stabilité, ces groupes constituent ce que l'on pourrait appeler des « noyaux durs » autour desquels va s'organiser le sens adjectival d'*ancien*.

6. Représentation arborescente.

Nous pouvons aussi construire une représentation arborescente à partir des valeurs des indices de similitudes que nous avons calculées ; ce type de représentation synthétique permet d'apprécier l'organisation des sens d'un élément lexical, et de se rendre compte de la validité du modèle proposé ; en effet, sur ce modèle de représentation, la distance qui sépare deux « feuilles » différentes est proportionnelle à la valeur de leur distance[11] dans le tableau de données qui sert à la construction de l'arbre ; voir pour cela J.P. BARTHÉLEMY et A. GUÉNOCHE (1988).

Sur la figure 7 est représenté l'arbre que nous avons construit avec les 18 synonymes provenant des « noyaux de sens » mentionnés au paragraphe précédent[12] ; nous y observons 6 groupes qui se détachent les uns des autres, et nous pouvons remarquer que celui qui s'éloigne le plus du centre de la représentation est constitué d'*antérieur* et de *précédent*, ce qui montre une fois encore que ce sens « modal[13] » de l'adjectif *ancien* est sémantiquement distinct du sens qualificatif illustré par les autres synonymes.

Il faut signaler toutefois la position excentrée de *haut* qui, comme nous l'avons signalé, est synonyme d'*ancien* dans des contextes particuliers ; ajoutons à cela que dans ce cas, *haut* possède un comportement spécial, perceptible dans l'exemple ci-dessous :

C'est un vase de la haute époque.
** C'est un vase de l'époque qui est haute.*

---

[10] Lors de nos travaux sur l'adjectif *propre*, nous avons constaté une forte similitude entre *apte* et *adéquat*, alors que manifestement ces deux adjectifs ne s'appliquent pas à la même classe de substantifs ; voir J. FRANÇOIS et J.L. MANGUIN (à paraître).

[11] La distance choisie étant égale à 1 moins la similitude.

[12] Nous avons choisi de présenter cette version réduite, car celle qui contient tous les synonymes serait ici d'une lisibilité nettement moins bonne.

[13] Le terme de «modal » convient certes mieux à des adjectifs comme *véritable*, *faux*, *éventuel* ou *possible*, mais dans la mesure où *ancien* permet non seulement de situer de manière temporelle un référent, mais aussi parfois de l'exclure d'un ensemble, comme dans « *un ancien cheminot* », nous l'avons employé ici.

A cette constatation, il faut ajouter que même dans cette acception, *haut* peut accepter la gradation, comme dans cet exemple que nous fournit Frantext :

> « Quelle que soit la bonne interprétation, il est frappant de constater que ce trou de suspension existe dans tous les cachets de ces très hautes époques et dans tous les cylindres-sceaux. » (*L'Histoire et ses méthodes*, p. 396).

Le fait de ne pas trouver d'exemple où *haut* serait cette fois postposé au même substantif nous laisse penser que nous avons affaire ici à un adjectif qui s'inscrit dans le continuum entre adjectifs « modaux » et « qualificatifs » envisagé par M. SALLES (2001).

7. Conclusion et perspectives.

La méthode que nous avons présentée ici permet donc, comme nous l'avions annoncé en introduction, d'aboutir à une quantification de la relation de synonymie ; la validité de cette quantification se juge à la qualité des regroupements de synonymes qu'elle permet d'obtenir. De plus, les représentations arborescentes qui en découlent présentent certainement un intérêt dans l'apprentissage guidé des langues.

Nous avons montré d'autre part que les indices de similitude s'avèrent utiles pour mettre en évidence la polysémie intrinsèque d'un item lexical, ainsi que pour qualifier la désambiguïsation que peut apporter un synonyme en cas de substitution. Cette information peut s'avérer importante dans le cas, par exemple, de la recherche documentaire[14].

Au point de vue dictionnairique, les indices de similitudes sont, pour la recherche et l'examen des synonymes, un outil exploratoire qui donne une profondeur supplémentaire aux dictionnaires de synonymes ; en effet, la définition même de ces indices donne accès directement à une « transitivité » de la relation et peut, dans certains cas, faire émerger parmi les « synonymes d'ordre deux » des unités qui s'avèrent être très proches de l'item de départ. Cette constatation peut éventuellement conduire à une révision de certaines entrées du dictionnaire, si les synonymes qui apparaissent ainsi révèlent une lacune inacceptable.

Au point de vue lexicologique, il est clair que l'étape suivante du travail consiste à intégrer la dimension co-textuelle et à déterminer les moyens à employer pour combiner l'aspect syntagmatique à l'étude paradigmatique que nous avons décrite ici. Nous avons déjà œuvré à cet élargissement et présenté certains résultats encourageants lors de notre intervention de novembre 2002 ; l'espace nous étant malheureusement compté, et l'étude n'ayant été effectuée que sur un seul item (*curieux*) qui n'entre pas dans la classe des adjectifs abordés dans le présent volume, nous nous bornerons à décrire les principes de la méthode.

La première étape consiste à dresser une liste des cooccurrants de l'élément lexical étudié par un relevé sur corpus ; dans l'exemple d'un adjectif, il peut s'agir de relever tous les substantifs dont le mot étudié se trouve être épithète. Ensuite, à partir de cette liste et de celle des synonymes du mot étudié, on cherche toutes les cooccurrences entre un synonyme et un cooccurrant ; le résultat peut se résumer en un tableau rectangulaire, dont les lignes portent les synonymes, et les colonnes les cooccurrants ; à l'intersection d'une ligne et d'une colonne, on trouve le nombre de cooccurrences de l'en-tête de la ligne avec l'en-tête de la colonne. Comme il s'agit de comparer pour chaque synonyme la proportion d'emploi avec les cooccurrants, il est nécessaire de pondérer les données sur chaque ligne. Après cela, nous

[14] Cela dit, il est important dans ce genre d'application de prendre en compte le taux de polysémie présent dans les textes avant de mettre en action des procédés et des ressources parfois coûteux ; voir C. DE LOUPY (2002).

pouvons calculer les indices de similitude entre les différentes lignes[15] (autrement dit les synonymes), et nous obtenons un tableau comparable à celui que nous avons décrit au paragraphe 5.

Nous possédons alors deux tableaux de similitudes que par commodité nous appellerons « similitudes paradigmatiques » pour celles obtenues à partir des relations mentionnées dans les dictionnaires, et « similitudes syntagmatiques » pour celles obtenues par l'analyse de corpus. Dans l'étude que nous avons faite sur *curieux*, nous avons observé que les similitudes syntagmatiques conduisaient par classification à des regroupements de synonymes parfois incohérents, preuve que les données syntagmatiques seules ne permettent pas d'accéder à la totalité de l'information paradigmatique. Mais l'intérêt est de pouvoir combiner les deux tableaux : en faisant le produit (case par case) des deux similitudes, puis en effectuant une classification sur le résultat, nous obtenons des regroupements pertinents ; les différences entre les « similitudes combinées » et les similitudes paradigmatiques s'avèrent être un tremplin pour des études complémentaires. Elles peuvent ainsi révéler des choix lexicographiques dépassés ou contestables dans la confection des dictionnaires, mettre en relief des différences diachroniques (si l'on dispose de corpus de différentes époques), ou encore faire émerger des différences sémantiques suivant la syntaxe des cooccurrences relevées (par exemple suivant l'emploi anté- ou postposé de l'adjectif).

Bibliographie.

BAILLY, René (1946) : *Dictionnaire des synonymes*. Paris, Larousse.

BARTHÉLEMY, Jean-Pierre et Alain GUÉNOCHE (1988) : *Les arbres et les représentations des proximités*. Paris, Masson.

BÉNAC, Henri (1956) : *Dictionnaire des synonymes*. Paris, Hachette.

BERGE, Claude (1958) : *Théorie des graphes et ses applications*. Paris, Dunod.

BERTAUD DU CHAZAUD, Henri (1971) : *Nouveau dictionnaire des synonymes*. Paris, Robert.

BRODDA, Benny et Hans KARLGREN (1969) : « Synonyms and synonyms of synonyms », *SMIL*, 5, p. 3-17. Stockholm.

BORGATTI, S.P., M.G. EVERETT et L.C. FREEMAN (1996) : *UCINET IV version 1.68*. Natick MA, Analytic Technologies.

BOURIGAULT, Didier (2002) : « UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », *Actes de la conférences TALN 2002*, p. 75-84. Nancy.

BOUROCHE, Jean-Marie et Gilbert SAPORTA (1980) : *L'analyse des données*. Paris, P.U.F., collection « Que sais-je ? ».

DE LOUPY, Claude : « Évaluation des taux de synonymie et de polysémie dans un texte », *Actes de la conférences TALN 2002*, p. 225-234. Nancy.

FRANÇOIS, Jacques et Jean-Luc MANGUIN (à paraître) : « La polysémie adjectivale entre synonymie et sélection contextuelle : le cas de *propre* », *Cahiers de l'Institut de Linguistique de Louvain*. Louvain.

[15] Les indices utilisés dans ce cas ne sont plus de même nature, puisque les données sont quantitatives ; toutefois, la transposition de la formule est très simple, comme dans P. LEGENDRE et L. LEGENDRE (1998).

Frantext, base textuelle catégorisée (1999) : CNRS, ATILF (Analyse et traitement informatique de la langue française), UMR CNRS-Université Nancy2, http://www.inalf.fr/atilf .

GAUME, Bruno, Karine DUVIGNAU, Olivier GASQUET et Marie-Dominique GINESTE (2002) : « Forms of meaning, meaning of forms », *Journal of Experimental and Theoretical Artificial Intelligence (JETAI)* , V. 14 N. 1 , p. 61-74. Binghampton, Taylor & Francis.

GOES, Jan (1998) : *L'adjectif entre nom et verbe*. Paris-Bruxelles, Duculot.

GUIZOT, François (1864) : *Dictionnaire Universel des synonymes de la Langue Française*. Paris, Didier (7ème édition).

KAHLMANN, André (1975) : *Traitement automatique d'un dictionnaire de synonymes*. Stockholm.

LAFAYE, Pierre-Benjamin (1858) : *Dictionnaire des synonymes de la Langue Française*. Paris, Hachette.

LAFOURCADE, Mathieu et Violaine PRINCE (2001) : « Synonymies et vecteurs conceptuels », *Actes de la conférence TALN 2001*, p. 233-242. Tours.

Grand Larousse de la Langue Française (1971) : Paris, Larousse.

LEGENDRE, Pierre et Louis LEGENDRE (1998) : *Numerical Ecology*, Amsterdam, Elsevier.

LENEPVEU, Véronique (2002) : « Adjectifs et adverbes : une corrélation syntactico-sémantique », *Le Français Moderne*, 70-1, p.45-70 . Paris, CILF.

Lexis, *Larousse de la langue française* (2002) : sous la dir. de J. DUBOIS. Paris, Larousse.

NOAILLY, Michèle (1999) : *L'adjectif en français*. Gap-Paris, Ophrys.

Le Grand Robert, dictionnaire de la langue française (1985) : sous la dir. d'A. REY. Paris, Dictionnaires Le Robert.

Le Petit Robert, dictionnaire de la langue française (2001) : sous la dir. de J. REY-DEBOVE et A. REY. Paris, Dictionnaires Le Robert.

PLOUX, Sabine et Bernard VICTORRI (1998) : « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *Traitement automatique des langues*, 39-1, p. 161-182. Paris, ATALA.

SALLES, Mathilde (2001) : « Hypothèse d'un continuum entre les adjectifs « modaux » et les adjectifs qualificatifs », *L'Information grammaticale*, 88, p.23-27. Paris.

Langue française, 136, (2002) sous la dir. de C. SCHNEDECKER. Paris, Larousse.

Trésor de la Langue Française informatisé (2001) : CNRS, ATILF (Analyse et traitement informatique de la langue française), UMR CNRS-Université Nancy2, http://www.inalf.fr/tlfi.

VESSEREAU, André (1947) : *La statistique*. Paris, P.U.F., collection « Que sais-je ? ».

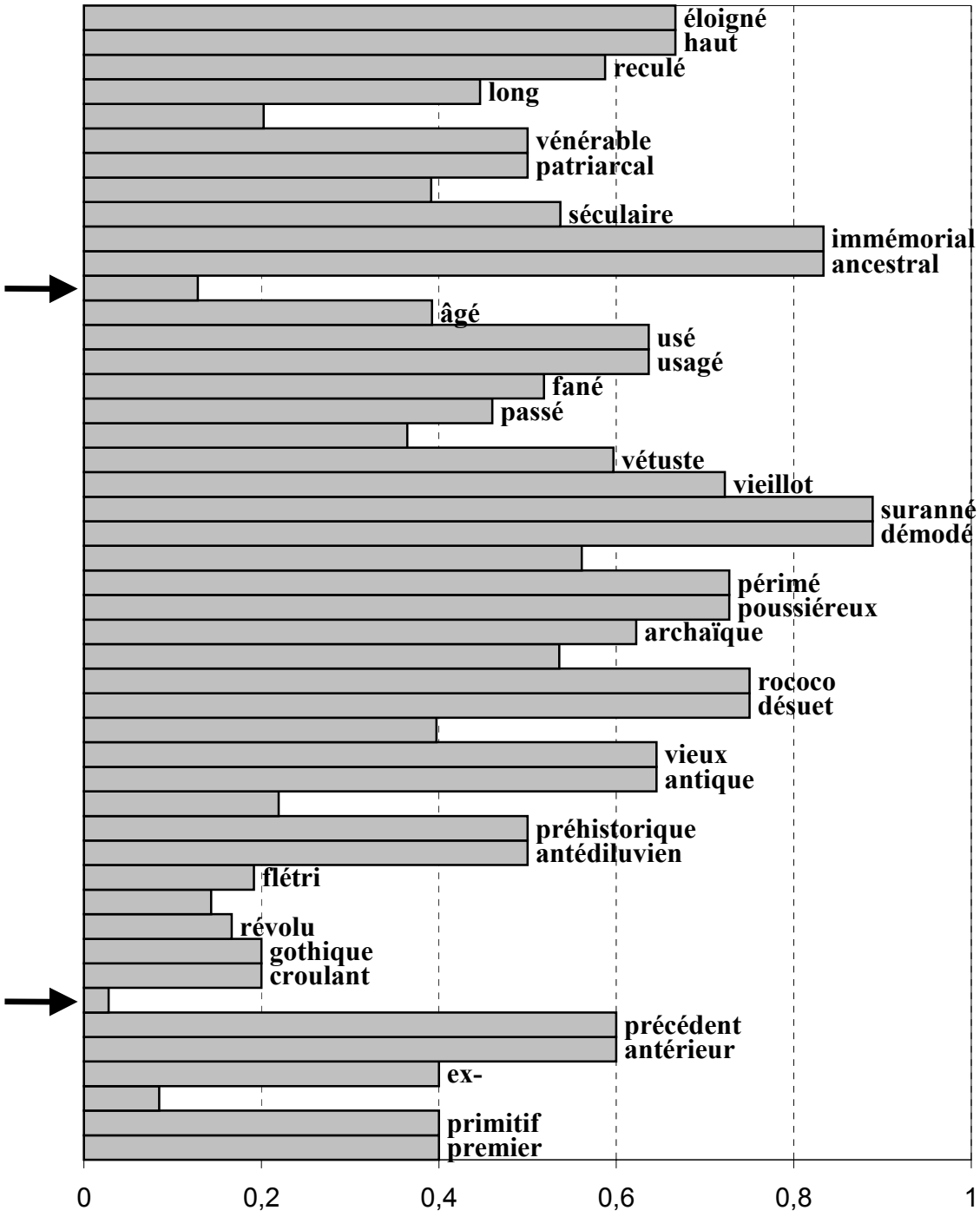

Figure 5 : classification des synonymes par lien moyen.

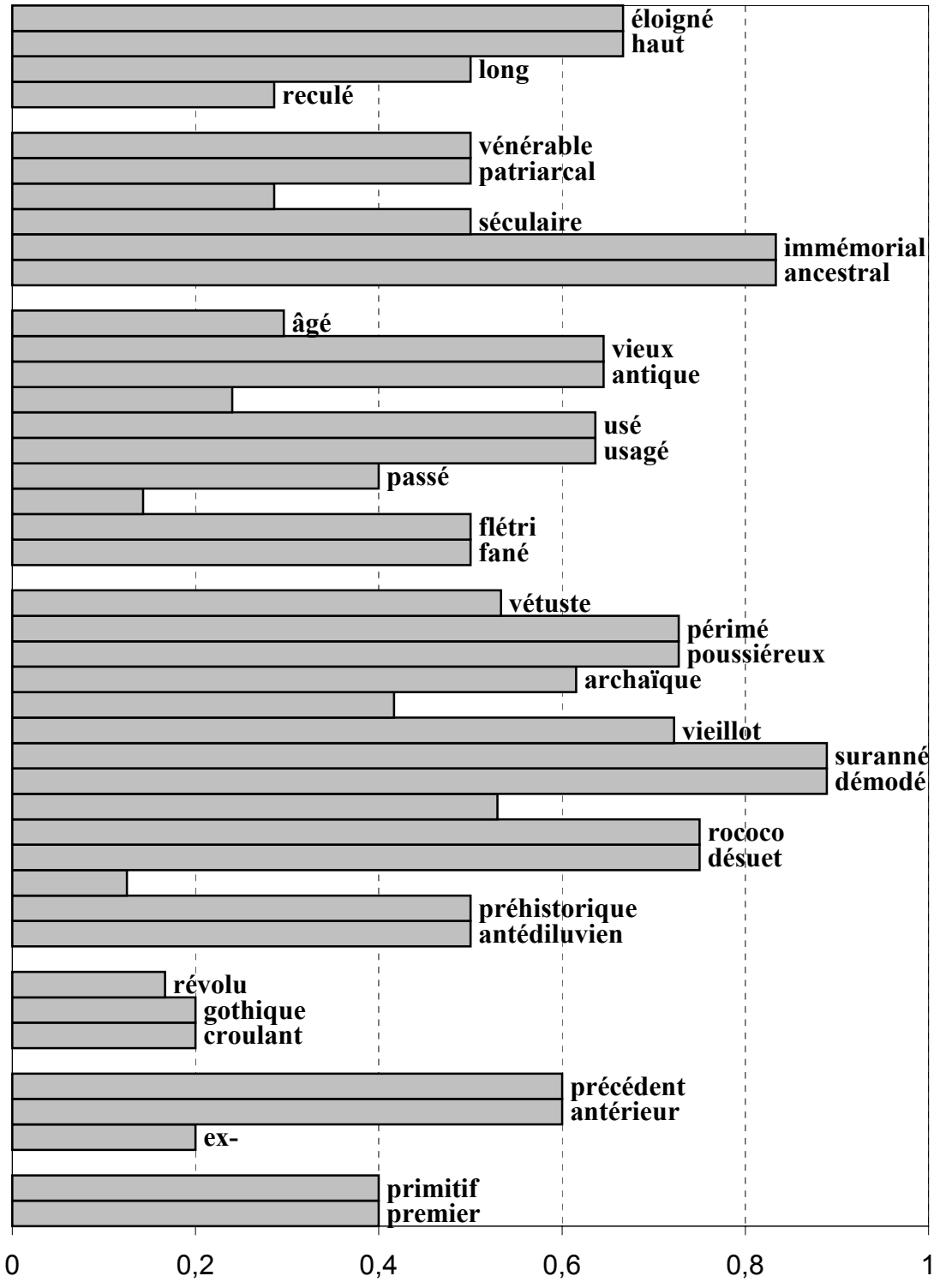

Figure 6 : classification des synonymes par lien complet.

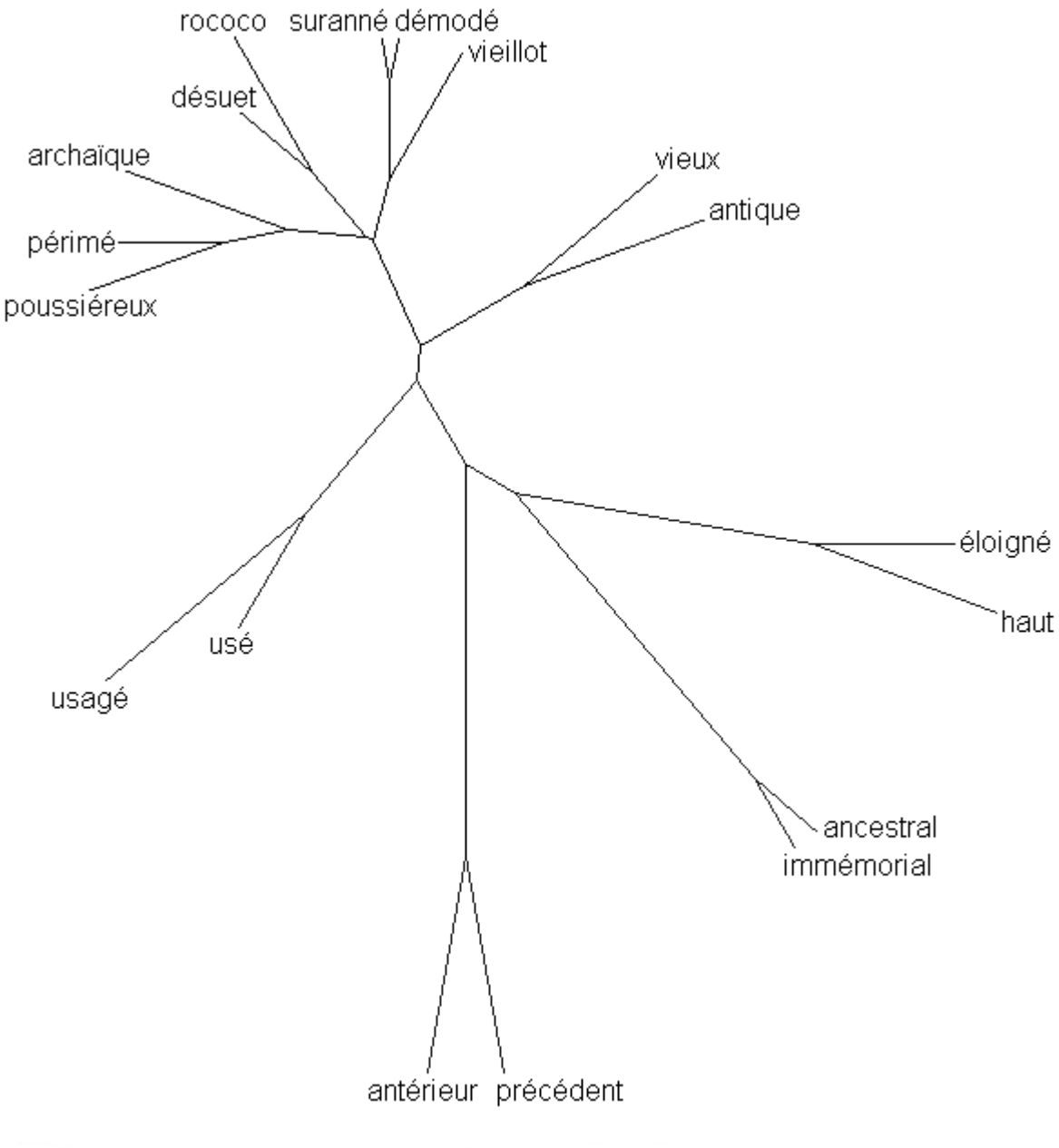

Figure 7 : représentation arborescente des distances sémantiques.