

Approche évolutive des notions de base pour une représentation thématique des connaissances générales

Alain Joubert, Mathieu Lafourcade, Didier Schwab

Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier
{alain.joubert ; lafourcade ; schwab}@lirmm.fr

Résumé

Dans le domaine du Traitement Automatique du Langage Naturel, pour élaborer un système de représentation thématique des connaissances générales, des méthodes s'appuyant sur des thésaurus sont utilisées depuis une quinzaine d'années. Un thésaurus est constitué d'un ensemble de concepts qui définissent un système générateur d'un espace vectoriel modélisant les connaissances générales. Ces concepts, souvent organisés en une hiérarchie arborescente, constituent un instrument fondamental, mais totalement figé. Même si les notions évoluent (nous pensons par exemple aux domaines techniques), un thésaurus ne peut quant à lui être modifié que lors d'un processus particulièrement lourd, car nécessitant la collaboration d'experts humains. C'est à ce problème que nous nous attaquons ici. Après avoir détaillé les caractéristiques que doit posséder un système générateur de l'espace vectoriel de modélisation des connaissances, nous définissons les « notions de base ». Celles-ci, dont la construction s'appuie initialement sur les concepts d'un thésaurus, constituent un autre système générateur de cet espace vectoriel. Nous abordons la détermination des acceptions exprimant les notions de base, ce qui nous amène naturellement à nous poser la question de leur nombre. Enfin, nous explicitons comment, s'affranchissant des concepts du thésaurus, ces notions de base évoluent par un processus itératif au fur et à mesure de l'analyse de nouveaux textes.

Mots clés : thésaurus, vecteurs conceptuels, notions de base, évolutivité.

Abstract

In the field of Natural Language Processing, in order to arrive at a thematic representation system of general knowledge, methods leaning on thesaurus have been widely used for about fifteen years. A thesaurus consists of a set of concepts defining a system generating a vector space to model general knowledge. These concepts, often organized as a hierarchy, constitute a fundamental, but fixed tool. When concepts evolve (as in technical fields), a thesaurus can evolve, but only as a result of an arduous process, requiring the collaboration of human experts. After detailing the desired characteristics of knowledge modelling systems, we define the « basic notions ». Their construction is initially based on the concepts of a thesaurus. They constitute another generating system of this vector space. We discuss the establishment of these basic notions, which naturally leads us to the question of their number. Lastly, we clarify how, being freed from the concepts of the thesaurus, the basic notions evolve progressively in an iterative process as new texts are being analysed.

Keywords: thesaurus, conceptual vectors, basic notions, evolution.

1. Introduction : représentation thématique des connaissances générales

Lafourcade et Sandford (1999) puis Lafourcade (2001), à la suite de Chauché (1990), ont développé un système de représentation thématique¹ des connaissances générales. Celui-ci est

¹ Nous ne considérons donc que les relations thématiques ; les relations « transversales » entre objets n'appartenant pas au même sous arbre de la hiérarchie ne sont pas prises en compte. Pour ne citer qu'un exemple illustrant notre propos, le lien entre *Harry Potter* et sa chouette *Hedwig* ne se fera qu'au travers des notions de *littérature fantastique* et de *sorcellerie*, ainsi que des concepts qui leur sont associés. Ce lien ne sera pas plus fort

fondé sur une représentation vectorielle qui repose initialement sur la hiérarchie de concepts du thésaurus *Larousse* (Larousse, 1999). C'est ce thésaurus qui, initialement, définit l'étendue des connaissances générales, par opposition aux connaissances spécialisées spécifiques à un domaine particulier : nous considérons qu'un terme fait partie des connaissances générales lorsqu'il appartient à l'un des dictionnaires généralistes (ex. : *Larousse*, *Robert*, ...). L'utilisation d'un espace vectoriel pour la modélisation existe depuis longtemps en Recherche d'Informations (cf. par exemple Salton et MacGill 1983). Le thésaurus utilisé possède une structure arborescente, hiérarchisée en 4 niveaux, plus la racine constituée du concept universel ; il comporte 873 concepts feuilles de niveau 4. Ces concepts forment un système générateur de l'espace vectoriel noté C_{873} ; cet espace constitue une modélisation de notre représentation thématique des connaissances générales. Cette approche vectorielle, s'appuyant sur un ensemble de concepts prédéterminé, est celle préconisée par Chauché (1990).

Les concepts définissant le système générateur constituent les éléments fondamentaux de notre représentation des connaissances. Le but d'un système générateur est donc de pouvoir, en un minimum de vecteurs, représenter le maximum de connaissances générales. Les 873 concepts du thésaurus utilisé, donnés *a priori*, constituent-ils le « meilleur » système générateur ? Il est d'autant plus légitime de se poser cette question que d'autres thésaurus généralistes ont été publiés. Le plus ancien est probablement le Roget's Thesaurus (Kipfer, 2001) dont la première édition date du milieu du XIX^e siècle, organisé en une structure arborescente qui compte dans sa version actuelle 1075 concepts feuilles regroupés en 15 classes. Ce thésaurus a servi de base à différents travaux, par exemple Yarowsky (1992).

Les 873 concepts utilisés ne constituent pas une base de l'espace vectoriel C_{873} . En effet, il est à remarquer qu'il est possible de trouver des relations entre certains de ces 873 vecteurs générateurs ; par exemple, entre les concepts 1 *EXISTENCE* et 2 *INEXISTENCE* existe une relation d'antonymie ; même s'il serait naïf de considérer que leurs vecteurs associés sont opposés (Schwab, 2005), il est manifeste que la dimension de l'espace C_{873} est donc inférieure à 873. Toutefois, pour des raisons de simplicité, la décomposition d'une acception quelconque reste unique dans le thésaurus considéré. Cette interdépendance entre concepts a parfois été exploitée, comme par exemple dans le modèle LSA (Deerwester *et al.* 1990).

Chaque acception, c'est-à-dire chaque sens d'un terme, se traduit donc sur C_{873} par un vecteur unique, appelé vecteur conceptuel (Schwab, 2005) : chacune des 873 composantes d'un vecteur conceptuel représente l'intensité d'une des 873 idées génératrices. Les vecteurs conceptuels sont construits actuellement à partir de la hiérarchie d'un thésaurus, donc totalement figée. Cette construction s'effectue par apprentissage automatique à partir de différentes sources : dictionnaires, listes de synonymes ou d'antonymes, etc. Par exemple, dans le cas d'une définition d'une acception, on combine les vecteurs conceptuels des différentes acceptions que l'on rencontre dans le texte de cette définition pour former le vecteur conceptuel de l'acception définie. Pour obtenir une pertinence de la base, Schwab *et al.* (2004) ont montré la nécessité du caractère permanent de l'apprentissage.

Les 873 concepts du thésaurus *Larousse* constituent un système générateur. Ce dernier n'est pas unique ! Chaque individu humain possède son propre système générateur qui lui permet

que celui reliant n'importe quel élève de Poudlard avec n'importe lequel de leurs animaux. Pour tenir compte de cette différence, il faudrait considérer les co-occurrences des termes *Harry Potter* et *Hedwig* dans les textes concernés, éventuellement pondérées par les distances les séparant dans les arbres d'analyse morpho-syntaxique des phrases où on les rencontre.

de se faire sa propre représentation des connaissances ; de plus, nos connaissances personnelles s'accroissent au cours de la vie et donc ce système individuel évolue dans le temps. On peut considérer que le système générateur du thésaurus *Larousse* correspond en fait à celui d'un individu type. Le système générateur individuel ne correspond vraisemblablement jamais à celui du thésaurus utilisé (aucun de nous n'est véritablement réduit à un individu type).

En considérant uniquement les connaissances générales, c'est-à-dire en faisant abstraction de toute spécialisation, peut-on trouver un « meilleur » système générateur de l'espace vectoriel modélisant la représentation thématique des connaissances que celui défini par les concepts du thésaurus utilisé ? Quelle est la signification de ce mot « meilleur » ? Même s'il paraît délicat, voire impossible, de mesurer la pertinence d'un ensemble de concepts définissant un système générateur, quels sont les critères que doit vérifier un tel ensemble ?

2. Quels sont les critères d'un « bon » système générateur ?

Il paraît prétentieux d'affirmer ce que serait un « bon » système générateur d'un espace vectoriel modélisant notre représentation thématique des connaissances générales. Il est toutefois possible d'explicitier un certain nombre de caractéristiques minimales que doit posséder un tel système. Ceci est à rapprocher des travaux de Wilks (1977).

2.1. Étendue : c'est un système générateur

Un système générateur doit couvrir l'ensemble du domaine généraliste (puisque c'est ici le domaine considéré). Qu'entend-on par « l'ensemble du domaine généraliste » ? Même s'il est impossible de définir de manière précise les limites du domaine, on peut considérer que les termes appartenant aux dictionnaires généralistes courants en donnent une idée raisonnable.

On ne peut cependant pas en conclure que toute nouvelle acception qu'il serait possible d'exprimer en fonction d'acceptions déjà existantes pourrait être considérée comme faisant partie du domaine. Pour donner un exemple : le *Quidditch* est un jeu de balle volante qui se joue entre deux équipes d'élèves apprentis dans les écoles de sorcellerie. Le terme *Quidditch* ne fait pas (encore) partie du vocabulaire généraliste.

2.2. Représentativité : c'est un système générateur proche d'une base

Si l'on souhaite conserver le caractère généraliste du système générateur, il ne faut pas y inclure automatiquement toute nouvelle acception. Pour revenir sur l'exemple ci-dessus, le terme *Quidditch* ne fait pas partie du *Petit Larousse*.

Un système générateur doit donc permettre, en un minimum de concepts de différencier un maximum de notions « courantes » : nos expérimentations montrent qu'un millier de concepts semblent suffisants pour cela. Le système générateur, même si ce n'est pas une base de l'espace de représentation (peut-on vraiment le vérifier ?), doit en être relativement proche.

2.3. Évolutivité : ce n'est pas un système figé

La principale critique faite à un thésaurus est son caractère figé. Un système générateur doit pouvoir facilement prendre en compte l'évolution des notions. De plus, un système générateur, tout comme un thésaurus, ne peut pas *a priori* être exhaustif : il paraît difficile, voire impossible, de définir les limites du domaine considéré. Cela montre la nécessité, en fonction des besoins, du caractère évolutif du système générateur. Il faudra toutefois faire attention à ce que cette évolutivité n'œuvre pas à l'encontre de la représentativité.

Un système générateur doit-il nécessairement être organisé en une structure hiérarchique arborescente, comme le sont le thésaurus Larousse ou le *Roget's* ? Dans l'état actuel de nos travaux, nous n'utilisons pas cette *caractéristique*. La question est toutefois abordée à la section 3.3.

3. Méthode : définition et évolution des notions de base

3.1. Définition des notions de base

De façon incrémentale, au fur et à mesure de l'analyse de textes, on rencontre des mots. Soit k le nombre de termes différents rencontrés (pour la langue française, la plupart des dictionnaires généralistes actuels possèdent environ 80.000 entrées). La grande majorité des termes étant polysémiques, ces k termes différents possèdent k' acceptions et génèrent donc k' vecteurs sur l'espace vectoriel C_{873} (dans l'état actuel de nos expérimentations, k' dépasse 400.000 (Schwab *et al.* 2004)). De plus, il est nécessaire de pondérer ces k' vecteurs en fonction de leur fréquence d'apparition dans les textes étudiés. En effet, certaines acceptions de termes se rencontrent beaucoup plus fréquemment que d'autres et elles jouent donc un rôle plus important dans notre connaissance². Il semble naturel que cette fonction de pondération, nécessairement croissante, soit une fonction logarithmique de la fréquence d'apparition ; effectivement, il est nécessaire de réaliser un amortissement de l'impact de cette fréquence sur la norme du vecteur d'acception correspondant. Il paraît également raisonnable de pondérer chaque occurrence de ces k' vecteurs en fonction de la profondeur de l'acception correspondante dans l'arbre d'analyse syntaxique du texte étudié ; en effet, il est logique de penser que plus un mot est « perdu » dans les profondeurs d'une phrase, moins il est important pour le sens global de cette phrase³. Il semble naturel que cette fonction de pondération, nécessairement décroissante, puisse être une exponentielle négative de la profondeur.

De manière plus formelle, si nous appelons p la profondeur d'un terme t dans l'arbre d'analyse syntaxique du texte étudié, la norme du vecteur v représentant l'acception correspondante (après désambiguïsation entre les éventuelles différentes acceptions de t) sera :

$$\|v\| = e^{-\alpha p} \quad \text{où } \alpha \text{ est un coefficient de pondération.}$$

La i^{e} occurrence de cette acception a du terme t sera représentée sur C_{873} par un vecteur $v_{a,i}$ dont la norme sera :

$$\|v_{a,i}\| = e^{-\alpha p_{a,i}} \quad \text{où } p_{a,i} \text{ désigne la profondeur dans l'arbre d'analyse de la } i^{\text{e}} \text{ occurrence de l'acception } a.$$

Afin de tenir compte de la fréquence d'apparition des acceptions, il paraît raisonnable d'envisager une sommation (logarithmique) des différents vecteurs représentant chaque occurrence d'une même acception. Ainsi, le vecteur v_a représentant l'acception a sur l'ensemble des textes traités aura pour norme :

$$\|v_a\| = \text{Log} (\sum_i f(\|v_{a,i}\|)),$$

² En fonction du corpus de textes étudié, en particulier s'il s'agit de dictionnaires, il faudra toutefois se méfier de termes généraux très fréquemment utilisés dans les libellés des définitions, tels que « action », « partie de » ... qui, selon Schwab *et al.* (2004) appartiennent au métalangage et qu'il ne faut évidemment pas comptabiliser ici.

³ Bien évidemment, ceci est une généralité. L'importance d'un mot dépend également de sa fonction dans la phrase, et la profondeur dans l'arbre d'analyse syntaxique n'est qu'un critère de l'importance du mot considéré.

car, comme cela est expliqué plus haut, nous souhaitons que $\|v_a\|$ soit fonction du logarithme de la fréquence d'apparition de l'acceptation a .

$\|v_a\|$ étant une norme, elle doit en vérifier les propriétés :

$$1^\circ / \|v_a\| \geq 0$$

$$2^\circ / \|v_a\| = 0 \Leftrightarrow v_a \text{ est le vecteur nul}$$

$$3^\circ / \|v_{a,i} + v_{a,j}\| \leq \|v_{a,i}\| + \|v_{a,j}\|$$

qui se traduit ici par : $\|v_a\| \leq \sum_i \|v_{a,i}\|$.

S'il n'y a qu'une seule occurrence de l'acceptation a , alors $\|v_a\| = \|v_{a,1}\|$.

S'il y a plusieurs occurrences de l'acceptation a , alors $\|v_a\| < \sum_i \|v_{a,i}\|$.

Ces différentes conditions conduisent à envisager :

$$\|v_a\| = \text{Log} \left(\sum_i e^{\|v_{a,i}\|} \right).$$

Dans un but de simplification, en tenant compte de la proximité thématique⁴, les k vecteurs v_a obtenus peuvent se regrouper en n nuages, avec $n \ll k$. Les n vecteurs barycentres de ces n nuages forment un système générateur d'un espace vectoriel B_n . Nous regroupons ainsi les termes thématiquement proches d'un même concept. Bien que la méthode soit différente, notre objectif est à rapprocher de celui développé par Landauer et Dumais (1997) pour la méthode Latent Semantic Analysis (LSA). En considérant un nombre n relativement grand, probablement de plusieurs centaines à quelques milliers, cet espace B_n peut être quasiment confondu avec C_{873} , à condition que les textes étudiés ne se restreignent pas à un domaine spécifique et balayent l'ensemble des connaissances générales.

À quoi correspondent ces n vecteurs ? Ce sont les « notions de base » déduites de l'analyse des textes. La figure 1 explicite le principe de notre méthodologie.

3.2. Expression des notions de base

L'espace B_n dépend fortement des textes rencontrés ; il peut correspondre en fait à l'espace de représentation thématique des connaissances pour un individu humain. Chacun d'entre nous possède ses propres références et ses propres définitions : les espaces B_n , ainsi que leurs systèmes générateurs, des différents individus se ressemblent, mais ne sont pas nécessairement absolument identiques.

Pour chacune de ces n notions de base, notée b_i , il est possible de trouver le terme le plus proche, en utilisant la notion de distance thématique définie sur C_{873} . Sous la condition qu'il existe un terme « suffisamment » proche de b_i , celui-ci exprimera cette notion de base.

⁴ Il est possible de mesurer cette proximité thématique à l'aide de la distance angulaire entre vecteurs conceptuels : celle-ci s'appuie sur la notion de similarité entre vecteurs, souvent utilisée en recherche d'information, par exemple Baeza-Yates et Ribeiro-Neto (1999).

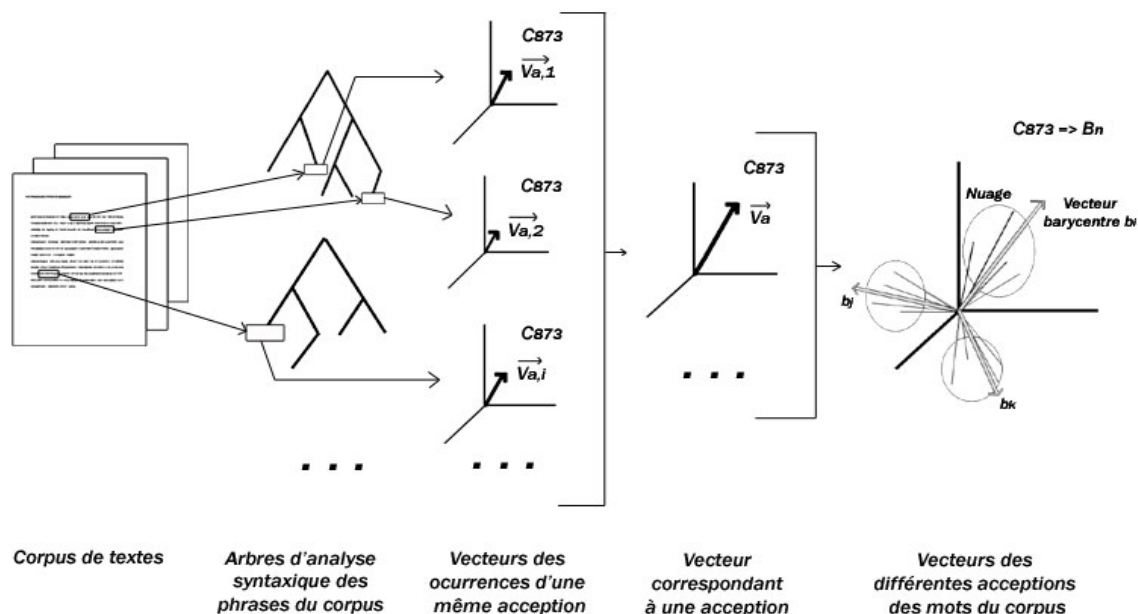


Figure 1. Ce schéma illustre la méthode utilisée pour définir les notions de base

L'étude des textes permet d'obtenir les arbres d'analyse morpho-syntaxique des phrases les constituant. Cette analyse est réalisée grâce à l'outil SYGFRAN (pour le français) développé avec SYGMART (Chauché, 1984). Une analyse sémantique telle que décrite dans Schwab (2005) est alors réalisée et chaque occurrence d'une acception se traduit par un vecteur sur C_{873} dont la norme est fonction de sa profondeur dans l'arbre. En sommant (logarithmiquement) les vecteurs des différentes occurrences d'une même acception, on obtient le vecteur correspondant à cette acception sur C_{873} . Les vecteurs des différentes acceptions peuvent se regrouper en nuages dont les barycentres définissent les notions de base.

En fait, il s'agit non pas de trouver le terme le plus proche, mais l'acception de terme la plus proche de b_i . Il peut se poser alors la question de son expression afin d'éviter tout risque d'ambiguïté. De plus, il est indispensable de prendre en compte la fréquence de l'acception candidate pour exprimer la notion de base : une acception trop peu fréquente peut-elle être un « bon » concept ? (par exemple, n'est-elle pas trop désuète ? ou trop spécialisée ?) L'acception la plus fréquente, ou plus exactement celle dont le vecteur a la norme la plus grande dans le nuage (voir la figure 1), ne constitue pas nécessairement le « meilleur » concept, si elle est relativement éloignée du barycentre du nuage de vecteurs qu'elle est censée représenter.

Mais alors, comment avoir une idée de la taille du domaine concerné par b_i ? Cette différenciation b_i - b_j dépend majoritairement de la valeur de n et donc de la définition des nuages, c'est-à-dire des paramètres de la discrétisation des nuages. Comme nous l'avons vu plus haut, il est bien évident que plus on voudra de précision, plus les nuages auront une étendue réduite, et plus ils seront nombreux : plus les notions de bases seront fines et précises, plus elles seront nombreuses (c'est le classique compromis précision-simplicité d'une modélisation). Si les nuages sont nombreux, chacun d'eux aura un poids relativement faible ; ils seront donc susceptibles d'évoluer plus rapidement lors de l'analyse de nouveaux textes (c'est le compromis précision-stabilité). En conséquence, plus les notions de base seront

nombreuses, plus il risque d'être difficile pour les exprimer de trouver des termes non ambigus qui en sont « suffisamment » proches.

3.3. Quel nombre de notions de base faut-il considérer ?

En n'effectuant aucun regroupement de termes thématiquement proches et donc en considérant chaque acception comme un concept, chaque vecteur d'acception est un vecteur générateur ; cela revient à poser $n = k'$. Naïvement, il pourrait être alors envisageable de tout représenter sans ambiguïté. Mais, d'une part, un espace vectoriel de dimension 400.000 paraît volumineux dans l'état actuel de la technologie des ordinateurs au niveau volume de stockage, et il pourrait conduire à des temps de traitement difficilement acceptables. D'autre part, est-ce vraiment « sans ambiguïté » ? Le fait que toute acception ne soit « décomposable » que sur un seul concept ne permet pas, en particulier, de mettre en œuvre la notion de distance angulaire entre vecteurs. Entre deux vecteurs concepts-acceptations quelconques, la distance angulaire serait invariablement égale à $\pi/2$. Il est manifeste que cette solution revenant à considérer que les distances entre acceptations sont toutes identiques ne correspond en rien à la réalité : intuitivement, s'il est possible d'établir une relation thématique entre deux acceptations, leur distance semble inférieure à celle qui sépare deux acceptations qui ne possèdent aucun point commun ; par exemple, s'il existe une relation de co-hyponymie (exemple : entre *CHAT* et *CHIEN*) ou une relation d'hyponymie/hyperonymie (exemple : entre *CHAT* et *MAMMIFERE*), la distance entre les acceptations est manifestement inférieure à celle qui sépare deux acceptations pour lesquelles aucune relation n'apparaît clairement (exemple : entre *CHAT* et *INEXISTENCE*, dont les champs sémantiques n'ont apparemment que peu de points communs).

Il serait toutefois envisageable de définir une distance ultramétrique entre concepts (Schwab, 2005) qui est calculée en fonction du chemin minimal entre les deux concepts considérés dans la structure arborescente du thésaurus. Cette distance pourrait même être améliorée par une pondération tenant compte de l'ordre des concepts dans leurs fratries, dans les cas où un tel ordre représente une information pertinente. Mais il faudrait pour cela disposer d'un thésaurus. Or, en construisant B_n , nous voulons nous affranchir de C_{873} et de l'aspect contraint et figé d'un thésaurus fixe défini *a priori*. La solution pourrait consister à considérer que l'espace B_n n'est pas homogène et donc à découper l'espace B_n en régions. Chacun des n vecteurs appartenant à une région, la distance entre deux vecteurs dépendrait de leur appartenance ou non à la même région. Il pourrait même être envisagé de regrouper certaines régions en super-régions⁵ ; seule l'expérience pourrait montrer si une telle structuration en plusieurs niveaux se manifeste clairement.

À l'opposé, il est manifeste qu'une valeur trop faible du nombre des concepts limite de façon draconienne la qualité de la discrétisation. Pour donner un exemple s'appuyant sur le thésaurus *Larousse*, en remontant simplement d'un niveau dans sa structure hiérarchique, et en ne considérant donc que les 95 concepts de niveau 3 (les 873 concepts générateurs de C_{873} sont de niveau 4), tous les animaux sont regroupés sous le concept $C3:ANIMAUX$, et il serait alors difficile de les discriminer, même en considérant leurs caractéristiques ou leurs comportements. Une trop grande simplification des concepts, et donc une réduction trop importante de leur nombre, conduit à un maillage beaucoup trop lâche de l'espace, généraliste ou spécialisé, que l'on souhaite modéliser.

⁵ À titre d'illustration, ceci pourrait être comparé à la structure spatiale de l'Univers. Les galaxies se regroupent en amas de galaxies, eux-mêmes se regroupant en super-amas : notre Galaxie appartient à l'Amas Local, lui-même compris dans le super-amas de la Vierge.

Cette solution qui conduit à une représentation thématique trop imprécise dans le cas d'une analyse fine, pourrait toutefois être utilisée en classification de textes. En effet, pour une phrase ou un paragraphe donné, elle permet de manière rapide de déterminer le domaine thématique concerné.

Il n'existe pas de valeur optimale du nombre de notions de base ; seule l'expérience peut en donner un ordre de grandeur. Dans le cadre des connaissances générales, il semble logique, au vu des thésaurus existants, que cet ordre de grandeur soit d'environ un millier. Les travaux menés par Lafourcade *et al.* (2002) semblent montrer qu'un nombre de quelques milliers conduirait, sans trop alourdir le système, à de meilleurs résultats.

3.4. Évolutivité du système

Au fur et à mesure de l'analyse de nouveaux textes, les nuages de vecteurs évoluent : c'est « l'évolution des notions », avec éventuellement des phénomènes de différenciation ou de regroupement. Ceci est à opposer à la définition des 873 concepts de base qui sont immuables. Il est tout de même à remarquer que ces notions de base évolutives sont initialement exprimées en fonction des 873 concepts fixes (n'est-il pas rassurant d'avoir des repères fixes pour se laisser la possibilité de vérifier une certaine pertinence des vecteurs ?).

À partir du système générateur des concepts fixes, nous avons construit le système générateur des notions de base. Toute acception peut à présent s'exprimer dans l'espace des notions de base ; par conséquent, toute nouvelle acception s'exprimant en fonction d'acceptions déjà existantes (ce qui est le cas des définitions dictionnariques) peut s'exprimer dans l'espace des notions de base.

Les concepts du thésaurus servent donc pour une initialisation du processus. Après une première définition des notions de base, il est possible de s'affranchir des concepts, puisque, comme nous venons de le voir, toute nouvelle acception est décomposable sur l'espace généré par les notions de base.

Il est toutefois à remarquer que le processus d'évolution est nécessairement relativement lourd. En effet, il s'effectue en deux étapes :

1. évolution des notions de base, c'est-à-dire évolution du système générateur :

Chaque introduction d'une nouvelle acception modifie le nuage auquel elle appartient. Le barycentre de ce nuage est donc déplacé, ce qui entraîne une modification du vecteur de la notion de base correspondante. Il se peut même, dans certains cas, que l'introduction d'une nouvelle acception ait une influence telle sur le nuage que celui-ci puisse se morceler en nuages plus restreints, ou à l'opposé se regroupe avec un nuage voisin ;

2. modification des coordonnées des vecteurs d'acception pour lesquels les composantes sur les vecteurs des notions de base modifiées est non nulle :

Après modification du vecteur d'une notion de base (éventuellement plusieurs), il est indispensable de parcourir l'ensemble des vecteurs d'acception existants et de modifier les coordonnées de ceux pour lesquels la projection sur le (les) vecteur notion de base modifié est non nulle : c'est un classique changement de repère. On se rappellera que la base de données vectorielles est en apprentissage permanent et la révision des notions de base pourra être réalisée simultanément à cet apprentissage. En effet, lors de la révision d'un vecteur d'acception, nous révisons non seulement l'intensité de ses composantes, mais également ses composantes elles-mêmes, à savoir les notions de base sur lesquelles il s'appuie.

La figure 2 qui explicite ce processus évolutif montre clairement son principe itératif : chaque introduction de nouvelle acception génère une itération conduisant à une évolution du système générateur de l'espace des notions de base.

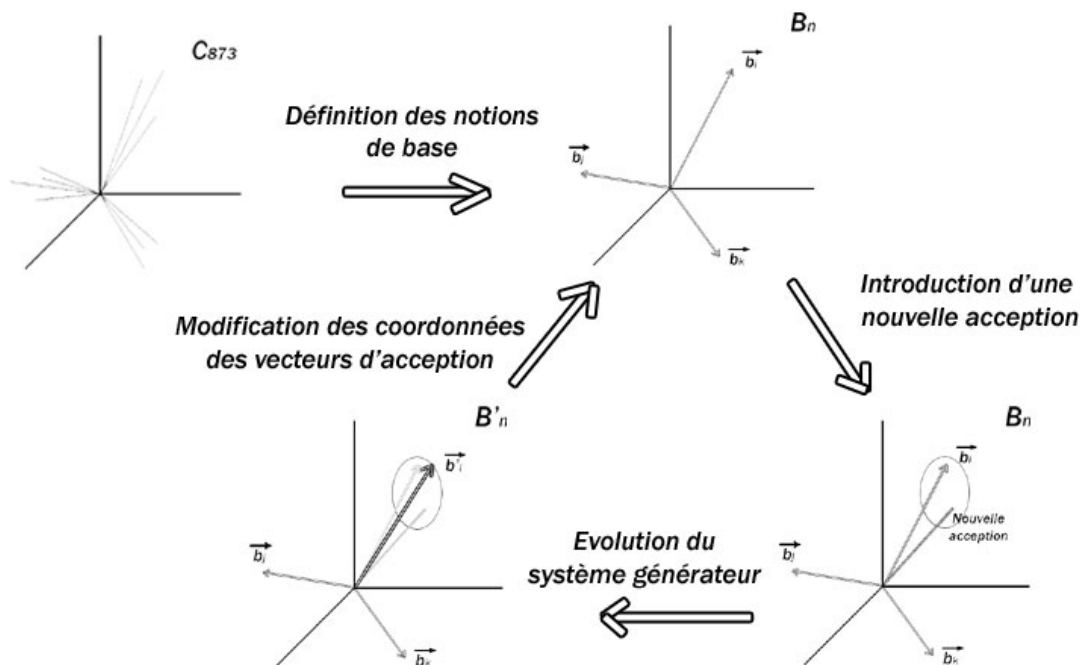


Figure 2. Ce schéma qui rappelle de manière succincte la définition des notions de base, illustre la structure itérative permettant leur évolution après introduction d'une nouvelle acception

Cette mise à jour, même si elle peut être relativement longue est totalement automatique. Ce n'est pas le cas avec un thésaurus fixe défini *a priori* qui ne peut être modifié que par un expert humain. Dans ce dernier cas, la première étape, entièrement « manuelle », est effectivement beaucoup plus lourde ; la deuxième étape, automatique, est quant à elle analogue à celle réalisée dans l'espace des notions de base.

4. Conclusion

La modélisation par un espace vectoriel pour la représentation thématique des connaissances générales est utilisée depuis une quinzaine d'années. La définition du système générateur de cet espace de modélisation repose couramment sur la hiérarchie d'un thésaurus : c'est une structure fondamentale, mais totalement figée. L'évolution des connaissances nécessite de disposer d'un système générateur qui puisse évoluer. Nous avons donc défini les notions de base. Leur construction s'appuie initialement sur l'ossature que constituent les concepts d'un thésaurus. Au fur et à mesure de l'analyse de nouveaux textes, l'apparition d'acceptions permet l'évolution du système générateur de l'espace de représentation des connaissances. Cette définition évolutive des notions de base s'affranchit totalement de la hiérarchie figée des concepts. Elle permet d'envisager des applications qui suivront l'évolution des notions, en particulier dans les domaines spécialisés où les notions évoluent souvent plus rapidement que dans le domaine généraliste.

Références

- BAEZA-YATES R., RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. ACM Press, New York.
- CHAUCHÉ J. (1984). « Un outil multidimensionnel de l'analyse du discours ». In *Proceedings of the 22nd conference on Association for Computational Linguistics*. Stanford California : 11-15.
- CHAUCHÉ J. (1990). « Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance ». In *TA Information* 31 (1) : 17-24.
- DEERWESTER S., DUMAIS S., LANDAUER T., FUMAS G., HARSHMAN R. (1990). « Indexing by latent semantic analysis ». In *Journal of the American Society of Information Science* 416 (6) : 391-407.
- KIPFER B.A. (2001). *Roget's International Thesaurus*. Sixth edition, Harper Resource (first edition : 1852).
- LAFOURCADE M., SANDFORD E. (1999). « Analyse et désambiguïsation lexicale par les vecteurs sémantiques ». In *Actes de TALN 1999*. Cargèse : 351-356.
- LAFOURCADE M. (2001). « Lexical sorting and lexical transfert by conceptual vectors ». In *Proceedings of the First International Workshop on Multimedia Annotation (MMA'2001)*. Tokyo.
- LAFOURCADE M., PRINCE V., SCHWAB D. (2002). « Vecteurs Conceptuels et Structuration émergente des Terminologies ». In *Traitement Algorithmique des Langues* 43 (1) : 43-72.
- LANDAUER T., DUMAIS S. (1997). « A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge ». In *Psychological Review* 104 (2) : 211-240.
- LAROUSSE (1999). *Thésaurus Larousse – des idées aux mots, des mots aux idées*. Larousse, Paris.
- SALTON G., MACGILL M.J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.
- SCHWAB D., LAFOURCADE M., PRINCE V. (2004). « Hypothèses pour la construction et l'exploitation conjointe d'une base lexicale sémantique basée sur les vecteurs conceptuels ». In *Actes des JADT 2004*. Presses universitaires de Louvain, Louvain-la-Neuve : 1008-1018.
- SCHWAB D. (2005). *Approche hybride – lexicale et thématique – pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de textes*. Thèse de doctorat, Université de Montpellier II.
- WILKS Y. (1977). « Good and bad arguments about semantic primitives ». In *Communication and Cognition* : 181-221.
- YAROWSKY D. (1992). « Word-sense disambiguation using statistical models of Roget's categories trained on large corpora ». In *Proceedings of COLING'92*. Nantes : 454-460.