# Analyse et transfert par vecteurs sémantiques


## Résumé

Cet article introduit une représentation sémantique basée sur des vecteurs de notions. Ces vecteurs sémantiques ont pour but de rendre compte de l'ensemble des idées évoquées dans un segment textuel. Ce type de représentation en conjonction avec une analyse morpho-syntaxique classique permet d'effectuer dans de nombreux cas une désambiguisation lexicale efficace, mais permet surtout d'effectuer un tranfert lexical sans dictionnaire de transfert, ni approche par pivot.


## 1. Introduction

Dans le cadre du traitement automatique de la langue naturelle (TALN), nous nous intéressons à la désambiguïsation lexicale sémantique couplée à une analyse morphosyntaxique (Rastier 1987). En effet, nombreuses sont les approches maintenant classiques en recherche d'information. Le traitement peut être statistique (Salton et MacGill 1983), dépourvu de syntaxe et à base de mot-clés (Salton 1988), basé sur différentes logiques (Sowa 1994), ou encore sur des taxonomies (Resnik 1995).

À partir d'un espace de notions élémentaires, il est possible de construire des vecteurs (dit sémantiques) et de les associer à des mots. Les mots ambigus fusionnent les multiples vecteurs correspondant aux différents sens. Cette approche par vecteurs repose sur des bases mathématiques classiques et bien connues. Il est ainsi possible d'effectuer des manipulations formellement bien fondées auxquelles nous attachons des interprétations (linguistiques) raisonnables. Nous définissons ainsi les notions de cuisson de vecteurs ainsi qu'une fonction de distance vectorielle traduisant la proximité sémantique.

Nous définissons en premier lieu ce que sont les vecteurs sémantiques. Ces vecteurs sont attachés à des termes et sont directement issus d'informations que l'on peut trouver dans des thésauri. Afin d'affiner la représentation, nous tenons compte des voisinages de notions correspondant à un ensemble de familles de sens. Cette opération de « compilation » d'ensembles de vecteurs sémantiques (appelés lexiques sémantiques) est itérative et convergente. Ces vecteurs constituent ensuite la base de l'analyse sémantique dont le résultat (les vecteurs sémantique « recuits ») permet des désambiguïsations et des transferts lexicaux.

## 2. Vecteurs sémantiques

### *2.1 Généralités*

La définition de « vecteurs sémantiques » a pour but obtenir une représentation sémantique de segment textuel indépendante du niveau syntaxique considéré (mots, syntagmes, phrases et textes). Plus précisément, il s'agit d'une projection des idées exprimées dans un tel segment sur un jeu de notions prédéterminé. L'ordre de grandeur du jeu de notions que nous considérons est d'environ 900. Il s'agit donc d'un n-uplet à 900 valeurs. Le fait de qualifier de tels vecteurs « sémantiques » vient plus du fait qu'ils contiennent une représentation des sens de termes évoqués dans un segment textuel que par la signification globale du segment textuel.

Chaque composante d'un vecteur - une dimension de l'espace vectoriel - représente une notion issue d'un ou de plusieurs thesauri. Ces derniers constituent les sources d'information nécessaires à la création de l'espace vectoriel. La description formelle des notions dans le thesaurus rentre dans le cadre de la théorie des champs sémantiques. Ces champs de sens se chevauchent sans qu'il soit possible pour l'instant de les différencier (Baylon et Mignot 1995)

ou de donner une relation calculable entre ces champs. Par exemple, l'antonymie apparaissant entre les notions ACCORD et DESACCORD du thesaurus pourrait supposer une opposition entre ces deux dimensions dans notre espace vectoriel. Leur coefficient serait donc respectivement à 1 et –1. Nous avons délibérément écarté cette approche car la négation d'un concept ne modifie pas la référence sémantique à celui-ci. L'antonymie reste une relation sémantique forte.

Il est important de noter que nous ne disposons pas d'une base vectorielle de l'espace vectoriel engendré par les vecteurs sémantiques. Nous disposons par contre d'un ensemble générateur composé des vecteurs des notions choisies.

Le domaine de définition d'une dimension d'un vecteur sémantique est réel dans [0, 1]. Cette valeur représente l'intensité de la notion dans le vecteur sémantique (0 aucune activation, 1 activation maximale et unique). La norme des vecteurs est toujours égale à un entier fixé (mathématiquement 1).

Les vecteurs sémantiques subissent un certain nombre de prétraitements avant d'être incorporés à une analyse syntaxique classique (Chauché 1984). Cette dernière est alors qualifiée de « sémantique » (plus pour la clarté du discours que par une réelle nature sémantique). À l'issue de l'analyse sémantique, les vecteurs présents dans l'arbre d'analyse ont été transformés, avec comme perspective à chaque niveau syntagmatique, l'émergence des notions pertinentes.

Nous avons défini trois états possibles pour un vecteur sémantique : cru, cuit et recuit. Un vecteur sémantique cru est la traduction directe des informations des thesauri. Un vecteur sémantique cuit est un vecteur sémantique brut augmenté des notions voisines. Un vecteur sémantique cuit est manipulé lors de l'analyse sémantique selon l'arbre d'analyse et le contexte afin d'obtenir un vecteur sémantique recuit (la métaphore avec le recuit-simulé est ici forte).

### *2.2 Vecteurs sémantiques crus*

Un vecteur cru ne contient que des références à des notions immédiatement reliées. Ces informations sont directement issues des thesaurus ayant servis à l'indexation d'un terme et de son vecteur. Ce travail d'indexation est essentiellement manuel.

Par exemple, le nom féminin *friction* se projette sur les notions suivantes : FROTTEMENT, SOINS DU CORPS, DESACCORD, CONFLIT et de COIFFURE (qui correspondent respectivement aux dimensions 228, 393, 429, 649, 867). Il peut se traduire sous forme de vecteur :

| 1 … | **228** | | **393** | | **429** | | **649** | | **867** | … 873 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 … 0 | $1/\sqrt{5}$ | 0 … 0 | $1/\sqrt{5}$ | 0 … 0 | $1/\sqrt{5}$ | 0 … 0 | $1/\sqrt{5}$ | 0 … 0 | $1/\sqrt{5}$ | 0 … 0 |

La valeur $1/\sqrt{5}$ est issue de la normalisation du vecteur sémantique, la norme étant égale à 1. Toutes les autres notions sont à 0. En effet, la création du vecteur cru consiste à affecter la valeur 1 pour chaque notion atteinte, puis à normer ce vecteur. Si un terme atteint *n* notions, chaque dimension respective se verra affectée la valeur $1/\sqrt{n}$.

Le fait de les normer permet d'avoir tous les vecteurs sémantiques à la même distance de l'origine et cet ensemble se projette sur une sphère (à 873 dimensions). Il est ainsi facile de comparer deux vecteurs à partir de leur angle. Une métaphore possible est celle de l'observation des étoiles. En effet, pour un spectateur naïf, toutes les étoiles semblent à la même distance et seule l'angle entre deux étoiles (et le spectateur) constitue une mesure.

### *2.3 Vecteurs sémantiques cuits*

Après une rapide expérimentation (sur un texte de mille mots), il nous est apparu que l'utilisation de vecteurs sémantiques bruts ne donnait pas de résultat particulièrement satisfaisant. Nous avons identifié les raisons suivantes :

- Alors qu'intuitivement deux notions ont clairement un rapport au moins indirect (par exemple MALADIE et MEDECINE), ce rapport n'est pas explicite dans les vecteurs

sémantiques crus. Il s'agit d'un manque de profondeur des notions. Certaines notions sont voisines dans une des hiérarchies d'un thesaurus, mais ce voisinage n'est pas exprimé dans le calcul des vecteurs sémantiques.

- La deuxième raison est le pendant de la première. Si certaines notions manquent, d'autres peuvent alors interférer pour choisir un sens inapproprié. Il nous faut donc augmenter les notions pertinentes de leur voisinage pour diminuer l'intensité des notions non pertinentes.

- Enfin, les notions d'un thesaurus sont classées de façon hiérarchique selon une ontologie (par une relation « général » vers « spécialisé »). Le groupe de notions dans lequel apparaît une notion de notre vecteur sémantique est alors un voisinage immédiat possible de la notion. Les notions dans ce voisinage correspondent aux notions les plus proches sémantiquement les unes des autres. Entre autres, les antonymes font partie du même voisinage. Les vecteurs crus ne rendent pas compte de ce voisinage.

Le voisinage permet d'augmenter un vecteur sémantique des notions proches de chaque notion présente dans ce vecteur. Nous appelons « cuisson » le processus d'augmentation d'un vecteur en fonction de son voisinage. Nous pouvons définir itérativement la fonction de cuisson comme suit :

$$V^{i+1}(x) = V^{i}(x) + \sum_{k=1}^{\dim(V)} V_k^i(x) \times V^0(tête_k)$$

- $V^i(x)$ est le vecteur sémantique cuit à l'itération i. Le vecteur cru est noté $V^0(x)$. Ce dernier est directement issu de l'indexation.

- dim(V) est la dimension d'un vecteur sémantique. Ici, il s'agit d'une constante (de valeur 873).

- La notation $tête_k$ correspond au mot représentant la notion d'indice k dans un vecteur V. Par exemple $tête_{649}$ = « conflit »

- $V_k(x)$ correspond à la valeur de la $k^{ieme}$ dimension de V(x).

Cuire un vecteur V quelconque consiste à considérer tour à tour chaque dimension k, et pour chacune d'elle à additionner à V le vecteur cru correspondant au terme (le mot-tête) lié à la notion. L'intensité de ce vecteur est proportionnelle à la valeur de V sur la dimension k.

La cuisson d'un terme débute à partir de son vecteur cru (i = 0). On peut effectuer un nombre de « tours de cuisson » arbitraire. Le vecteur est normé entre chaque tour. Plus le nombre de tours est important, plus le vecteur cuit rendra fidèlement compte du voisinage.

Ce processus est convergent. Plus précisément, au bout d'un nombre fini d'itérations, nous avons : $dist(V^{i+1}(x), V^i(x)) < \varepsilon$ avec ε aussi petit que l'on veut. *dist(V, V')* est une fonction de distance entre deux vecteurs *V* et *V'*.

Par exemple, pour le terme *friction*, nous avons (la norme des vecteurs est 1000) :

$V^0$ : `frottement/447 soins du corps/447 désaccord/447 conflit/447 coiffure/447`

$V^1$ : `frottement/472 conflit/444 désaccord/365 soins du corps/349 coiffure/344 guerre/187 énergie/180 mécanique/177 musique/173 tribunal/167 inimitié/138 poli/127 adversité/116 dessus/101 machine/65 révolution/64 défiance/57 mouvement/54 conseil/50 guérison/46 divorce/45 …`

$V^2$ : `conflit/387 désaccord/312 guerre/296 frottement/278 soins du corps/258 tribunal/254 coiffure/249 énergie/221 mécanique/219 musique/214 inimitié/201 adversité/197 dessus/150 machine/138 révolution/134 mouvement/123 poli/118 politique/116 conseil/116 manœuvre/104 passe-temps/99 force/90 défiance/87 détection/82 enseignement/75 guérison/73 discordance/73 divorce/70 transport par route/69 son/68 médecine/66 dissemblance/64 paix/61 profane/60 jeu/53 maladie/51 vigueur/47 difficulté/46 main-d'œuvre/43 choc/43 repos/38 immunité/38 …`

### *2.4 Vecteurs sémantiques recuits*

Les vecteurs sémantiques recuits sont ceux issus de l'analyse sémantique. Avant l'analyse sémantique, les nœuds de l'arbre de l'analyse morpho-syntaxique correspondant aux termes du segment textuel sont affectés de vecteurs sémantiques cuits. L'analyse sémantique consiste à effectuer des propagations et des fusions des différents vecteurs (voir 3.2). Les vecteurs recuits sont le résultat de ce processus.

Soit la phrase suivante :

(1) Après la publication *des notes d'examen*, on a pu constater une grande effervescence chez les étudiants.

En sortie de notre système, le syntagme prépositionnel (2) « *des notes d'examen* » a le vecteur sémantique donné ci-dessous. Afin de garder l'exemple simple, ce vecteur a été obtenu avec des vecteurs sémantiques bruts. Le même type de vecteur peut être obtenu à partir des vecteurs sémantiques cuits, avec cependant beaucoup plus de dimensions activées et de façon plus fine. La liste des notions activées et leur intensité dans le vecteur sémantique du syntagme (2) est la suivante :

Vecteur sémantique de "*notes d'examen*" après analyse sémantique:

| # | Notion | Intensité | # | Notion | Intensité |
|---|---|---|---|---|---|
| 97 | ADJONCTION | 250 | 751 | DISCOURS | 165 |
| 391 | MEDECINE | 165 | 762 | ECRITURE | 250 |
| 412 | RECHERCHE | 165 | 766 | PRESSE | 250 |
| 414 | ENSEIGNEMENT | 415 | 770 | CORRESPONDANCE | 250 |
| 419 | QUESTION | 165 | 795 | SALAIRE | 250 |
| 426 | CONTROLE | 165 | 825 | PAIEMENT | 250 |
| 536 | PREPARATION | 165 | 831 | PRIX | 500 |
| 726 | COMMUNICATION | 250 | | | |

Pour l'exemple, nous avons renormé et arrondi les intensités à des valeurs entières. La norme du vecteur est ici de 1000. Les intensités sont calculées par sommation des coefficients en fonction de la structure syntaxique de la phrase.

## 3. Exploitation

### *3.1 Lexique sémantique*

Le lexique sémantique est un dictionnaire associant des mots à des vecteurs sémantiques. Un tel lexique est indexé manuellement - pour l'instant - à partir de thesauri et de dictionnaires. L'indexation du lexique sémantique consiste à saisir la liste des notions attachées à un terme. La compilation du lexique consiste à cuire les vecteurs crus.

Un lexique sémantique L contenant N termes permet trois opérations de base :

(a) trouver à partir d'un mot *m* (potentiellement augmenté d'informations morphosyntaxiques comme la catégorie, le genre, le nombre, etc.) son vecteur sémantique, si $m \in$ L ;

(b) l'opération inverse : à partir d'un vecteur sémantique V, trouver le mot *m* dont le vecteur semantique est dans L le plus proche du vecteur V. C'est-à-dire $m_i$ tel que :

$$m_i \in L \mid dist(V,V(m_i)) \leq dist(V,V(m_k)) \quad \forall k = 1\ldots N,\ m_k \in L$$

(c) à partir d'un vecteur sémantique, trouver l'ensemble de mots $\{ m_i \}$ tel que le vecteur sémantique soit suffisement proche de V. C'est à dire $\{ m_i \}$ tel que :

$$\{m_i\} \subseteq L \mid dist(V,V(m_i)) \leq seuil$$

Les mots polysémiques sont stockés d'une manière relativement différente des mots monosémiques dans le lexique sémantique. Les mots monosémiques sont liés à un seul vecteur sémantique. Les mots polysémiques sont liés à plusieurs sens dans le lexique sémantique et chacun de ceux-ci à un vecteur cru. Tous les sens sont énumérés sous le mot polysémique et le vecteur sémantique du mot est alors le vecteur somme (normé) de tous les vecteurs sémantiques des sens de ce mot.

Par exemple, le mot *friction* a trois sens. Ces trois sens sont étiquetés par des gloses les plus proches du sens du mot décrit : un des sens de *friction* est donné par le terme *frottement*. Ces gloses n'ont qu'une valeur informative pour l'indexeur. Le sens *frottement* est lié à la notion FROTTEMENT du thesaurus. Par contre le verbe *frictionner* n'est pas polysémique. Nous avons donc :

```
("friction" :N :FEM
      ("frottement" 228)
      ("massage" 393 867)
      ("conflit" 429 649))
```

```
("frictionner" :V 228 380 854 867)
```

FROTTEMENT / PROPRETE / NETTOYAGE / COIFFURE

À chacun des trois sens, est associée une liste de numéro de notion. On obtient les trois couples (*frottement*, 228), (*massage*, (393, 867)) et (*conflit*, (429, 649)). Pour l'analyse sémantique, nous regroupons les notions données par les sens des mots ; ce qui nous donne la liste de notions (FROTTEMENT /228, SOINS DU CORPS/393, DESACCORD/429, CONFLIT/649, COIFFURE/867). C'est ce vecteur cru qui est ensuite cuit.

La liaison entre les vecteurs sémantiques et les mots se fait par les critères morphosyntaxiques des mots.C'est-à-dire que chaque mot sera codé en fonction des critères issus de l'analyse syntaxique du texte traité. Notons alors que les entrées du lexique sont un tuple (préfixe, lemme, catégorie syntaxique, paramètres *Pi*…) où les *Pi* correspondent à des paramètres différents suivant que l'on ait un nom, un verbe, un adverbe ou un adjectif.

Les critères syntaxiques généraux sont les suivants et sont appliqués pour tout mot, qu'il soit un nom commun, un verbe, un adjectif ou un adverbe : le *préfixe* du mot et le *lemme* du mot. Le nom *table* sera donné comme ayant un préfixe nul (ε mot vide) et un lemme *tabler* comme le verbe *tabler*. Pour discriminer les noms, les verbes, les adjectifs et les adverbes, une suite de critères est donnée. Les mots sont regroupés dans le lexique en fonction de ces critères. Les noms outre le préfixe et le lemme, ont comme critères :

- la catégorie des noms : N pour la *table* ;
- le morphème de dérivation : E qui nous donne *table* à partir du lemme *tabler* ;
- la sous-catégorie des noms communs : NCOM pour la *table* ;
- le genre : féminin (FEM) pour la *table*.

Les verbes outre le préfixe et le lemme, doivent être de catégorie V, être de type pronominal (PRONOM) ou actif (0), transitif ou intransitif (TRANS, INTR).

Les adjectifs et les adverbes sont codés de la même manière avec outre le préfixe et le lemme, le morphème de dérivation comme le nom, la catégorie syntaxique ADJOINT, les adjectifs étant alors de sous-catégorie ADNOM (adjoint du nom) et les adverbes de sous-catégoire ADVERB.

Voici un exemple de notre lexique sémantique pour un nom commun, un verbe, un adjectif et un adverbe :

1) L'entrée du nom féminin *table* est présentée ainsi dans le lexique sémantique dans la partie réservée aux noms (CAT=N) :

```
0 'tabler' E NCOM FEM : ADVAL='K5'; VECTEUR_SEMANTIQUE = %INDICES(ARITH:
828, 850, 855, 856, 851, 855, 232).
" : UL = 'tableau'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 828).
" : UL = 'meuble'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 850, 855, 856).
" : UL = 'couvert'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 851).
" : UL = 'cuisine'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 855).
" : UL = 't_astronomie'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 232).
```

Il présente cinq sens, *tableau*, *meuble*, *couvert*, *cuisine* et *terme d'astronomie*, avec chacun leur vecteur sémantique et le vecteur sémantique du nom *table* est alors la somme des vecteurs sémantiques des sens.

2) Le verbe *rentrer* est codé de la manière suivante dans la partie réservée aux verbes (CAT=V) :

```
0 'rentrer' VB 0 : ADVAL='K3'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 131,
203, 205, 727, 227).
" : UL = 'pénétrer'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 131, 203, 205).
" : UL = 'garder_pour_soi'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 727).
" : UL = 'heurter'; VECTEUR_SEMANTIQUE = %INDICES(ARITH: 227).
```

Il a trois sens, *pénétrer*, *garder pour soi* et *heurter*, et de la même manière, le vecteur sémantique du verbe *tabler* est la somme des vecteurs sémantiques des trois sens.

3) L'adjectif *autre* est codé de la façon suivante dans la partie réservée aux adjoints (CAT=ADJOINT) :

```
0 'autre' 0 ADNOM : VECTEUR_SEMANTIQUE = %INDICES(ARITH: 16, 22, 23).
```

4) L'adverbe *vraiment* est codé ainsi dans la partie réservée aux adjoints :

```
0 'vrai' EMENT ADVERB : VECTEUR_SEMANTIQUE = %INDICES(ARITH: 409).
```

### *3.2 Analyse sémantique*

Les textes sont d'abord traités par une analyse morphosyntaxique. Celle-ci calcule de façon classique les entités comme les mots, les syntagmes, les phrases et les textes auxquels nous associons les vecteurs sémantiques.

L'analyse sémantique correspond à une étape de cuisson supplémentaire des vecteurs sémantiques des mots du texte traité. Elle est faite non pas en fonction des notions du thesaurus mais en fonction de la structure syntaxique du texte traité. Les vecteurs sémantiques sont alors *recuits*.

La structure syntaxique du texte ordonne par niveaux les points suivants : le texte correspond à la racine de la structure syntaxique. Elle a comme fils, les points représentant les phrases des textes. Les points correspondant aux syntagmes – groupe nominal, groupe prépositionnel ou groupe verbal – se trouvent sous les points des phrases. Au niveau le plus bas de la structure, se trouvent alors les points – les feuilles – correspondant aux mots du texte.

Les fonctions *Fi* de cuisson des vecteurs sémantiques peuvent être variées, mais tiennent compte de cette structure syntaxique. Le but est de calculer en fonction du contexte du texte ou de la phrase, les vecteurs sémantiques recuits des mots.

La fonction *F1* de cuisson, consiste à ramener les vecteurs sémantiques des mots vers la racine de la structure pour pouvoir calculer l'ensemble du contexte sémantique du texte puis de raffiner les vecteurs sémantiques des mots en fonction de ce contexte.

Une deuxième fonction *F2* possible consiste à calculer le vecteur sémantique du mot au fur et à mesure de la construction de gauche à droite. Elle correspond à une construction séquentielle du contexte.

Une troisième fonction *F3* calcule les vecteurs sémantiques des syntagmes puis des phrases et du texte. Mais au lieu d'effectuer une comparaison directe des vecteurs sémantiques des mots avec les phrases et/ou le texte comme en *F1*, elle redescend vers le bas de la structure en affinant les vecteurs sémantiques des syntagmes et enfin, les vecteurs sémantiques des mots. Cette opération peut s'effectuer une ou plusieurs fois. Cela correspond à l'heuristique du *Recuit-Simulé.*

Dans les expérimentations actuelles, nous utilisons une combinaison de ces fonctions. L'objectif est d'intégrer à la structure syntaxique préexistante, les représentations sémantiques aux points de la structure syntaxique. Chaque point de la structure s'enrichit de la représentation sémantique vectorielle des points dont il est voisin structurellement (ascendant, descendant, frère, etc.) dans la structure. Les segments en incise ne perturbent donc pas les calculs de voisinage.

On intègre les vecteurs sémantiques cuits aux feuilles de la structure. L'augmentation sémantique de la structure syntaxique se fait ainsi d'abord du bas vers le haut, éventuellement de façon transversale (gauche-droite et droite-gauche) et enfin en redescente. La remontée des notions vers le haut de la structure s'effectue par une fusion des notions dans les vecteurs. Cette fusion est généralement une combinaison linéaire. Cette sommation dépend des fonctions syntaxiques des mots dans l'arborescence. Le vecteur obtenu à la racine de la structure est alors le vecteur représentant le contexte sémantique de l'ensemble du texte.

Certains problème se posent pour les lexèmes complexes non figés. Un tel lexème est un groupe de mots dont l'acception peut être différent suivant le contexte : il peut être le regroupement des sens de chaque mot le composant ou non. Dans ce dernier cas, son sens n'est pas calculable. Le groupe « avoir le bras long » a un sens différent suivant la phrase : « Il a le bras long et se sert de cette influence. » ou la phrase « Il a le bras long et bronzé. »

Deux choses sont à considérer à ce niveau : le point peut-il intégrer ou non le vecteur sémantique et si oui, comment l'intègre-t-il ? Le vecteur sémantique importé du lexique sémantique, est intégrable au point considéré si tous les descendants du même point sont marqués. Dans le cas où cette intégration peut se faire, elle est constituée par une combinaison linéaire des vecteurs sémantiques des points-fils plus le résultat de la lecture du lexique sémantique. Cette lecture se justifie par le fait qu'un lexème complexe non figé peut être indexé dans le lexique. Ce résultat est soit (1) nul si aucun schéma n'est reconnu appartenant au lexique sémantique, soit (2) le vecteur sémantique si le schéma est reconnu.

Dans le cas où un élément non marqué $x$ a son vecteur sémantique intégrable, la valeur de ce vecteur sémantique V est alors :

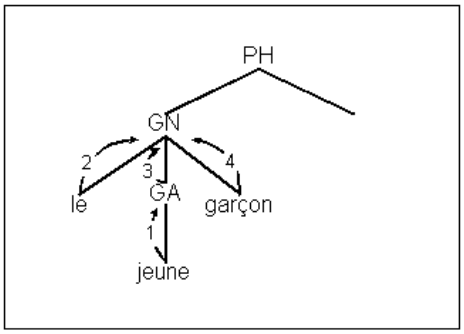
$$V = Lecture(x, L) + \sum_{y \in D} k \times V(y)$$

où :

- *Lecture* est une fonction qui nous renvoie le tuple correspondant au vecteur sémantique à intégrer dans l'étiquette vectorielle du point $x$ de l'arborescence considérée. Ce tuple est lu dans le lexique sémantique L ;

- D = Descendants(x) ;

- $k = 1$ si la fonction syntaxique du point associé au vecteur est gouverneur du groupe considéré et $\alpha$ ($\alpha$<1) sinon. En effet, un point particulier, dans un syntagme nominal, a une fonction privilégiée, celle de *gouverneur*. Les autres sont *gouvernés*. Le vecteur sémantique *V* représentant le groupe les contenant, doit donc tenir compte de cette particularité. Le vecteur sémantique est ensuite normé.




La structure syntaxique *se sémantise* selon la propagation des vecteurs. Cela constitue l'augmentation lexicale sémantique. Cette étape est illustrée par la figure ci-dessus pour un exemple de groupe nominal (GN) dans une phrase (PH). Les flèches numérotées de 1 à 4 signifient l'ordre de remontée des vecteurs sémantiques des mots dans le groupe nominal "*le jeune garçon*". Le vecteur sémantique du mot *le* est nul car c'est un déictique. Le vecteur sémantique de l'adjectif *jeune* est affaibli de $\alpha$ dans le groupe nominal par rapport au vecteur du mot *garçon*. La redescente est similaire. Ce processus peut être itéré. Le coefficient $\alpha$ peut varier.

### *3.3 Désambiguïsation sémantique*

La recherche du sens d'un mot polysémique est la sélection des sens indexés dans le lexique sémantique. L'idée est de comparer chaque sens indexé avec le vecteur recuit et aussi les contextes de la phrase (contexte local) ou du texte (contexte global). Ce sont les vecteurs sémantiques associés aux points représentant le texte et la phrase.

Nous souhaitons que cette comparaison soit exprimée par une distance mathématique dans notre espace vectoriel *E*. Plusieurs choix de fonctions de distances s'offrent à nous. Outre la distance euclidienne (dont l'interprétation est difficile), nous pouvons aussi calculer la distance en termes de longueurs de chemins dans la hiérarchie du thesaurus (dont l'interprétation est encore plus douteuse et surtout dépend de l'ontologie).

Une bonne fonction de distance doit être liée à la proximité sémantique. Celle-ci peut raisonnablement être représentée par le cosinus de l'angle de deux vecteurs. Plus deux entités – mot, syntagme, phrase ou texte – évoquent les mêmes idées, plus leurs vecteurs associés sont colinéaires. C'est-à-dire que $\forall\ V1, V2 \in E$, $cosinus(V_1, V_2)$ tend vers 1. Or le *cosinus* dans notre espace vectoriel n'est pas une distance au sens mathématique (car *cosinus*(x, x) $\neq$ 0).

Deux remarques sont à considérer sur notre espace vectoriel *E* : tous les vecteurs sont normés à 1 et tous les coefficients de nos vecteurs sémantiques sont positifs ou nuls (mais jamais négatifs). Par conséquent, l'angle que forment deux vecteurs sémantiques quelconques est compris entre 0 et $\pi/2$. La fonction *ArcCosinus* constitue alors une distance au sens mathématique dans *E*. La fonction de distance *dist* que nous utilisons est donc définie ainsi :

$$\forall V_1,\ V_2 \in E,\ \mathrm{dist}(V_1,\ V_2) = \mathrm{ArcCosinus}(V_1,\ V_2)$$

Le but est de pouvoir comparer chaque vecteur sémantique d'un point avec le contenu des lexiques sémantiques, et ainsi de pouvoir sélectionner le sens le plus proche.

### *3.4 Transfert lexical sémantique*

Le transfert lexical consiste à choisir le mot de la langue cible en fonction du mot de la langue source. Cette opération est généralement faite par l'utilisation de dictionnaires de correspondance dont l'indexation est lourde en termes de codage manuel. Nous voulons en fait nous affranchir de cet effort énorme en temps et en compétence. Pour cela, il est nécessaire de disposer d'une structure pivot entre les deux langues qui puisse nous permettre

de calculer le mot correspondant en langue cible. L'espace vectoriel que nous avons construit pour l'analyse sémantique constitue une telle structure pivot.

Ceci nous conduit à la réflexion suivante sur les champs notionnels des thesauri des deux langues : existe-t-il une base notionnelle commune à plusieurs langues ? Ou encore, peut-on passer d'une base de notions pertinentes pour une langue à une autre base dans une autre langue (Spark Jones 1986) ? Mathématiquement on peut répondre par l'affirmative, il s'agit d'une application linéaire entre deux espaces vectoriels. Linguistiquement, nous laisserons cette question en suspens. Toutefois, nous basons notre expérience sur l'hypothèse que l'espace vectoriel de notions de la langue source peut approximer de façon satisfaisante une base de notions pour la langue cible.

Nous avons besoin pour cela de disposer d'un lexique sémantique dans la langue cible pour pouvoir en sélectionner les termes. Un jeu restreint de mots monosémiques et polysémiques a été constitué dans un lexique sémantique de la langue cible (anglais) à des fins expérimentales. Il semble toutefois possible d'envisager l'automatisation partielle de la création de ce lexique sémantique de la langue cible à partir d'un noyau générateur et de bases de textes analysés en langue source accompagnés de leur analyse et traduction en langue cible.

L'idée de notre transfert lexical est de sélectionner un terme dans la langue cible dont le vecteur minimise la fonction de distance *dist* (définie dans la section 3.3). Si la distance calculée excède un seuil de confiance, on peut raisonnablement douter de la validité de la sélection. Il se peut alors que le terme source soit culturellement très spécifique ou que le lexique sémantique cible est incomplet.

Quoi qu'il en soit, la fonction *dist* respecte la notion de *proximité sémantique* entre les mots qu'ils soient de la langue source ou de la langue cible, puisqu'ils sont tous dans le même espace vectoriel.

La méthode de recherche du terme équivalent repose sur la comparaison du vecteur sémantique du mot de la langue source avec les vecteurs sémantiques de la langue cible. Plutôt que calculer la distance du mot de la langue source avec tous les mots de la langue cible, il vaut mieux calculer la distance avec certains mots qui serviront de borne de recherche. La méthode pour retrouver le vecteur sémantique du mot cible repose sur une recherche rapide dichotomique.

## 4. Premiers résultats, perspectives et conclusion

L'approche proposée ici est en cours d'évaluation et les lexiques sémantiques sont indexés conjointement. Les premiers résultats sont particulièrement encourageants. En effet, nous constatons que la sélection d'un terme en langue cible est plus satisfaisante que dans le cas d'un lexique de transfert classique. Il ne pose pas de problème pour les termes monosémiques. Pour les termes polysémiques, soit le contexte est suffisant pour sélectionner le bon équivalent, soit le résultat revient à choisir au hasard. De plus, chaque lexique sémantique étant indépendant, ils peuvent être utilisés pour une rétrotraduction ou pour d'autres applications comme la traduction automatique.

Nous avons testé notre approche sur un corpus très réduit de phrases contenant des termes typiquement ambigus (du type de celles présentées dans cet article). Dans l'état actuel de l'indexation, nous arrivons à sélectionner le sens et le terme correct dans 2/3 des cas. Avec un lexique de transfert classique, la sélection se serait faite au hasard.

Les cas typiques d'échec (voire de contresens) sont ceux où ce n'est pas le contexte lexical qui permet de faire la sélection, mais une interprétation liée un à modèle du monde. Le contexte lexical peut aussi renforcer l'erreur. Un exemple typique serait « L'intérêt qu'il retire de son étude sur les banques est évident ». Pour le terme « intérêt », le sens de « profit/somme d'argent » se trouve particulièrement renforcé à cause de la présence de « banques ».

Nous comptons poursuivre et affiner notre étude. D'une part, l'indexation du lexique sémantique du français et de l'anglais doit se poursuivre. L'évaluation doit être poursuivie afin de correctement identifier les causes d'échec et y remédier si possible. D'autre part, nous

cherchons à exploiter cette voie pour l'aide à la résolution d'anaphores, à la sélection de préférences syntaxiques et à l'identification automatique de locutions.

Nous avons présenté dans cet article, une approche de TALN basée sur des vecteurs sémantiques. De tels vecteurs représentent avec finesse l'ensemble des idées évoquées dans un segment textuel. Il est ainsi souvent possible, en cas d'ambiguïté lexicale de sélectionner le sens correct d'un terme. Dans le cas de la traduction automatique, le transfert lexical s'en trouve amélioré et ce sans avoir à développer de lexique de correspondance langue source/langue cible. Cette approche partage les qualités des systèmes par pivot tout en s'affranchissant des difficultés de la définition et du calcul des formes pivots.

## Références

Baylon C. et Mignot X. (1995) *Sémantique du language : initiation*. Editions Nathan, Paris.

Chauché J. (1984) *Un outil multidimensionnel de l'analyse du discours*. Dans Proceedings of the International Conference COLING'84, Standford, July 1984.

Kirkpatrick B. (1986) *Roget's Thesaurus of English Words and Phrases*. Penguin Books, 1997.

Larousse (1992) *Thésaurus Larousse – des idées aux mots – des mots au idées*. Larousse, ISBN 2-03-320-148-1, 1992.

Rastier, F. (1987) *Langages : Sémantique et Intelligence Artificielle*. Bernard Willerval Jouve, Septembre 187.

Resnik P. (1995) *Using Information content to evaluate semantic similarity in a taxonomy*. Dans Proceedings of IJCAI-95, 1995.

Salton G. (1988) *Term-Weighting Approaches in Automatic Text Retrieval*. McGraw-Hill computer science serie. McGraw-Hill, Volume 24, 1988.

Salton G. et M. J. MacGrill (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill computer science serie. McGraw-Hill, New-York, 1983.

Sowa J. F. (1984) *Conceptual Structure*. Addison Wesley, 1994.

Sparck Jones K. (1986) *Synonymy and Semantic Classification*. Edinburgh Information Technology Serie, 1986.