Application des vecteurs sémantiques à la fouille de texte

Jacques Chauché

LIRMM-CNRS et Université Montpellier 2
161 rue Ada, 34395 Montpellier cedex 5
chauche@lirmm.fr
http://www.lirmm.fr/ chauche

Résumé. L'approche présentée ici se base sur un traitement du contenu syntaxico-sémantique par un analyseur du Français, le système SYGFRAN(SYGFRAN) , pour retrouver un ensemble de phrases appartenant à différents discours du président François Mitterrand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Ce traitement se fait par calcul de vecteurs sémantiques de phrases (méthodologie définie dans l'article) et par la définition d'une relation de similitude décrivant l'inclination de vecteurs dont l'inclinaison, ou distance angulaire, est proche. A l'aide de cette relation, des phrases sont attribuées par le système à l'un ou l'autre des auteurs, et l'article indique des F-mesures obtenues sur le premier corpus, dit d'apprentissage, légèrement supérieures à 80%.

1 Présentation

Le défi DEFT 2005 organisé pour le congrès annuel TALN consiste à retrouver un ensemble de phrases appartenant à différents discours du président François Mitterrand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Les phrases introduites traitent d'une thématique distincte de la thématique retenue pour les phrases des discours de Jacques Chirac. Il est donc possible d'aborder ce problème par la recherche d'une distinction sémantique. L'approche présentée ici se fonde sur un traitement du contenu par opposition aux traitements habituels basés sur des approches statistiques. Le vocabulaire utilisé par l'un ou l'autre n'aura d'importance qu'à travers les idées qu'il véhicule. L'originalité de cette approche vient du fait qu'elle s'appuie sur la structure syntaxique du texte. L'analyse syntaxique produit cette structure syntaxique pour les groupes, les phrases et le texte. L'obtention de cette structure est très difficile et le taux de construction complète se situe autour de 30 %. Pour qu'une analyse soit complète il est nécessaire que la ou les structures (en cas d'ambiguïté) aient une seule racine. Dans les autres cas l'analyse produit une structure partielle qui permet néanmoins le calcul d'un vecteur sémantique avec une légère erreur. Dans ce cas le vecteur sémantique est obtenu à partir des groupes fonctionnels. Le taux de reconnaissance des groupes fonctionnels et des fonctions syntaxiques se situe autour de 70 % d'après les résultats de l'évaluation EASY. L'exposé de cette approche comprend trois parties : la méthode d'analyse syntaxique, la construction d'un vecteur sémantique associée à une structure et enfin l'utilisation de la suite des vecteurs pour identifier les passages rechercher.

L'analyse morphosyntaxique produit pour chaque texte une structure. L'analyse sémantique déduit de cette structure un vecteur sémantique. Le calcul décisionnel s'appuie alors sur ces vecteurs pour définir une appartenance. La méthode d'analyse qui est à la base comprend trois parties : l'analyse morphologique, l'analyse syntaxique et l'analyse sémantique. Ces trois analyses s'enchaînent dans cet ordre.

2 État de l'art

La méthode mise en oeuvre dans le cadre de DEFT'05 a des caractéristiques linguistiques qui se distinguent des méthodes statistiques utilisées pour des problématiques liées à la classification de textes, et également d'autres méthodes principalement fondées sur la prépondérance lexicale (importance des mots).

Pour les méthodes statistiques les plus connues, citons principalement les suivantes :

- La méthode Rocchio(Rocchio 1971) utilise un vecteur représentant le document, ce vecteur est directement issu du vocabulaire (habituellement de type TF-IDF de Salton).
- La méthode Naïve Bayes est basée sur un modèle probabiliste axé également sur le vocabulaire.
- Enfin la méthode SVM(T. Joachims 2002) (machine à vecteur de support) utilise un changement dereprésentation du texte avec une lemmatisation possible de ce dernier (introduction des catégories grammaticales)

Il en existe certes bien d'autres, comme les arbres de décision (Breiman et al. 1984) ou le clustering (Romesburg 2004). Si ces méthodes font leurs preuves quand il s'agit d'une classification globale en classification d'images ou de séquences de gènes, ou quand il faut s'appuyer sur une hiérarchie pré-ordonnée de concepts ou des ontologies, elles négligent le fait essentiel qu'un document en langage naturel est un objet construit pour l'esprit, et qu'il comprend une structuration forte, dénotée par la syntaxe. Or dans toutes ces méthodes, tout se passe comme si un texte était un 'sac de mots' (et le terme anglais *words bag* est d'ailleurs utilisé comme fondement hypothétique), ce qui est loin d'être le cas. Dans (Chauché et al. 2003) nous avons montré que le réel **contenu** d'un texte ne pouvait être obtenu que par une prise en compte des informations sémantiques véhiculées par les constructions syntaxiques, et que ce contenu était un gage de bonne classification lorsque l'on s'attaque aux textes d'une certaine taille et d'une certaine complexité.

Il existe certes des méthodes issues soit de la recherche documentaire, soit du TALN (Traitement Automatique du Langage Naturel) qui tentent de tenir compte d'aspects morphologiques et syntaxiques de manière locale, se fondant sur une analyse de surface, mais elles restent majoritairement couplées à des approches statistiques comme (Manning 1999). La syntaxe est utilisée comme un filtre de représentation (Besançon 2002), et on en trouve quelques traces en ce qui concerne les groupes nominaux, prépositionnels ou adjectivaux, dans les tâches liées à l'extraction ou à l'indexation terminologiques (Bourrigault 1993), (Daille 1995).

Le problème de l'usage de la syntaxe est profondément lié aux performances des analyseurs syntaxiques. En effet, si les analyseurs morphologiques (*POS Taggers*) atteignent depuis une dizaine d'années une certaine stabilité et une qualité non négligeables, les analyseurs syntaxiques restent à l'état de produits instables, compte tenu de la réelle complexité de l'analyse en profondeur. Ayant développé un analyseur possédant la capacité de produire les arbres d'analyse en constituants et en dépendances, nous nous sommes appuyés sur cette particu-

larité pour produire des informations complémentaires en provenance de la structure. C'est pourquoi, ce travail a pour but de faire apparaître à la fois les avantages et les inconvénients d'une approche qui tient compte de toutes les informations qu'un texte peut contenir, et pas seulement d'une approche fondée sur la prépondérance lexicale.

3 Analyse morphologique

L'analyse morphologique a pour but de déterminer toutes les caractéristiques d'un mot donné, c'est-à-dire d'une chaîne de caractères isolée par des espaces, tabulation ou changement de ligne. (Nous ne traitons pas les césures). Le principe de l'analyse morphologique repose sur la notion de segment : prefixe,infixe,suffixe. L'analyse énumère toutes les décompositions possibles d'un mot en segments définis dans un dictionnaire de segments. Des règles d'accords associées aux segments permettent de limiter la combinatoire et surtout les fausses interprétations. Ainsi avec le mot 'changes' nous aurons les solutions suivantes :

- 'chang' (racine verbale) suivit de 'es' (suffixe verbal conjugaison du présent 2^{eme} personne du singulier) → verbe conjugué
- 'chang' (racine verbale) suivit de ' e' (infixe de dérivation nominale) suivit de 's' (suffixe nominal de marque du pluriel) → nom commun

Les racines verbales sont extraites du "Bescherelle 1 La Conjugaison" avec une factorisation spécifique. Dans ce cas nous aurons une racine verbale 'change' qui ne produira pas de solution car elle impose certains suffixes comme 'ons', 'ais', ... Le resultat de l'analyse morphologique donne donc pour chaque mot l'ensemble des interprétations syntaxiques possibles. A ce stade la désambiguïsation sémantique n'est pas nécessaire et un mot comme 'gratin' n'aura qu'une seule solution : nom commun.

4 Analyse syntaxique

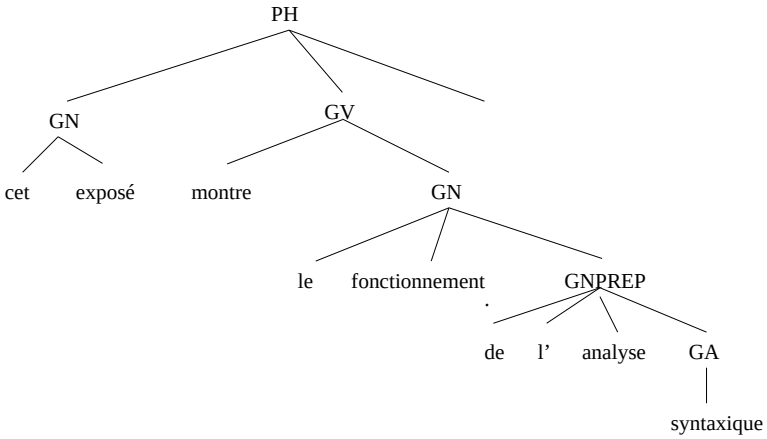
L'analyse syntaxique doit construire la structure syntaxique d'une phrase puis d'un texte. L'analyse doit être robuste, c'est-à-dire qu'elle doit fournir un résultat, même partiel, dans tous les cas. La campagne d'évaluation sur les analyseurs syntaxiques concomitante au defi Deft05 a montré que la construction d'un analyseur syntaxique est loin d'être achevée. L'analyseur syntaxique utilisé n'est pas défini à partir d'une grammaire générative (catégorielle ou syntagmatique). La principale raison est d'ordre algorithmique. En définissant une grammaire générative l'auteur de cette grammaire doit inclure dans ses règles des contraintes spécifiques à l'environnement d'application de chaque règle. Etant donné la diversité naturelle de la langue, nous ne pensons pas qu'une telle démarche soit réaliste. L'approche définie ici correspond à une construction algorithmique d'un filtre constructiviste de structures partielles. Cette construction s'effectue par la définition de règles de réécriture sur des ensembles de structures. Ainsi nous ne cernons pas la construction de la langue mais définissons une fonction qui sur la langue doit produire le résultat attendu. Le modèle théorique des systèmes de réécriture sur des mots a été défini par les algorithmes de Markov. Ces systèmes ont la capacité calculatoire maximale et toute fonction Turing calculable est exprimable par un algorithme de Markov(Mendelson 1964). L'extension apportée n'augmente donc pas la capacité de calcul mais apporte une plus grande aisance dans l'écriture des algorithmes. L'analyseur utilisé est

écrit en "SYGMART" et comprend environ 200 grammaires et 10000 règles. Une grammaire est la traduction d'un algorithme de Markov et les grammaires sont composées à l'intérieur d'un réseau. Les éléments manipulés sont donc des structures de même nature que le résultat souhaité.

Un élément manipulé par le système est un triplet $\{\mathfrak{A}, \mathfrak{E}, \mathfrak{f}\}$ appelé élément structuré. $\mathfrak{A}$ est un ensemble fini d'arborescences neutres, c'est-à-dire sans étiquetage particulier des points. $\mathfrak{E}$ est un ensemble d'étiquettes, chaque étiquette étant une collection de couples attributs/valeurs. $\mathfrak{f}$ est une fonction de l'ensemble des points de $\mathfrak{A}$ dans $\mathfrak{E}$. Cette fonction est partout définie sur les points de $\mathfrak{A}$ et n'est pas obligatoirement injective (deux points de $\mathfrak{A}$ peuvent faire référence à la même étiquette) ni surjective (du fait qu'une étiquette de $\mathfrak{E}$ peut faire référence à une autre étiquette de $\mathfrak{E}$ une étiquette peut être atteinte en dehors de la fonction $\mathfrak{f}$). Une structure syntaxique correspond donc bien à une forme particulière de ces objets.

Le résultat de l'analyse morphologique est placé dans un élément structuré : chaque mot correspond à un point d'une arborescence linéaire et chaque solution associée à un mot correspond à une étiquette distincte dont un point dépendant du mot lui fait référence par la fonction $\mathfrak{f}$. Souvent par abus de langage on identifie le point et l'étiquette qui lui est associée par la fonction $\mathfrak{f}$.

L'analyse s'effectue par la recherche de schéma suivi de la transformation de ce schéma. Par exemple l'analyse de la phrase "Cet exposé montre le fonctionnement de l'analyse syntaxique." produit la structure suivante :



les différentes étapes de cette construction sont données en annexe.

5 Analyse sémantique

Les vecteurs sémantiques sont une modélisation de la notion linguistique de champ sémantique. Les vecteurs sémantiques sont définis par la partie positive d'un espace vectoriel (le contraire d'une idée est également une idée et ne pourrait être associée à une valeur négative ou opposée. Avec deux idées antinomiques, laquelle se verrait attribuer une valeur positive et laquelle une valeur négative ?). En sémantique lexicale, un champ sémantique est constitué d'un ensemble d'unités lexicales que l'on considère comme doté d'une organisation structurelle

sous-jacente (A.J. Greimas et J. Courtes 1993). Cette notion se complète par la charge sémantique qui est constituée des investissements sémantiques susceptibles d'être distribués, lors de la réalisation dans une langue naturelle, sur les différents éléments constitutifs de l'énoncé linguistique.

La définition des vecteurs sémantiques correspond donc à la projection dans un espace vectoriel de la notion de champ sémantique. Cette projection correspond à l'investissement sémantique qui est la procédure par laquelle une structure syntaxique donnée se voit attribuée des valeurs sémantiques préalablement définies. Le problème devient alors celui de la définition de l'espace. Cette espace doit avoir une généralité maximale afin de couvrir l'ensemble du langage. Il doit également être défini par un ensemble fini d'éléments constitutifs. Il n'est bien sûr pas possible de définir une base de cet espace. Aussi l'espace sémantique sera approché par une famille génératrice. Chaque terme sera alors défini dans cet espace.

5.1 Vecteur de terme

Définition
Un vecteur sémantique projette un terme donné dans un espace sémantique dont une famille génératrice correspond à un ensemble d'idées.

L'ensemble des idées nécessaires pour former une famille génératrice peut être défini par un thésaurus.

La procédure est la suivante : on projette la totalité des lexies du dictionnaire sur un espace défini à partir d'une famille de concepts "à la Roget" (Roget 1852). Pour le Français, les lexicologues du Larousse ont défini une famille de 873 concepts hiérarchisés en 4 niveaux (Larousse 1992). Sur un plan vectoriel, cela produit un espace de dimension inférieure ou égale à 873. Les approches à la "Roget" sont relativement nombreuses depuis quelques années, dans la littérature anglo-saxonne, (Yarowsky 1992), (Ellman et Tait 1999). En Français, l'indexation automatique à partir du thésaurus a été proposée à l'origine par nous-mêmes, (Chauché 1990), mais on la retrouve aujourd'hui utilisée dans de nombreux travaux (Crestan et al. 2003).

Formellement, on considère que tout terme t du dictionnaire est représenté par un vecteur $\vec{t}$ dans l'espace vectoriel considéré, que l'on nommera $\vec{\mathcal{V}}$. On suppose qu'il existe une application qui plonge l'espace lexical linguistique dans l'espace vectoriel engendré par la famille de concepts du thésaurus.

Dans cet espace, le vecteur nul correspond à l'abscence de mot, ce qui en pratique ne sert à rien puisque l'on ne s'intéresse pas à la sémantique du mot vide.

Le deuxième aspect correspond à l'interprétation des vecteurs et par conséquence la forme calculatoire. Une idée exprimée par un mot ou une phrase sera d'autant plus associée à une idée génératice que leurs vecteurs respectifs seront colinéaires. L'importance doit donc être portée sur les distances angulaires et non pas sur leurs normes respectives.

Les calculs sur les vecteurs feront appel à des combinaisons linéaires des différents vecteurs représentant le texte. Mais l'importance d'un élément doit être donnée par la fonction syntaxique qu'il représente et non pas par l'importance de sa composition. Aussi tous les termes d'une combinaison linéaire entrant dans le calcul d'un élément devront avoir la même intensité.

Afin de simplifier les calculs nous utiliserons une version normée $\vec{t}_{nor}$ de chaque vecteur et nous produirons toujours un résultat normé. Comme on ne traite que de vecteurs normés, par convention, on écrira $\vec{t}$ pour désigner le vecteur normé du terme t. Pour cela, on introduit

une norme euclidienne sur l'espace vectoriel sémantique. Le traitement s'effectuera toujours sur l'hypersphère unité.

La majorité des mots étant polysémique, chacun renvoie à une multiplicité d'idées, ou concepts du thésaurus.

Les idées associées au mot *calcul* sont par exemple : Calcul, Opération arithmétique, Maladie et Intention.

L'emploi de ce mot simplement ne permet donc pas de définir sa signification : par exemple, *calcul arithmétique* ou *calcul biliaire,* ou *Il m'a aidé par calcul*.

Cela signifie que le terme doit être représenté, non seulement par la manière dont il est indexé dans le thésaurus, mais aussi par ses différentes significations, qui elles, ont un sens lorsque le mot est utilisé dans une construction (groupe ou phrase).

Le calcul sémantique sur une phrase doit donc incliner le sens du mot c̈alculv̈ers une des significations possibles.

5.2 Vecteur sémantique d'une phrase

Définition

On dira que l'on représente toute *phrase* construite, par un vecteur produit comme une combinaison linéaire de vecteurs sémantiques des *groupes* qui la composent.

On dira que l'on représente tout *groupe* construit, par un vecteur produit comme une combinaison linéaire de vecteurs sémantiques des *termes* qui le composent.

Pour cela on introduit les opérations suivantes :

Somme normée : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ représentant les vecteurs (normés) de deux termes t_1 et t_2.

$$\overrightarrow{(t_1 + t_2)_{nor}} = \frac{\vec{t_1}+\vec{t_2}}{\|\vec{t_1}+\vec{t_2}\|} \tag{1}$$

Remarque : la somme normée n'est pas associative :

$\overrightarrow{(t_1 + t_2 + t_3)_{nor}}$ n'est pas égal à $(\overrightarrow{(t_1 + t_2)_{nor}} + \vec{t_3}))_{nor}$. Par convention, on ne retiendra comme opération de somme que la somme normée, et on omettra dorénavant l'indice 'nor'.

Combinaison normée Pour le calcul d'un groupe nous utiliserons une combinaison linéaire avec normalisation du résultat. Si $\vec{t_i}$ sont les vecteurs associés aux composants du groupe et i sont les coefficients associés aux fonctions des t_i alors :

$$combi(\vec{t_i}) = \frac{\sum \alpha_i * \vec{t_i}}{\| \sum \alpha_i * \vec{t_i}\|} \tag{2}$$

Cette fonction constitue l'opération élémentaire du calcul sémantique. Pour deux vecteurs et des coefficients égaux à 1, cette opération correspond à la somme normée.

Multiplication par un scalaire : Soit un vecteur $\vec{t}$ normé. Soit λ un scalaire. Le vecteur $\vec{\lambda t}$ est égal à $\lambda * \vec{t}$. Cela signifie que toutes les composantes du vecteur sont multipliées par le scalaire.

Remarque : cette multiplication a pour objectif de renforcer la "présence" du vecteur dans une combinaison linéaire, et ne s'utilise en principe jamais isolément.

Produit terme à terme : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ normés. Le produit terme à terme des deux vecteurs se définit comme :

$$\overrightarrow{(t_1 * t_2)_{nor}} = \frac{\vec{t_1} * \vec{t_2}}{\|\vec{t_1} * \vec{t_2}\|} \tag{3}$$

où si $a_{p,i}$ est la ième composante de $\vec{t_1} * \vec{t_2}$, et $a_{1,i}$ et $a_{2,i}$ respectivement celles de $\vec{t_1}$, et $\vec{t_2}$, on a :

$$\forall i \in [1, 873], a_{p,i} = a_{1,i} * a_{2,i} \tag{4}$$

Par convention, on omettra l'indice nor et on appellera par défaut $\overrightarrow{(t_1 * t_2)}$ le produit terme à terme normé.

L'interprétation sémantique du produit est évidente : elle permet d'incliner le sens d'un mot vers une sens majoritaire du texte tout en ne concervant, pour ce mot, que les idées initiales :

Si x_i et y_i sont deux composantes non nulles d'un vecteur sémantique du terme t dans le texte T. Initialement ces deux composantes ont la même valeur. Ces composantes seront nécessairement non nulles dans le vecteur sémantique calculé pour le texte T. Si les concepts associés aux composantes x_i et y_i sont présents avec des poids différents dans le texte (ils apparaîtront dans un nombre de mots différents ou sur des éléments ayant des fonctions syntaxiques différentes), alors le produit du vecteur texte par le vecteur du terme t aura pour effet : de ne pas ajouter de concepts sous-jacent à t (les composantes correspondantes étant nulles, le résultat est nul) et de renforcer la composante la plus importante dans ce texte. Ainsi le sens du terme t sera actualisé par rapport à son environnement textuel.

Distance "angulaire" : La distance selon Salton, servant de mesure de similarité est calculée comme le *cosinus* de l'angle de deux vecteurs.

$$sim(\vec{t_1}, \vec{t_2}) = \cos \widehat{\vec{t_1}, \vec{t_2}} = \frac{\vec{t_1}.\vec{t_2}}{\|\vec{t_1} * \vec{t_2}\|} \tag{5}$$

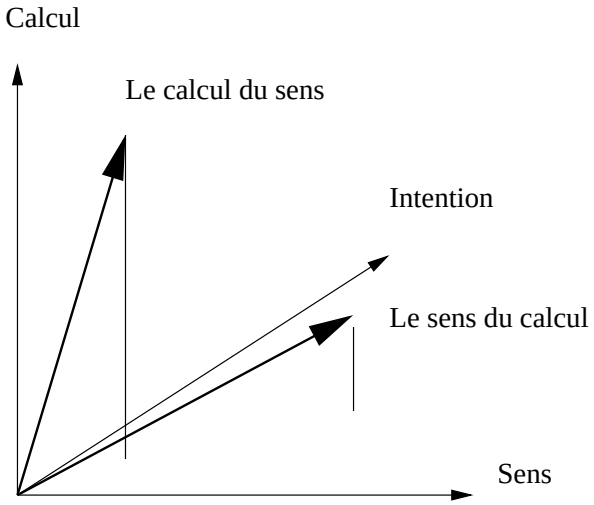
où "." est le produit vectoriel classiquement défini. La distance que nous utilisons correspond à une mesure relative à l'angle $\widehat{\vec{t_1}, \vec{t_2}}$. Comme nous ramenons tous les angles considérés à l'espace $[0, \frac{\pi}{2}]$, alors la mesure que nous proposons se calcule par :

$$\delta(\vec{t_1}, \vec{t_2}) = 1 - \cos \widehat{\vec{t_1}, \vec{t_2}} \tag{6}$$

Remarques : Ramener les valeurs de δ à $[0, 1]$ est plus pratique que de mesurer des valeurs entre 0 et $1,67$ radiants. Lorsque deux vecteurs sont totalement divergents (intersection vide), leur angle est de $\frac{\pi}{2}$, et le cosinus vaut 0 : leur distance est maximale et vaut 1. Lorsque ces vecteurs sont très proches, leur angle tend vers 0 , le cosinus tend vers 1 et la distance, vers 0. Tous les vecteurs ont un angle forcément compris entre 0 et $\frac{\pi}{2}$, par construction, et appartiennent au même espace vectoriel.

5.3 Vecteur de groupe

La deuxième propriété du calcul sémantique correspond à une définition différenciée d'un groupe suivant sa structure. Ainsi le sens du groupe "le calcul du sens" est distinct du sens du groupe "le sens du calcul",ces deux groupes ayant rigoureusement les mêmes éléments (le langage naturel n'étant pas commutatif). Comme le mot ësensëst très riche sémantiquement (une vingtaine de sens justement) nous prendrons pour l'exemple de la représentation l'idée associée : Sens. L'idée est différente du terme, selon les lexicologues, en ce qu'elle étiquette un champ sémantique. Le terme peut appartenir ou relever de plusieurs champs, en raison de sa polysémie. Dans le sous-espace ayant comme axe *Calcul*, *Intention* et *Sens* les vecteurs associés aux deux groupes précédents seront :



Le vecteur associé à un groupe est obtenu par une combinaison linéaire des vecteurs associés aux éléments de ce groupe. Les coefficients de cette combinaison linéaire dépendent de la fonction syntaxique de l'élément : gouverneur du groupe, sujet, objet, etc...

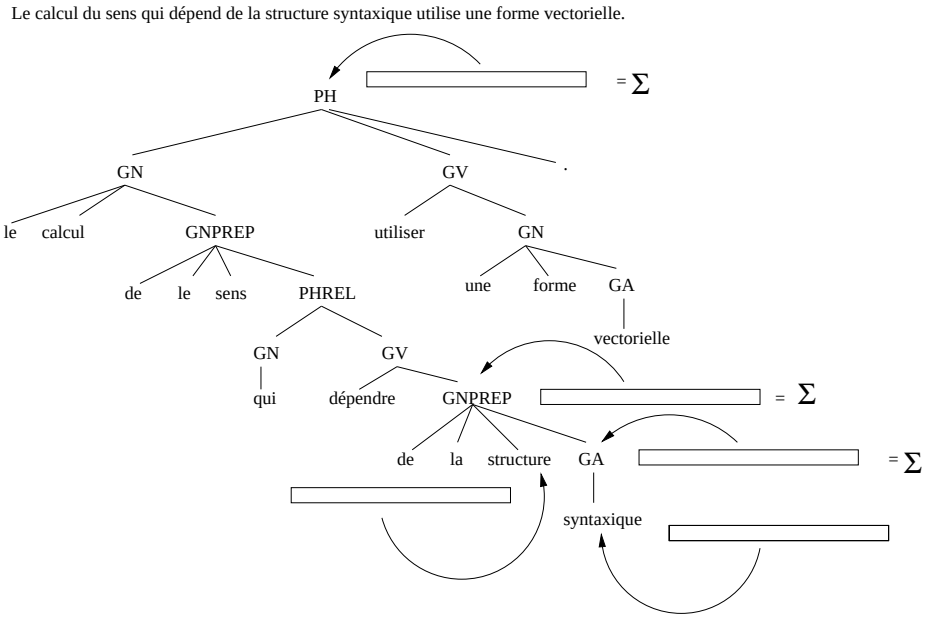
5.4 Calcul du vecteur de phrase

Le calcul d'un vecteur de phrase s'effectue (sur une) en plusieurs étapes à partir de la structure syntaxique :

– La première étape consiste à associer à chaque feuille un vecteur sémantique issu de la lecture d'un dictionnaire (vecteur de terme)

Si un élément a plusieurs sens ou interprétations possibles, le vecteur associé correspond au *centroïde* de l'ensemble des vecteurs associés à chaque interprétation (somme normée de tous les vecteurs indexant ce terme).

– La deuxième étape consiste à calculer récursivement le vecteur associé à chaque groupe.



– La troisième étape actualise les vecteurs associés aux feuilles. Cette actualisation consiste à effectuer un produit terme à terme du vecteur à actualiser avec le vecteur obtenu du texte.

Cette actualisation terminée un nouveau calcul est effectué. La convergence est très rapide et deux itérations suffisent pour obtenir un vecteur significatif.

5.5 Propriétés du modèle

Le classement s'effectue à partir des vecteurs sémantiques de phrases.

La valeur intrinsèque de la norme d'un vecteur n'est pas significative. Seule compte l'inclinaison de ce vecteur par rapport à une idée ou un autre vecteur donné (la distance angulaire). Aussi tous les calculs se termineront par la normalisation des vecteurs. On ne considérera donc que les points de la sphère unité.

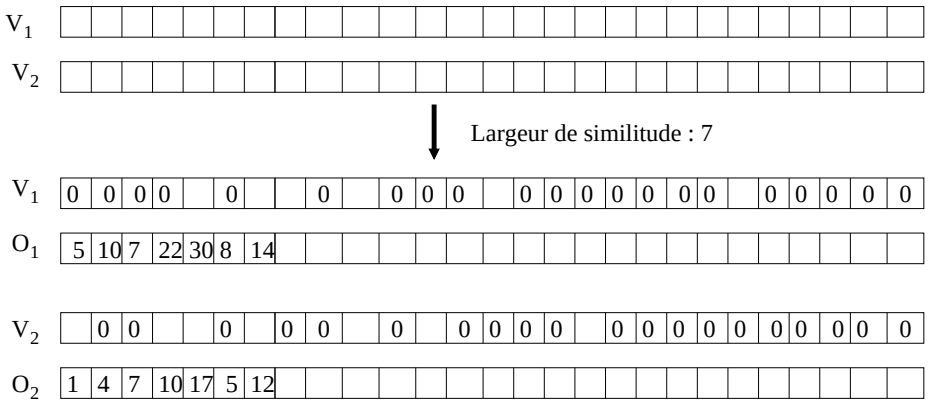
5.5.1 Comparaison d'inclinaison entre vecteurs

La première mesure de comparaison sera donc la valeur de l'arc séparant deux vecteurs sur cette sphère (Cette mesure sera donc naturellement donnée par la fonction *arcosinus* ($\vec{V_1}, \vec{V_2}$)). Comme toutes les composantes de tous les vecteurs sont positives ou nulles nous utiliserons seulement le produit scalaire. Dans ce contexte un élément sera plus proche d'un autre par rapport à un troisième si le produit scalaire de cet élément avec le troisième a une valeur supérieure au produit scalaire du deuxième avec le troisième.

Le produit scalaire ne rend que très imparfaitement **l'inclination** d'un élément par rapport à l'autre. En effet, l'inclination comprend *l'inclinaison*, mais indique jusqu'à quel point un vecteur "s'assimile" à un autre. Aussi le produit scalaire sera completé par une mesure de *similitude* tenant compte de l'importance relative de chaque idée à l'intérieur de chaque vecteur.

5.5.2 Similitude entre vecteurs d'inclinaison proche, ou mesure d'inclination

Le calcul de la similitude s'effectue sur une largeur donnée. On associe un *vecteur d'indices* à chaque vecteur opérande. Ce vecteur est trié de façon que sa lecture donne un ordre décroissant des composantes du vecteur auquel il est associé. Les composantes du vecteur pour lesquelles l'indice ne se trouve pas dans les premiers éléments du vecteur d'indices sont annulées. Une fois cette opération terminée le nouveau vecteur est renormé. Ensuite la valeur de la similitude correspond à la somme des produits des composantes pondérées par l'écart relatif existant dans les vecteurs d'indices.



$$\alpha_5 = V_1[5] \times V_2[5] \times \frac{1}{1 + \beta \times (1-6) \times (1-6)}$$

$$Sim(V_1, V_2) = \sum \alpha_i$$

Nous avons bien évidemment pour tout vecteur $\vec{V}$ non nul $sim(\vec{V}, \vec{V}) = 1$ et du fait que toutes les composantes sont positives ou nulles :

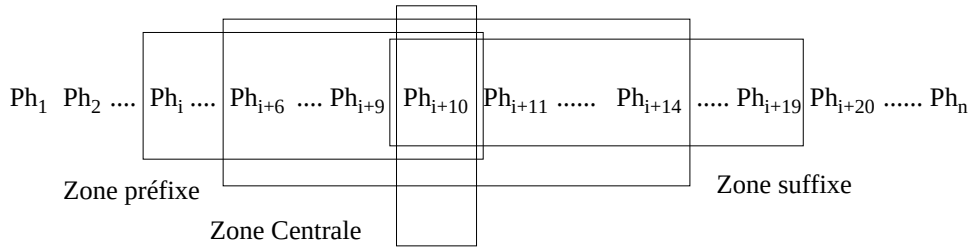
5.5.3 Propriétés de la similitude

- pour tous vecteurs $\vec{V_1}$ et$\vec{V_2}$ orthogonaux : $sim(\vec{V_1}.\vec{V_2}) = 0$.
- la similitude est également symétrique :

$$sim(\vec{V_1}.\vec{V_2}) = sim(\vec{V_2}.\vec{V_1}) \tag{7}$$

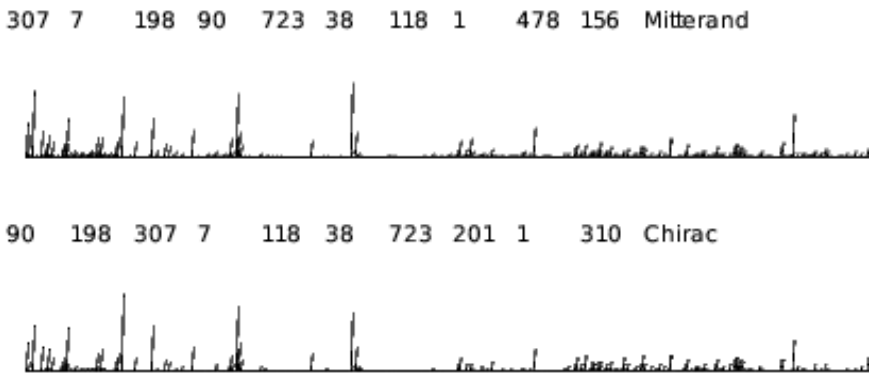
6 Fouille de texte

Le calcul des vecteurs sémantiques s'effectue sur chaque phrase du texte. Le principe de décision pour la sélection d'une phrase est construit sur le calcul moyen des vecteurs associés aux phrases situées à l'intérieur d'une fenêtre. Pour une décision à propos de la phrase Ph_{i+10} les vecteurs concernés seront les centroïdes des vecteurs associés aux trois zones *préfixe,centrale* et *suffixe* définies ci-après.



Le classement des phrases s'effectue par comparaison des différents vecteurs avec des vecteurs spécifiques $\overrightarrow{V_{Chirac}}$et $\overrightarrow{V_{Mitterrand}}$, que nous symboliserons par $\vec{V_C}$ et $\vec{V_M}$ respectivement.

> Ces deux vecteurs ont été obtenus en calculant le centroïde des vecteurs des phrases associées à chacun d'eux dans le corpus d'apprentissage. Pour le calcul de ce vecteur, chaque vecteur est affecté d'un poids proportionnel à sa taille (le coefficient utilisé est égal au millième du carré de la longueur en octets).



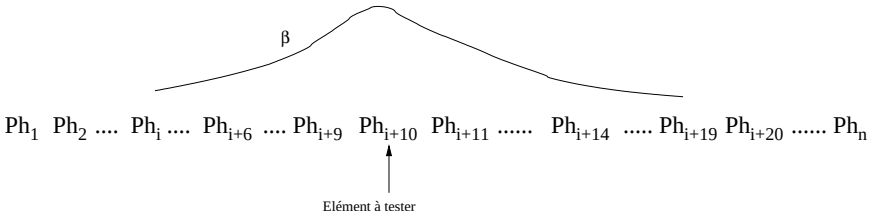
> Les indices correspondent aux concepts majoritaires dans l'ordre décroissant.

Soient :
- $\vec{V_{ZP}}$ le vecteur de la zone préfixe
- $\vec{V_{ZC}}$ le vecteur de la zone centrale
- $\vec{V_{ZS}}$ le vecteur de la zone suffixe

Le premier filtre compare les produits scalaires <$\vec{V_{ZP}}.\vec{V_C}$> et <$\vec{V_{ZP}}.\vec{V_M}$>, <$\vec{V_{ZC}}.\vec{V_C}$> et <$\vec{V_{ZC}}.\vec{V_M}$> et <$\vec{V_{ZS}}.\vec{V_C}$> et <$\vec{V_{ZS}}.\vec{V_M}$>.

> Pour qu'une phrase soit candidate au classement comme phrase appartenant au discours de Mitterrand il est nécessaire qu'au moins un produit scalaire ($<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_M}>$ ou $<\vec{V_{ZS}}.\vec{V_M}>$) soit supérieur à son correspondant ($<\vec{V_{ZP}}.\vec{V_C}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ ou $<\vec{V_{ZS}}.\vec{V_C}>$).

Dans le cas où une phrase est candidate au classement un score d'appartenance est évalué. Ce score correspond au nombre de produits scalaires $<V_{Ph_i}.\vec{V_M}>$ supérieurs aux produits scalaires $<V_{Ph_i}.\vec{V_C}>$. Dans le calcul du score on fait intervenir un coefficient indiquant la proximité avec la phrase candidate :



Si le score atteint un certain seuil (dépendant de β) et que deux produits scalaires au moins ($<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_M}>$ ou $<\vec{V_{ZS}}.\vec{V_M}>$) dont le central sont supérieurs aux produits scalaires correspondants ($<\vec{V_{ZP}}.\vec{V_C}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ ou $<\vec{V_{ZS}}.\vec{V_C}>$) on effectue un calcul de similitude :

> Le calcul de similitude compare seulement cinq phrases : les deux précédentes, la phrase sélectionnée et les deux suivantes :
>
> pour la phrase sélectionnée :
>
> $$\text{Sim}(\vec{V_M}, V_{Ph_{i+10}}) \text{ et } \text{Sim}(\vec{V_C}, V_{Ph_{i+10}})$$

Si la similitude de la phrase testée par rapport au vecteur représentant le discours de Mitterrand est supérieure à la similitude du vecteur testé par rapport au vecteur représentant le discours de Chirac et qu'il en va de même soit pour les deux phrases précédentes soit pour les deux phrases suivantes la phrase testée est attribuée au discours de Mitterrand (Nous supposons donc que les parties insérées du discours de Mitterrand comportent au moins quatre phrases consécutives).

La recherche des phrases associées au discours de Mitterrand se termine par un petit correctif éventuel pour tenir compte de la propriété : *Il n'y a pas de phrase isolée associée au discours de Mitterrand.* Ainsi si l'on a une configuration comme CCMCC où C désigne une phrase associée au discours de Chirac et M une phrase associée au discours de Mitterrand, la associée au discours de Mitterrand est désélectionnée. Il en va de même pour l'inverse. Une configuration comme MMCMM sélectionne la centrale comme appartenant au discours de Mitterrand.

7 RésultatsObtenus

Le corpus d'apprentissage comprenait 7523 phrases appartenant au discours de Mitterrand. L'extraction a donné 6479 textes dont 5782 correctement trouvés. Soit une précision de 0.89,

un rappel de 0.76 et un Fscore de 0.82. Nous pouvons en déduire que les discours correspondants étaient relativement bien séparés sémantiquement.

Sur le corpus de test le rappel s'est effondré : 0.15. La précision a faibli dans une moindre proportion : 0.77. Les discours correspondants se distinguent nettement moins du point de vue sémantique. Ce phénomène peut facilement s'expliquer par le fait que les discours ne sont pas toujours bien catégorisés sémantiquement entre politique étrangère et politique nationale et donc que la recherche de la distinction uniquement par le sens n'était pas la mieux adaptée.

8 Conclusion

Dans le problème que nous avons eu à traiter l'importance de la stylistique est évidente et pourrait faire l'objet de travaux plus approfondis. Néanmoins ce fut la première expérience de l'utilisation d'un analyse syntaxico-sémantique sur un corpus conséquent. La robustesse de l'analyse a bien été mise en évidence et ces résultats sont un encouragement à poursuivre l'utilisation d'analyseurs robustes dans des travaux de fouilles de textes.

Références

[Besançon 2002] Besançon R., Rajman M. (2002). Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents. *TALN 2002 : 9ème conférence Internationale sur le Traitement Automatique du Langage Naturel* . Nancy.

[Bourrigault 1993] Bourrigault D. (1993) Analyse locale pour le repérage des termes complexes dans les textes. *revue Internationale sur le Traitement Automatique des Langues*, Numéro Spécial sur le Traitement Automatique de la Composition Nominale ; vol. 34,n¡2 . pp 105-118.

[Breiman et al. 1984] Breiman L., Friedman J., Olshen R.A., Stone C.J. (1984), *Classification and regression trees*. Wadsworth.

[Chauché 1984] Chauché J. (1984) Un outil multidimensionnel de l'analyse du discours. *International conference on computational linguistics Stanford California 1984* .

[Chauché 1990] Chauché J. (1990) Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TA Information* vol 1/1, p 17-24.

[Chauché et al. 2003] Chauché J., Prince V. , Jaillet S., Teisseire M. (2003) Classification Automatique de Textes à partir de leur Analyse Syntaxico-Sémantique . *TALN'03 : 10 ème Conférence Internationale sur le Traitement Automatique du Langage Naturel* , pp. 55-65

[Daille 1995] Daille B. (1995), Repérage et extraction de terminologie par une approche mixte statistique et linguistique, *Revue Internationale du Traitement Automatique des Langues*, ATALA, 36(1-2) :101-118.

[A.J. Greimas et J. Courtes 1993] Greimas A.J., Courtés J. (1993) Sémiotique dictionnaire raisonné de la théorie du langage *Hachette Université* Hachette Livre 1993 ISBN 2-01-020648-7.

[Crestan et al. 2003] Crestan E. , El-Bèze M., de Loupy Claude (2003) Peut-on trouver la taille de contexte optimale en désambiguisation sémantique ? *Actes de TALN200*311-14 juin, Batz-sur-Mer. Vol 1. Pp 85-94.

[Ellman et Tait 1999] Ellman J., Tait, J. (1999) Roget's thesaurus : An additional Knowledge Source for Textual CBR ? *Proc. of 19th SGES Int. Conf. on Knowledge-Based and Applied AI*. Springer-Verlag. pp 204 Ð 217.

[T. Joachims 2002] Joachims T. (2002) Kluwer Academic Publishers. *Learning to Classify Text Using Support Vector Machines* .

[Larousse 1992] Larousse (1992) *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Paris.

[Manning 1999] Manning C. D., Schütze H. (1999) *Foundations of Statistical Natural Language Processing*, The MIT Press, ISBN 0-262-13360-1.

[Mendelson 1964] Mendelson E. (1964) *Introduction to mathematical logic*. D. VAN NOSTRAND COMPANY. New York.

[Rocchio 1971] Rocchio J. (1971) Relevance feedback in information retrieval. *SMART Retrivial System : Experiments in Automatic Document Processing*.

[Roget 1852] Roget P. (1852) Thesaurus of English Words and Phrases *Longman*, London.

[Romesburg 2004] Romesburg H. C. (2004) *Cluster Analysis for Researchers*, 340 pp. ISBN 1411606175 reprint of 1990 edition published by Krieger Pub. Co.

[Yarowsky 1992] Yarowsky D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proc. of COLING92* .

[SYGFRAN] *http ://www.lirmm.fr/~ chauche* .

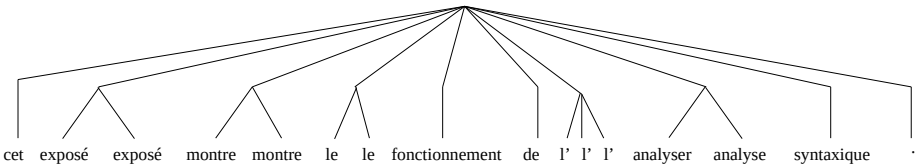
Annexe

L'analyse (très simplifiée) de la phrase :

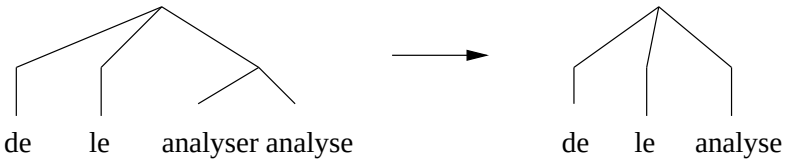
"Cet exposé montre le fonctionnement de l'analyse syntaxique."

suit les étapes :

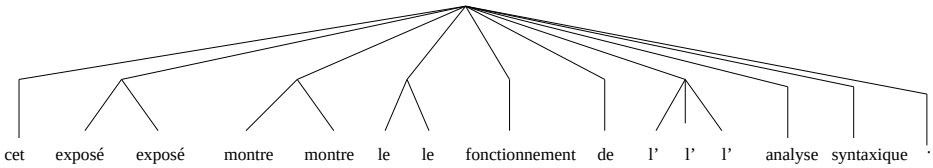
– Sortie de l'analyse morphologique :



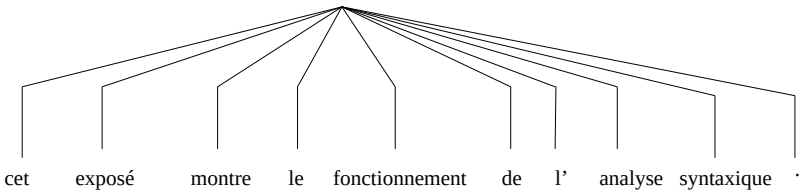
– Étapes de la construction syntaxique :
Une série de filtres permet d'éliminer les cas de suites impossibles : par exemple ' de le analyser annalyse ' → 'de le analyse' correspond à la règle :



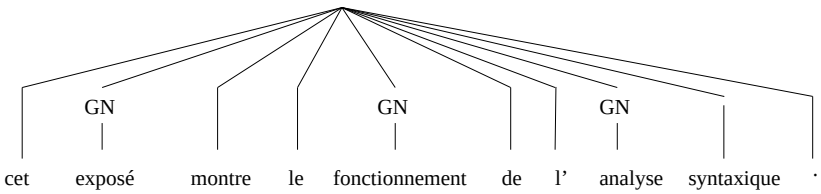
La recherche du schéma correspond à une sous-arborescence de l'élément transformé. Ansi dans ce cas il y a une sélection par rapport aux différentes solutions de l'. Cette règle n'élimine pas les différents points complémentaires et le résultat de l'application de cette règle donne :



Comme on peut le constater il n'y a pas de notion d'analyse gauche/droite dans les transformations. Seules les configurations applicables sont réalisées en respectant l'ordre des occurrences et la priorité des règles. Dans la définition des algorithmes de Markov, la priorité d'application par rapport à l'ordre des occurrences s'effectue de gauche à droite. En SYGMART c'est l'inverse, la priorité est de droite à gauche. C'est-à-dire que si une règle est applicable sur plusieurs occurrences, l'occurence la plus à droite sera d'abord considérée pour l'application. Une autre différence fondamentale se situe dans les pas de transformations : Dans un algorithme de Markov, à chaque pas, une seule occurence est transformée. En SYGMART, à chaque pas, toutes les occurences possibles sont transformées simultanément.
Le résultat de l'application de ces filtres produit la structure :

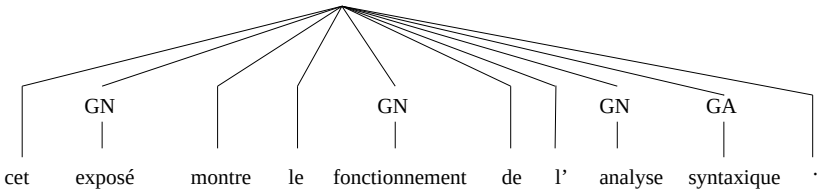


Les noms permettent de construire des groupes nominaux :

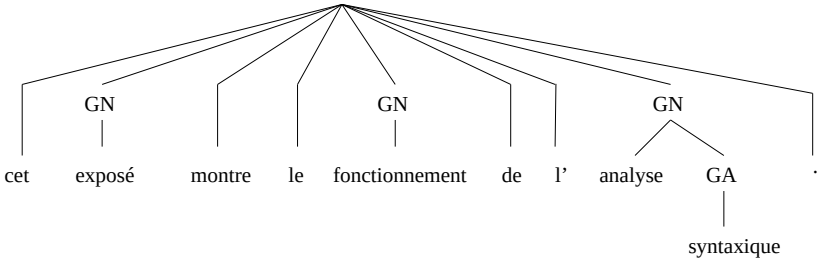


Les adjectifs forment des groupes adjectivaux :

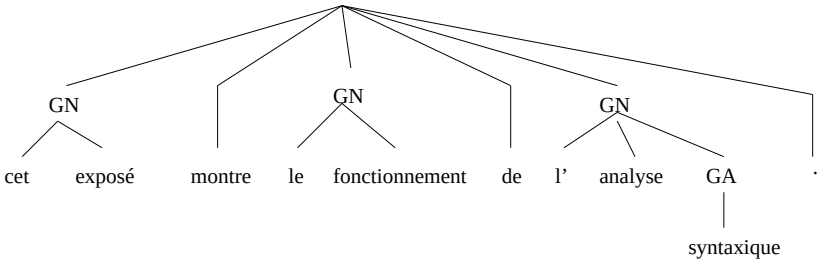
Application des vecteurs sémantiques à la fouille de texte



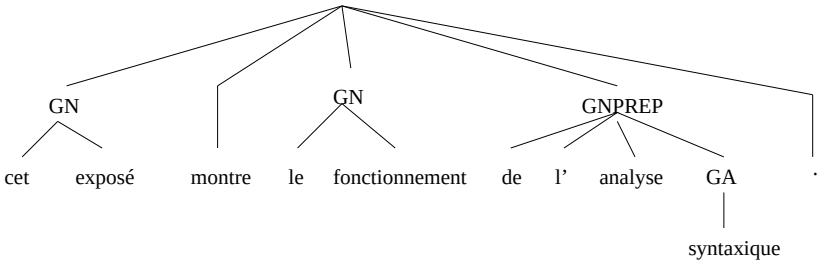
Les adjectifs sont rattachés aux groupes nominaux :



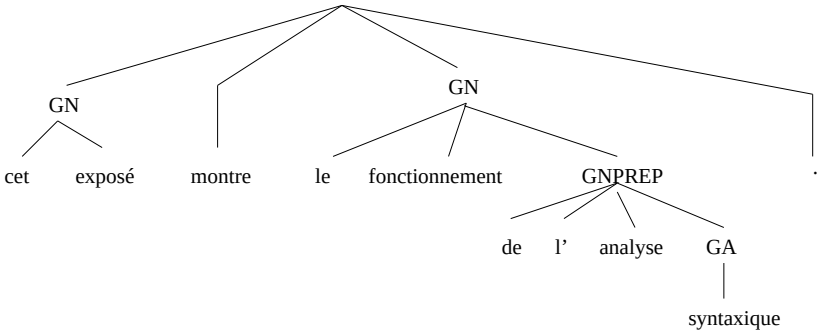
Les déterminants sont ajoutés aux groupes nominaux :



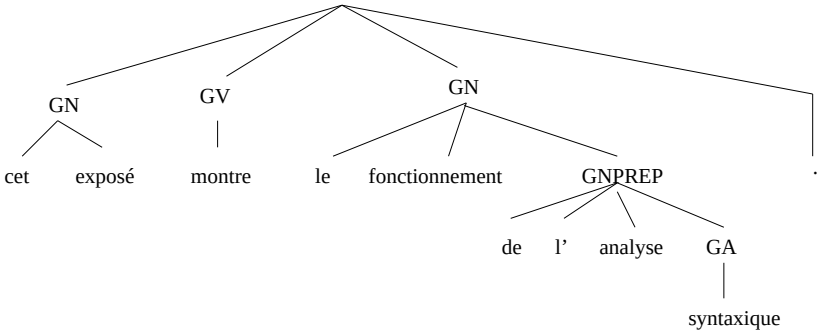
Les groupes prépositionnels sont alors construits :



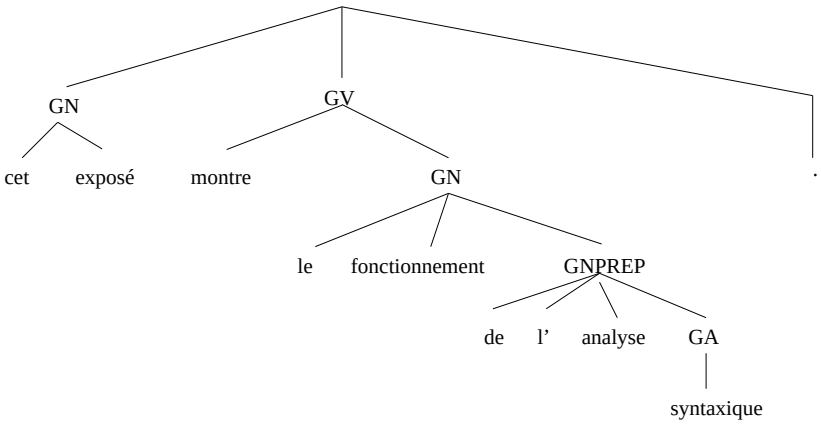
Le groupe prépositionnel est rattaché au groupe nominal :



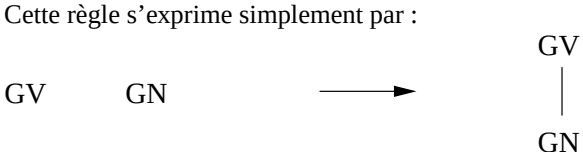
Le verbe détermine un groupe verbal :



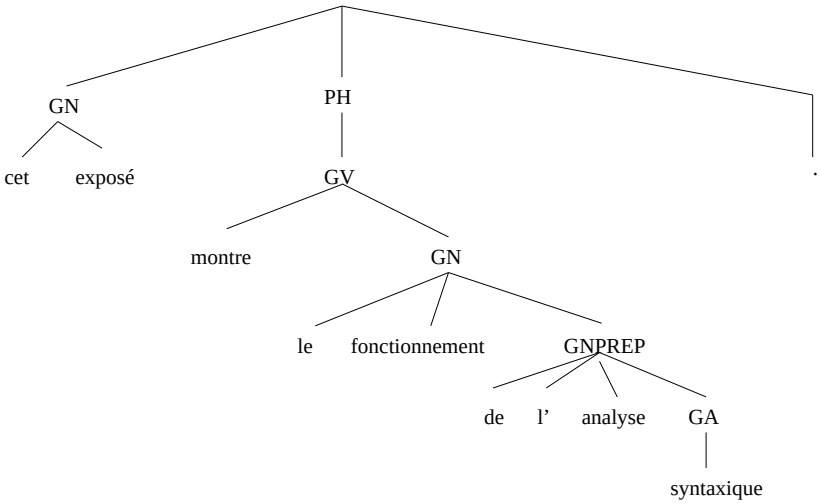
Le complément d’objet direct est alors attaché au groupe verbal :



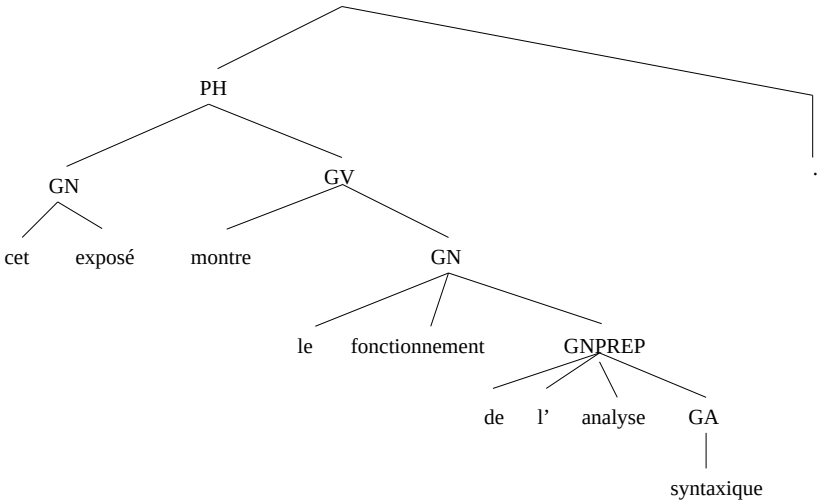
Cette règle s’exprime simplement par :



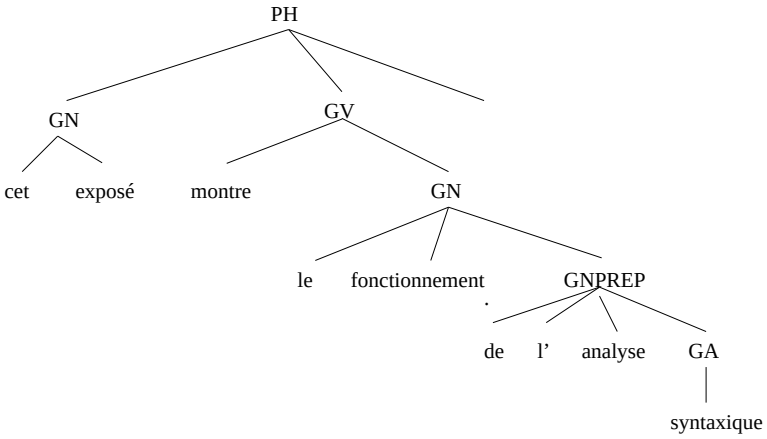
des contraintes de continuïté doivent être bien sûr respectées dans ce cas.
Le groupe verbal constitue un groupe :



Le sujet est alors introduit dans le groupe :



Enfin la ponctuation termine la construction :



Summary

The approach presented here is based on a treatment of the syntactico-semantics contents by an analyzer of French, system SYGFRAN(SYGFRAN), to find a group of sentences belonging to various speeches of president François Mitterrand mixed with a group of sentences belonging to various speeches of president Chirac. This treatment is done by a calculation of semantic vectors of sentences (methodology defined in the article) and by the definition of a relation of similarity describing the inclination of vectors to which the slope, in angular distance, is close. Using this relation, sentences are allotted by the system to one or the other of the authors, and the article indicates the F-measurements obtained on the first corpus (also called training corpus) slightly higher than 80%.