Dépendances syntaxiques et méthodes de détection de passages pour une segmentation sur le locuteur et le thème

Loïc Maisonnasse, Caroline Tambellini

Laboratoire CLIPS IMAG – Université Joseph Fourier
385, rue de la bibliothèque - BP 53
38041 Grenoble cedex9
loic.maisonnasse@imag.fr
caroline.tambellini@imag.fr

Résumé. Nous montrons ici l'intérêt des dépendances syntaxiques et des méthodes de détection de passages pour la détection du locuteur dans les discours. Nous nous appuyons sur les méthodes développées dans le cadre de la campagne de fouilles de texte DEFT'05 dans le but de distinguer le discours de François Mitterrand de celui de Jacques Chirac. Nous évaluons l'utilisation des dépendances syntaxiques en tant qu'unité caractérisant les différences entre les deux discours. Ces unités, obtenues par traitement linguistique, constituent donc les descripteurs retenus pour la représentation des discours, et sur lesquels nous appliquons un apprentissage. Les deux présidents ayant des discours sur des thématiques différentes nous combinons cet apprentissage avec l'utilisation d'une méthode de segmentation thématique pour détecter les changements de locuteur.

1 Introduction

1.1 Principes

DEFT'05 a pour objectif la détection des phrases de F. Mitterrand dans des allocutions de J. Chirac. Pour ce faire, nous disposons d'un corpus composé d'allocutions de J. Chirac au sein desquelles des portions d'allocutions de F. Mitterrand ont été insérées.

La réussite de la tâche réside dans l'utilisation des éléments qui distinguent au mieux les deux discours. Pour cela, nous partons du principe que le vocabulaire des discours n'est pas le seul élément mettant en évidence la différence entre les discours des deux orateurs. Cette différence s'exprime également au niveau des tournures et des constructions de phrases propres à chaque auteur.

Globalement, nous montrons l'intérêt de l'utilisation d'éléments représentatifs de la syntaxe pour déterminer l'origine d'un discours : notre méthode utilise les dépendances syntaxiques pour la détection du locuteur.

Les discours de F. Mitterrand ayant une thématique différente des allocutions de J. Chirac, nous évaluons la complémentarité d'une détection des différences au niveau syntaxique avec une approche plus thématique. Pour cela, nous combinons notre première approche avec la piste des changements thématiques dans le but d'améliorer la détection des allocutions de F. Mitterrand.

1.2 Proposition

Pour réaliser cette tâche, nous créons deux visions du corpus (une par homme politique). Une vision pour un homme politique fournit pour chaque segment du document la pertinence qu'il soit issu de son propre discours.

Ce sont ces deux visions que nous fusionnons ensuite et qui nous permettent de déterminer les allocutions de chacun des présidents (figure 1). Cette fusion est effectuée en sélectionnant la vision la plus pertinente pour chaque segment.

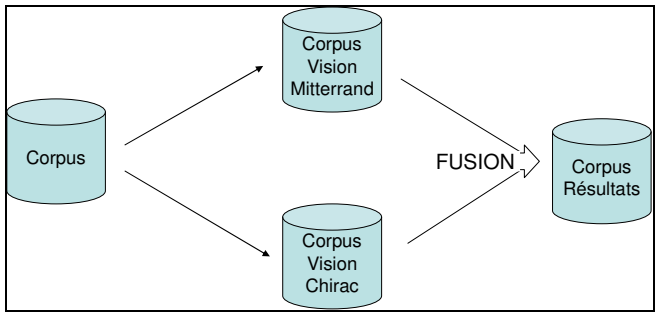


Fig. 1 – *Méthode*

Nous établissons tout d'abord pour chaque président un profil. Ce profil est issu de différentes analyses du corpus d'apprentissage (figure 2) : trois segmentations du corpus et différentes fonctions de pondérations nous permettent de créer le profil le plus adéquat. Les segmentations du corpus correspondent soit au découpage du corpus en deux blocs un pour chaque président, soit à son découpage par allocution et par président ou encore par phrase et par président.

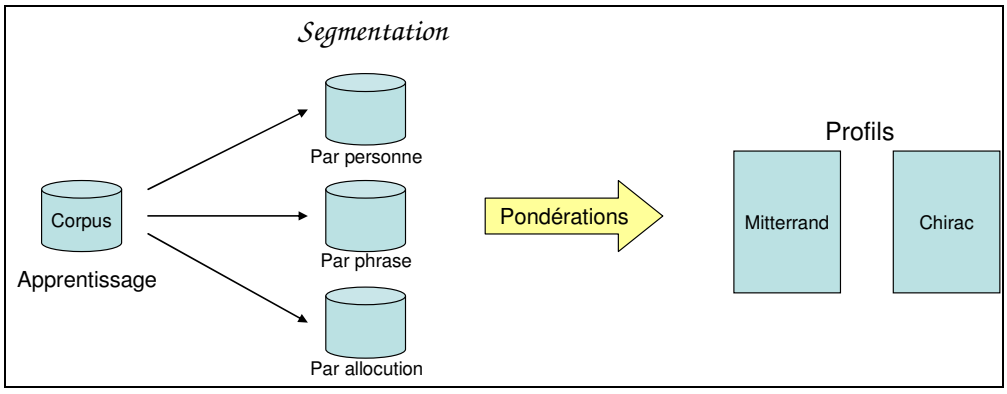


Fig. 2 – *Elaboration du profil de chaque président selon différents découpages*

Le corpus d'évaluation est ensuite évalué selon le prisme de chacun des deux profils (figure 3). Le corpus est analysé en fonction de chaque profil afin de produire les visions du

corpus selon chaque président. Ce sont ces deux corpus que nous fusionnons afin d'attribuer chaque segment à un homme politique.

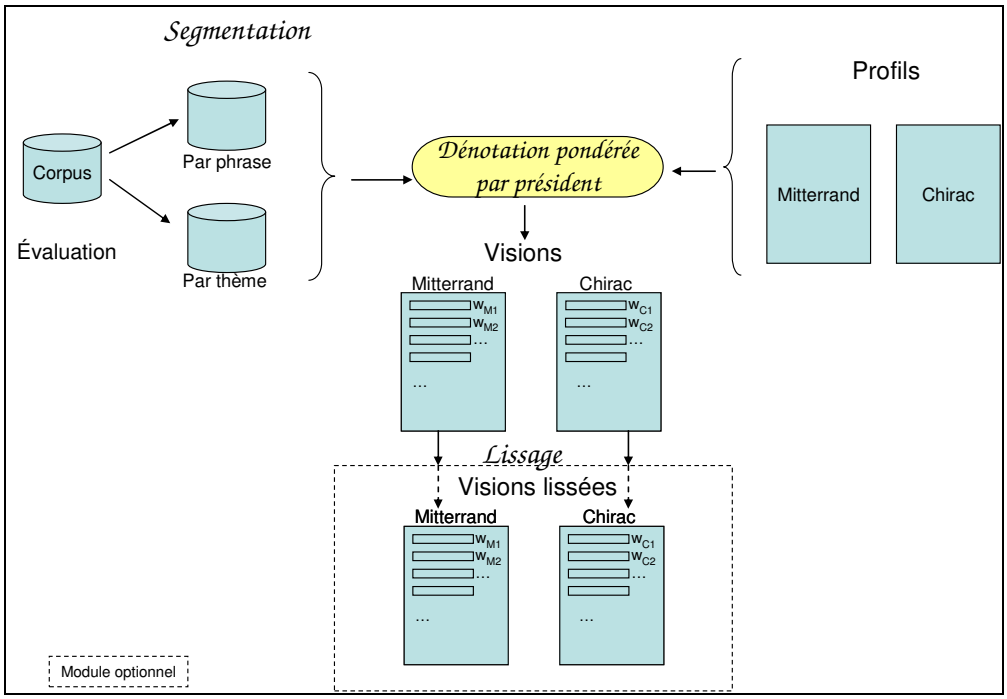


FIG. 3 – *Création des visions selon chaque homme politique*

Dans la suite de l'article, nous présentons tout d'abord les dépendances syntaxiques et leur application à notre contexte. Ensuite, nous traitons de la création des profils, basé sur ces dépendances syntaxiques. Nous abordons la création des visions du corpus par président. Nous verrons plus précisément la segmentation thématique et le lissage. Nous présentons les résultats obtenus lors de la campagne d'évaluation DEFT'05. En conclusion, nous discutons les limites et les perspectives de notre méthode.

2 Utilisation des dépendances syntaxiques

2.1 Les dépendances syntaxiques

2.1.1 Définition

Les dépendances décrivent les liens entre les mots au sein d'une phrase, elles peuvent être de différents types (logique, syntaxique ou sémantique). Chaque dépendance est étiquetée par la fonction qui lie ses mots ; au niveau syntaxique, ce sont par exemple des fonctions telles que sujet ou objet. L'ensemble des dépendances d'une phrase forme une structure appelée arbre de dépendances.

2.1.2 Utilisation en recherche d'information

Des travaux en recherche d'information ont déjà porté sur l'utilisation des structures de dépendances pour accéder au contenu des documents. Ces travaux s'appuient sur l'analyse de dépendances pour améliorer la représentation des documents. Ils s'articulent majoritairement autour de deux axes : soit les dépendances sont utilisées pour extraire des syntagmes, soit la structure de dépendance est utilisée comme index et une fonction de correspondance adaptée à cette structure est employée.

Dans les travaux de Strzalkowski (Strzalkowski et al., 1994), les auteurs produisent une représentation proche d'un arbre de dépendances à l'aide d'un analyseur. Ils sélectionnent ensuite un certain nombre de paires candidates à la formation de termes composés à l'aide de patrons sur l'arbre de dépendances. Ces nouveaux termes sont ensuite ajoutés dans l'index des documents. Un schéma de pondération sous forme de tf-idf, pondérant les termes par rapport à leur fréquence à l'intérieur du document (tf) et par rapport à l'inverse de leur fréquence dans tout le corpus (idf)[1], est adapté pour donner plus d'importance à l'idf des termes composés. Les auteurs notent une augmentation de l'ordre de 20% de la précision moyenne avec l'utilisation des mots composés. Cette étude ne permet cependant pas de conclure directement si l'amélioration est due à l'utilisation des dépendances. Celles-ci servent ici à l'extraction de mots composés, on ne sait pas si le gain provient de la dépendance entre les mots ou si le gain provient des mots composés qui peuvent être extraits par d'autres méthodes.

Partant de l'hypothèse que la conversion des structures de dépendances en syntagmes entraîne une perte d'information due à la linéarisation des constituants, des recherches ont porté sur l'utilisation directe de ces structures. Ainsi, Metzler (Metzler *et al.*, 1989), extrait des arbres de dépendances binaires sur des phrases en anglais. Ces arbres sont extraits des documents à l'aide de l'analyseur COP (Constituent Object Parser). Dans ce système, lors de l'interrogation, l'utilisateur doit déterminer les termes pertinents pour sa requête et indiquer les dépendances entre ces termes. Le système évalue alors les documents pertinents pour la requête en effectuant plusieurs types de correspondances entre les dépendances de la requête et celles contenues dans les arbres des documents.

Pour sa part, Smeaton (Smeaton, 1999) propose un modèle proche mais utilise une analyse qui conserve les ambiguïtés syntaxiques les plus courantes de l'arbre de dépendances. Les résultats obtenus par ce modèle restent cependant inférieurs à ceux obtenus en appliquant une pondération simple sur les syntagmes représentés par les arbres.

2.1.3 Synthèse et proposition

Dans notre méthode, nous nous positionnons à un niveau intermédiaire. Dans le but de manipuler une structure moins complexe que l'arbre de dépendances et de conserver une partie de la structure, nous considérons le résultat de l'analyse comme un ensemble de dépendances. Chacune de ces dépendances est caractérisée par son type et la liste des lemmes qu'elle relie. Ainsi, la phrase *'le chat mange la souris'* est représentée par : *'{DETERM(le,chat), DETERM(le,souris), SUBJ(chat,manger), OBJ(souris,manger)}'* où DETERM correspond à la relation de déterminant entre 'le' et 'chat' et où SUBJ explicite la

1. idf = log(N/n) où N est le nombre de documents dans le corpus, et n ceux qui contiennent le terme.

relation de sujet entre ‘manger’ et ‘chat’. Ce sont ces dépendances syntaxiques que nous allons utiliser dans les profils.

2.2 Les dépendances syntaxiques dans notre proposition

2.2.1 Principe général

Dans notre approche, les dépendances syntaxiques sont utilisées pour établir les profils des présidents et également pour analyser le corpus d’évaluation. Pour les profils, les dépendances produites servent alors de descripteur et un poids représente leur pouvoir discriminant. Ce poids est déterminé à l’aide d’un apprentissage.

Pour le corpus d’évaluation, nous analysons en dépendances chaque segment du corpus. Nous établissons un profil pour chaque segment à l’intérieur duquel chaque dépendance est pondérée par sa fréquence. L’ensemble de ces profils du corpus d’évaluation est alors comparé aux profils des présidents, ce qui permet de donner un score à chaque segment et pour chaque président. Cela permet donc d’établir une vision du corpus par président.

2.2.2 Mise en œuvre

Pour permettre l’évaluation de notre approche, le corpus d’entraînement utilisé dans le cadre de la campagne DEFT05 a été divisé en deux parties. La première (de l’allocution 100 à l’allocution 5) est utilisée pour l’apprentissage, le reste est utilisé pour l’évaluation.

Nous utilisons l’analyseur ‘Xerox Incremental Parser’ (XIP) (Aït-Mokhtar *et al*., 2002). A partir des résultats de cette analyse, la liste des dépendances syntaxiques de chaque phrase est produite. De manière à couvrir le plus de phénomènes linguistiques, nous conservons, dans un premier temps, toutes les dépendances produites par l’analyseur syntaxique (tableau 1). La liste des dépendances pouvant être produite par cet analyseur est présentée en annexe. Certaines des dépendances produites distinguent plus les tournures de phrase, il s'agit notamment des dépendances telles que *NEGAT* qui décrivent la forme négative d’un verbe, ou encore COREF qui lient une anaphore et son antécédent. D’autres reflètent plus le vocabulaire verbal et les collocations, c’est notamment le cas des dépendances telles que *NMOD*, *VMOD* qui lient respectivement un nom et un verbe avec un modificateur.

ADJARG::absent::de::Louvre
COREF::Bruxelles::qui
COREF::Espagne::qui
DETERM::le::suppression
NEGAT::veiller
NMOD::monsieur::allègre
SUBJ::aider::famille
VMOD::hisser::engager

TAB. 1 – *Exemple de dépendances extraites du profil de J. Chirac*

3 Génération des profils

Afin de générer le meilleur profil possible, nous avons établi différents profils selon trois segmentations (par personne, par phrase, par allocution) et par différentes pondérations. Puis nous avons sélectionné le meilleur d'entre eux.

3.1 Principe

Les différents profils dépendent d'une part des segmentations du corpus d'apprentissage et d'autre part de la pondération servant de base à l'apprentissage. Trois segmentations du corpus d'apprentissage sont mises en place. La première se base sur le regroupement de l'ensemble des phrases d'un président : apprentissage par personne. Le deuxième sur la segmentation du corpus selon les phrases : apprentissage par phrase. Enfin le dernier se base sur un segmentation par allocution : apprentissage par allocution.

Pour chacune de ces segmentations, par l'utilisation de deux pondérations différentes, nous obtenons deux types de profils. D'autres profils sont établis en combinant des profils provenant de différentes segmentations.

Pour évaluer ces différents profils, nous utilisons uniquement le découpage par phrase du corpus d'évaluation. Pour établir le meilleur profil, chacune des segmentations finales est comparée à la vérité terrain en utilisant la fonction de F-score proposée par les organisateurs de DEFT'05.

3.1.1 Apprentissage par personne

Dans notre première méthode d'apprentissage, pour acquérir les profils nous concaténons l'ensemble des phrases de chaque politicien au sein de deux documents. Les dépendances extraites de ces deux documents sont stockées sous la forme de deux vecteurs représentant les profils de chaque président. Le poids des dépendances dans chaque profil est calculé selon deux pondérations ltc[2] et lnc[3] présentées ci-dessous et habituellement utilisées en recherche d'information pour pondérer les descripteurs des documents :

2 Pondération composée d'une pondération locale (tf) et d'une pondération globale (idf) ainsi que d'une normalisation

3 Pondération composée d'une pondération locale (tf) et d'une normalisation, sans pondération globale

ltc	$w_{i,j} = \frac{\ln(f_{i,j}+1) * \log(2/df_i)}{\sqrt{\sum_i f_{i,j}^2}}$
lnc	$w_{i,j} = \frac{\ln(f_{i,j}+1)}{\sqrt{\sum_i f_{i,j}^2}}$

TAB. 2 – *Pondération pour l'apprentissage par personne*

où : $w_{i,j}$ est la pondération finale de la dépendance i pour le président j,
$f_{i,j}$ est la fréquence de la dépendance i dans le discours de j,
$d_{f,i}$ est le nombre de documents contenant i (ici 1 ou 2).

3.1.2 Apprentissage par phrase ou par allocution

Nous utilisons une deuxième méthode dans laquelle nous avons calculé le poids des dépendances à l'aide d'un apprentissage basé sur l'une ou l'autre des segmentations par phrase ou par allocution. Le poids d'une dépendance résulte alors de sa répartition dans les unités pertinentes et non pertinentes pour un profil. Ce calcul est effectué soit par la formule de *Rocchio* soit par la formule utilisée dans (Brouard, 2002) (*N*). Ces deux formules sont présentées ci-dessous :

Rocchio	$w_{i,j} = \alpha \frac{\lvert Qi \cap Pj \rvert}{\lvert P_j \rvert} - \beta \frac{\lvert Qi \cap \overline{Pj} \rvert}{\lvert \overline{P_j} \rvert}$, α = β =1
N	$w_{i,j} = \frac{\lvert Q_i \cap P_j \rvert}{\lvert P_j \rvert} * \frac{\lvert Q_i \cap P_j \rvert}{\lvert Q_i \rvert}$

TAB. 3 – *Pondération pour l'apprentissage par phrases, par allocution*

Où : P_j : ensemble des unités pertinentes pour le président j
Q_j : ensemble des unités contenants la dépendance i

3.1.3 Combinaison de profils

Nous avons également combiné les profils de différentes granularités de manière à bénéficier des informations extraites à chacun des niveaux. Nous avons pour cela regroupé l'apprentissage par personne avec l'apprentissage sur les allocutions. L'apprentissage par personne est utilisé comme poids initial dans le nouvel apprentissage.

3.2 Evaluation des profils

Nous avons appliqué les méthodes présentées sur notre corpus d'apprentissage. Pour les différentes méthodes, les profils obtenus ont été utilisés pour extraire les phrases de F. Mitterrand du corpus d'évaluation. Lors de cette évaluation les résultats suivants ont été obtenus :

	Pondération	F-score
Apprentissage par personne	*ltc*	0,3138
	lnc	0,2400
Apprentissage par phrase	*N*	0,1843
	Rocchio	0,1814
Apprentissage par allocution	*N*	**0,3286**
	Rocchio	0,2411

TAB. 4 – *Résultats des différents apprentissages*

Le meilleur résultat est celui obtenu avec une pondération N sur les allocutions avec un F-score proche de 0,33. Les moins bons résultats sont ceux obtenus à l'aide de l'apprentissage sur les phrases. Cela semble provenir du fait que les phrases prises seules constituent de trop petits éléments d'apprentissage. L'apprentissage par personne, quant à lui donne de bons résultats notamment par l'utilisation de la pondération *ltc*. La différence de cette pondération par rapport à la pondération *lnc* est qu'elle donne un poids nul aux dépendances qui apparaissent dans les deux profils. Les dépendances communes aux deux profils représentent 8 % des descripteurs extraits, les autres appartenant uniquement à l'un des profils. Le vocabulaire des dépendances est donc fortement spécifique à chaque politicien et cette spécificité capture en partie la différence entre les deux discours.

Cependant si la majorité des dépendances communes sont effectivement non pertinentes pour déterminer un profil, la pondération actuelle ne permet pas de faire de nuances entre les dépendances qui apparaissent beaucoup dans un profil et peu dans l'autre.

Les résultats obtenus par combinaison de profils sont présentés dans le tableau suivant :

	Cœfficient	F-score
Apprentissage par personne (Ltc) et Apprentissage par allocution (N)	1	0,3195
Apprentissage par personne (Ltc) et Apprentissage par allocution (Rocchio)	1	**0,3340**

TAB. 5 – *Regroupement apprentissage par personne et apprentissage par allocutions*

La pondération globale ltc combinée avec l'apprentissage basé sur la pondération N sur les allocutions fournit des résultats inférieurs à ceux obtenus par la simple formule N. Ces deux formules ne sont donc pas complémentaires. A contrario, l'apprentissage par personne ltc combiné avec un apprentissage sur les allocutions de type Rocchio améliore les résultats

de base et dépasse les résultats obtenus à l'aide de l'apprentissage N sur les allocutions. Le résultat peut même être légèrement amélioré en utilisant un α de 4/9 et un β de 5/9 dans la formule de Rocchio (tableau 3).

Globalement ces premiers résultats sont faibles par rapport à la moyenne de DEFT05, cela provient du fait qu'à cette étape, les phrases sont considérées de manière indépendante. La dépendance entre les phrases sera abordée lors de la création des visions par président (partie 4). Les étapes suivantes de notre modèle dépendant de la création des profils, nous avons évalué différentes variations sur ces profils.

3.3 Evaluation de l'apport des dépendances syntaxiques

Il est intéressant de comparer les dépendances que nous avons utilisées au sein des profils avec d'autres descripteurs habituellement utilisés pour une telle tâche. Nous effectuons la comparaison entre l'utilisation des mots simples, de la forme lemmatisée de ces mots et des dépendances en tant que dimension de nos profils. Les résultats obtenus sont présentés dans le tableau ci-dessous :

		Dépendance	Mot	Lemme
Apprentissage par personne	ltc	0,3138	0,2312	0,1534
	lnc	0,2400	0,1814	0,1908
Apprentissage par allocution	Rocchio	0,1814	0,1803	0,1814
Apprentissage par personne et Apprentissage par allocution	ltc et Rocchio[4]	**0,3340**	**0,2403**	**0,2155**

TAB. 6 – *comparaison des différents descripteurs (F-score)*

Le meilleur apprentissage, tous descripteurs confondus, est celui obtenu par combinaison de profils, l'apprentissage par personne, lui, est légèrement inférieur. L'apprentissage sur les allocutions basé sur la pondération de Rocchio donne les résultats les plus faibles et ses résultats sont similaires pour les trois descripteurs. Indépendamment de l'apprentissage, les dépendances donnent des résultats supérieurs à tous les autres descripteurs. Elles capturent donc mieux les différences entre le discours de J. Chirac et de F. Mitterrand que les mots simples ou les lemmes. On remarque aussi que l'utilisation de la forme lemmatisée des mots donne de moins bons résultats que l'utilisation des mots bruts. Ce résultat confirme le fait que d'importantes différences entre les deux discours ne se situent pas au niveau du vocabulaire. En effet, parmi les descripteurs utilisés ce sont les lemmes qui représentent le plus le vocabulaire. Or d'après les résultats le passage du mot au lemme supprime de l'information permettant de distinguer les deux discours. Cette information est essentiellement l'information syntaxique représentée par les flexions des mots, ce type d'information semble important pour différencier les discours.

Les dépendances donnent globalement de meilleurs résultats que les deux autres descripteurs, or les dépendances sont constituées de lemmes. Les résultats des dépendances étant meilleurs, c'est donc l'information contenue par la relation entre les mots qui permet d'obtenir une meilleure distinction des profils.

[4] Regroupement apprentissage par personne (Ltc) et apprentissage par allocutions (Rocchio)

3.4 Evaluation des fonctions syntaxiques

L'analyseur syntaxique que nous utilisons extrait plusieurs types de dépendances syntaxiques, nous nous sommes intéressés à évaluer quelles sont les fonctions syntaxiques qui permettent de différencier les discours des deux politiciens. Pour cela, nous avons étudié pour chaque fonction syntaxique l'impact de la suppression, dans les profils, des dépendances étiquetées par cette fonction. Si la suppression d'un type de fonction syntaxique entraîne une baisse du F-score alors ce type contient des éléments qui permettent de distinguer les deux discours. Au contraire si la suppression d'un type de fonction syntaxique entraîne une amélioration du F-score, alors cette fonction syntaxique n'est pas caractéristique des différences entre les deux discours. Les résultats obtenus sur l'apprentissage par personne en ltc sont présentés dans le tableau suivant :

Fonction syntaxique supprimée	F-score	Variation du F-score	Répartition dans le corpus
base	0,3138		
ADJARG	0,3140	0,07%	0,34%
AUXIL	0,3142	0,13%	2,79%
COORDITEMS	0,3134	-0,14%	0,53%
COREF	0,3178	1,28%	3,35%
DETERM	0,3058	-2,55%	23,78%
NEGAT	0,3127	-0,34%	1,11%
NMOD	0,2917	-7,03%	24,33%
NN	0,3093	-1,43%	0,91%
NPR	0,3138	0,00%	0,01%
OBJGM	0,3138	0,00%	0,00%
REFLEX	0,3126	-0,38%	1,80%
SEQNP	0,3135	-0,09%	0,73%
SUBJ	0,3047	-2,89%	12,71%
SUBJCLIT	0,3124	-0,45%	0,14%
VARG	0,3053	-2,70%	12,08%
VMOD	0,2922	-6,87%	15,35%
NARG	0,3131	-0,21%	0,02%
NGM	0,3138	0,00%	0,02%

TAB. 7 – *Etude des fonctions syntaxiques sur l'apprentissage ltc*

Les résultats obtenus montrent qu'à l'exception de trois fonctions syntaxiques (ADJARG, AUXIL, COREF) la suppression d'une fonction syntaxique entraîne une baisse du F-score. La fonction COREF (qui lie une anaphore et son antécédent) est celle dont la suppression entraîne la plus forte amélioration, elle est donc la moins pertinente. Or, cette dépendance est fortement liée à la structure de la phrase.

Nous remarquons aussi que les fonctions syntaxiques qui semblent avoir le plus d'intérêt pour différencier les deux discours sont VMOD (qui lie un verbe avec un de ces modificateurs) et NMOD (qui lie un nom avec un de ces modificateurs). Ces deux fonctions syntaxiques sont fortement représentatives des collocations verbales et nominales.

Pour cette tâche, les informations représentées par ces collocations sont celles permettant de distinguer les profils. Ce résultat s'explique par le fait que les discours politiques sont rédigés à l'avance, certaines tournures de phrases sont donc communes aux deux discours.

4 Création des visions selon chaque président

4.1 Principe

Pour interroger les profils, nous utilisons différentes segmentations sur le corpus d'évaluation (figure 3). Cette interrogation crée des visions du corpus selon chaque profil. Plusieurs segmentations du corpus sont envisageables : une segmentation par phrase (format du corpus), et une segmentation thématique (à mettre en œuvre). Une méthode de lissage est proposée pour prendre en compte le contexte des segments. Ces deux aspects sont explicités dans la suite de l'article.

4.2 Segmentation thématique

Le corpus de DEFT'05 est composé d'allocutions officielles de J. Chirac dans lesquelles des allocutions de F. Mitterrand ont été insérées. Chacun des deux discours des présidents abordent des thèmes différents. En effet, au sein des allocutions de J. Chirac évoquant la politique internationale, des phrases de F. Mitterrand issues de discours traitant de politique nationale sont introduites. Ainsi, la rupture thématique peut être une des manières de détecter les phrases issues du corpus de F. Mitterrand. Nous avons donc voulu étudier la piste des changements thématiques pour déterminer les allocutions de F. Mitterrand. Différentes méthodes peuvent être utilisées pour faire de la recherche de passage (Passage Retrieval) sur les documents. Wilkinson (Wilkinson et al., 1994) utilise la structure logique du document : phrases, paragraphes, etc pour découper le document en passages. Callan (Callan et al., 1994) et Zobel (Zobel et al., 1994) utilisent une fenêtre fixe de mots pour déterminer les passages. Salton (Salton et al., 1993) utilise ce qu'il appelle les thèmes de textes. Cette méthode utilise des cartes de relations de textes pour déterminer des thèmes de textes. A partir des thèmes détectés, il est possible de faire des regroupements de parties du document pour créer des passages. Nous avons choisi d'utiliser la méthode du TextTiling de Hearst (Hearst, 1997) qui permet d'effectuer un découpage thématique en recherchant les ruptures thématiques contenues dans le document.

4.2.1 Principe du TextTiling

La méthode du TextTiling (Hearst, 1997) recherche les ruptures de thèmes et les identifie lorsqu'un bloc du document présente un moins grand nombre de mots traitant du thème. La méthode du TextTiling (figure 4) découpe tout d'abord (1) le document en blocs composés d'un nombre fixe de phrases (3 à 5 phrases généralement). Ensuite, (2) toutes les paires des blocs adjacents de textes sont comparées et une valeur de similarité leur est attribuée. (3) La

suite résultante des valeurs de similarités, après être mise sous forme de graphes et aplanie, est examinée pour déterminer les pics et les vallées sur le graphique. (4) Des valeurs de similarités élevées, impliquant que les blocs adjacents se suivent de façon logique, sont susceptibles de former des pics, tandis que des valeurs de similarités faibles, indiquant une potentielle limite entre les blocs, créent des vallées. Un pic correspond donc à deux blocs fortement liés thématiquement alors qu'une vallée correspond à une rupture de thèmes. Chaque vallée est donc considérée comme une rupture de thèmes et correspond à une limite entre deux blocs thématiquement différents.

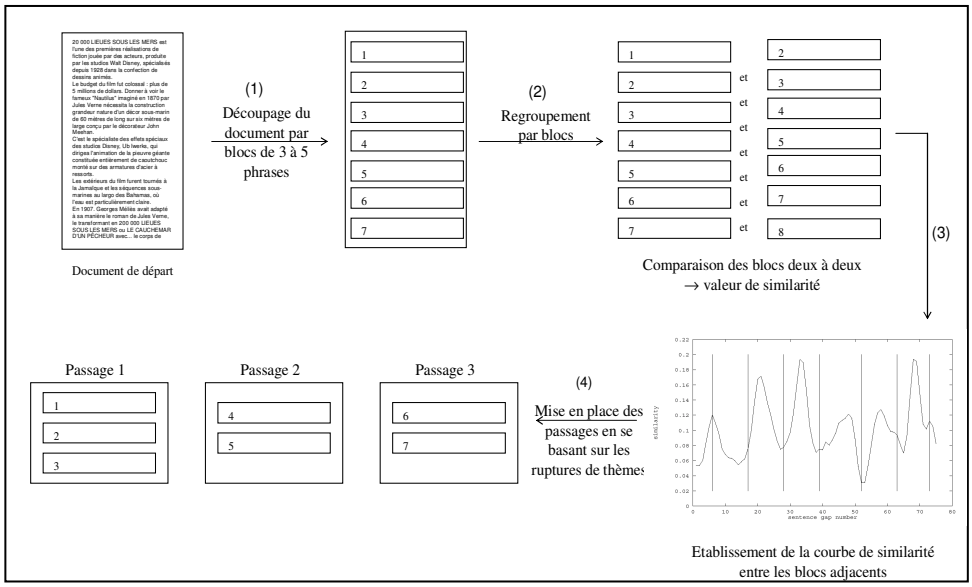


FIG. 4 – *Méthode du TextTiling*

Pour obtenir de bons résultats avec ce type d'analyse, il est préférable d'avoir des textes dont les termes caractéristiques des thèmes développés ne possèdent pas de synonymes.

4.2.2 Adaptation au contexte DEFT'05

Afin de déterminer les changements thématiques, nous avons implémenté une adaptation de la méthode du TextTiling. Nous avons gardé le principe du TextTiling et nous l'avons adapté au contexte des tours de paroles. Nous détaillons ici les principales étapes du processus (figure 4) :

- Découpage du document en blocs (1) ;
- Calcul de similarité des blocs pris 2 à 2 (2) ;
- Détermination du seuil permettant d'identifier les ruptures thématiques (3) ;
- Détermination des ruptures (4) ;

Le nombre de phrases constituant les blocs est de 15 tours de parole. Cette valeur a été déterminée suite à différentes expérimentations avec des blocs composés de 10 à 20 tours de paroles. Il s'est avéré que les meilleurs résultats ont été obtenus avec des blocs de 15 tours de paroles. Cette valeur se justifie car les allocutions de F. Mitterrand ont une longueur moyenne de 19 tours de paroles et la majorité des allocutions ont une longueur proche de 15 tours de paroles. Une fois ces blocs de 15 tours de paroles formés (1), la similarité entre les blocs pris deux à deux est calculée (2) :

$$sim(a,b) = \frac{\sum_{t=1}^{n} w_{t,a} w_{t,b}}{\sqrt{\sum_{t=1}^{n} w_{t,a}^2 \sum_{t=1}^{n} w_{t,b}^2}}$$

où t varie pour tous les termes du document et $w_{t,a}$ est le poids tf.idf[5] assigné au terme t dans le bloc a.

Une fois cette similarité calculée, il faut fixer le seuil qui permettra d'identifier les ruptures. Nous nous sommes mis dans l'optique d'utiliser ce découpage thématique comme base d'un lissage. Pour déterminer le seuil à appliquer, nous avons effectué plusieurs expérimentations et nous avons obtenu les meilleurs résultats en prenant un seuil correspondant aux 50 % des valeurs de similarités les plus faibles. Plus précisément, nous calculons les valeurs de similarité des blocs pris 2 à 2 (soit 1810 valeurs de similarité dans notre cas), puis nous les ordonnons par ordre décroissant, on se base alors sur la valeur médiane de similarité (soit la valeur de similarité à la position 905, une fois les valeurs ordonnées par ordre décroissant). Cette valeur correspond au seuil (3). Enfin, pour chaque valeur inférieure au seuil, on considère qu'il existe une rupture thématique entre les deux blocs correspondant à cette valeur de similarité (4). Si la valeur de similarité entre un bloc A et un bloc B est inférieure au seuil, la première phrase du bloc B est renvoyée dans un fichier résultat pour être ensuite utilisée dans le processus de détermination des phrases de F. Mitterrand (figure 5). La première phrase du bloc B sert donc de délimiteur de passages.

Les passages ainsi obtenus sont utilisés pour modifier le score des phrases. Le score de chaque passage est calculé en effectuant la moyenne des scores des phrases qu'il contient. Le score final d'une phrase résulte de la somme pondérée entre son score initial et le score du passage dans lequel elle se situe. Nous utilisons donc le TextTiling comme base d'un lissage plus que comme un réel outil de découpage thématique.

5 **tf.idf =** $\frac{\text{nombre d'apparitions du terme dans le bloc}}{\text{nombre de blocs contenant le terme}}$

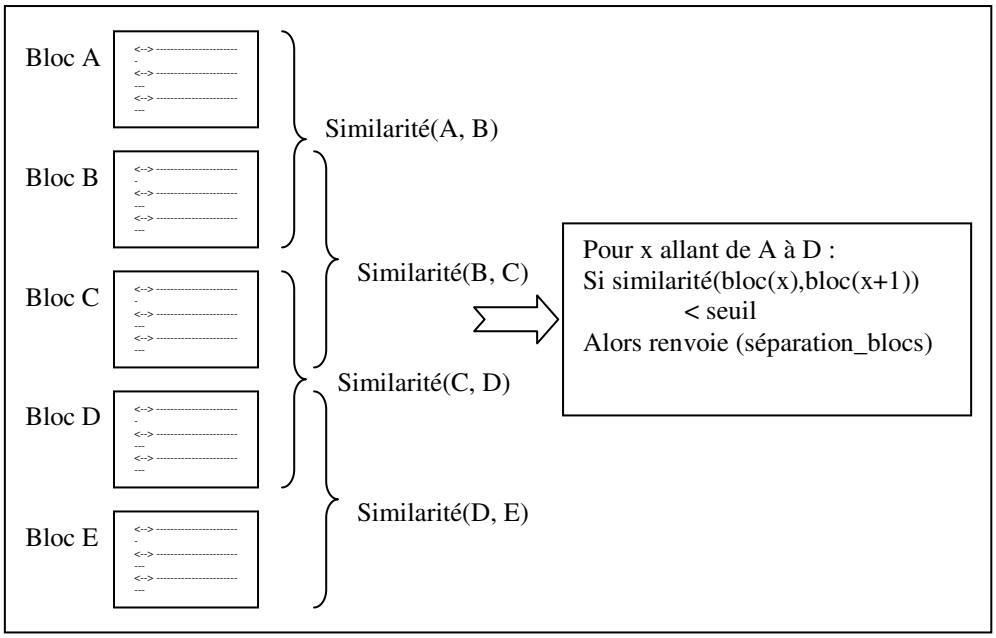


FIG. 5 – *Principe de détermination des ruptures*

4.2.3 Limites de la méthode

Cette méthode de découpage thématique connaît quelques limites. Tout d'abord, pour avoir de bons résultats, il faut que les allocutions des différents présidents soient thématiquement bien différentes. Or bien que les allocutions des deux présidents soient différentes sur la portée du thème (politique internationale contre politique nationale), cette différence est peut être insuffisante pour la méthode du TextTiling. En effet, mis à part le vocabulaire propre aux discours nationaux ou aux discours internationaux, les deux discours sont finalement très similaires car traitant tous les deux du thème plus général : la politique. C'est pourquoi, même si les deux discours sont thématiquement différents, ils peuvent être très similaires au niveau du vocabulaire utilisé dans l'ensemble. D'autre part, compte tenu du fait que l'on compare des blocs, la rupture thématique ne pourra être déterminée qu'entre deux blocs, or celle-ci peut avoir effectivement lieu au milieu d'un bloc. La méthode du TextTiling ne nous permet pas de le détecter. On constate donc que le choix de la taille du bloc est un problème important dans le bon fonctionnement du système. Il serait intéressant d'envisager une méthode basée sur des fenêtres glissantes afin de déterminer au mieux les limites entre les blocs (Callan et al., 1994). Les méthodes de fenêtres glissantes nécessitent une adaptation du calcul de l'idf. Enfin, une comparaison des blocs non contigus pourrait également être envisagée en vue d'améliorer les résultats.

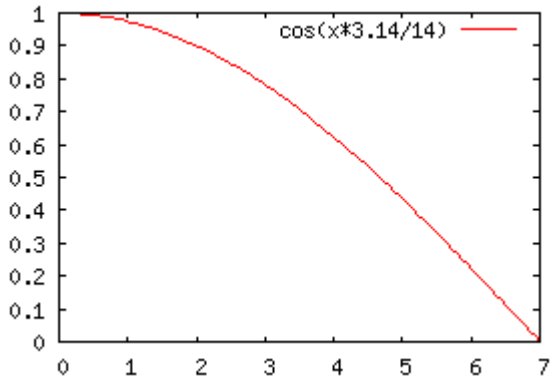
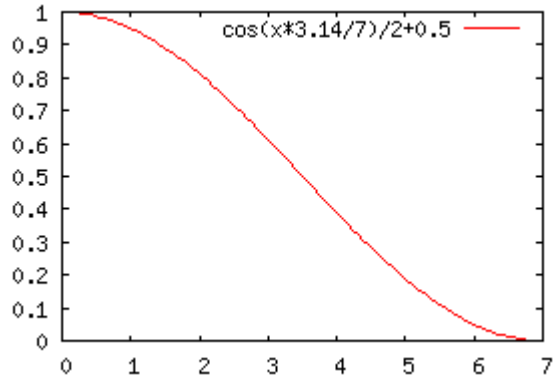
4.3 Lissage

Au sein de l'évaluation DEFT'05, les phrases de F. Mitterrand insérées dans les allocutions de J. Chirac sont regroupées ; les phrases ne sont pas indépendantes. Il est intéressant de prendre en compte cette caractéristique en tenant compte du contexte de chaque phrase. Notre module de lissage est basé sur une méthode qui prend en compte le score relatif au locuteur des phrases voisines dans le calcul du score d'une phrase. Le score d'une phrase est ainsi diffusé sur les phrases qui lui sont voisines au sein d'une même vision.

4.3.1 Méthode

Une fois l'ensemble des phrases évalué par l'une des méthodes d'apprentissage, le score de chaque phrase est recalculé par rapport aux scores des phrases voisines à l'aide de fonctions de lissage. Le calcul s'effectue à l'aide de l'équation suivante pour laquelle nous avons testé deux fonctions de lissage différentes (*A* et *B*) basées sur des cosinus :

$$St_i = S_i + \sum_{j \in [1,N]} f(j) * (S_{i-j} + S_{i+j})$$

A	$f(j) = \cos(\frac{j * \pi}{2N})$	
B	$f(j) = \frac{\cos(\frac{j * \pi}{N})}{2} + 0.5$	

TAB. 8 – *Fonctions de lissage*

Où St_i : score final de la i-éme phrase (par ordre de lecture) pour un président
S_i : score de la i-éme phrase pour un président obtenu par apprentissage
N : distance des phrases voisines à prendre en compte

4.3.2 Evaluation

Pour tester les fonctions de lissage et la taille de la fenêtre, nous avons utilisé les résultats obtenus à l'aide de la combinaison Rocchio et ltc (F-score de 0,3340). Les résultats de cet apprentissage ont donc été recalculés à l'aide des fonctions précédentes et en faisant varier la taille de la fenêtre utilisée. Les insertions de F. Mitterrand ayant une taille moyenne de 19 phrases dans notre corpus d'apprentissage, nous sélectionnons des distances de lissage qui soient inférieures à 9 phrases, de manière à ce que le lissage corresponde aux insertions.

	Distance des phrases voisines			
Fonction de lissage	5	6	7	8
A	0,6664	0,6736	0,6691	-
B	-	0,6696	**0,6742**	0,6731

TAB. 9 – *Résultats du lissage (F-score)*

La fonction de lissage B donne de meilleurs résultats car elle privilégie les phrases proches au sein de la fenêtre. La distance de lissage qui semble la plus adaptée correspond à une fenêtre de 7 phrases. L'utilisation du lissage donne une amélioration de 100% des résultats obtenus à l'aide des profils et permet de dépasser le F-score moyen des participants à DEFT05.

5 Evaluation

5.1 Les différentes tâches de DEFT'05

DEFT'05 est composée de trois tâches qui correspondent au traitement de trois variations du corpus. La tâche 1 correspond à l'étude du corpus dans lequel les années et les noms de personnes sont remplacés par les balises <*date*> et <*nom*>. Dans la tâche 2 seules les années sont remplacées par la balise <date>. La tâche 3 conserve toutes les informations du corpus. Tous les résultats précédemment cités ont été obtenus sur le corpus de la tâche 3.

5.2 Processus

Notre approche n'ayant été entraînée que pour la tâche 3, les mêmes types d'exécutions ont été soumis pour les trois tâches (tableau 10).

- L'exécution 1 consiste en un apprentissage par personne en ltc couplé avec un apprentissage sur les allocutions de type Rocchio (voir partie 2). Sur cet

apprentissage, un lissage basé sur la fonction B avec une taille de fenêtre 7 est appliqué ;
- L'exécution 2 est similaire à la première mais l'apprentissage de type Rocchio est effectué sur les phrases ;
- L'exécution 3 est similaire à celle de la tâche 1, à la différence près que les résultats de l'apprentissage sont d'abord modifiés en prenant en compte le poids global des passages détectés. Le lissage n'est effectué qu'après cette étape ;

Exécution 1	Exécution 2	Exécution 3
Apprentissage sur les présidents (Ltc) + apprentissage sur les allocutions (Rocchio) + lissage	Apprentissage sur les présidents (Ltc) + apprentissage sur les phrases (Rocchio) + lissage	Apprentissage sur les présidents (Ltc) + apprentissage sur les allocutions (Rocchio) + découpage thématique + lissage

TAB. 10 – *Récapitulatif des trois exécutions*

5.3 Résultats

Les résultats des trois tâches sont comparés aux moyennes de l'évaluation DEFT'05 comme présentés dans le tableau suivant :

Tâche	Exécution	Précision	Rappel	F-score
1	moyenne	-	-	0,6229
	1	0,7477	**0,7549**	**0,7513**
	2	0,9265	0,4216	0,5795
	3	**0,9415**	0,2669	*0,4159*
2	moyenne	-	-	0,6738
	1	0,7533	**0,7563**	**0,7548**
	2	**0,9246**	0,4300	0,5871
	3	0,7943	0,6725	0,7283
3	moyenne	-	-	0,6902
	1	0,7534	**0,7568**	**0,7551**
	2	**0,9268**	0,4337	0,5909
	3	0,7923	0,6725	0,7275

TAB. 11 – *Résultats d'évaluation DEFT'05*

Les meilleurs résultats obtenus dans les trois tâches sont ceux obtenus à l'aide de l'exécution 1. L'utilisation des phrases à la place des allocutions donne un F-score largement inférieur, mais la précision obtenue est améliorée. La troisième exécution donne des résultats intermédiaires sauf pour la première tâche où le résultat est faible. Il est intéressant de

remarquer que nos résultats sont stables sur les trois tâches. Cette stabilité peut s'expliquer par le fait que notre système se base plus sur la forme des phrases que sur les informations telles que les noms et les dates et par conséquent celui-ci est peu affecté par leur suppression.

Nous remarquons aussi que les résultats obtenus à l'aide du corpus final sont globalement supérieurs à ceux obtenus sur notre corpus d'entraînement. Cela peut venir du fait que le nombre de dépendances possibles est grand, cela est visible lors de l'interrogation des profils où un nombre important de dépendances n'est pas trouvé dans aucun des deux profils. L'augmentation de la taille du corpus d'apprentissage permet d'avoir une meilleure couverture des dépendances pouvant être utilisées par les présidents et par conséquent améliore les résultats.

Enfin nous soulignons l'intérêt de notre deuxième exécution, en effet si cette expérimentation ne fournit pas les meilleurs F-score, lors de l'évaluation elle donne de très bonnes performances en terme de précision. Elle est située sur le front de Pareto sur les trois tâches, notamment elle donne les meilleurs résultats en terme de précision sur les tâches 2 et 3. Cette amélioration de la précision est essentiellement due à la fonction de lissage, sur notre corpus d'évaluation, avant le lissage la précision obtenue est de 23%, après lissage celle-ci passe à 79%.

6 Conclusion et perspectives

Notre participation à DEFT'05 nous a permis d'évaluer l'intérêt d'une approche basée sur des éléments représentatifs de la syntaxe pour la détection de phrases ainsi que l'intérêt de la détection de passages dans un contexte de fouilles de texte. Les résultats obtenus montrent que dans ce cas les dépendances syntaxiques permettent d'obtenir de bonnes extractions, qui sont notamment meilleures que les extractions basées sur les lemmes ou les mots. Les évaluations des différentes dépendances montrent que celles qui représentent les collocations nominales ou verbales sont celles qui capturent le mieux la différence entre les deux discours. Cependant les allocutions des deux présidents étant issues de discours rédigés à l'avance, il se peut que des tournures de phrases soient communes aux deux discours. Dans un contexte de discours spontané, les tournures et expressions propres à chacun sont marquées, ce qui nous laisse penser que notre méthode basée sur les dépendances donnerait de meilleurs résultats dans ce contexte.

Les résultats présentés dans l'article sont obtenus à l'aide d'un unique analyseur syntaxique (XIP) une partie des résultats dépend donc de la qualité de cet analyseur et des informations qu'il produit. La méthode proposée quand à elle est indépendante de l'analyseur, tester cette méthode avec un autre analyseur permettrait d'évaluer la généricité de l'approche.

Les algorithmes d'apprentissage que nous avons utilisés sont assez simples, étudier l'impact d'un descripteur tel que les dépendances sur des apprentissages plus performants tels que ceux basés sur des Modèles de Markov Cachés amélioreraient sûrement les performances du système.

On peut remarquer également que la présence ou non d'informations telles que des dates ou des noms a peu d'impact sur une telle méthode. Toutefois, la prise en compte de telles informations dans un module supplémentaire permettrait d'améliorer notre système. La détection de passages permet d'améliorer la précision du système. Une détection plus précise

des ruptures augmenterait cette précision. Une première solution serait l'utilisation d'un prétraitement du corpus utilisant un anti-dictionnaire.

Références

Aït-Mokhtar S., Chanod J.P., Roux C. (2002) Robustness beyond shallowness : Incremental Deep Parsing. *the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data*, Cambridge University Press, 121-144.

Brouard C. (2002) RELIEFS : un système d'inspiration cognitive pour le filtrage adaptatif de documents textuels, *Revue des Sciences et Technologies de l'Information*, vol7, no1/2, 157-182.

Callan (1994) Passage-level evidence in document retrieval. *Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hearst M.A. (1997) TextTiling: Segmenting Text into Multi-paragraphe Subtopic Passage, Actes de *Computational Linguistics*, 33-64.

Metzler D.P., Haas S.W. (1989) The constituent object parser: syntactic structure matching for information retrieval, *ACM Transactions on Information Systems*, vol. 7, n°3, 296-316.

Salton, Allan, Buckley (1993) Approaches to passage retrieval in full text information systems, *ACM SIGIR Conference on Research and Development in Information Retrieval*

Smeaton A.F. (1999) Using NLP or NLP Resources for Information Retrieval Tasks, *Natural Language Information Retrieval*, T. Strzalkowski (Ed.), Kluwer Academic Publishers, 99-111.

Strzalkowski T., Carballo J.P., Marinescu M. (1994) Natural Language Information Retrieval: TREC-3 Report. *Overview of the Third Text REtrieval Conference TREC 1994*, 39-54.

Wilkinson (1994), Effective retrieval of structured models, *ACM SIGIR conference on Research and development in information retrieval,* 311-317.

Zobel, Moffat, Wilkinson, Sacks-Davis (1994), Efficient retrieval of partial documents, *Information Processing and Management,* vol. 31, n°3, 361-377.

Annexe

Liste des dépendances syntaxiques de XIP :

Les relations entre le sujet et le verbe	
SUBJ(X,Y)	Y est sujet de X
SUBJCLIT(X,Y)	Y est sujet de X mais c'est un pronom redondant ex : jean dort-il?

Le verbe et ces compléments	
VARG(X,Y)	Y est argument de X
VMOD(X,Y)	Y est un complément du verbe ex : 'pierre vient pour parler': VMOD(vient,parler)
VMOD(X,Y,Z)	Z et un syntagme prépositionnel qui complète X et qui est introduit par la préposition Y
AUXIL(X,Y)	Y est auxiliaire de X
le nom et ces compléments	
NARG(X,Y,Z)	Z complète X est introduit par la préposition Y
NARG(X,Y)	Y complète X
NMOD(X,Y,Z)	Z est un modificateur du nom X introduit par la préposition Y
NMOD(X,Y)	Y est un modificateur du nom X
NN(X,Y)	Y est un syntagme nominal apposée à X
l'adjectif et ces compléments	
ADJARG(X,Y)	Y est argument de X
La coordination d'éléments	
COORDITEMS(X,Y,Z)	X et Z sont deux éléments coordonnée par Y
Déterminant	
DETERM(X,Y)	X déterminant du syntagme nominal Y
antécédent	
COREF(X,Y)	X est l'antécédent de Y
forme	
NEGAT(X)	X : verbe utiliser sous une forme négative
REFLEX(X,Y)	Structure réflexive Y pronom réflexif
SEQNP(X,Y)	Séquence de NP
dépendances spécialisées	
NPR(X,Y)	Parenthèse
OBJGM(X,Y)	Guillemets
NGM(X,Y)	Guillemets

Summary

In this article, we show the interests of syntactic dependencies and passage detection methods to identify a speaker in a discourse. We based our analysis on the methods we used for DEFT'05 evaluation campaign where the aim is to differentiate Francois Mitterrand discourse from Jacques Chirac one. In this campaign we first evaluated the use of syntactic dependency as unit characterizing differences between the two discourses. These units extracted by natural language processing are the selected descriptor for representing the discourse; consequently we used a learning based on these descriptors. At last, as the two presidents' discourses belong from different theme, we combine the dependency based learning with topic segmentation in order to improve sentence selection.