Identification de thème et reconnaissance du style d'un auteur pour une tâche de filtrage de textes

Michèle Jardino*, Martine Hurault-Plantet*, Gabriel Illouz*

*LIMSI-CNRS, BP 133, 91403 ORSAY Cedex
{Michele.Jardino, Martine.Hurault-Plantet, Gabriel.Illouz}@limsi.fr
http://www.limsi.fr/Scientifique/lir

Résumé. Pour résoudre une tâche de filtrage des textes d'un auteur, insérés dans les textes d'un autre auteur, nous avons utilisé à la fois le style de l'auteur et la structure thématique du texte. Nous caractérisons le style d'un auteur par un modèle de langage n-grammes de mots ou de caractères entraîné sur un corpus d'apprentissage. Nous appliquons ensuite les modèles sur chaque phrase du corpus de test pour en calculer l'auteur le plus probable. Un algorithme de lissage transforme ensuite les résultats en segments continus pour chaque auteur. Parallèlement, nous avons élaboré une méthode d'identification du thème de chaque auteur dans un document. Nous déterminons d'abord les segments de texte de plus grande densité, pour chaque mot du document, par chaînage lexical. Puis, nous recherchons les chaînes lexicales principales des deux thèmes, par hypothèse celles dont les segments respectifs sont les plus étendus et se recouvrent le moins. Les résultats des deux méthodes sont finalement fusionnés.

1 Introduction

La tâche soumise à évaluation dans l'atelier DEFT'05 de la conférence TALN 2005 consistait à séparer les allocutions respectives de deux hommes politiques dans le même document. L'atelier d'évaluation s'est déroulé en deux temps : une phase d'entraînement pendant laquelle nous avons disposé d'un corpus d'apprentissage, puis une phase de test avec un nouveau corpus sur lequel l'évaluation proprement dite a été faite. Chaque corpus est composé d'allocutions de Jacques Chirac dans lesquelles des segments d'allocutions de François Mitterrand ont été glissés. Le nombre de phrases de Chirac est, dans chaque corpus, nettement plus important que le nombre de phrases de Mitterrand. L'évaluation de la tâche se fait par rapport à la reconnaissance des phrases de Mitterrand. La tâche ainsi définie est du filtrage de textes dans le sens où les phrases d'un auteur donné doivent être filtrées dans le flux des textes d'un autre auteur. Dans l'atelier d'évaluation, trois tâches étaient proposées pour trois versions différentes du corpus de test : une version avec des étiquettes à la place des noms propres et des dates, une version avec des étiquettes à la place des dates, et une version intégrale comportant les noms propres et les dates.

Deux critères peuvent aider à séparer les textes des deux auteurs dans chaque allocution : l'ensemble des caractéristiques d'écriture, propre à chaque auteur et qui représente son style, et les thèmes abordés. En effet, le thème de chaque allocution de Chirac a été choisi différent de celui du fragment d'allocution de Mitterrand qu'il contient. Nous avons donc essayé de

filtrer les phrases de Mitterrand en utilisant d'une part le style de l'auteur avec une méthode par apprentissage, et d'autre part les thèmes du texte avec une méthode sans apprentissage.

La reconnaissance automatique du style d'un auteur (stylométrie) s'appuie sur différents indicateurs comme les mots les plus fréquents ou les mots-outils (Holmes, 1998), les syllabes et les catégories morpho-syntaxiques (Beaudoin et Yvon, 2004) ou encore les caractères (Peng *et al.*, 2003). Dans les premières études, ces indicateurs était utilisés de manière simple, par exemple en termes de présence ou absence dans les textes à identifier. Dans les travaux plus récents, des modèles de langages markoviens de type n-grammes (Jelinek, 1998) ont été expérimentés avec succès pour reconnaître des auteurs à l'aide des caractères (Teahan, 2000, Kmelev *et al.*, 2001, Peng *et al*., 2003) ou pour caractériser des styles dans des pièces écrites en vers à partir des syllabes métriques (Beaudoin et Yvon, 2004). Au vu des performances obtenues il nous a semblé intéressant d'appliquer ces modèles de Markov à la tâche DEFT'05. Nous avons dans un premier temps expérimenté des modèles n-grammes de mots (Hurault-Plantet *et al.*, 2005) puis des modèles n-grammes de caractères (Jardino, 2006) qui se sont révélés plus performants. Nous construisons pour chaque auteur un modèle de langage probabiliste de type n-grammes à partir du corpus d'apprentissage. Chacun des modèles est ensuite appliqué sur chacune des phrases à tester, l'auteur reconnu est alors celui dont le modèle donne la plus grande probabilité à la phrase. Le système attribue ainsi à chaque phrase l'un des deux auteurs. Un algorithme est ensuite utilisé pour reconstituer des segments de texte continus pour chaque auteur.

Parallèlement, nous avons élaboré une méthode d'identification de thème basée sur la notion de chaîne lexicale. Une chaîne lexicale relie toutes les occurrences d'un même mot, et éventuellement des mots qui lui sont sémantiquement liés, dans un texte. La chaîne s'interrompt lorsque deux occurrences consécutives du mot considéré sont trop distantes l'une de l'autre. Cette notion a d'abord été utilisée pour mesurer la cohésion lexicale d'un texte (Morris et Hirst, 1991). En segmentation thématique, les chaînes lexicales sont utilisées pour détecter les ruptures thématiques soit en détectant les endroits de forte concentration de fins et débuts de chaîne dans le texte (Stokes *et al*., 2002), soit en utilisant les similarités entre phrases calculées sur les chaînes actives (Sitbon et Bellot, 2004). Les mots utilisés pour construire les chaînes lexicales sont en général les mots *porteurs de sens*, excluant donc les mots *vides* et les verbes. Nous considérons que chaque auteur développe dans son allocution un thème central représenté par un ensemble de mots. Dans chaque allocution, nous recherchons deux thèmes distincts, un pour chaque auteur. Lorsque nous ne pouvons pas identifier deux thèmes distincts, nous supposons qu'il n'y a qu'un seul thème et donc un seul auteur, et qu'il n'y a donc pas d'insertion de phrases de Mitterrand dans l'allocution de Chirac. Pour identifier ces thèmes, nous utilisons les liens de cooccurrence entre chaînes lexicales sur l'ensemble du texte.

Les évaluations présentées dans les sections qui suivent ont été faites sur la version intégrale du corpus de test de DEFT'05, c'est-à-dire celle comportant les noms propres et les dates. Ce choix est motivé par le fait que les résultats obtenus pour les trois versions du corpus sont très peu différents (Alphonse *et al*., 2005), les noms propres et les dates étant de fréquence faible dans chaque texte. Les mesures utilisées sont la précision, le rappel et le

F-score[1]. La section 2 est consacrée à la méthode de reconnaissance de style. Dans la section 3, nous présentons la méthode d'identification du thème de chaque auteur dans une allocution, puis dans la section 4, la méthode de fusion des résultats. Nous conclurons sur un bilan de ces méthodes et une comparaison avec les méthodes utilisées par les autres participants à l'atelier DEFT'05.

2 Reconnaissance du style d'auteur avec des modèles de langage n-grammes

Les modèles de langage n-grammes de mots ont été initialement développés dans les systèmes de reconnaissance de la parole (Jelinek, 1998) avec des valeurs typiques de *n* égales à 3 ou 4. Ils sont maintenant également utilisés dans des systèmes de traitement automatique de la langue comme la recherche d'information (Alvarez et al, 2004; Ponte et Croft, 1998), avec une portée réduite à 1. Les modèles n-grammes ont été appliqués à d'autres entités que les mots, par exemple les caractères, *n* variant de 2 à 6, comme signature d'auteur (Markov, 1913; Peng et al, 2003) ou de langue (Shannon, 1951) ou les parties du discours pour identifier le genre des textes (Illouz et Jardino, 2001) ou encore les syllabes métriques pour caractériser le style de pièces écrites en vers (Beaudoin et al, 2004). Comme ce sont des modèles probabilistes, leur apprentissage est facilité par le grand nombre de textes électroniques disponibles aujourd'hui. La connaissance capturée par ces modèles est contextuelle, elle permet de prédire un item connaissant les *n-1* items qui le précédent (modèle de Markov).

En utilisant des n-grammes de mots incluant la ponctuation nous espérions capturer le style de l'auteur puisque le modèle n-grammes de mots mélange des informations de différentes natures comme le lexique de l'auteur, la syntaxe qu'il emploie localement ou encore les signes qui ponctuent ses phrases. C'est ce que nous décrivons dans les paragraphes 2.1 à 2.4 et qui a été mis en œuvre pour l'évaluation DEFT'05. Dans une expérience postérieure décrite au paragraphe 2.5 nous montrons que cette information est mieux prise en compte par un modèle n-grammes de caractères.

Dans le paragraphe 2.1 une expérience préliminaire de partitionnement en deux des phrases nous permet de choisir un ensemble de mots pour bâtir les modèles n-grammes de mots. Dans la partie 2.2 nous construisons les modèles pour chaque auteur pour différentes valeurs de n et déterminons la valeur optimale du nombre de mots n à prendre en compte. Puis nous utilisons ces modèles pour reconnaître les phrases de Mitterrand insérées dans les allocutions de Chirac. La partie 2.3 montre le pouvoir discriminant des modèles. La partie 2.4 décrit deux méthodes de lissage des résultats obtenus avec les n-grammes de mots pour prendre en compte les caractéristiques de la tâche DEFT'05, à savoir le repérage d'au plus une suite de phrases de Mitterrand dans chaque allocution de Chirac, le nombre minimal des phrases de Mitterrand étant 2. La partie 2.5 décrit le modèle n-grammes de caractères et présente ses performances.

[1] Si *N* est le nombre de phrases de Mitterrand dans le corpus, *A*, le nombre de phrases attribuées à Mitterrand par notre système, et *B* le nombre de phrases de Mitterrand dans A, on a :

Précision = B / A Rappel = B / N F-score(1) = 2 x Précision x Rappel / (Précision + Rappel)

2.1 Sélection de mots pour reconnaître un auteur

Nous avons réalisé un ensemble d'expériences préliminaires sur le corpus d'apprentissage décrit ci-dessous pour rechercher l'ensemble de mots le plus discriminant pour séparer les phrases de Chirac et de Mitterrand. Un mot est une suite de caractères séparée du mot précédent et du mot suivant par un espace[2].

2.1.1 Corpus d'apprentissage

Le corpus d'apprentissage comprend 587 discours de Chirac dont 400 contiennent une séquence de phrases de Mitterrand. Nous avons formaté ce corpus en remplaçant toutes les majuscules par des minuscules et en conservant tous les signes de ponctuation. Quelques statistiques du corpus sont rassemblées dans le tableau 1.

	Nb phrases	Nb mots		
		lexique	total	moyen/phrase
Chirac	49 890	27 069	1 254 924	25 (16)
Mitterrand	7 523	13 858	246 552	33 (24)

TAB. 1 – *Statistiques du corpus d'apprentissage. Les mots incluent les signes de ponctuation. Entre parenthèses sont indiqués les écarts-type.*

La plupart des 13 858 mots employés par Mitterrand appartiennent au lexique de Chirac, soit 10 707 mots en commun. Les 3 151 mots du lexique de Mitterrand qui ne sont pas employés par Chirac sont majoritairement des hapax (2 636 mots). L'information sur le recouvrement des lexiques à la manière de l'indice de Jaccard est donc clairement insuffisante pour reconnaître les phrases de Mitterrand de celles de Chirac. Par contre il nous a paru intéressant de tester si l'information capturée par la distribution de mots ou d'un sous-ensemble de mots dans les phrases permettait une séparation du corpus d'apprentissage en deux, c'est ce que nous décrivons dans le paragraphe suivant.

2.1.2 Partition automatique des phrases pour différents ensembles de mots

Nous avons partagé de manière automatique toutes les phrases du corpus d'apprentissage en deux classes avec un algorithme de classification non supervisée (Jardino, 2000) pour différents ensembles de mots.

La mesure utilisée pour la classification prend en compte la distribution des mots dans chaque phrase, c'est-à-dire leur fréquence relative dans la phrase. L'algorithme de classification tend à trouver pour chacune des deux classes de phrases une distribution moyenne des mots la plus proche possible de celles des phrases contenues dans la classe. Cette méthode appliquée à un corpus bien constitué (le Brown Corpus) avait permis de retrouver automatiquement des partitions en genre faites manuellement soit 4 classes étiquetées fiction, presse, non-fiction et divers, ceci avec des taux de rappel et précision proches de 100% (Illouz et Jardino, 2001).

[2] Les ponctuations ont été préalablement mises entre deux espaces.

Le tableau 2 reporte les valeurs de précision, rappel et F-score obtenues sur le corpus d'entraînement initial, pour différents ensembles de mots. Ces ensembles de mots correspondent à des choix de fréquence des mots dans le corpus d'apprentissage. Ils permettent d'expérimenter différentes zones de la courbe de Zipf comme les zones de haute fréquence où se regroupent les mots-outils, et les zones de moyenne fréquence généralement associées aux mots porteurs de sens. Ils incluent les signes de ponctuation. Nous avons également expérimenté une représentation réduite au point et à la virgule.

Ensemble de mots	Rappel	Précision	F-score
Tous les mots	0,66	0,19	0,30
Mots de fréquence > 500	0,50	0,18	0,26
Mots de fréquence < 500	0,59	0,17	0,26
Mots tels que 10 < fréquence < 500	0,57	0,16	0,25
Point et Virgule	0,50	0,12	0,19

TAB. 2 – *Rappel et précision des phrases de Mitterrand pour une partition non supervisée des phrases en deux classes pour différents ensembles de mots, ponctuation incluse*.

Les scores montrent que pour le corpus de DEFT'05 la séparation automatique en deux classes est loin de coïncider avec les deux classes réelles constituées respectivement des phrases de Chirac et de Mitterrand. On retrouve plus de 50% des phrases de Mitterrand mais le taux de précision est très faible car beaucoup de phrases de Chirac se retrouvent dans la classe où les phrases de Mitterrand sont majoritaires. Néanmoins, on observe que les meilleurs résultats sont obtenus en conservant tous les mots et la ponctuation. En conséquence, pour toutes les expériences suivantes prenant en compte l'ordre des mots dans la phrase, nous avons conservé tous les mots ainsi que la ponctuation.

En corollaire de cette expérience, on peut remarquer que si la distribution des mots dans les textes nous avait permis de reconnaître un genre avec un F-score proche de 100% (Illouz et Jardino, 2001), elle ne suffit pas ici pour discriminer les phrases de Mitterrand de celles de Chirac car les phrases de Chirac et de Mitterrand du corpus DEFT'05 partagent un même genre, celui du discours politique.

2.2 Modèles de langage n-grammes pour identifier l'auteur de chaque phrase

Nous avons utilisé le logiciel de CMU (Clarkson, 1997) qui permet de construire et d'évaluer des modèles de langage n-grammes. Pour une valeur de *n* donnée, ces modèles permettent de calculer les probabilités d'obtenir un mot connaissant les *n-1* mots qui le précèdent (modèles de Markov d'ordre *n*). Les probabilités sont calculées à partir des fréquences des suites de mots observées dans le corpus d'apprentissage. Pour les évènements non observés, qui deviennent de plus en plus nombreux quand l'ordre *n* augmente, on se replie sur les probabilités de n-grammes d'ordre inférieur pondérées à la manière de Witten-Bell (Witten et Bell, 1991) qui prend en compte le nombre de contextes dans lesquels sont observés les n-grammes dans le corpus d'apprentissage.

Nous avons partagé le corpus d'apprentissage en deux sous-ensembles : un pour les 49 890 phrases de Chirac et un pour les 7 523 phrases de Mitterrand. À partir de chaque sous-ensemble, nous avons construit plusieurs modèles de langage n-grammes en faisant varier *n* de 1 à 8.

Ces modèles ont été appliqués aux phrases du corpus de test de DEFT'05 pour permettre d'identifier les phrases de Mitterrand insérées dans les allocutions de Chirac. Ces phrases ont été formatées comme les phrases du corpus d'apprentissage, le tableau 3 en donne quelques caractéristiques. Le nombre moyen de mots par phrase est comparable à celui du corpus d'apprentissage.

Nb allocutions	Nb phrases	Nb mots	Nb mots moyen/phrase
294	27 162	703 653	26 (17)

TAB. 3 – *Statistiques du corpus de test. Entre parenthèses sont indiqués les écarts-type.*

On calcule pour chaque phrase du corpus de test deux probabilités données respectivement par le modèle Mitterrand et par le modèle Chirac. La probabilité d'une phrase est le produit des probabilités de chaque mot de la phrase étant donné les mots qui le précèdent. En représentant une phrase de longueur L par la succession de ses mots "$m_1 \dots m_i \dots m_L$", la probabilité de cette phrase P_A (phrase) est calculée à partir du modèle n-grammes de mots pour l'auteur A selon l'équation (1).

$$P_A(\text{phrase}) = \Pi_i \, p_A (m_i / m_{i-n+1} \dots m_{i-1})$$

où $p_A (m_i / m_{i-n+1} \dots m_{i-1})$ est la probabilité donnée par le modèle de l'auteur A que le mot m_i succède à la chaîne de n-1 mots $m_{i-n+1} \dots m_{i-1}$. S'il n'y a pas de mots précédents (mots en début de phrase) ou s'il n'existe pas de prédiction du mot dans le contexte du test, on prédit celui-ci selon un contexte réduit par la méthode du repli évoquée plus haut. La phrase est affectée à l'auteur dont le modèle donne la plus grande probabilité.

Nous avons utilisé ces modèles du plus simple : l'unigramme de mots qui fait partie de la famille des représentations de type *sac de mots*, jusqu'à des représentations octogrammes de mots. La valeur de *n* dans n-grammes détermine le choix de la taille du passé, elle dépend de la tâche et en particulier de la taille des données. Nous l'avons évalué en terme de F-score sur un corpus de test. Pour la tâche DEFT'05, nous avions partagé le corpus d'entraînement étiqueté en deux parties de taille différente : 90% consacrée à l'apprentissage des modèles n-grammes et 10% à leur évaluation. Nous avions trouvé un maximum pour n=3 avec F=0,67 (Hurault-Plantet et al, 2005). Après l'atelier d'évaluation nous avons recommencé ces expériences en prenant le corpus d'entraînement initial étiqueté en entier pour l'apprentissage et le corpus de test pour l'évaluation. Le tableau 4 montre que le modèle n-grammes donne des résultats assez plats à partir de n=3 avec F=0,48.

n	1	2	3	4	6	8
Rappel	0,83	0,62	0,58	0,59	0,60	0,60
Précision	0,21	0,36	0,40	0,41	0,40	0,39
F-score	0,34	0,46	**0,48**	**0,48**	**0,48**	0,47

TAB. 4 – *Scores de reconnaissance des phrases de Mitterrand insérées dans des allocutions de Chirac avec des modèles n-grammes de mots pour n variant de 1 à 8, sur le test intégral.*

Le contraste entre les résultats donnés par les deux corpus de test montre que le premier corpus de test que nous avions sélectionné n'était pas assez représentatif de la tâche. Une estimation moyenne sur 10 textes tests, obtenus en fractionnant le texte d'apprentissage et en utilisant à chaque fois les neuf dixièmes restants pour l'apprentissage des modèles, aurait donné des résultats vraisemblablement plus proches de ceux obtenus sur le « vrai » test.

2.3 Pouvoir discriminant des modèles

Pour comprendre les résultats moyens obtenus avec des n-grammes de mots, nous avons fait une analyse statistique des probabilités des phrases, en nous inspirant des travaux de Shannon sur l'entropie de l'anglais (Shannon, 1951). Nous avons voulu vérifier dans quelle mesure on pouvait trouver une entropie liée spécifiquement à un auteur à partir de l'entropie des phrases du test de DEFT'05. Il existe un lien direct entre la probabilité d'une phrase P_A (phrase) et son entropie H_A (phrase) qui est :

$$H_A \text{ (phrase)} = - (1/L) \log (P_A).$$

L'entropie H_A (phrase) est une grandeur moyenne qui permet de s'affranchir de la longueur des phrases et qui varie entre 0 et la taille du lexique, en sens inverse de la probabilité de la phrase. Nous avons relevé pour chaque phrase du corpus de test les valeurs d'entropie données par les deux modèles 6-grammes de mots de Chirac et Mitterrand et compté combien de phrases de Chirac puis de Mitterrand avaient la même entropie avec un pas de 0.01. Ces valeurs sont représentées sur la figure 1.

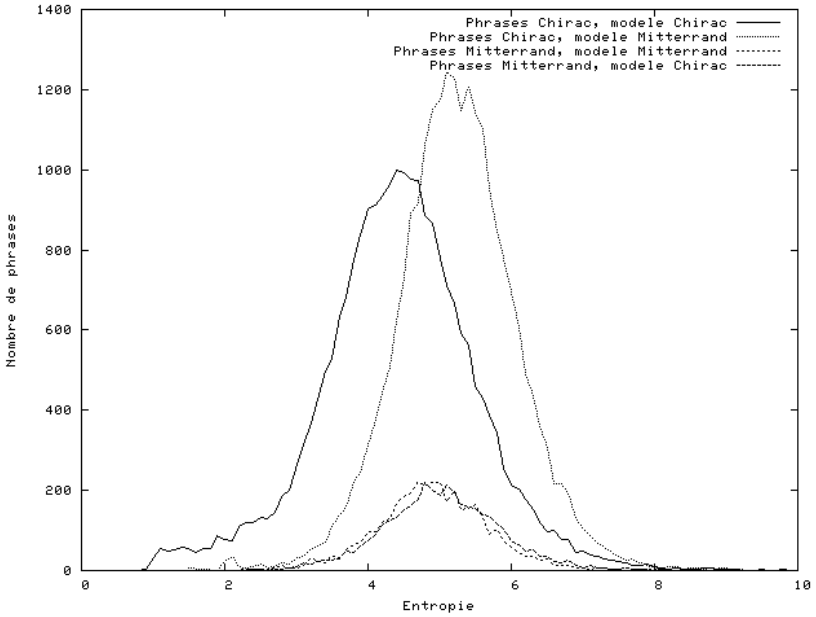

FIG. 1 – *Nombres de phrases de Chirac et de Mitterrand du corpus de test en fonction des valeurs d'entropie de ces phrases calculées avec les deux modèles 6-grammes de mots de Chirac et Mitterrand.*

Les deux courbes les plus hautes sont associées aux phrases de Chirac, la plus à gauche est donnée par le modèle de Chirac, la valeur moyenne est 4,42. La courbe à droite donnée par le modèle Mitterrand a une valeur moyenne d'entropie plus forte, 5,24 et se distingue nettement de l'autre courbe. Par contre les deux courbes plus basses associées aux phrases de Mitterrand se distinguent très peu, les entropies moyennes sont 4,91 et 5,03. On peut en déduire que le modèle Chirac sépare beaucoup mieux les phrases de Chirac de celles de Mitterrand, d'ailleurs si on calcule le F-score de reconnaissance des phrases de Chirac on obtient F = 0,89 , à comparer à F = 0,48 pour la reconnaissance des phrases de Mitterrand.

Il parait donc difficile d'attribuer une entropie d'auteur à partir de l'entropie des phrases au vu de ces courbes. Plusieurs facteurs peuvent expliquer la difficulté du modèle Mitterrand à séparer les phrases de Mitterrand de celles de Chirac. Comme signalé précédemment le lexique de Mitterrand (13 858 mots) est beaucoup plus petit que celui de Chirac (27 069 mots), les deux lexiques ayant 10 708 mots en commun. L'écart entre le lexique commun et les lexiques de chacun indique que le modèle Chirac permet de prédire une grande partie des phrases de Mitterrand, ce qui n'est pas le cas du modèle de Mitterrand qui donne une probabilité faible aux mots qui ne sont pas dans son lexique. Un autre facteur est la faible taille des phrases à identifier (Juola, 1997). On peut s'en rendre compte en calculant les entropies des 294 allocutions de Chirac et des 199 ensembles de phrases de Mitterrand insérées dans ces allocutions. La distribution de ces valeurs est représentée sur la figure 2.

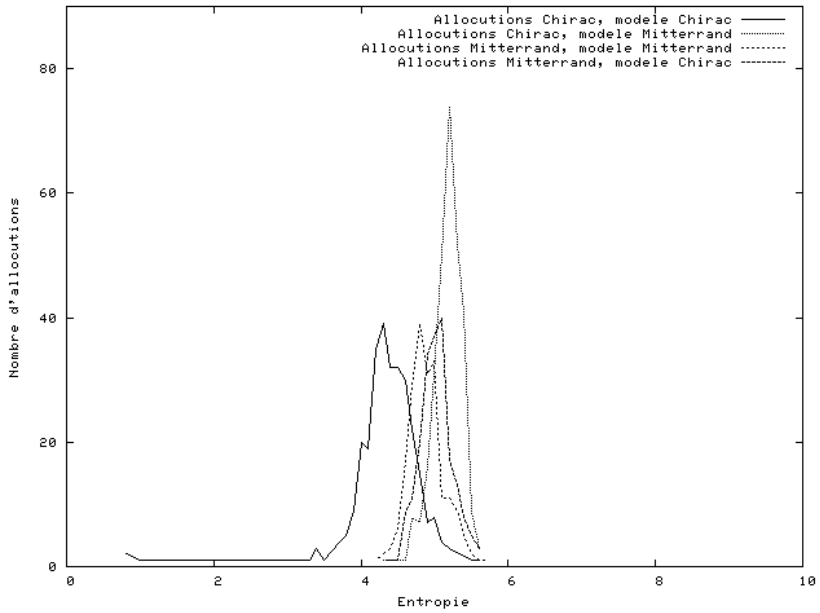

FIG. 2 – *Nombres d'allocutions de Chirac et de Mitterrand du test intégral en fonction des valeurs d'entropie calculées avec les modèles 6-grammes de mots pour Chirac et Mitterrand.*

On constate une meilleure séparation des ensembles de phrases de Mitterrand et un resserrement des courbes qui indiquent de moindres variations statistiques autour des moyennes.

2.4 Construction des ensembles continus de phrases d'un même auteur

La méthode de reconnaissance que nous venons de décrire attribue un auteur à chaque phrase, ce qui induit des passages discontinus de phrases de François Mitterrand dans les phrases de Jacques Chirac. Or, la tâche DEFT'05 consiste à trouver dans chaque allocution de Chirac l'insertion, si elle existe, d'une suite continue de phrases de Mitterrand, cette insertion étant d'au moins 2 phrases. Pour tenir compte de ces contraintes, nous avons d'abord utilisé une méthode simple décrite dans la première partie suivante. Dans la seconde partie, nous décrivons comment ces contraintes sont prises en compte de manière plus efficace avec l'algorithme de Viterbi, produisant un gain de 0,20 du F-score.

2.4.1 Prise en compte de paquets de phrases successives

Chaque fois qu'un ensemble de 1 à *k* phrases de Jacques Chirac est détecté entre deux phrases de Mitterrand, on leur attribue l'étiquette Mitterrand. Le réglage du paramètre *k* a été calculé sur 10% des corpus étiquetés. La meilleure valeur est $k = 4$. Cette méthode donne un taux de rappel élevé mais elle n'assure pas qu'une seule succession de phrases de Mitterrand soit obtenue dans un discours de Chirac. Les valeurs de F-score sont améliorées d'en moyenne 0,08 pour les trois versions du corpus de test de 0,48 à 0,56.

2.4.2 Algorithme de Viterbi

Plusieurs compétiteurs de DEFT'05 (El-Bèze *et al.*; Labadié *et al.*; Rigouste *et al.*, 2005) ont utilisé l'algorithme de Viterbi (Manning, 1999) avec de très bons résultats. Nous l'avons mis en œuvre pour voir comment il pouvait améliorer les résultats que nous avons obtenus avec les modèles n-grammes de mots (Hurault-plantet, 2005).

On considère la suite des étiquettes (attribution d'un auteur à une phrase) de chaque allocution comme une chaîne de Markov d'ordre 1 à laquelle on applique les contraintes de la tâche sous forme de probabilités de succession entre états. Nous avons considéré 4 états : 2 états Mitterrand M1 et M2 pour prendre en compte le fait qu'au moins 2 phrases de Mitterrand sont incluses dans chaque allocution, et 2 états pour les phrases de Chirac, un pour les phrases du début d'allocution et un pour les phrases de fin d'allocution pour tenir compte de l'inclusion des phrases de Mitterrand.

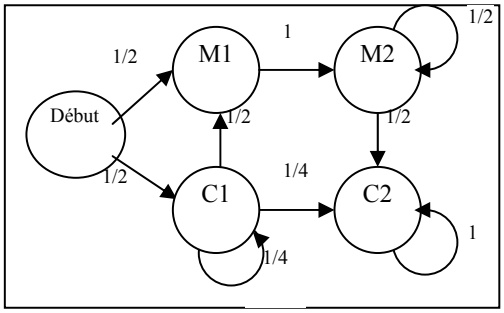

FIG. 3 – *Probabilités de transition entre les quatre états possibles des phrases d'une allocution, M1 et M2 pour les phrases de Mitterrand et C1 et C2 pour les phrases de Chirac.*

Étant donné les observations, qui sont les étiquettes que nous avons trouvées avec les modèles n-grammes et les probabilités de transition entre états représentées sur la figure 3, l'algorithme de Viterbi recherche parmi tous les chemins possibles entre états (4 par phrase) celui qui génère l'ensemble des observations avec la plus grande probabilité en ne conservant en mémoire pour chaque phrase que l'état de la phrase précédente qui l'a rendu accessible avec la plus grande probabilité.

Le tableau 5 regroupe les résultats de cet algorithme obtenus sur le corpus du test intégral. On constate comme l'ont déjà remarqué d'autres compétiteurs qu'une amélioration notable des scores de reconnaissance est obtenue avec l'algorithme de Viterbi, d'environ 0,20 pour le F-score avec l'effet de plateau pour *n* variant de 2 à 6.

n	1	2	3	4	6	8
Rappel	0,91	0,65	0,60	0,60	0,61	0,61
Précision	0,29	0,72	0,79	0,79	0,78	0,74
F-score	0,44	**0,68**	**0,68**	**0,68**	**0,68**	0,67

TAB. 5 – *Scores de reconnaissance des phrases de Mitterrand insérées dans des discours de Chirac, avec des modèles n-grammes de mots pour n variant de 1 à 8 complétés par un algorithme* de *Viterbi prenant en compte les contraintes de DEFT'05, sur le corpus de test intégral.*

2.5 Modèles de langage n-grammes de caractères pour identifier l'auteur de chaque phrase

Bien que très nettement améliorés par le lissage précédent, nos résultats restent moyens. Au vu de l'étude du caractère discriminant de nos modèles, nous pensons que l'apprentissage s'appuie sur des données statistiques trop pauvres, et, pour pallier cet effet et toujours dans l'esprit de modèles simples, nous avons pensé que l'utilisation de n-grammes de caractères pouvait être une piste intéressante, d'ailleurs déjà expérimentée par d'autres pour identifier un auteur (Teahan, 2000, Kmelev *et al.*, 2001, Peng *et al*., 2003). Cette étude est détaillée dans (Jardino, 2006). Nous en extrayons ici les points principaux.

2.5.1 Corpus d'apprentissage

Nous avons séparé tous les caractères du texte que nous avions initialement prétraité pour les n-grammes de mots. Les statistiques sur les caractères du corpus d'apprentissage sont rassemblées dans le tableau 6.

	Nb phrases	Nb caractères		
		lexique	total	moyen/phrase
Chirac	49 890	80	5 504 100	110 (69)
Mitterrand	7 523	64	1 024 536	136 (100)

TAB. 6 - *Statistiques du corpus d'apprentissage. Les caractères incluent les signes de ponctuation. Entre parenthèses sont indiqués les écarts-type.*

Les caractères comportent les lettres en minuscule (incluant les lettres accentuées même peu fréquentes), les chiffres de 0 à 9 et les signes de ponctuation. Les différences de taille entre les lexiques de caractères de Chirac et Mitterrand proviennent de lettres accentuées peu fréquentes. Nous avons préféré garder ces informations pour minimiser les interventions sur le corpus initial.

Les tableaux 1 et 6 montrent une redondance moyenne des caractères 1000 fois supérieure à celle des mots aussi bien pour Chirac que pour Mitterrand ce qui devrait entraîner une meilleure fiabilité statistique des modèles fondés sur les caractères comparée à celle des modèles fondés sur les mots. Le nombre moyen de caractères par mot est de 4 avec un écart-type de 3 aussi bien pour Chirac que pour Mitterrand.

2.5.2 Modèles n-grammes de caractères

En représentant une phrase de L caractères par leur succession : $c_1 \ldots c_i \ldots c_L$, la probabilité de cette phrase, P_A^C (phrase), calculée à partir du modèle n-grammes de caractères de l'auteur A, est :

$$P_A{}^C \text{ (phrase)} = \Pi_i \; p_A (c_i / c_{i-n+1} \ldots c_{i-1})$$

où $p_A (c_i / c_{i-n+1} \ldots c_{i-1})$ est la probabilité que le caractère c_i succède à la chaîne de n-1 caractères $c_{i-n+1} \ldots c_{i-1}$. Cette probabilité est calculée à partir des fréquences relatives de la chaîne $c_{i-n+1} \ldots c_{i-1}$ ci-n+1 …ci-1 c_i dans les phrases de l'auteur A et avec une méthode de repli vers les n-grammes d'ordre inférieur pour les suites de caractères non observées dans le corpus d'apprentissage.

Nous procédons ensuite de la même manière que pour les n-grammes de mots. C'est à dire que nous affectons chaque phrase à l'auteur dont le modèle aura donné la plus grande probabilité. Le tableau 7 représente les résultats obtenus sur le corpus de test dont les caractères ont été séparés.

n	1	2	3	4	5	6
Rappel	0,54	0,66	0,70	0,65	0,53	0,44
	0,48	0,72	0,78	**0,71**	0,52	0,38
Précision	0,24	0,31	0,37	0,47	0,56	0,60
	0,52	0,62	0,74	**0,87**	0,91	0,91
F-score	0,34	0,42	0,49	0,55	0,55	0,51
	0,50	0,66	0,75	**0,78**	0,66	0,54

TAB. 7 – *Scores de reconnaissance des phrases de Mitterrand insérées dans des discours de Chirac, avec des modèles n-grammes de caractères pour n variant de 1 à 6 sur le corpus de test intégral. La deuxième ligne de chaque rangée correspond à ces mêmes modèles complétés par un algorithme* de *Viterbi prenant en compte les contraintes de DEFT'05.*

Les modèles 4-grammes de caractères ont un score d'identification des phrases de Mitterrand supérieur à celui des 4-grammes de mots, 0,55 contre 0,48. Ce score est spectaculairement amélioré par l'algorithme de Viterbi, passant de 0,55 à 0,78 et se situant au 4ème rang de l'évaluation DEFT'05 où le meilleur score est 0,88.

2.5.3 Analyse du modèle n-grammes de caractères

On peut remarquer que la valeur de n = 4 correspond à la longueur moyenne des mots. Nous avons extrait les n-grammes de caractères les plus probables pour Chirac et Mitterrand. Ils sont rassemblés dans le tableau 8, en ordre décroissant des probabilités.

Chirac		Mitterrand	
-0.0001	è r e	-0.0002	s q u
-0.0001	t q u	-0.0002	e q u
-0.0001	t i q u	-0.0004	t q
-0.0001	t i o n	-0.0004	c e q u
-0.0001	i è r e	-0.0005	ê m e
-0.0001	e s q u	-0.0005	è r e
-0.0001	c e q u	-0.0005	o t r e
-0.0001	c ' e s	-0.0005	m ê m e
-0.0001	, q u	-0.0005	, q u
-0.0002	ê m e	-0.0006	t i o n
-0.0002	é q u	-0.0006	r q u
-0.0002	m ê m e	-0.0006	i q u
-0.0002	m i q u	-0.0006	e s q u
-0.0002	i q u	-0.0007	ê t r e
-0.0003	è m e	-0.0007	t i q u
-0.0003	x q u	-0.0007	i è r e
-0.0003	u x q u	-0.0007	i s q u

TAB. 8 – *N-grammes de caractères les plus probables dans les modèles Chirac et Mitterrand obtenus à partir du corpus d'apprentissage intégral. Le nombre devant chaque graphie correspond au logarithme népérien de la probabilité de la graphie.*

Ce sont principalement des trigrammes et des quadrigrammes dont il est difficile de donner une interprétation en termes de suffixes, préfixes ou lexèmes puisque la plupart d'entre eux comme « tqu » ou « cequ » se situent à cheval sur deux mots. Les expériences précédentes qui ont été faites sans caractère « ESPACE » ont été reproduites avec un caractère « ESPACE ». On obtient les mêmes résultats pour les valeurs de n de 1 à 4. On constate ensuite seulement une décroissance plus lente des performances quand n augmente au-delà de 4.

En conclusion les modèles n-grammes permettent de reconnaître des auteurs même sur des textes courts, pourvu qu'il y ait assez de données d'apprentissage. Ceci implique d'adapter les items sur lesquels s'appuyer pour obtenir des statistiques fiables. Pour la tâche DEFT'05 les caractères sont des items statistiquement plus fiables et plus discriminants que les mots.

L'approche markovienne décrite dans ce chapitre 2 identifie des auteurs phrase par phrase. Nous avons vu que le rôle des mots à ce niveau était moindre que celui des caractères. On peut en déduire que l'information thématique véhiculée par les mots n'est pas ou peu représentée au niveau de la phrase, elle s'argumente davantage au fil des phrases.

C'est cet aspect, complémentaire du précédent, qui est pris en compte dans le chapitre 3 suivant.

3 Identification de thème

Un texte développe en général un thème central qui lui donne sa cohésion. Halliday et Hasan (1976) ont étudié en particulier les relations lexicales – répétition, synonymie, hyperonymie, cooccurrence – qui marquent cette cohésion. Morris et Hirst (1991) ont ensuite développé la notion de chaîne lexicale, formée au long d'un texte par les relations lexicales, pour rendre compte de la continuité de sens du texte et en déterminer la structure. Les relations lexicales que nous considérons pour construire une chaîne lexicale pour chaque mot d'un texte sont la répétition, la dérivation adjectivale, et la cooccurrence[3]. Nous effectuons d'abord une sélection des mots puis nous construisons une chaîne lexicale pour chaque mot sélectionné de l'allocution suivant cette approche : chaque chaîne lexicale est composée d'un ou plusieurs segments de texte, chaque segment comportant au moins une phrase et correspondant à une zone plus dense du texte pour le mot considéré, c'est-à-dire une zone où le mot considéré est plus fréquent. Les segments respectifs de deux chaînes différentes peuvent être complètement disjoints, et les chaînes correspondantes ne seront pas cooccurrentes. Ou bien les segments peuvent se recouvrir partiellement ou complètement, et les deux chaînes correspondantes seront cooccurrentes. L'hypothèse que nous utilisons pour séparer les textes des deux auteurs est la disjonction des segments de texte des deux chaînes lexicales respectivement associées au mot le plus fréquent de chaque thème.

Le paragraphe 3.1 est consacré à la construction des chaînes lexicales pour chaque allocution et à la construction des segments associés, et fournit une évaluation de l'apport des relations lexicales utilisées. Dans le paragraphe 3.2, nous décrivons la méthode de séparation des thèmes des deux auteurs et l'attribution d'un thème à chaque auteur, puis nous testons la validité de l'hypothèse sous-jacente de disjonction des chaînes lexicales associées au mot le plus fréquent de chaque thème.

3.1 Construction des chaînes lexicales

3.1.1 Sélection des mots

Les critères de sélection des mots ont été déterminés sur le corpus d'apprentissage intégral de DEFT'05. Les résultats présentés ici ont été obtenus sur le corpus de test associé qui comporte les noms propres et les dates. Les noms propres et les dates étant de fréquence faible dans chaque texte, les résultats obtenus pour les trois versions du corpus sont très peu différents (Alphonse, 2005).

Le principe général de sélection est d'éliminer certains mots peu discriminants puis de regrouper sous un même mot des mots qui lui sont sémantiquement liés par les relations lexicales précisées dans l'introduction du paragraphe 3.

[3] Nous n'avons pas utilisé la synonymie et l'hyperonymie qui, dans ce contexte de séparation de textes, risquent d'introduire trop de bruit.

Pour déterminer les mots que nous allons utiliser pour former les chaînes lexicales, nous effectuons d'abord un prétraitement des phrases. Les mots de chaque phrase sont lemmatisés par le TreeTagger de Schmid (1999), et nous ne conservons pour indexer la phrase que les lemmes (ou le mot lui-même lorsqu'il est inconnu du logiciel) des substantifs, des adjectifs, des noms propres et des abréviations, ainsi que les dates[4].

Certains mots comme *France*, présent dans 286 allocutions sur 294, *grand* (282 / 294), *pays* (278 / 294), *français* (264 / 294), *monde* (253 / 294), sont particulièrement fréquents dans les allocutions du corpus et utilisés par les deux auteurs. Nous avons pris les treize[5] mots les plus fréquents du vocabulaire du corpus d'apprentissage, qui ont donc un poids *idf*[6] faible, et nous les avons ajouté à la liste de mots vides[7] que nous utilisons. Nous avons comparé à une autre approche classique qui consiste à utiliser les pondérations *tf*idf*[8] des mots. Nous avons testé deux méthodes : la première consiste à classer les mots d'une allocution par poids *tf*idf* décroissant avant de rechercher l'identifiant du thème de chaque auteur. La deuxième méthode consiste à supprimer les mots de poids *tf*idf* trop faible dans chaque allocution avant d'effectuer le classement par fréquence décroissante pour la recherche des identifiants des thèmes.

Le tableau 9 montre que l'approche qui consiste à supprimer les mots trop fréquents dans l'ensemble du corpus obtient de meilleurs résultats que les approches utilisant la pondération *tf*idf*.

Approche utilisée	précision	rappel	F-score
Suppression des mots les plus fréquents (*idf* faible)	**0.52**	**0.55**	**0.53**
Classement par tf*idf décroissants	0.43	0.48	0.45
Suppression des mots de tf*idf inférieur à 0.1	0.41	0.45	0.43
Suppression des mots de tf*idf inférieur à 0.2	0.42	0.45	0.44

TAB. 9 – *Comparaison des approches « idf » et « tf*idf » sur le corpus d'apprentissage, version intégrale.*

Nous effectuons ensuite une reconnaissance automatique des adjectifs dont le substantif est dans l'allocution, et nous remplaçons ces adjectifs par les substantifs correspondants. Pour effectuer la reconnaissance automatique des adjectifs, le système utilise une liste de terminaisons d'adjectifs[9] qui lui permet de déterminer des adjectifs candidats et leurs racines respectives. Le système recherche ensuite les substantifs qui commencent par ces racines. Le bruit généré par cette méthode est limité par deux contraintes : la racine doit avoir une taille minimum (au moins deux caractères), et la taille de l'adjectif doit être supérieure à celle du substantif. Ces deux contraintes sont directement issues d'expérimentations sur le corpus d'apprentissage. Nous avons fait une évaluation manuelle de cette méthode sur une petite

[4] Nous avons constaté que les verbes n'étaient pas discriminants dans notre approche.

[5] La courbe des fréquences ne présentant pas de décrochement net, nous avons pris ce seuil après un examen des mots les plus fréquents.

[6] *Inverse document frequency* : égal à l'inverse du nombre de documents du corpus contenant ce mot.

[7] http://www.idi.ntnu.no/emner/tdt4215/resources/frenchST.txt

[8] *term frequency * inverse document frequency*

[9] http://www.protic.net/profs/martin/1et2/francais/derivation.html

partie du corpus. Le tableau 10 montre les résultats obtenus : la supériorité de la précision sur le rappel découle des contraintes assez fortes utilisées, en particulier la contrainte sur la plus grande longueur de l'adjectif qui supprime de nombreuses dérivations comme *pauvre-pauvreté*. La méthode développée utilise une approche morphologique très basique et peut être améliorée. Nous n'avons pas trouvé, dans la littérature, d'évaluation d'outils de dérivation adjectivale.

Corpus testé	précision	rappel	F-score
25 premières allocutions	0.60	0.37	0.46
50 premières allocutions	0.54	0.34	0.42

TAB. 10 – *Évaluation manuelle de la reconnaissance des adjectifs dérivés de noms.*

Pour augmenter le nombre de mots associés dans la construction d'une chaîne lexicale, nous utilisons une méthode supplémentaire basée sur la recherche des cooccurrences fréquentes[10] maximales dont l'algorithme est décrit par Grahne *et al.* (2003). La recherche des cooccurrences fréquentes maximales est utilisée en fouille de texte pour construire des règles d'association entre éléments, étant donné un corpus de transactions[11]. Il s'agit de déterminer les ensembles maximaux d'éléments (ici les mots) cooccurrents dans un ensemble de transactions (ici les phrases), ayant un support (fréquence) minimum.

Le système recherche d'abord les cooccurrences les plus fréquentes de mots dans une allocution, puis génère la règle suivante d'équivalence entre les mots des ensembles trouvés : le mot le plus fréquent de chaque ensemble est utilisé pour indexer chacun des autres mots de l'ensemble. Le système ré-indexe alors chaque phrase de l'allocution suivant ces règles. Les mots qui sont en forte cooccurrence avec un mot de plus forte fréquence sont alors représentés par ce mot. Pour limiter le bruit qui peut en résulter nous ne recherchons les ensembles cooccurrents que sur les mots les plus fréquents[12]. Nous éliminons enfin les mots de trop faible fréquence[13]. Les seuils de fréquence utilisés sont fonction du rang de la fréquence et varient donc pour chaque allocution. Ils ont été déterminés sur le corpus d'apprentissage.

Le tableau 11 montre l'indexation des phrases par les mots retenus pour l'allocution 242 du corpus de test.

[10] En anglais *frequent itemset*. Ce terme est aussi traduit par *motifs fréquents* dans Bastide et al. (2002).

[11] Les règles d'associations ont été utilisées à l'origine en marketing pour détecter des associations préférentielles entre les produits achetés chez de grands distributeurs.

[12] Nous prenons comme seuil la cinquième plus forte fréquence dans le document considéré.

[13] Nous avons choisi un seuil égal à la vingt-cinquième plus forte fréquence dans le document considéré, ou à défaut, un seuil de fréquence égal à 2.

Auteur	N° de la phrase	Mots sélectionnés
C	1	présidente flamme cas
C	2	émotion vie
C	3	présidente mental_handicap madame mental handicap accueil
C	4	combat obstacle digne
C	5	flamme droit
C	6	droit handicap
C	7	dignité droit
M	8	sujet général mondial attention négociation commerce débat raison
M	9	liberté général commerce échange
M	10	sujet rencontre mot
M	11	général obstacle accord
M	12	tiers accord
M	….	….
M	23	général accord
M	24	besoin
M	25	mondial
M	26	
M	27	traitement
M	28	droit
M	29	vrai part
M	30	échange
C	31	dignité attention droit handicap
C	32	regard
C	……	…….
C	69	handicap raison
C	70	vrai

TAB. 11 – *Les mots sélectionnés de l'allocution 242. Les zones de textes attribuées à Chirac et à Mitterrand par l'identification des thèmes sont respectivement marquées en gris clair et en gris foncé.*

Pour cette allocution, le système a trouvé un ensemble de cooccurrences fréquentes maximales, *mental_handicap,* ainsi que deux dérivations adjectivales, *handicap-handicapé* et *nation-national*. Le système remplace donc, dans toutes les phrases de l'allocution, *mental* et *handicapé* par *handicap*, et *national* par *nation*.

3.1.2 Construction des segments des chaînes lexicales

Le système associe ensuite à chaque mot, sélectionné conformément au paragraphe précédent, les segments du texte de l'allocution où il apparaît. Pour trouver ces segments, le système fait un chaînage lexical en repérant les occurrences du mot sur des blocs consécutifs de cinq phrases[14], en commençant par la première phrase de l'allocution. La chaîne s'arrête lorsqu'un bloc ne comporte pas le mot, et reprend au premier bloc où il apparaît à nouveau. Un segment est un ensemble de blocs consécutifs où le mot apparaît, d'où on élimine ensuite les phrases, en début et fin du segment, qui ne contiennent pas le mot. Le tableau 12 rassemble les mots les plus fréquents de l'allocution 242 et leurs segments.

[14] Pour les allocutions courtes (moins de 35 phrases), nous utilisons une distance maximale de trois phrases.

Mot	Fréquence	Segments (n° des phrases)
handicap	12	3-6 ; 31-33 ; 42-69
dignité	7	7-7 ; 31-38 ; 48-62
droit	6	5-7 ; 16-16 ; 28-31
vie	6	2-2 ; 37-37 ; 49-65
accord	5	11-23
combat	4	4-4 ; 38-40

TAB. 12 – *Les segments de texte des mots les plus fréquents de l'allocution 242.*

3.1.3 Évaluation de l'apport des relations lexicales dans la construction des chaînes

Les différentes relations lexicales utilisées facilitent le chaînage lexical car elles augmentent le nombre d'occurrences dans le texte des mots sélectionnés. Mais elles apportent aussi du bruit par les erreurs qu'elles produisent. Nous avons donc évalué les résultats de notre méthode de filtrage de textes avec et sans l'utilisation des différentes méthodes facilitant le chaînage lexical – lemmatisation, mise en minuscules, dérivation adjectivale et cooccurrence –, et en ne conservant à chaque fois que les noms, noms propres, adjectifs, et abréviations.

Nous voyons dans le tableau 13 que, globalement, le renforcement du chaînage lexical apporte une amélioration de la méthode, malgré le score très moyen de la reconnaissance de la dérivation adjectivale. Chaque regroupement sémantique – lemmatisation, dérivation adjectivale et ensembles cooccurrents – apporte une petite amélioration.

Renforcement du chaînage lexical	précision	rappel	F-score
Lemmatisation, mise en minuscule, dérivation adjectivale, et ensembles cooccurrents	**0.52**	**0.55**	**0.53**
Sans mise en minuscule	0.51	0.53	0.52
Sans les ensembles cooccurrents	0.50	0.53	0.51
Sans les ensembles cooccurrents ni dérivation adjectivale	0.44	0.51	0.47
Sans lemmatisation, sans mise en minuscule	0.46	0.47	0.47

TAB. 13 – *Évaluation des diverses techniques de renforcement du chaînage lexical.*

3.2 Séparation des thèmes des deux auteurs

3.2.1 Détermination des deux thèmes

Nous identifions tout d'abord un thème par son mot le plus fréquent, que nous appelons *identifiant* du thème. Nous prenons comme identifiant du premier thème dans une allocution celui dont la chaîne lexicale est la plus longue, la longueur de la chaîne étant le nombre de phrases de l'ensemble de ses segments. Dans l'exemple de l'allocution 242, l'identifiant du premier thème est *handicap*. Le système cherche ensuite l'identifiant de l'autre thème, c'est-à-dire le mot dont la chaîne lexicale est la première, dans l'ordre des fréquences décroissantes, dont les segments n'ont pas d'intersection avec les segments de la chaîne lexicale de l'identifiant du premier thème. Si le système ne trouve pas de chaîne disjointe, il n'identifie qu'un seul thème et donc un seul auteur, Chirac.

Une fois les deux identifiants de thème trouvés, le système regroupe autour de chacun d'eux les mots dont les chaînes lexicales sont en cooccurrence avec la sienne. Pour cela, le système agrège chacune des chaînes restantes avec la chaîne de l'identifiant avec laquelle elle possède un segment, ou une partie de segment, en commun. Si une chaîne a un segment en commun avec les chaînes des deux identifiants à la fois, elle est considérée comme chaîne commune aux deux thèmes et n'est pas agrégée. Nous obtenons donc deux thèmes représentés chacun par un ensemble de mots. Dans l'exemple de l'allocution 242, les mots *dignité*, *vie*, *combat*, *vrai*, *regard*, *besoin* sont regroupés avec l'identifiant de thème *handicap*, et les mots *général*, *mondial*, *négociation*, *nation*, *justice*, *échange* sont regroupés avec l'identifiant de thème *accord*. Les mots *droit*, *raison*, *obstacle*, *rencontre*, *liberté* sont des mots communs aux deux identifiants, ils ne sont donc pas regroupés.

3.2.2 Attribution d'un auteur à chaque thème

Sachant que l'allocution commence et finit par une phrase de Chirac, le système attribue à l'auteur Chirac celui des deux thèmes qui débute l'allocution et la termine. Il attribue l'autre thème à l'auteur Mitterrand. Si l'un des thèmes débute l'allocution et l'autre la termine, nous considérons que les thèmes ont été mal détectés et aucune phrase n'est attribuée à Mitterrand. On voit dans le tableau 14 que, pour l'allocution 242, le thème principal *handicap* est attribué à Chirac et le thème principal *accord* à Mitterrand. Dans le tableau 14, les zones de textes attribuées à Chirac et à Mitterrand sont en grisés différents.

Identifiant	Segments	Thème	Segments du thème	Auteur
handicap	3-6 ; 31-33 ; 42-69	handicap, dignité, vie, combat, vrai, regard, besoin	1-7 ; 24-24 ; 29-29 ; 31-40 ; 42-70	Chirac
accord	11-23	accord, général, mondial, négociation, nation, justice, échange	8-23 ; 25-25 ; 27-27 ; 30-30	Mitterrand

TAB. 14 – *Les thèmes et leurs segments pour l'allocution 242.*

3.2.3 Évaluation de l'hypothèse de disjonction des identifiants des thèmes des deux auteurs

Nous avons voulu vérifier si notre hypothèse de disjonction des identifiants des thèmes des deux auteurs était vérifiée dans le corpus de test. Pour cela, nous avons séparé dans chaque allocution le texte de Chirac et la partie insérée du texte de Mitterrand, et nous avons pris comme identifiant de thème pour chaque texte la chaîne lexicale la plus longue. Nous avons ensuite regardé si ces identifiants sont ceux qui sont détectés par notre système. Les résultats sont donnés dans le tableau 15.

Nous voyons tout d'abord que 95 allocutions sur 294 ne comportent pas d'insertion de texte de Mitterrand, et que, dans ce cas, le pourcentage d'allocutions entièrement attribuées à Chirac est de 45%. Dans les 55% de cas restants, on trouve donc une chaîne lexicale disjointe de la chaîne lexicale la plus longue, bien que le texte garde à la lecture sa cohésion. En effet, cette chaîne est l'identifiant d'un autre thème, sorte de digression, sémantiquement lié au thème central du texte par un contexte commun qui n'apparaît pas explicitement. Quant aux 199 allocutions de Chirac qui comportent effectivement l'insertion d'un texte de Mitterrand, dans 132 cas les identifiants des thèmes des deux auteurs ont des segments disjoints, mais dans 67 autres cas, les segments se recoupent, l'identifiant du thème de l'un constituant un simple composant du thème de l'autre.

Identifiants	0 identifiant reconnu	1 identifiant reconnu	Les 2 identifiants reconnus	Total
Segments séparés	5 allocutions 57% auteur M reconnu	32 allocutions 49% auteur M reconnu	95 allocutions 75% auteur M reconnu	132
Segments sécants	0	67 allocutions 30% auteur M reconnu	0	67
Un seul identifiant (auteur C)	0	95 allocutions 45% auteur C reconnu	-	95
Total	5	194	95	294

TAB. 15 – *Nombre d'allocutions pour lesquelles les identifiants des thèmes sont reconnus par notre système, suivant que les segments respectifs de chaque identifiant sont séparés ou sécants, et pourcentage de phrases de Mitterrand retrouvées dans ce cas.*

Notre hypothèse de disjonction des identifiants des thèmes des deux auteurs se trouve donc vérifiée dans 66% des cas. Nous constatons que dans les cas où cette hypothèse se vérifie, les deux identifiants des thèmes des auteurs sont mieux reconnus (72% des cas, contre 0% si les identifiants sont sécants), et les phrases de Mitterrand sont mieux extraites (68% des cas contre 30% si les identifiants sont sécants). Dans les rares cas pour lesquels aucun identifiant n'est reconnu (à peine 2%), nous avons constaté que les thèmes des deux auteurs avaient un composant fréquent en commun (*europe* ou *organisation* par exemple), composant minoritaire pour chacun des deux thèmes mais dont la fréquence sur l'ensemble du texte est la plus élevée.

3.3 Lissage des segments

Les segments obtenus pour chaque auteur forment rarement des blocs bien délimités. Deux schémas apparaissent fréquemment : ce sont d'une part l'entrelacement de phrases isolées des deux auteurs d'une part, et d'autre part l'existence de groupes de phrases non attribuées entre deux segments d'auteurs différents. Cela est dû à la présence dans les phrases de mots communs entre les auteurs qui ont été supprimés (voir paragraphe 3.2.1), ou de mots trop peu fréquents qui n'ont pas été sélectionnés.

Nous appliquons des heuristiques simples pour délimiter plus clairement les segments attribués à chacun des auteurs. Nous supprimons les segments ne comportant qu'une phrase isolée, car ils sont peu significatifs comparés aux segments s'étendant sur plusieurs phrases, et nous complétons les segments attribués à Mitterrand en les étendant jusqu'aux limites des segments attribués à Chirac. Nous considérons en effet que l'identifiant du thème de Chirac est en général mieux reconnu que celui du thème de Mitterrand, la partie insérée étant en général plus petite que l'allocution de Chirac dans laquelle elle s'insère. Cette heuristique affaiblit la précision mais augmente le rappel, et au final, améliore le Fscore.

Ces heuristiques conduisent, dans l'exemple de l'allocution 242, à attribuer à Mitterrand les phrases numérotées de 8 à 30.

3.4 Discussion

D'autres participants à DEFT'05 ont utilisé la notion de thème. C'est le cas de Labadié *et al*. (2005) et de Maisonnasse et Tambellini (2005), qui ont utilisé un logiciel de segmentation thématique, associé à des méthodes d'apprentissage. La segmentation thématique effectue un découpage de chaque allocution. Labadié *et al*. (2005) utilisent ensuite un modèle bayésien appris sur le corpus d'entraînement pour attribuer à chaque segment l'un des deux auteurs. L'algorithme de Viterbi permet enfin de lisser les résultats obtenus. La combinaison de ces méthodes mène à un F-score maximum de 0.74. Maisonnasse et Tambellini (2005) attribuent d'abord à chaque phrase un score issu d'un modèle d'apprentissage sur les dépendances syntaxiques. Un calcul du score moyen de chaque segment permet alors de modifier le score de chaque phrase. Un lissage est ensuite effectué à l'aide d'un modèle de diffusion. La segmentation thématique n'apporte rien au modèle initial d'apprentissage qui obtient un meilleur F-score (0.75 au lieu de 0.72) sans la modification du score de chaque phrase apportée par la segmentation thématique.

Rigouste *et al.* (2005) ont utilisé avec succès un modèle de mélange multi-thématique, en obtenant un F-score de 0.87. Les paramètres correspondant aux thèmes des deux auteurs sont estimés sur le corpus d'apprentissage, et sont utilisés ensuite par l'algorithme de Viterbi sur le corpus de test pour estimer l'auteur le plus probable de chaque phrase. Cette méthode s'est révélée très efficace, les distributions thématiques s'avérant différentes pour chacun des deux auteurs. L'apprentissage des distributions thématiques de chaque auteur semble donc mieux adaptée à la résolution de ce problème de filtrage que la segmentation thématique sans apprentissage.

Bien que les thèmes de Chirac et de Mitterrand dans un même document aient été choisis différents, ils ont en commun une partie du vocabulaire. C'est un choix des organisateurs de DEFT'05 qui ont cherché à maximiser le nombre de mots communs à une allocution et à sa partie insérée (Alphonse *et al.,* 2005). La méthode de séparation des thèmes des auteurs que nous avons mise en œuvre et qui recherche des zones de plus grande densité pour des mots associés à un thème semble avoir été pénalisée par le bruit engendré par ce choix.

4 Fusion des résultats

Le résultat à produire pour l'atelier d'évaluation était la liste des phrases de Mitterrand dans l'ensemble des allocutions du corpus de test. La méthode basée sur la reconnaissance du style de l'auteur et la méthode basée sur l'identification de thème produisent chacune un ensemble de phrases attribuées à Mitterrand. Nous pouvions espérer que l'intersection des deux ensembles produirait une meilleure précision, la reconnaissance d'un auteur par une méthode étant alors attestée par l'autre méthode. Nous avons donc expérimenté sur le corpus d'apprentissage une simple intersection des deux ensembles, ne retenant que les phrases attribuées à Mitterrand par les deux méthodes. Cette intersection produisait effectivement une bonne précision mais un faible rappel, et faisait donc baisser le F-score. En effet, le nombre d'allocutions pour lesquelles des phrases de Mitterrand sont trouvées est plus faible par la méthode d'identification des thèmes que par la méthode de reconnaissance du style de l'auteur (voir tableau 16). Nous n'avions donc une intersection possible que pour un petit nombre d'allocutions.

Ne disposant d'aucun moyen permettant de faire une pondération des résultats qui serait comparable pour les deux méthodes, nous avons donc décidé d'adopter une stratégie mixte qui consiste à retenir d'une part les phrases attribuées à Mitterrand par les deux méthodes

afin d'augmenter la précision, et d'autre part les phrases attribuées à Mitterrand par la méthode de reconnaissance de style lorsque la méthode d'identification de thèmes ne renvoie aucune phrase de Mitterrand. Dans le tableau 16, on peut voir que la méthode utilisant la reconnaissance du style trouve des phrases de Mitterrand dans 281 allocutions sur 294, alors que la méthode d'identification des thèmes n'en trouve que dans 222 allocutions. On constate surtout une très grande différence entre les nombres de phrases identifiées comme appartenant à Mitterrand par chacune des deux méthodes.

Méthode de filtrage des phrases de Mitterrand	Précision	Rappel	F-score	Nombre de phrases ramenées	Nombre d'allocutions
Reconnaissance du style	0.40	**0.88**	0.55	8 271	281 / 294
Identification des thèmes	0.52	0.55	0.53	4 009	222 / 294
Intersection simple	**0.80**	0.46	**0.58**	2 185	167 / 294
Intersection élargie	0.54	0.61	**0.57**	4 273	250 / 294

TAB. 16 – *Résultats obtenus par les différentes méthodes sur le test intégral DEFT'05*

Sur le corpus d'apprentissage, l'intersection simple obtenait un score inférieur à l'intersection élargie, car, si la précision était bonne, le rappel était beaucoup plus faible. Sur le corpus de test, les résultats ont été différents, et le tableau 16 montre qu'on obtient des scores comparables avec les deux méthodes même si l'intersection simple permet effectivement d'améliorer nettement la précision.

5 Conclusion

Notre participation à DEFT'05 nous a placé au 7^{ième} rang sur 11 participants, avec un score moyen de 0.56 sur les trois tâches (voir la section 1), les scores des participants s'étendant de 0.88 à 0.18. Nous avons développé en parallèle deux méthodes très différentes, l'une basée sur la reconnaissance du style des auteurs par un modèle de langage appris sur le corpus d'entraînement et appliqué phrase par phrase, et l'autre basée sur la séparation des thèmes des deux auteurs, identifiés sans apprentissage sur chaque allocution. La simple fusion des résultats des deux méthodes améliore notablement la précision mais très peu le F-score. Nous avons testé un autre type de coopération entre les deux méthodes, qui consiste d'abord à attribuer un auteur à chaque chaîne lexicale de l'allocution par le modèle de reconnaissance du style de l'auteur, puis à calculer l'auteur de chaque phrase suivant les chaînes lexicales qu'elle contient. Le F-score obtenu est légèrement inférieur à celui obtenu par simple fusion.

Plusieurs raisons déterminent les résultats moyens obtenus par la méthode de séparation des thèmes des deux auteurs dans chaque allocution. Tout d'abord la présence de mots de forte fréquence communs aux textes des deux auteurs et bruitant la détection de zones de plus forte densité pour le thème (voir le paragraphe 3.4). L'analyse des thèmes de chaque allocution (voir le paragraphe 3.2.3) nous montre en effet que le tiers des allocutions de Chirac qui comportent des insertions de phrases de Mitterrand, a pour identifiant de thème un mot utilisé aussi, avec une fréquence plus faible néanmoins, par l'autre auteur. Ensuite, le fait de ne pas utiliser de connaissances préalables produit des confusions entre une digression de thème et un thème séparé. Enfin, pour cette même raison, une confusion est aussi possible entre digression et thème séparé lorsque le texte inséré est de petite taille. Mais par ailleurs,

nous trouvions intéressant de tester une méthode n'utilisant pas d'apprentissage pour une telle tâche, car il arrive qu'on ne puisse pas disposer d'un corpus d'apprentissage ayant des caractéristiques très proches du corpus à étudier.

En revanche, la méthode de reconnaissance du style a pu être largement améliorée à la fois par l'utilisation de l'algorithme de Viterbi et par l'utilisation de caractères à la place de mots dans le modèle n-grammes. Le F-score est alors à 0.78 (voir le paragraphe 2.5). L'algorithme de Viterbi a été utilisé par plusieurs participants à l'atelier d'évaluation, et c'est son utilisation, avec une attribution préalable d'un auteur, ou d'un thème, à chaque phrase par un modèle probabiliste appris sur corpus, qui donne les meilleurs résultats (El Bèze *et al.*, 2005 ; Rigouste *et al.*, 2005).

Références

Alphonse E., A. Amrani, J. Azé, T. Heitz, A-D. Mezaour et M. Roche (2005). Préparation des données et analyse des résultats de DEFT'05. *Actes de TALN 2005*, Dourdan, France, 2:99-111.

Alvarez C., P. Langlais et J.Y. Nie (2004). Mots composés dans les modèles de langue pour la recherche d'information. *Actes de TALN 2004*, Fès, Maroc, 11-16.

Bastide, Y., R. Taouil, N. Pasquier, G. Stumme et L. Lakhal (2002). Pascal, un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, 21:65-95.

Beaudouin V. et F. Yvon (2004).Contribution de la métrique à la stylométrie. *Actes de JADT 2004*, Louvain, Belgique, 108-117.

Clarkson P.R., R. Rosenfeld (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. *Proceedings of ESCA Eurospeech*, Rhodes, Grèce, 1:2707-2710.

El-Bèze M., J-M. Torres-Moreno et F. Béchet (2005). Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac. *Actes de TALN 2005*, *Dourdan, France*, 2:125-134.

Grahne G. et J. Zhu (2003). Efficiently Using Prefix-trees in Mining Frequent Itemsets. Proceedings of the *First IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, Melbourne, FL.

Halliday M. et R. Hasan (1976). *Cohesion in English*. Longman Group.

Holmes D.I. (1998). The evolution of Stylometry in Humanities Scholarship. *Library and Linguistic Computing*, 13(3):111-117.

Hurault-Plantet M., M. Jardino et G. Illouz (2005). Modèles de langage n-grammes et segmentation thématique pour une tâche de filtrage de textes. *Actes de TALN 2005,* Dourdan, France, 2:135-144.

Illouz G. et M. Jardino (2001). Analyse statistique et géométrique de corpus textuels, *T.A.L., Traitement automatique des langues et linguistique de corpus*, 42(2):501-516.

Jardino M. (2000). Unsupervised non-hierarchical entropy-based clustering*, Data Analysis, Classification and Related Methods.* Eds. H.-H.Bock, W.Gaul, M.Schader. Springer, 29-35.

Jardino M. (2006). Identification des auteurs de textes courts avec des n-grammes de caractères. *Actes de JADT 2006,* Besançon, France, 2:543-549.

Jelinek F. (1998). *Statistical Methods for Speech Recognition*. MIT Press.

Juola P. (1997). What can we do with small corpora? Document categorization via cross-entropy. *Proceedings of an interdisciplinary workshop on similarity and categorization*, Edinburgh, UK.

Khmelev D.V. et J.T. Tweedie (2002). Using Markov Chains for Identification of Writers. *Literary and Linguistic Computing*, 16(4):299-307.

Labadié A., Y. Romero et L. Sitbon (2005). Segmentation et classification : deux politiques complémentaires. *Actes de TALN 2005,* Dourdan, France, 2:183-192.

Lebart L., A. Morineau et M. Piron (2000). *Statistique exploratoire multidimensionnelle*. Dunod.

Manning C.D. et H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Markov A.A. (1913). An example of Statistical Study on the Text of Eugene Onegin illustrating the linking of events to a chain. *Titre traduit du russe. Izvestija Imp. Akademii nauk, serija,* I(3):153-162.

Morris J. et G. Hirst (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21-48.

Peng F., D. Schuurmans, V. Keselj et S. Wang (2003). Language Independendent Authorship Attribution using Character Level Language Models. *Proceedings of ACL 2003,* Sapporo, Japon, 267-274.

Ponte, J.M. et W.B. Croft, (1998). A language Modeling Approach to Information Retrieval. *Proceedings of SIGIR 1998,* Melbourne, Australie, 275-281.

Rigouste L., O. Cappé et F. Yvon (2005). Modèle de mélange multi-thématique pour la fouille de textes. *Actes de TALN 2005,* Dourdan, France, 2:193-202.

Schmid H. (1999). Improvements in Part-of-Speech Tagging with an Application To German. In Armstrong, S., Chuch, K. W., Isabelle, P., Tzoukermann, E. & Yarowski, D. (Eds.), *Natural Language Processing Using Very Large Corpora*. Dordrecht : Kluwer Academic Publisher.

Shannon C.E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50-64.

Silber H.G. et K.F. McCoy (2000). Efficient text summarization using lexical chains. *Proceedings of the 5th international conference on Intelligent User Interfaces*, New Orleans, Etats-Unis, 252-255.

Sitbon, L. et P. Bellot (2004). Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. *Actes de TALN 2004*. Fès, Maroc, 441-450.

Stokes N., Carthy J., et Smeaton A. (2002). Segmenting broadcast news streams using lexical chains. *Proceedings of Starting Artificial Intelligence Researchers Symposium (STAIRS 2002),* 145-154.

Teahan W.J., (2000). Text classification and segmentation using minimum cross-entropy. *Proceedings of RIAO 2000*, Paris, France, 943-961.

Witten I.H. et Bell T.C. (1991). The zero-frequency problem : estimating the probabilities of novel events in adaptative text compression. *IEEE Transactions on Information Theory*, 37(4):1085-1094.

Summary

We have used both the style and the theme structure in texts for authorship attribution. The challenge DEFT'05 was to detect contiguous sentences of Mitterrand's speeches inserted in Chirac's speeches. The author's style has been defined with word-based or character-based n-gram language models. One model per author has been built using a train corpus. Each model has been applied to each sentence of a test corpus in order to detect the most probable author. Results have been fitted according to the task's constraints. At the same time we have developped a topic detection system using lexical chains. We have assumed that both authors could be differenciated according to the topic they talk about in each speech. One topic is defined by the main lexical chain which is the longuest one with a minimum overlap with the other chains. Then results of both methods have been merged.