Actes du sixième défi fouille de texte, DEFT2010, Montréal, Canada, 23 juillet 2010.
Proceedings of the Sixth DEFT Workshop, DEFT2010, Montréal, Canada, 23rd July 2010.
Pages 57-68

TALN 2010, Montréal, 19–23 juillet 2010

Classification de textes en comparant les fréquences lexicales

Michel Généreux
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

Résumé. Cet article fait état de travaux menés dans le cadre de la campagne DEFT 2010 concernant la classification de textes selon leur décennie ou leur origine. Nous détaillons d'abord l'approche adoptée ainsi que les ressources utilisées. On compare les fréquences des termes du lexique extrait d'un corpus d'apprentissage avec celles du lexique de référence, obtenant ainsi une liste de termes discriminants ou saillants pour chaque classe nous permettant d'attribuer un score à chaque document comme base de classification. Cette approche donne des résultats très compétitifs pour la classification selon l'origine et acceptable pour la classification diachronique. Nous utilisons aussi les lexiques de termes saillants servant de modèles pour la classification pour caractériser une classe de textes donnée.

Abstract. This article reports on work conducted under the tasks at DEFT 2010 concerning the classification of texts according to their decade or origin. First, we describe the approach and the resources used. We compare the frequencies of terms in the lexicon extracted from a training corpus with those extracted from a reference corpus, obtaining a list of discriminating or salient terms for each class allowing us to attribute a score to each document as a basis for classification. The approach gives very competitive results for the classification by origin and acceptable for the diachronic classification. We also use the glossaries of salient terms serving as models for the classification to characterize a given class of texts.

Mots-clés : Corpus comparables, Classification de textes, Analyse diachronique, Saillance, Correction orthographique.

Keywords: Comparable corpora, Classification of texts, Diachronic analysis, Salience, Spelling correction.

1 Introduction

Cette année le 6ième atelier *DÉfi Fouille de Texte (DEFT)* est consacré à la catégorisation de textes selon leur appartenance à une décennie (Tâche 1) ou selon leur origine (Tâche 2). Dans cet article nous présentons d'abord les méthodes que nous avons utilisées, en détaillant les ressources que nous avons mobilisées pour chacune de nos soumissions. Après quelques remarques sur la correction orthographique, nous faisons l'analyse, pour les deux tâches, des lexiques servant de modèles pour l'attribution d'un «score» à chaque texte, ce qui nous amène à discuter des termes les plus saillants pour chaque classe, des termes dont la variation diachronique est notable (incluant la disparition et l'apparition de termes) ainsi que des lexiques reliés au sport et à l'information. Finalement, nous présentons les résultats obtenus lors de la campagne et concluons.

2 Approche et Ressources Utilisées

Nous traitons l'ensemble des trois tâches comme un problème de classification. Notre approche est statistique mais contrairement à un bon nombre d'entre elles les modèles de classification que nous utilisons vont au-delà du sac de mots. L'idée de base est toutefois simple et a été utilisée dans des travaux sur le comportement diachronique d'expressions (Belica, 1996), ce qui s'apparente à la Tâche 1. Nous étendons et adaptons cette approche pour la classification de textes selon leur origine (Tâche 2). Dans cette approche, on génère une liste de termes saillants (i.e. des modèles pour chaque classe de textes) sur la base d'une comparaison fréquentielle entre les éléments lexicaux (1-grammes, 2-grammes et 3-grammes) d'un corpus d'apprentissage et d'un corpus de référence. Nous utilisons le *log odds ratio* (Baroni & Bernardini, 2004; Everitt, 1992) comme mesure statistique de la saillance d'un n-gramme. Le *log odds ratio* compare la fréquence d'occurrence de chaque n-gramme dans un corpus spécialisé (le corpus d'apprentissage) à sa fréquence d'occurrence dans un corpus de référence :

$$log\ odds\ ratio = ln(ad/cb) = ln(a) + ln(d) - ln(c) - ln(b)$$

où a est la fréquence du mot dans le corpus spécialisé, b est la taille du corpus spécialisé moins a, c est la fréquence du mot dans le corpus général et d est la taille du corpus général moins c. Une grande valeur de saillance positive indique une saillance forte, alors qu'une grande valeur négative indique un n-gramme sans importance pour la classe en question. Donc, à partir des corpus d'apprentissage, nous avons produit des modèles de classification pour chacune des classes des deux tâches (15 classes pour la Tâche 1 et 12 classes pour la Tâche 2). De plus, ces modèles se sub-divisaient en sous-classes, une pour chaque type de n-gramme (1, 2 et 3). Tous les textes ont été préalablement étiquetés morpho-syntaxiquement avec TreeTagger[1](Schmid, 1994), de telle sorte que chaque unité lexicale fût composée du lemme et de sa catégorie grammaticale (e.g. pomme_NOM).

Nous avons adopté une partie (75 Meg tokens) du corpus *frWAC*[2] (M. Baroni & Zanchetta., 2009) comme corpus de référence. FrWAC fait partie d'une collection de corpus de très grande taille récoltés sur Internet. Ce corpus est un bon candidat comme référence puisqu'il présente une diversification en thèmes et en genres. À titre illustratif, le tableau 1 présente le n-gramme le plus saillant pour chacun des modèles.

3 Tâche 1 : Classification selon la Décennie

Dans cette tâche, le corpus des décennies a été constitué à partir d'une «ocrisation» de journaux papiers, avec tout le bruit que cela implique, y compris au niveau des découpages de mots (e.g. dans le tableau 1, «der nier», «pro duire», etc.). Nous avons donc appliqué un pré-traitement à ce corpus visant à éliminer le plus possible les erreurs orthographiques. Nous avons d'abord construit un dictionnaire avec tous les mots de notre corpus de référence (847544 mots). Pour chaque mot du corpus de décennies, nous vérifions son orthographe de la manière suivante : s'il existe dans le dictionnaire, il est pris tel quel, sinon on essaie de le remplacer par le mot ayant la plus petite distance de Levenhstein avec un des mille mots les plus fréquents du dictionnaire. Pour limiter les temps de calcul, nous nous sommes aussi limité aux mots n'ayant pas plus ou moins de deux lettres de différences (en taille) et dont la distance de Levenhstein ne dépasse pas 2. De plus, puisque l'«ocrisation» coupait des mots en deux, nous avons remplacé toutes les paires de

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
[2] `http://wacky.sslmit.unibo.it/doku.php`

Classe	1-gramme	2-gramme	3-gramme
Décennies			
Décennie 1800	numéro	avoyer être	liste du émigré
Décennie 1810	numéro	il con	celui qui avoyer
Décennie 1820	numéro	pro mettre	être ainsi que
Décennie 1830	numéro	voix nombreux	vif de curiosité
Décennie 1840	numéro	der nier	ce der nier
Décennie 1850	leur	un même	quelque un même
Décennie 1860	numéro	li liberté	être ad mettre
Décennie 1870	numéro	der nier	on nous écrire
Décennie 1880	floquet	on mander	question du scrutin
Décennie 1890	floquet	pro duire	se avoir dresser
Décennie 1900	unioniste	nom do	sortir do ce
Décennie 1910	pagnon	on mander	température se être
Décennie 1920	loucheur	pro chainer	gouvernement de Empire
Décennie 1930	reichsmark	gouverne mentir	franc par action
Décennie 1940	numéro	bri tannique	communiquer du haut
Pays			
Québec Sports	Red	Red Wings	but sur balle
Québec Informations	Irak	monsieur Landry	premier ministre Jean
France Sports	Roland-Garros	Bernard Sainz	Internationaux de France
France Informations	Irak	Viktor Tchernomyrdine	guerre en Irak
Journaux			
Le Devoir Sports	Red	coupe Stanley	but sur balle
Le Devoir Informations	Irak	monsieur Landry	premier ministre Jean
La Presse Sports	Hurricanes	série éliminatoire	but sur balle
La Presse Informations	Irak	Bernard Landry	premier ministre Jean
Le Monde Sports	Roland-Garros	Bernard Sainz	Internationaux de France
Le Monde Informations	Irak	Viktor Tchernomyrdine	guerre en Irak
L'Est Répub. Sports	L1	Christophe Mengin	Ford Focus WRC
L'Est Répub. Informations	Irak	Ehud Barak	Roissy-Charles de Gaulle

TAB. 1 – Lemmas les plus saillants pour toutes les classes (modèles)

mots consécutives absents dans le dictionnaire par leur amalgame, s'il existait dans le dictionnaire. Au final, 0.82% (8856) des mots ont été corrigé, incluant 0.04% (434 mots) qui ont été ré-assemblé après un mauvais découpage. Cette correction orthographique n'avait pour but que de limiter le nombre d'erreurs introduites par l'«ocrisation», nous n'avons donc pas produit d'évaluation de cette correction. Le tableau 2 donne un aperçu de certaines corrections effectuées. Le tableau 3 dresse un portrait de la densité lexicale

10 premiers mots corrigés	**10 premières paires de mots raccordées**
caraclère → caractère	pira teries →pirateries
servloes → services	mouil lages →mouillages
térêt → intérêt	séné galais →sénégalais
Icns → dans	Pyré nées-Orientales → Pyrénées-Orientales
émment → comment	corré lative →corrélative
cieuse → cause	extraor dinaires →extraordinaires
msp → est	sémi nariste →séminariste
guel → quel	Arbu signy →Arbusigny
dâmes → mêmes	renou velés →renouvelés
mirait → serait	scru puleuse →scrupuleuse

TAB. 2 – Corrections orthographiques automatiques pour la Tâche 1 : corpus d'apprentissage

(Bacelar do Nascimento, 2000) du corpus des décennies, en tout point comparable à la densité lexicale du corpus de référence.

Classe	**Nb Doc.**	**Types**	**Tokens**	**Verbes**	**Adverbes**	**Noms**	**Adjectifs**
Référence	131090	847544	75836891	14.2%	4.5%	25.8%	7.0%
Décennie 1800	252	5707	61785	16.1%	5.9%	21.5%	5.9%
Décennie 1810	252	5637	60907	17.0%	6.5%	21.0%	5.8%
Décennie 1820	252	5630	62221	17.1%	6.5%	20.9%	5.8%
Décennie 1830	252	5526	63683	18.4%	6.8%	20.4%	5.5%
Décennie 1840	252	5676	62410	18.0%	6.2%	21.1%	5.4%
Décennie 1850	252	5956	60752	17.4%	6.1%	21.2%	6.0%
Décennie 1860	251	5845	59494	17.4%	5.9%	21.6%	6.1%
Décennie 1870	252	5755	60150	17.6%	6.2%	21.5%	5.9%
Décennie 1880	252	6181	59575	17.6%	6.5%	21.4%	6.0%
Décennie 1890	224	6028	53265	17.9%	6.5%	21.7%	6.1%
Décennie 1900	218	6008	51398	17.4%	6.1%	22.7%	6.5%
Décennie 1910	220	5925	51433	17.1%	5.5%	22.4%	6.6%
Décennie 1920	221	6232	50634	16.9%	6.2%	22.2%	6.5%
Décennie 1930	221	6163	50601	16.4%	5.7%	22.4%	6.9%
Décennie 1940	223	5905	49703	16.5%	5.6%	22.3%	7.1%

TAB. 3 – Densités lexicales pour la Tâche 1 : corpus d'apprentissage

Avant de présenter les résultats de la classification, nous faisons quelques observations intéressantes concernant le comportement diachronique de certaines expressions. Tout d'abord, le tableau 4 montre

des termes qui présentent une forte corrélation positive ou négative entre leur degré de saillance et les années qui passent, et ce sur toute la période couverte par la Tâche 1, soit 1800-1940. Une corrélation positive indique une utilisation de plus en plus prononcée avec le temps, alors qu'une corrélation négative indique une utilisation de moins en moins prononcée. Pour donner une illustration un peu plus parlante

Corrélation positive entre 1800 et 1940	**Corrélation négative entre 1800 et 1940**
constituer_VER	ouvrage_NOM
catholique_ADJ	roi_NOM
début_NOM	prouver_VER
con_NOM	former_VER
Etats-Unis_NAM	auteur_NOM
durée_NOM	point_ADV
conférence_NOM	art_NOM
façon_NOM	jugement_NOM
durer_VER	crainte_NOM
section_NOM	reste_NOM
non_ADV :seulement_ADV	avoir_VER :point_ADV
ce_PRO :qui_PRO :concerner_VER	projet_NOM :de_PRP :loi_NOM

TAB. 4 – Corrélations positive et négative durant la période 1800-1940

de ces processus de renforcement ou d'affaiblissement de l'utilisation d'un terme, nous présentons sur un diagramme en bâtons (voir figure 1) deux des termes dont l'utilisation va croissant et deux dont l'utilisation va décroissant (voir figure 2). Ainsi, les termes *catholique* et *Etats-Unis* sont de plus en plus utilisé entre 1800 et 1940, alors que *roi* et *projet de loi* le sont de moins en moins. Nous avons poussé un peu plus cette notion de progression ou régression de l'utilisation d'un terme en examinant ceux qui sont apparût ou disparût durant cette période. Un terme «apparaît» s'il n'existe pas durant au moins la portion 1800-1820 et au plus 1800-1910 et est utilisé au moins une fois durant chaque décennie du reste de la période couverte par la Tâche 1. À l'inverse, un terme «disparaît» s'il est utilisé au moins une fois durant chaque décennie de la portion 1800-1820 et au plus 1800-1910 et disparaît durant le reste de la période couverte par la Tâche 1. Des exemples illustratifs sont montrés dans le tableau 5.

Les résultats de la classification pour la Tâche 1 sont montrés et discutés à la section 5.

4 Tâche 2 : Classification selon l'Origine

Dans cette tâche, il s'agissait de classifier des articles de journaux plus récents selon leur origine national (*France* ou *Québec*) et selon leur source de publication (*Le Monde*, *L'Est Républicain*, *Le Devoir* ou *La Presse*). Chaque article appartenait à la rubrique sportive ou d'information, et ce détail nous était fournie. Le tableau 6 nous informe d'abord sur la densité du corpus de la Tâche 2, ici encore en tout point comparable à celle du corpus de référence. Nous donnons ici encore quelques informations sur des termes intéressants du corpus. Cette fois, le tableau 7 montre les termes tirés du corpus d'apprentissage ayant une saillance élevée pour toutes les articles de sport, ce qui peut représenter un *lexique sportif*. Le tableau 8 montre les termes tirés du corpus d'apprentissage ayant une saillance élevée pour toutes les articles

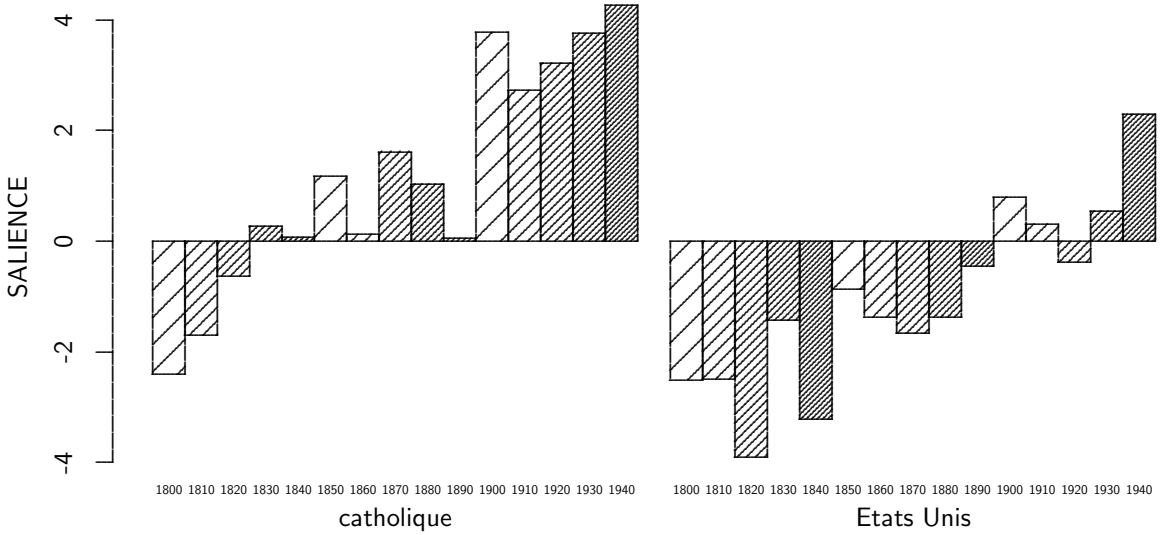

FIG. 1 – Comportement diachronique de *catholique* et *Etats-Unis*

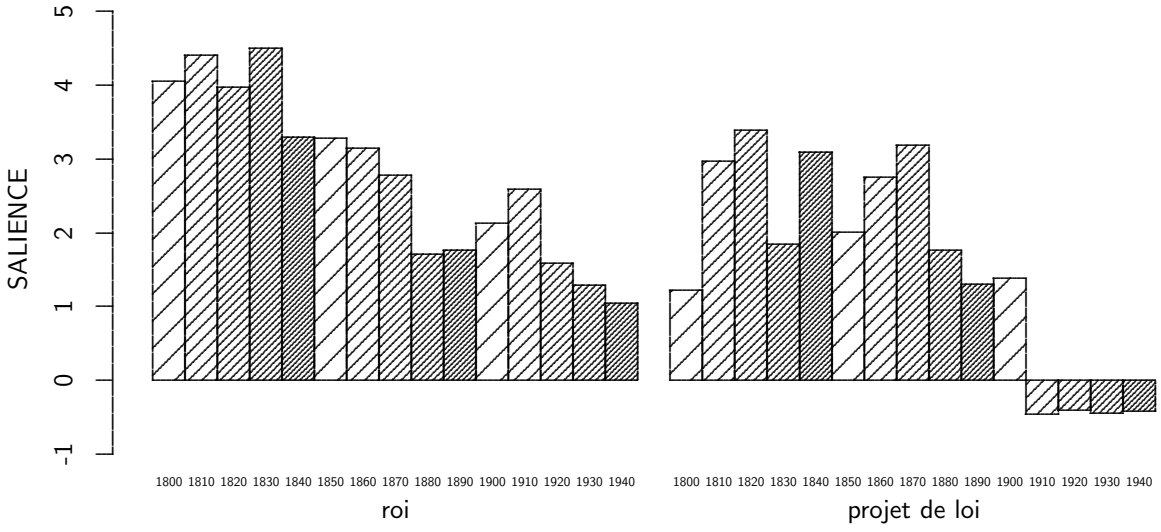

FIG. 2 – Comportement diachronique de *roi* et *projet de loi*

Apparition d'un terme durant une décennie	Disparition d'un terme durant une décennie
1830 chemin_NOM :de_PRP :fer_NOM	1830 caisse_NOM :de_PRP :amortissement_NOM
1840 clamer_VER	1830 Harpe_NAM
1840 Albert_NAM	1830 vendémiaire_NOM
1840 quelque_PRO :peu_ADV	1830 grand_ADJ :théâtre_NOM
1850 New-York_NAM	1830 partie_NOM :du_PRP :Monde_NAM
1860 tout_PRO :cas_NOM	1830 assemblée_NOM :constituant_ADJ
1870 commissariat_NOM	1830 dévoyer_VER
1870 défensif_ADJ	1830 avoyer_VER
1870 heure_NOM :actuel_ADJ	1830 étude_NOM :classique_ADJ
1880 télégramme_NOM	1830 faire_VER :le_DET :acquisition_NOM
1890 commerçant_NOM	1830 inimitié_NOM
1890 automobile_NOM	1830 division_NOM :militaire_ADJ
1900 tannique_NAM	1840 Saint-Domingue_NAM
1900 mentalité_NOM	1840 reprendre_VER :le_DET :discussion_NOM
1900 milieu_NOM :politique_ADJ	1860 avoir_VER :le_DET :malheur_NOM

TAB. 5 – Apparition et disparition d'un terme durant une décennie

Classe	Nb Doc.	Types	Tokens	Verbes	Adverbes	Noms	Adjectifs
Référence	131090	847544	75836891	14.2%	4.5%	25.8%	7.0%
Québec Sports	992	18625	418513	17.5%	6.1%	20.0%	5.1%
Québec Informations	999	22060	551252	16.2%	5.4%	22.7%	6.4%
France Sports	828	19588	356011	18.9%	7.4%	25.1%	7.0%
France Informations	900	21915	454721	15.4%	4.5%	23.2%	6.9%
Devoir Sports	475	12193	201127	17.8%	6.0%	20.1%	5.1%
Devoir Informations	501	15494	288188	16.0%	5.4%	22.6%	6.5%
Presse Sports	517	13700	217386	17.2%	6.2%	20.0%	5.1%
Presse Informations	498	15639	263064	16.4%	5.3%	22.7%	6.4%
Monde Sports	450	16956	304633	16.1%	6.3%	21.3%	6.0%
Monde Informations	450	15549	292073	15.7%	5.0%	22.6%	7.1%
Est Sports	378	11303	118575	15.4%	6.0%	20.5%	5.5%
Est Informations	450	13655	162648	14.7%	3.6%	24.2%	6.5%

TAB. 6 – Densités lexicales de la Tâche 2 : corpus d'apprentissage

Origine	1-gramme	2-gramme	3-gramme
France	Pro	Laurent Blanc	championnat de Europe
	OM	maillot jaune	quart de heure
	Nantes	club français	Coupe de Europe
	franc	Juventus Turin	Coupe de France
	Etats-Unis	Jacques Santini	million de franc
Québec	bâton	ce série	fin de semaine
	Montréal	être retirer	avoir mettre fin
	frappeur	avoir mentionner	avoir être retirer
	Robinson	six match	avoir marquer deux
	canadien	quatre coup	ne avoir donner
Devoir	s (seconde)	coupe Stanley	avoir faire savoir
	Hewitt	Lleyton Hewitt	Russe Marat Safin
	Lleyton	avoir franchir	tournoi de Wimbledon
	Seles	Monica Seles	savoir pas pourquoi
	Monica	faire savoir	Américaine Serena Williams
Presse	boulot	Brands Hatch	titan de Acadie-Bathurst
	Markov	avoir accomplir	Caroline du Nord
	Serguei	tu avoir	fin de saison
	CKAC	vendredi soir	se être emparer
	rocket	Kevin Weekes	Ligue du Champions
Monde	réputation	joueur français	ne se en
	responsabilité	cycliste international	ne pas être
	supporteurs	union cycliste	deux ou trois
	Zinedine	expliquer il	ne sembler pas
	rugby	Zinedine Zidane	union cycliste international
Est	hier	ski alpin	français du Jeux
	hier	Raymond Domenech	Ligue du Champions
	Peugeot	Britannique Tim	pays du Soleil
	10e	Justine Hénin	qui lui être
	rallye	titre olympique	dont il avoir

TAB. 7 – Lexique sportif

Origine	1-gramme	2-gramme	3-gramme
France	Blanche	Ben Laden	mise en examen
	Etats	François Hollande	mettre en examen
	Etat	parlement européen	département de Etat
	euro	tribunal correctionnel	million de euro
	Etats-Unis	maison Blanche	secrétaire de Etat
Québec	Québec	Québec avoir	attendre à ce
	État	ministre Jean	plan de action
	Montréal	gouvernement fédéral	il être aussi
	Ontario	Chrétien avoir	ce jour -ci
	Ottawa	comité exécutif	chef de accusation
Devoir	Devoir	avoir noter	conseil du ministre
	Lemieux	gouvernement québécois	plus ou moins
	caucus	se joindre	député du Bloc
	Parizeau	madame Pagé	union du municipalité
	compétition	Québec être	norme du travail
Presse	touriste	trois enfants	être encore plus
	Palestine	être signaler	nous avoir besoin
	musée	monsieur Netanyahu	dizaine de personnes
	Netanyahu	Ehud Barak	se être également
	heure	monsieur Barak	homme qui avoir
Monde	hôte	monsieur Bush	arme de destruction
	résumer	Pim Fortuyn	jouer un rôle
	Fortuyn	Viktor Tchernomyrdine	tout le monde
	Tchernomyrdine	nation uni	feuille de route
	uni	premier ministre	département de Etat
Est	hier	tribunal correctionnel	mettre en examen
	Premier	François Hollande	an avoir être
	hier	avoir requérir	million de franc
	hier	ben Laden	avoir être condamner
	correctionnel	trois homme	prendre le fuite

TAB. 8 – Lexique de l'information

d'information, ce qui peut représenter un *lexique d'information*. Les résultats de la classification pour la Tâche 2 sont montrés et discutés à la section suivante.

5 Résultats

Nous avons classifié chacun des articles du corpus de test en utilisant les ressources et la méthode décrites à la section 2. Ainsi, pour un texte donné, on compare la somme des valeurs de saillance de tous les termes présents dans les modèles. On pondère le choix des classes finales de la manière suivante : si les trois modèles «n-gramme» s'entendent[3] sur une même classe, cette classe est choisie avec un indice de confiance de un, si deux seulement s'entendent sur une classe alors celle-ci est choisie avec un indice de 0.7 et la classe unique reçoit un indice de 0.3. Finalement, si les trois classes diffèrent, alors la classe avec la somme des saillances la plus élevée reçoit un indice de 0.4 et les deux autres un indice de 0.3. Pour chaque tâche (Tâche 1, Tâche 2 - pays et Tâche 2 - journaux), nous avons produit trois soumissions. La première soumission incluait dans le calcul final les saillances de tous les termes issus des modèles produits à partir des fichiers d'apprentissage. La deuxième soumission excluait du calcul les saillances négatives et la troisième excluait du calcul les termes dits *hapax*. C'est la première soumission qui a obtenu les meilleurs résultats pour l'épreuve de classification avec les deux autres soumissions tout près derrière. Nous ne présentons ici que les détails des résultats reliés à la soumission 1. Le tableau 9 montre les résultats obtenus par l'ensemble des participants alors que les tableau 10 et 11 montrent les résultats que nous avons obtenus pour les deux tâches.

Statistique	**Tâche 1**	**Tâche 2 - Pays**	**Tâche 2 - Journaux**
Moyenne F-mesure	0.193	0.767	0.489
Médiane F-mesure	0.181	0.792	0.462
Écart-type F-mesure	0.098	0.1367	0.1887

TAB. 9 – Résultats Généraux

Pour la Tâche 2 dans son ensemble, nous faisons bonne figure, avec des F-mesure de 0.858 (Pays) et 0.630 (Journaux), comparativement à 0.767 (Pays) et 0.489 (Journaux) pour l'ensemble des participants. Les résultats sont plutôt faible pour la Tâche 1 si l'on regarde la F-mesure obtenue (0.183) mais «moyens» si l'on compare avec la F-mesure de l'ensemble des participants (0.193). Cependant, l'exactitude (0.167) reste bien au-delà de ce qu'on obtiendrait par chance (15 classes $\rightarrow$ 0.067). On remarque qu'il y a une corrélation marquée (0.53) entre le nombre de termes et le F-mesure, et une corrélation forte (0.81) entre la moyenne de la saillance et la F-mesure. L'approche est donc grandement dépendante du choix d'un corpus de référence approprié permettant de générer un nombre important de termes avec une saillance forte. Nous constatons aussi qu'il existe une forte corrélation négative entre la chronologie et la F-mesure : en d'autres termes, les résultats sont moins bons pour les décennies plus contemporaines, ce qui laisse supposer que le corpus de référence est plutôt construit à partir de documents récents, ce qui a pour conséquence de produire moins de termes saillants pour les décennies récentes. Un corpus de référence mieux réparti dans le temps permettrait sans doute d'éviter la dégradation des performances observées pour la classification des décennies plus récentes.

[3] Si la même classe obtient le score le plus élevé pour les 1-grammes, 2-grammes et 3-grammes.

Classe (Saillance moyenne, Nb de termes)	Rappel	Précision	F-mesure
1800 (8.19, 30343)	0.396	0.349	0.371
1810 (7.96, 30026)	0.181	0.212	0.195
1820 (7.95, 30387)	0.252	0.171	0.204
1830 (7.86, 31902)	0.496	0.131	0.207
1840 (7.73, 30614)	0.122	0.089	0.103
1850 (7.65, 29750)	0.211	0.134	0.164
1860 (7.59, 28722)	0.069	0.120	0.088
1870 (7.58, 29468)	0.121	0.127	0.124
1880 (7.51, 29484)	0.072	0.094	0.081
1890 (7.55, 27622)	0.130	0.137	0.133
1900 (7.73, 26901)	0.095	0.164	0.120
1910 (7.71, 26407)	0.130	0.179	0.151
1920 (7.48, 25696)	0.062	0.222	0.096
1930 (7.58, 26003)	0.053	0.227	0.086
1940 (7.66, 25078)	0.166	0.620	0.261
Exécution 1 Décennies (Exactitude = 0.167)	0.171	0.198	0.183

TAB. 10 – Résultats Exécution 1 - Tâche 1

Classe	Rappel	Précision	F-mesure
France	0.801	0.883	0.840
Québec	0.908	0.840	0.873
Exécution 1 Pays (Exactitude = 0.858)	0.854	0.861	0.858
La Presse	0.470	0.617	0.534
Le Devoir	0.598	0.543	0.569
Le Monde	0.926	0.568	0.704
L'Est Républicain	0.435	0.890	0.585
Exécution 1 Journaux (Exactitude = 0.606)	0.607	0.655	0.630

TAB. 11 – Résultats Exécution 1 - Tâche 2

6 Conclusion et Perspectives

Nous avons décrit les ressources utilisées et notre approche ainsi pour la classification de textes dans le cadre de la campagne DEFT 2010, soient une section du corpus frWAC, l'étiqueteur morpho-syntaxique TreeTagger et une méthode de classification basée sur une comparaison fréquentielle lexicale. Notons qu'il serait tout à fait possible et intéressant d'utiliser l'ensemble des textes d'apprentissage comme corpus de référence et de travailler directement sur les *lexis*, ce qui rendrait l'approche indépendante de toute ressource externe. La mesure de saillance deviendrait alors une mesure de distance sémantique entre un document à classifier et le «centre de gravité» du corpus d'apprentissage.

Cependant, l'approche décrite nous a non seulement permis d'obtenir des résultats très compétitifs pour la classification de textes selon leur origine mais aussi d'extraire des éléments lexicaux caractérisant une classe donnée ou subissant des changements diachroniques importants. Notre approche serait avantageu-

sement complétée par une contribution interdisciplinaire avec les sciences sociales (politique, histoire et sociologie) pour tirer le maximum d'information des termes saillants extraits.

Notre approche a obtenu des résultats moyens pour la classification selon la décennie, ce que nous expliquons par la faible variation diachronique du corpus de référence avec comme résultat un nombre moins important de termes saillants pour les périodes récentes. Néanmoins, cette approche comparative fût très productive et a produit des résultats intéressants, nous avons donc l'intention de l'appliquer à des textes d'autres périodes et provenant d'autres sources.

Références

BACELAR DO NASCIMENTO M. F. (2000). *Corpus de Référence du Portugais Contemporain*, In *Corpus, Méthodologie et Applications Linguistiques*, p. 25–30. Presses Univ. de Perpignan. Editor : M. Bilger.

BARONI M. & BERNARDINI S. (2004). Bootcat : Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, p. 1313–1316.

BELICA C. (1996). Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics*, **1**(1), 61–73.

EVERITT B. (1992). *The analysis of contingency tables*. London : Chapman and Hall.

M. BARONI, S. BERNARDINI A. F. & ZANCHETTA. E. (2009). The wacky wide web : A collection of very large linguistically processed web-crawled corpora. In *Language Resources and Evaluation*, volume 43, p. 209–226.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.