RegCan : une base de données textuelle sur la règlementation fédérale canadienne

Marc Tremblay-Faulkner
Louis-Robert Beaulieu-Guay
Éric Montpetit
Maxime Sainte-Marie
Jean-François Chartier
Vincent Larivière

Document remis à l'École de la fonction publique du Canada dans le cadre de l'appel d'offres 24062-19-096 (Perspectives sur les règlements générés par l'intelligence artificielle)

Centre interuniversitaire de recherche sur la science et la technologie

14 septembre 2018

Pour toute information supplémentaire en lien avec ces rapports ou pour toute communication ultérieure, veuillez contacter :

Louis-Robert Beaulieu-Guay
louis-robert.beaulieu-guay@umontreal.ca
514-754-6289
3150 Rue Jean-Brillant, C-3112
Montréal, Québec, H3T 1N8

Table des matières

Rapport d'analyse 1
Résumé et contributions principales 1
Objectifs du rapport 2
1.Présentation détaillée de l'équipe de recherche 3
2. Présentation de la base de données 5
3. Évoluation de l'activité règlementaire fédérale canadienne 7
Analyse du nombre de REIR produits 7
Analyse du nombre de mots des REIR 10
Analyse de spécificités lexicales 13
Analyse de la diversité lexicale et de la lisibilité du corpus 17
En Résumé 19
4. Agenda et contrainte règlementaire 20
Analyse thématique du corpus 20
La contrainte dans la règlementation fédérale canadienne 25
Conclusion et analyses futures 34
Références 36
Annexes 38
A. Indices de contrainte 38
B. Spécificités lexicales par ministère 40
C. CRSH d'Éric Montpetit 50
Rapport Méthodologique 57
1. Extraction des REIR 58
2. Nettoyage et prétraitement du corpus 58
3. Analyse de points de rupture 60
4. Diversité lexicale et lisibilité 61
5. Analyse thématique automatique: prétraitement et création du modèle 62
6. Analyse de spécificité 63
7. Analyse de correspondances 63
8. Analyse de clusters 66
Références 69

Rapport d'analyse

Notice: The dataset used for the analyses is from the English version of the Canada Gazette (part II). Subsequent work can be written in English.

Résumé et contributions principales

Dans ce rapport, nous présentons RegCan, une base de données textuelle inédite sur la règlementation fédérale canadienne dont l'unité d'analyse textuelle principale est le résumé d'étude d'impact de règlement (REIR). Le corpus regroupe 5843 REIR publiés dans la partie II de la Gazette du Canada entre 1998 et 2017. Les analyses exploratoires présentées dans ce rapport combine l'apprentissage machine non supervisée (analyse thématique automatique, analyse de *clusters*) à des techniques statistiques d'analyse textuelle (analyse de spécificités lexicales, analyse de correspondances). Les résultats de nos analyses démontrent la pertinence et la validité de RegCan pour répondre aux questionnements du Secrétariat du Conseil du Trésor et de l'École de la fonction publique du Canada.

Dans la première partie du rapport, nous montrons que le nombre de REIR produits a diminué au cours de la période 1998-2017, avec une baisse marquée dès 2006, alors que la moyenne du nombre de mots des REIR a augmenté significativement à partir de 2012. Nous émettons l'hypothèse qu'il s'agit du résultat de réformes administratives sur la règlementation. Par ailleurs, une analyse des spécificités lexicales du corpus montre que le vocabulaire utilisé dans les REIR change de manière significative et qu'on porte une attention nouvelle à la question des couts administratifs de la règlementation à partir de 2013. Bien que le corpus évolue dans le temps, une évaluation de la diversité lexicale et de la lisibilité indique que celui-ci est demeuré stable. Nous constatons cependant des variations importantes entre les ministères.

Dans la deuxième partie du rapport, nous cherchons à dresser un portrait à la fois précis et exhaustif de l'agenda règlementaire canadien en accordant une attention particulière à la question de la contrainte règlementaire. Pour ce faire, nous avons produit une analyse thématique automatique de l'activité règlementaire canadienne. Nos résultats indiquent que les principaux thèmes de la règlementation canadienne sont l'harmonisation procédurale, la sécurité intérieure et la clarification linguistique. Les thèmes en progression concernent les affaires autochtones, la règlementation sur les matières dangereuses et la sécurité intérieure. Nous croisons ensuite les différents thèmes du corpus avec une mesure originale de la contrainte règlementaire. Les résultats montrent que la contrainte dans l'activité règlementaire est demeurée globalement stable. Cependant, nous montrons que les formes de contrainte dans l'activité règlementaire varient, allant du contrôle des organisations, à la gestion des risques, à l'imposition d'interdictions.

Objectifs du rapport

Le présent rapport poursuit trois objectifs. Le premier objectif est de valider l'utilisation des résumés d'étude d'impact de règlement (REIR) pour analyser l'évolution règlementaire fédérale au Canada.

Le deuxième objectif est de fournir, à partir des REIR, une analyse centrée sur l'expression du concept particulier de la *contrainte* contenue dans la règlementation canadienne. Cette analyse est à la fois temporelle et thématique.

Enfin, nous souhaitons montrer à l'École de la fonction publique et au Secrétariat du Conseil du Trésor du Canada que notre équipe possède un bagage de connaissances fondamentales sur la règlementation, qu'elle a su développer une méthode originale pour l'analyser et qu'elle maitrise les méthodes et algorithmes de fouille de texte et d'apprentissage machine nécessaire pour faire progresser les connaissances sur le sujet.

Composée de spécialistes des politiques publiques, de l'analyse textuelle et de l'apprentissage machine, notre équipe compte aussi sur l'expertise et l'infrastructure du Centre interuniversitaire de recherche sur la science et la technologie (CIRST). Le CIRST a mis en place le Bureau des initiatives numériques (BIN) dont le mandat est de constituer, d'entretenir et de rendre public des bases de données textuelles. Le BIN se spécialise ainsi dans la constitution de corpus, le prétraitement des données textuelles, la modélisation, l'analytique automatique et semi-automatique, ainsi que dans la création, la visualisation, l'interprétation et l'évaluation de résultats.

1. Présentation détaillée de l'équipe de recherche

Tous les membres de l'équipe de recherche sont affiliés au CIRST. Marc Tremblay-Faulkner et Louis-Robert Beaulieu-Guay sont doctorants en science politique à l'Université de Montréal sous la direction d'Éric Montpetit. Leurs recherches portent respectivement sur rôle de l'expertise des groupes d'intérêt dans les forums de politiques publiques et l'impact de la consultation dans l'élaboration de la règlementation fédérale canadienne. Éric Montpetit est professeur de science politique à l'Université de Montréal et ses recherches portent sur l'impact des connaissances scientifiques sur les changements règlementaires au Canada, aux États-Unis et en Europe. Il est responsable de l'axe *Expertise, décision et débat public* au Centre interuniversitaire de recherche sur la science et la technologie (CIRST).

Jean-François Chartier est scientifique de données au Bureau des initiatives numériques du CIRST. Il est titulaire d'un doctorat en informatique cognitive et spécialiste des méthodes computationnelles de fouille de textes et d'apprentissage machine. Maxime Sainte-Marie est chercheur postdoctoral au CIRST et à la Chaire de recherche du Canada sur les transformations de la communication savante où il travaille avec Vincent Larivière. Formé au Canada, en Suède et en Belgique, il est titulaire d'un doctorat en informatique cognitive de l'Université du Québec à Montréal. Vincent Larivière est professeur agrégé à l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal, où il enseigne les méthodes de recherche en sciences de l'information et la bibliométrie. Titulaire de la Chaire de recherche du Canada sur les transformations de la communication savante, il est également membre du Collège de nouveaux chercheurs de la Société royale du Canada et du CIRST.

Notre équipe a d'importantes collaborations internationales sur le thème de la règlementation gouvernementale, notamment en Europe et aux États-Unis. Éric Montpetit, Louis-Robert Beaulieu-Guay et Marc Tremblay-Faulkner travaillent avec des chercheurs européens, David Aubin et Lise Freher, de l'Université catholique de Louvain, sur la comparaison des études d'impact règlementaire de l'Union européenne et du Canada.

Des discussions sont en cours avec Brendon Swedlow de la Northern Illinois University et Martin Lodge de la London School of Economics pour concilier à la leur notre méthode d'analyse de la règlementation. Ces derniers chercheurs s'intéressent tout particulièrement à la notion de risque dans la règlementation en Amérique du Nord et en Europe (Swedlow et al 2009; Hammit et al. 2013). Nous espérons pouvoir produire des analyses conjointes avec ce groupe dans un avenir proche.

Éric Montpetit est responsable d'un projet de recherche sur l'adaptation règlementaire financé par le Conseil de recherche en sciences humaines du Canada. Le projet porte sur le lien entre les avancées scientifiques/technologiques et les changements à la règlementation canadienne. Le projet repose sur

l'utilisation de techniques d'analyse textuelle liée à l'intelligence artificielle et vise la mise en place d'un site permettant de visualiser l'évolution règlementaire. Le projet est annexé à ce rapport.

2. Présentation de la base de données

Les données présentées dans ce rapport sont tirées d'un corpus de 5843 REIR que nous avons extrait des PDF de la partie II de la Gazette officielle après les avoir moissonnés du site web du gouvernement du Canada. Nous nous limitons à la période 1998-2017 pour des raisons techniques, car les textes règlementaires publiés avant celle période sont des images numérisées qui doivent faire l'objet d'une reconnaissance optique de caractère. Traiter ces documents demeure possible, mais pas dans les délais prescrits par l'appel d'offres. Nous n'étendons pas l'analyse au-delà de 2017, car nous ne prenons en compte que des années complètes. Les algorithmes de moissonnage de la Gazette que nous avons développés nous permettront toutefois de faire des mises à jour dès que de nouveaux REIR seront publiés dans la Gazette. La démarche d'extraction et de prétraitement du corpus est détaillée dans le rapport méthodologique.

Nous privilégions le REIR comme unité d'analyse et source textuelle principale du corpus plutôt que le texte du règlement[1]. Nous avons fait ce choix pour plusieurs raisons. Premièrement, il permet de circonscrire la base de données aux règlements susceptibles d'avoir un impact sur la population et les entreprises. N'incluant que les REIR, le corpus écarte les règlements qui ne sont accompagnés que d'une note explicative, soit les règlements dont la valeur est de nature plutôt symbolique. En effet, plusieurs règlements n'ont pour but que de confirmer certaines actions périrèglementaires, les décrets d'entrée en vigueur des lois ou les nominations en étant des exemples évocateurs.

Deuxièmement, le REIR est écrit en langage courant, plus accessible que le langage juridique des textes des règlements, ce qui contribue à accroitre la validité des analyses textuelles. Le langage juridique est moins varié ce qui, dans certains règlements, engendre la répétition de certains mots. Ces répétitions, dont le sens est souvent spécifique au contexte juridique, peuvent biaiser les analyses qui utilisent la fréquence des mots, la diversité lexicale et les matrices textuelles.

Troisièmement, le texte règlementaire omet des informations qui pourraient enrichir les analyses. Le contexte qui justifie le changement règlementaire, les objectifs que poursuit le ministère responsable du règlement et les sources d'information sur lesquelles repose le changement ne font généralement pas partie du texte du règlement, mais sont systématiquement inclus dans le REIR. Le REIR nous permet donc de produire des analyses sur une plus vaste gamme d'enjeux que ne le permet le texte du règlement.

Quatrièmement, le REIR est écrit sous un format relativement systématique. Bien que la forme du REIR ait évolué dans le temps, les variations des textes des REIR sont moindres que celles des textes règlementaires. En effet, certains textes règlementaires consignent plusieurs dizaines de pages de tableaux ou de listes alors que d'autres ne comprennent que des articles succincts. Les tableaux et listes, lorsqu'ils sont traitables, produisent des résultats aberrants. Les REIR sont produits sous la

[1] Notons que le texte règlementaire associé à un REIR n'est pas le texte du règlement complet, mais celui du changement règlementaire.

supervision du Secrétariat du Conseil du Trésor qui veille à leur uniformité relative. Les tableaux et les longues listes de certains règlements sont exclus des textes des REIR. Pour toutes ces raisons, nous préférons produire nos analyses à partir des textes des REIR plutôt que des textes des règlements. Cela étant dit, lorsque l'objectif s'y prête, nous sommes en mesure d'utiliser le texte règlementaire comme unité d'analyse, tel que le montrent certaines analyses présentées plus loin.

Il est à noter que dans le cadre de ce rapport, nous avons choisi de retirer la section « Consultation » des textes des REIR. Cette section documente la contribution de la société civile au changement règlementaire, alors que nous nous concentrons ici sur la règlementation en soi. Inclure la section Consultation pourrait donc biaiser nos résultats. Il va sans dire que la contribution de la société civile au changement règlementaire est un sujet qui mérite une analyse et notre corpus nous permettra éventuellement de réaliser une telle analyse.

Le Tableau 1 présente les données descriptives des termes (ou mots) du corpus des REIR et les compare avec les termes utilisés dans les textes règlementaires. Les deux distributions sont positivement asymétriques, avec une médiane beaucoup moins élevée que la moyenne. Tel qu'attendu, la variance des textes règlementaires est plus élevée que celle des REIR, avec un écart-type moyen de 4377 pour les textes règlementaires contre 1896 pour les REIR. La valeur élevée de l'écart-type observé pour les textes règlementaires confirme ce qui a été mentionné plus haut, à savoir que la taille des textes règlementaires varie fortement, allant de longues descriptions incluant plusieurs listes et tableaux à un petit nombre d'articles succincts. Dans l'ensemble, les données présentées par le Tableau 1 confirment le bienfondé de notre décision de nous concentrer sur les textes plus uniformes des REIR.

Tableau 1 : données descriptives du corpus non traité (n = 5843)

Nombre de mots (arrondi à l'unité)	**REIR**	**Texte règlementaire**
Moyenne	1443	1435
Médiane	852	451
Écart-type	1896	4098
Minimum	45	45
Maximum	34 827	138 497
Nombre total d'occurrences lexicales	9 332 647	10 739 985
Nombre total de types lexicaux	2 289 296	1 576 199

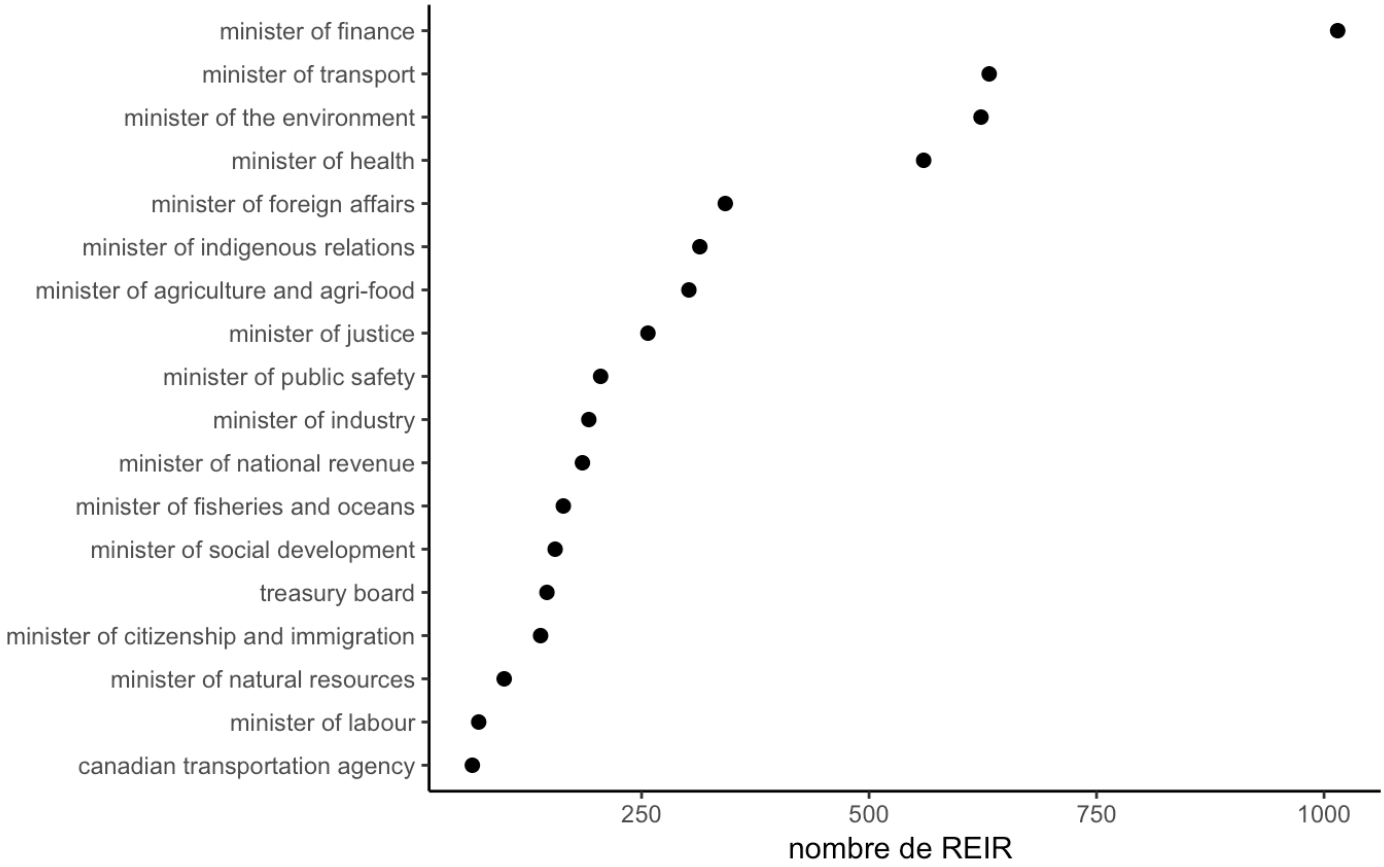
3. Évolution de l'activité règlementaire fédérale canadienne

Dans cette section, nous explorons la validité du corpus. Nous examinons d'abord l'évolution de l'activité règlementaire fédérale canadienne dans le temps et entre les ministères sur la base de deux indicateurs, soit le nombre de REIR produits et le nombre de mots contenus dans le REIR. Les résultats montrent que le nombre de REIR produits a diminué au cours de la période 1998-2017, avec une baisse marquée dès 2007, alors que la moyenne du nombre de mots des REIR a augmenté significativement à partir de 2012. Nous émettons l'hypothèse qu'il s'agit du résultat de deux réformes administratives, issues de la *Directive du Cabinet sur la rationalisation de la règlementation* (Canada 2007) et de la *Directive du Cabinet sur la gestion de la règlementation* (Canada 2012). Pour compléter la première série d'analyses, nous examinons les spécificités lexicales selon différentes périodes temporelles. L'analyse des spécificités montre que le vocabulaire utilisé dans les REIR change de manière significative et qu'on porte une attention nouvelle à la question des couts administratifs de la règlementation à partir de 2013. Bien que le corpus ait évolué, une analyse de la diversité lexicale et de la lisibilité indiquent que, sur ces aspects, le corpus est demeuré stable dans le temps. Nous montrons dans ce qui suit que, malgré quelques changements dans le temps, RegCan est valide pour mesurer l'activité règlementaire durant l'ensemble de la période à l'étude.

Analyse du nombre de REIR produits

La figure 1 montre le nombre total de REIR produits par chaque ministère ou agence responsable durant l'ensemble de la période. Les 5843 REIR de notre corpus ont été produits par un total de 47 ministères ou agences responsables distincts. À des fins de visualisation, la figure 1 ne présente que les ministères ou agences dont le nombre de REIR équivaut à au moins 1 % du corpus. Parmi ces différentes institutions, le ministère des Finances a été de loin la plus productive, avec 1015 REIR produits entre 1998 et 2017, soit 17,4 % des REIR du corpus. Cette activité règlementaire du ministère des Finances peut s'expliquer par son activité législative, dans la mesure où le ministère des Finances est responsable du quart de tous les projets de loi déposés au Parlement chaque année (Canada 2018). Les ministères des Transports, de l'Environnement et de la Santé suivent dans l'ordre, avec respectivement 632, 623 et 560 REIR. Les REIR de ces trois ministères, conjointement à ceux relevant du ministère des Finances, ont produit près de 50 % des changements règlementaires consignés dans la Gazette du Canada entre 1998 et 2017. Un autre groupement de ministères qui produit un nombre significatif de REIR inclut les Affaires mondiales (autrefois Affaires étrangères) (342), les Affaires autochtones (314), Agriculture (283) et Justice (257). L'Agence canadienne du transport, le ministère du Travail et le ministère des Ressources naturelles figurent au bas de la liste.

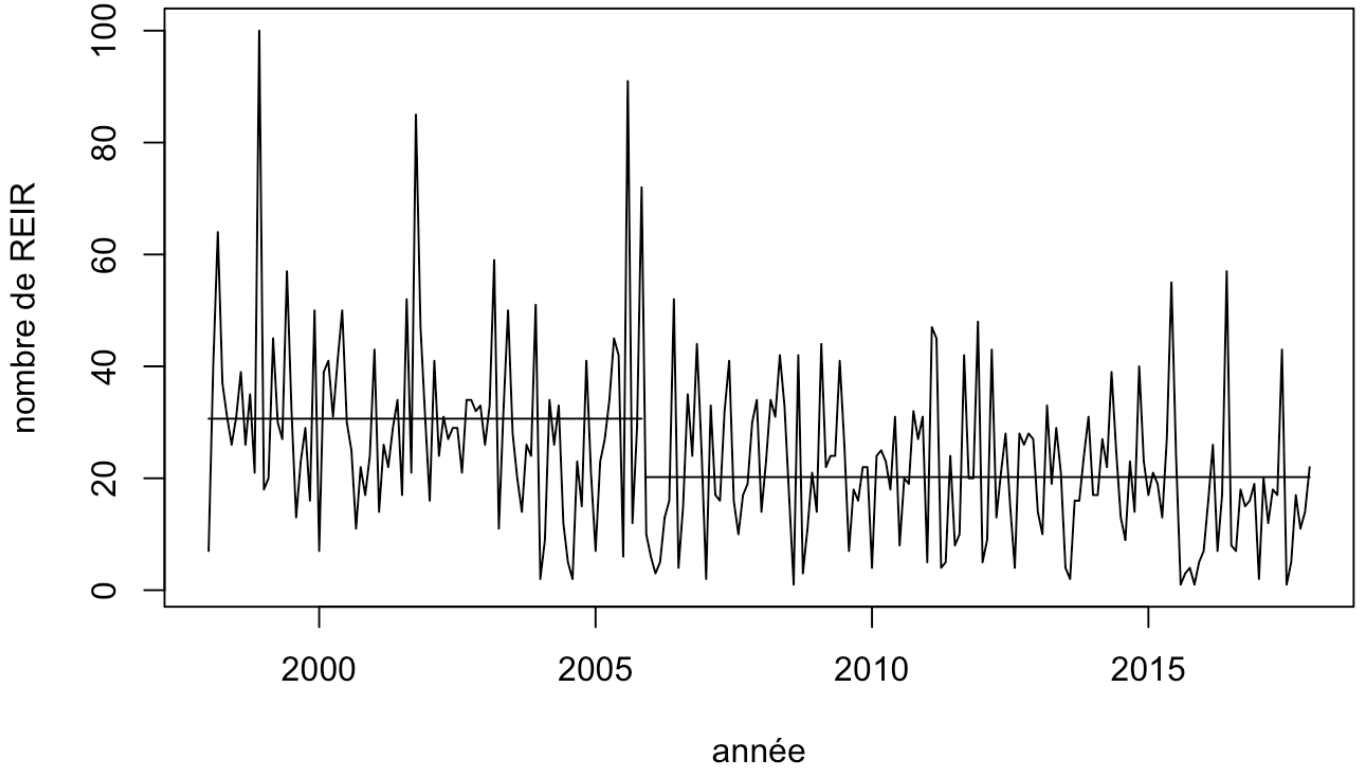
Figure 1 : nombre total de REIR produits par ministère ou agence responsable, 1998-2017



La figure 2 montre le nombre de REIR produits mensuellement durant la période observée. La baisse de l'activité règlementaire est significative. En agrégeant les données annuellement, on constate que le nombre de REIR passe de 457 en 1998 à 182 en 2017. Cette baisse pourrait s'expliquer par les réformes de 2007 et de 2012, lesquelles visaient à alléger le poids de la règlementation au Canada. Comme son nom l'indique, la *Directive du Cabinet sur la rationalisation de la règlementation* de 2007 visait à encourager la diversification des instruments de politiques publiques de manière à ce que le recours à la règlementation soit moins fréquent (Canada 2014). En 2012, le Cabinet récidive avec l'adoption d'une directive visant à mieux prendre en compte le poids d'un règlement sur les entreprises avant son adoption ainsi qu'à améliorer la coordination interministérielle de manière à éviter le chevauchement et la duplication (Canada 2012).

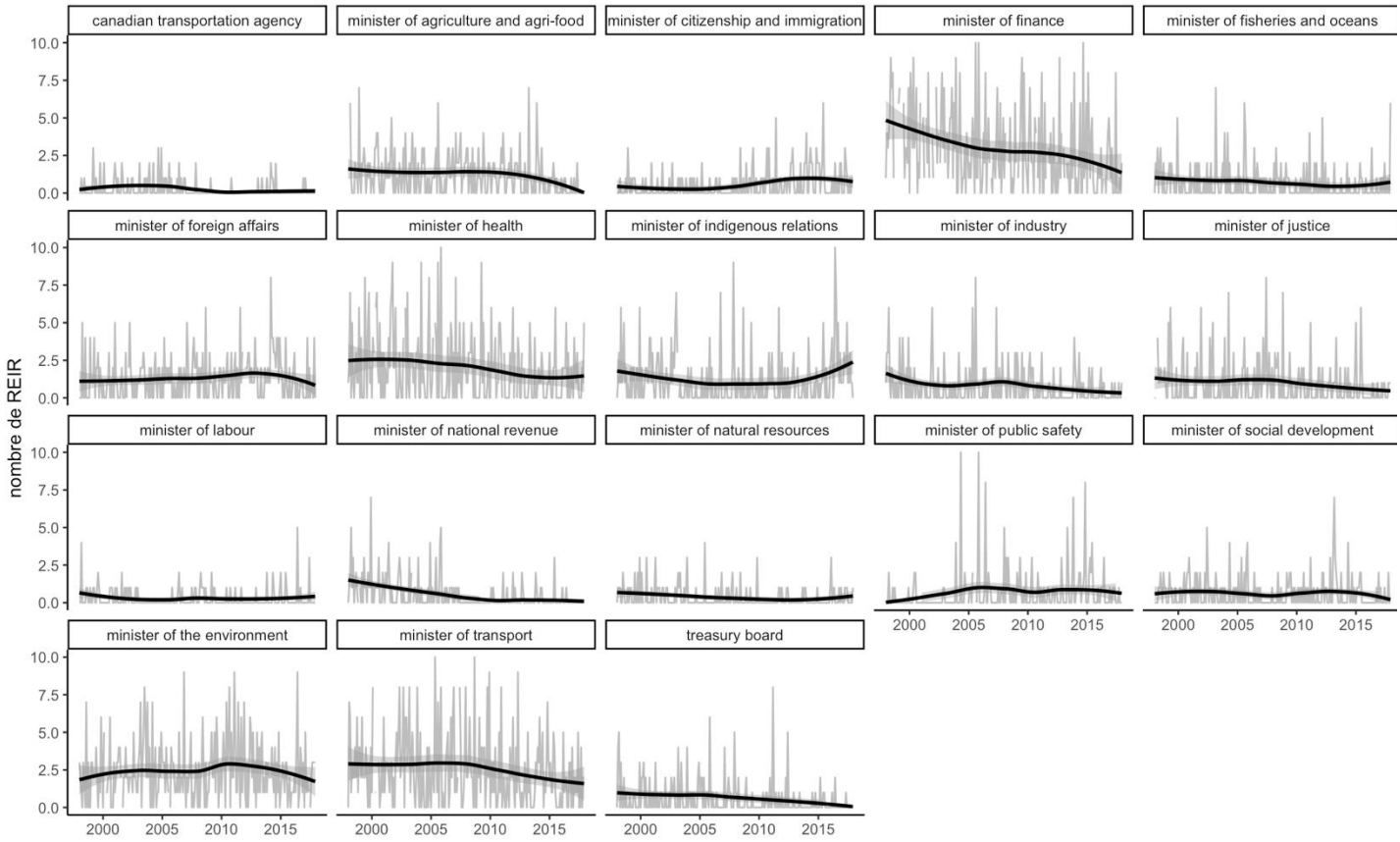
Pour mesurer l'impact de ces réformes administratives sur l'activité règlementaire, nous avons procédé à une analyse de points de rupture (*changepoint analysis*). Ce type d'analyse est utile pour détecter les changements de magnitude dans une série chronologique. La rupture des lignes horizontales de la figure 2 montre qu'il y a eu un changement significatif dans la moyenne de REIR produits. Ce changement de la moyenne, détecté à la fin de l'année 2006, est significatif. Voir le rapport méthodologique pour plus de détails sur cette technique.

Figure 2: nombre mensuel de REIR produits, 1998-2017



La figure 3 présente l'évolution du nombre de REIR produits au fil du temps pour chaque ministère ou agence responsable. La figure montre que la baisse globale de l'activité règlementaire est principalement attribuable aux ministères des Finances, de l'Industrie, de la Justice, du Transport et de la Santé. L'activité règlementaire est en faible hausse pour les ministères de l'Immigration et des Affaires autochtones ainsi que de la Sécurité publique dans une moindre mesure. Lorsque nous examinons les courbes par ministère, nous constatons que l'effet des réformes administratives sur la production de REIR n'a pas été uniforme d'un ministère à l'autre. En effet, alors que la rupture présentée à la figure 2 entre les périodes pré-2006 et post-2006 s'observe également pour les ministères des Finances et des Transports, il en va autrement dans le cas du ministère de l'Environnement, pour ne prendre qu'un exemple.

Figure 3 : nombre mensuel de REIR produits par ministère ou agence responsable, 1998-2017

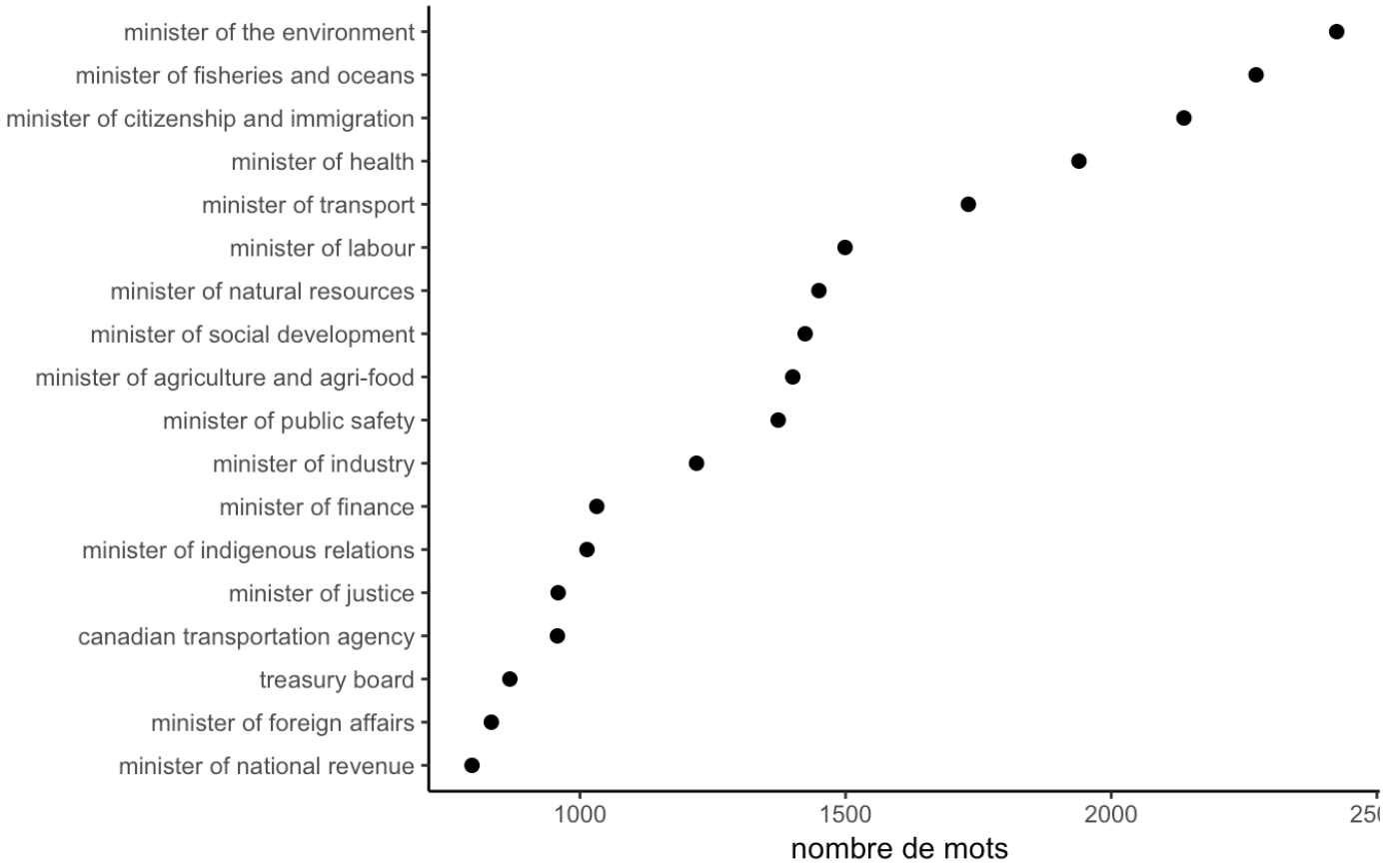


Analyse du nombre de mots des REIR

Après avoir mesuré l'activité règlementaire en termes de nombre de REIR produits, il est important d'analyser le contenu textuel du corpus. Pour ce faire, nous examinons le nombre de mots contenus dans les REIR. Tout comme pour l'analyse de l'activité règlementaire, l'objectif est d'analyser l'évolution du texte du REIR dans le temps et de vérifier la validité du corpus.

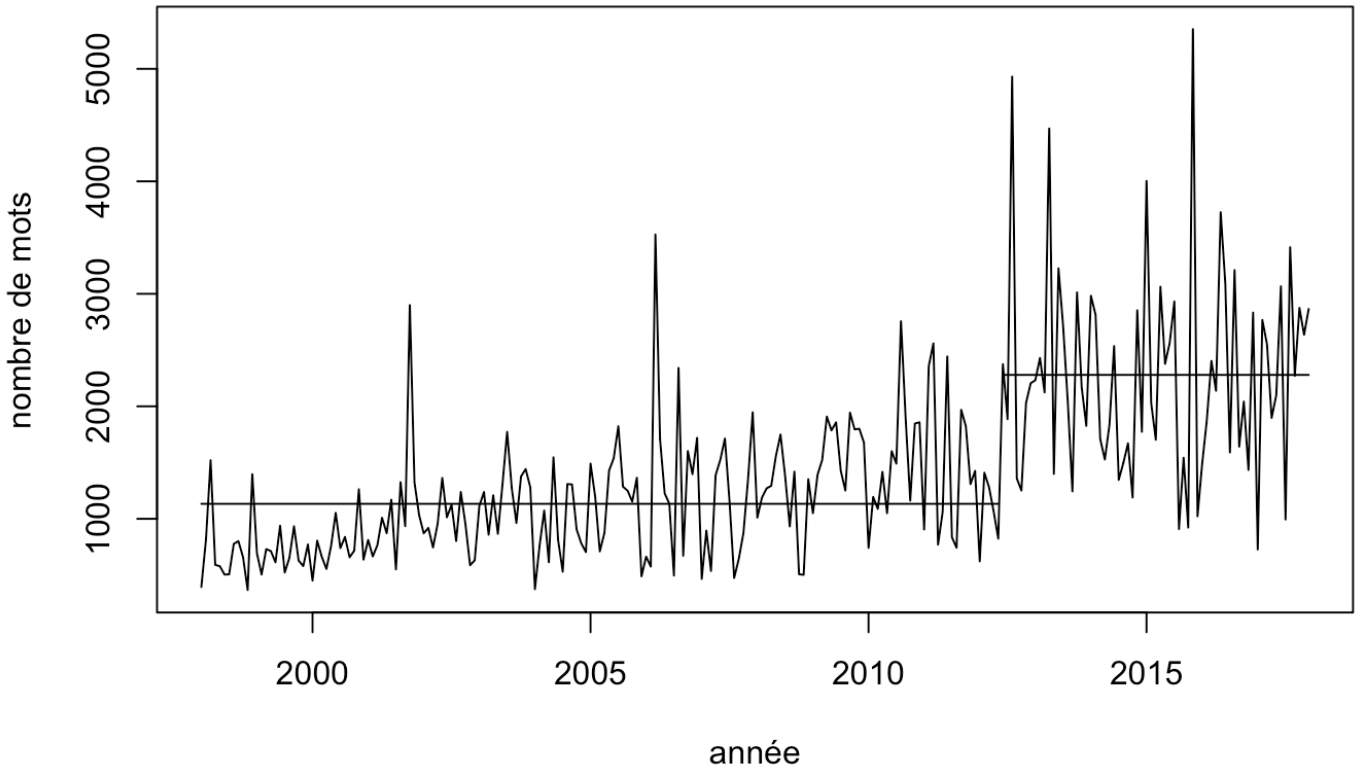
La Figure 4 montre le nombre de mots des REIR par ministère. Trois groupes de ministères se distinguent. En haut de la liste figurent les ministères de l'Environnement, des Pêches et Océans, de l'Immigration, du Transport et de la Santé. Notons qu'il s'agit de ministères qui produisent des règlements pour minimiser les risques encourus par des individus (une exposition à un produit toxique par exemple). Au bas de la liste figurent le Conseil du trésor, l'Office des transports du Canada, les ministères des Finances, des Affaires autochtones, de la Justice, des Affaires mondiales et du Revenu. Rappelons ici que la règlementation produite par ces ministères a souvent pour but de régir des relations entre les individus et les organisations (l'agence du revenu par exemple qui régit le versement des impôts des contribuables vers l'État). On peut donc supposer que les analyses d'impact sont plus volumineuses chez les ministères qui gèrent les risques que les autres.

Figure 4 : nombre total de mots des REIR par ministère, 1998-2017



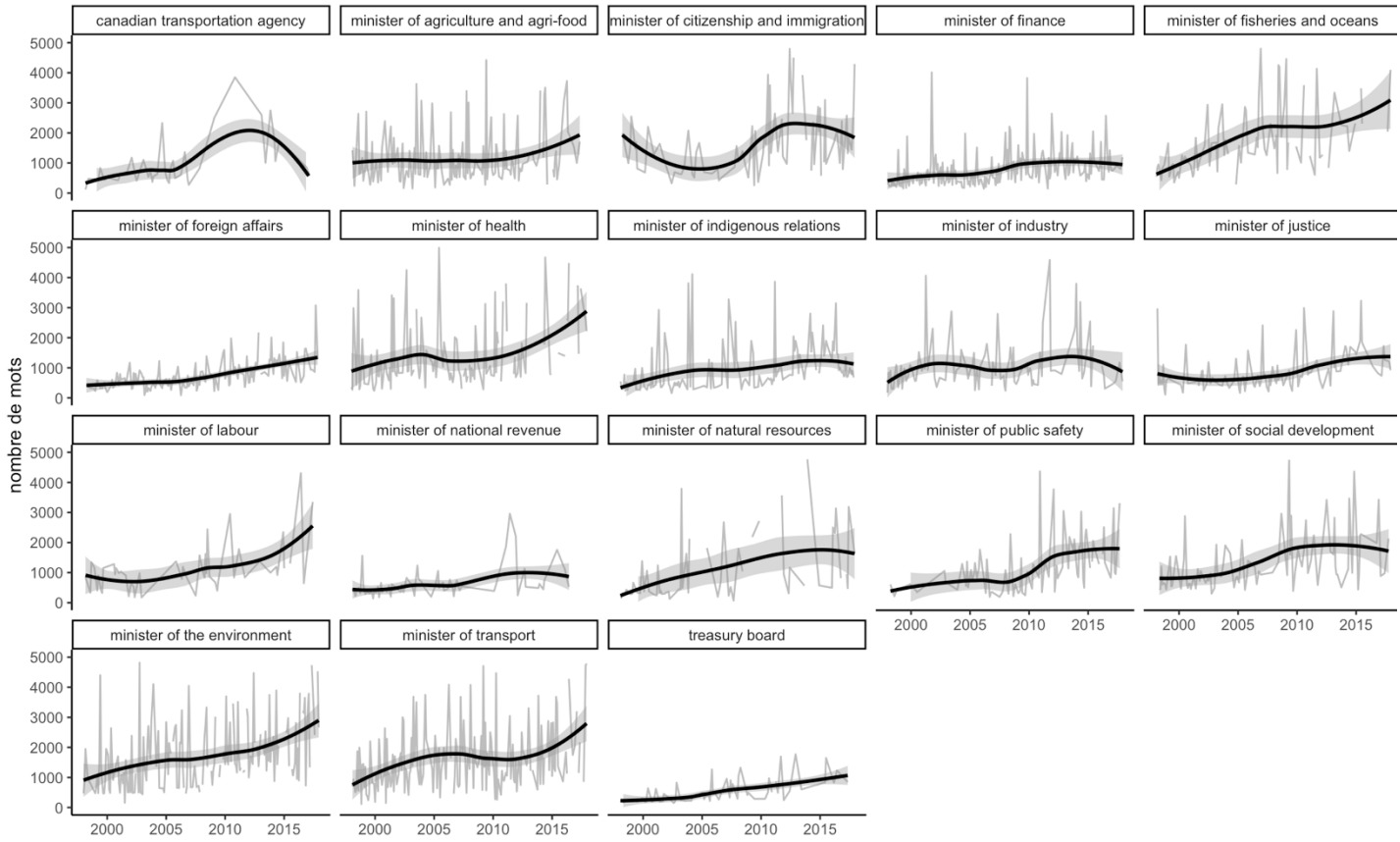
La figure 5 montre l'évolution mensuelle du nombre de mots des REIR produits entre 1998 et 2017. On constate une augmentation significative du nombre moyen de mots des REIR à partir du mois de septembre 2012. Comme pour le volume de l'activité règlementaire, nous soupçonnons que le changement soit le résultat des réformes administratives visant la règlementation fédérale canadienne. Bien que les réformes visent une rationalisation, elles promeuvent également la production de règlements sur la base de données probantes, lesquelles exigent la rédaction et présentation d'analyses couts-bénéfices plus étayées. À la fin des années 2000, les ressources du Centre d'expertise règlementaire du Secrétariat du Conseil du Trésor ont d'ailleurs été augmentées pour mieux accompagner les ministères et agences dans la production de ces analyses (Canada 2011). Aussi, la *Directive du Cabinet sur la gestion de la règlementation* de 2012 ajoute deux sections aux REIR, soit la section de la « règle du un pour un » et la section des « lentilles des petites entreprises » (Canada 2016a, Canada 2016b). Il est plausible que les nouvelles exigences en matière d'analyse couts-bénéfices et l'ajout de nouvelles sections aient contribué à l'augmentation du nombre de mots contenus dans les REIR.

Figure 5 : nombre mensuel de mots des REIR, 1998-2017



En distinguant cette évolution pour chaque ministère (figure 6), on constate que l'évolution du nombre de mots est plus ou moins généralisée. Hormis le ministère du Revenu, tous les ministères produisent aujourd'hui des REIR plus volumineux qu'en 1998.

Figure 6 : nombre mensuel de mots des REIR par ministère, 1998-2017



Analyse de spécificités lexicales

Pour compléter l'analyse quantitative des termes du corpus, nous avons procédé à une analyse de spécificités lexicales. Cette analyse poursuit deux objectifs. Elle nous informe d'abord sur la composition du corpus, et, ensuite, elle nous permet de comparer le vocabulaire des REIR en fonction de périodes distinctes. L'analyse de spécificité fait ressortir les collocations de mots (des mots qui cooccurrent l'un à la suite de l'autre) les plus et les moins spécifiques pour une période donnée, par rapport à l'ensemble des périodes précédentes. Elle peut ainsi nous informer sur l'évolution du vocabulaire utilisé dans les REIR.

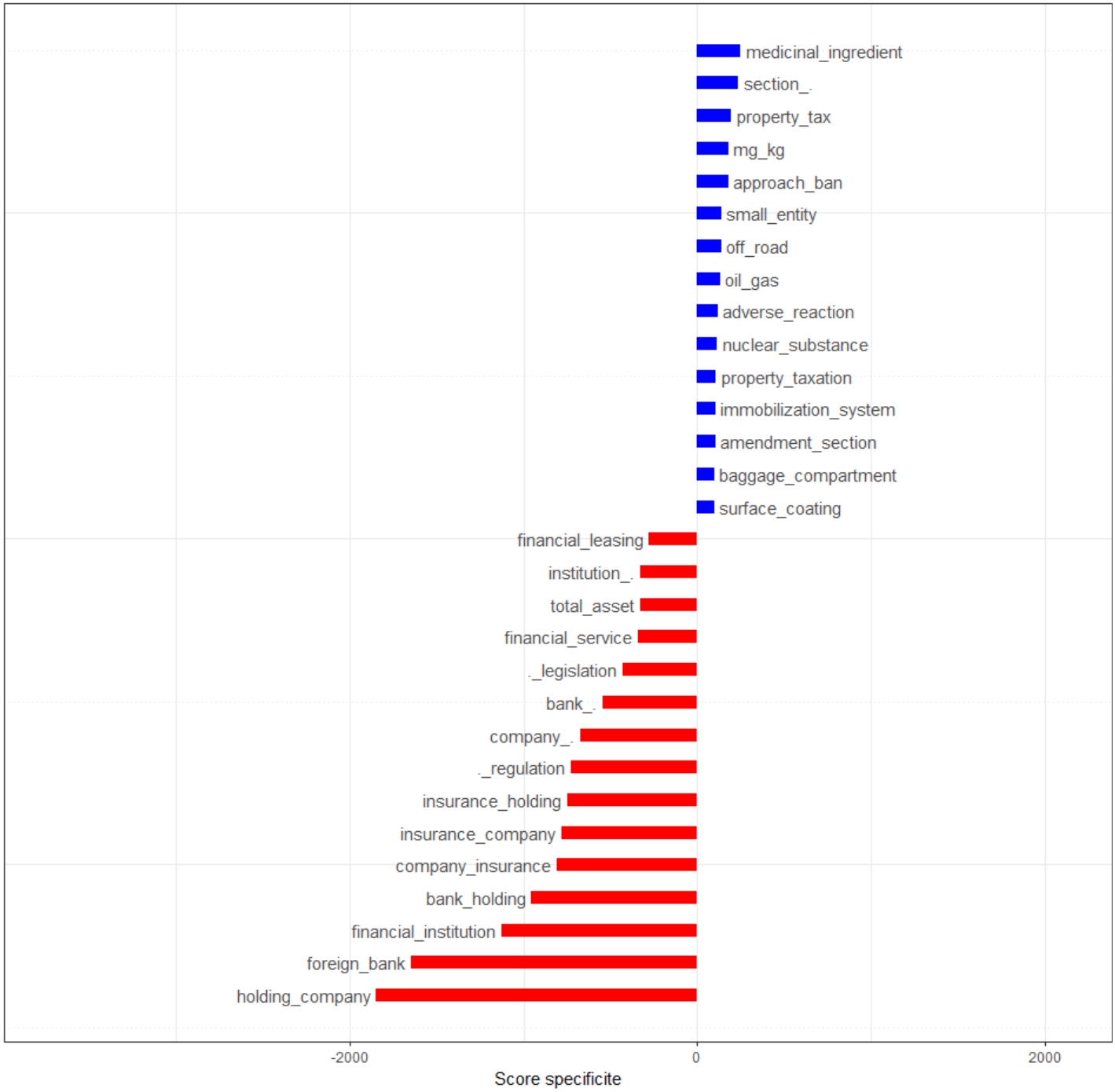
La figure 7 présente les résultats de l'analyse. Quatre périodes ont été choisies, soit de 1998 à 2002, de 2003 à 2007, de 2008 à 2012 et de 2013 à 2017. Ces périodes ont été définies de manière à ce qu'elles soient de durées comparables, tout en isolant les unes des autres les périodes antérieures et postérieures aux deux réformes administratives précédemment évoquées (respectivement mises en œuvre en 2007 et 2012). Les spécificités lexicales sont définies en fonction de la comparaison du vocabulaire de la période cible (en bleu) par rapport à ou aux périodes précédentes (en rouge). La période de 2003 à 2007 est comparée à celle de 1998 à 2002, celle de 2008 à 2012 est comparée aux

années allant de 1998 à 2007 et la période de 2013 à 2017 est comparée avec le vocabulaire des années 1998 à 2012. Ainsi, les collocations de mots en bleu, à l'intérieur de chaque période, peuvent être interprétées comme celles représentant une certaine nouveauté par rapport à la composition lexicale des années antérieures. Il est donc possible de mesurer l'évolution du vocabulaire de la règlementation à l'aide de ces figures. Une analyse équivalente, mais par ministère, est fournie en annexe.

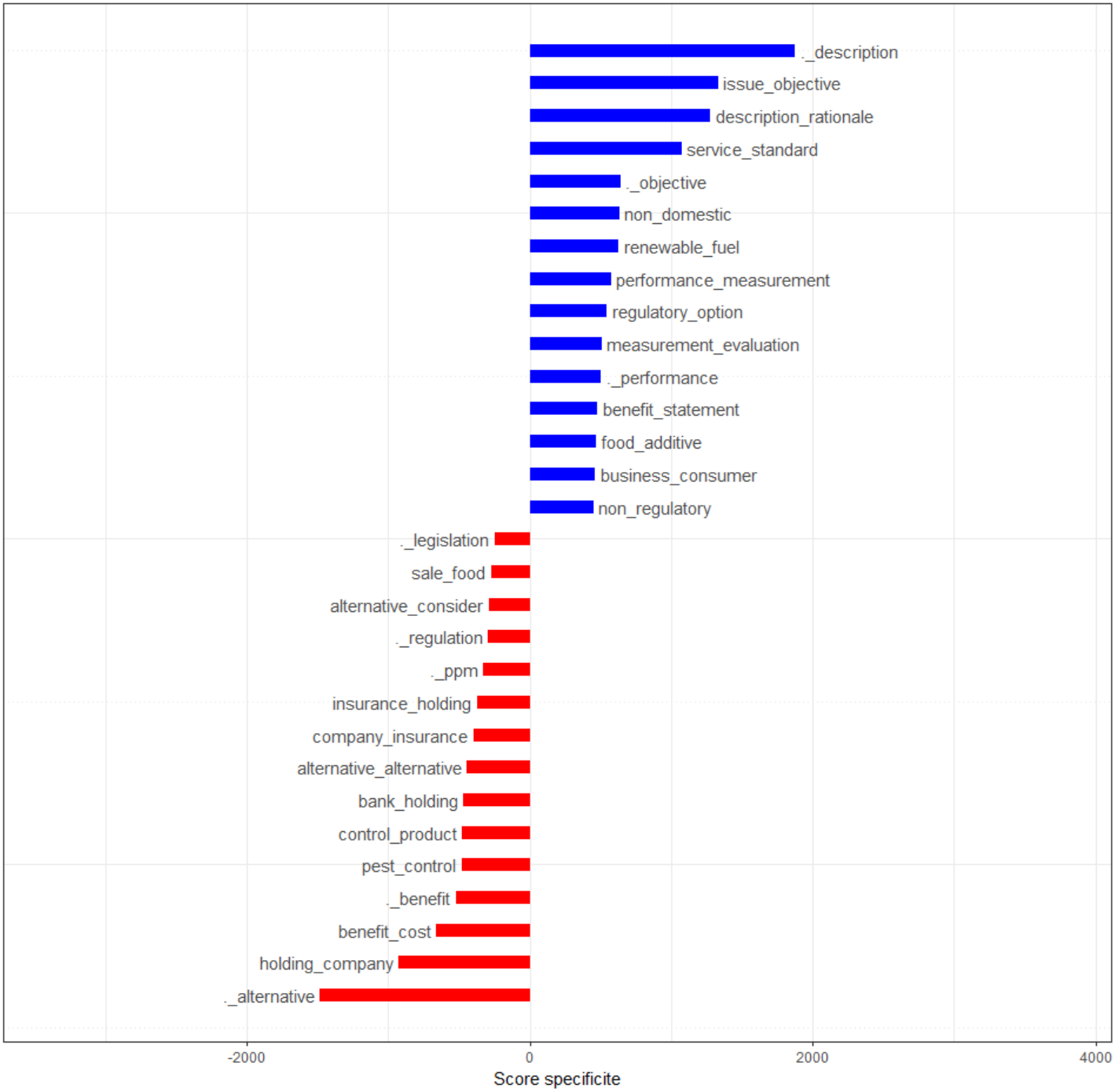
L'analyse de spécificité nous permet de faire les constats suivants : la règlementation antérieure à 2003 tournait autour du thème de la finance et des banques, alors que l'activité règlementaire pour les périodes allant de 2003 à 2007 et de 2013 à 2017 met respectivement l'accent sur les standards et les couts de la règlementation. Soulignons qu'il n'est pas étonnant que les collocations de mots *small_business* et *business_lens* soient spécifiques à la période 2013-2017. En effet, la *Directive du Cabinet sur la gestion de la règlementation* de 2012 ajoute une nouvelle section aux REIR nommée « lentilles des petites entreprises ». Au-delà de l'effet de la directive, l'analyse de spécificité montre une véritable préoccupation pour la lourdeur règlementaire lors de la période 2013-2017. En effet, cette période se distingue par la présence systématique des collocations de mots *administrative_cost*, *administrative_burden* et *cost_small* [entreprise]. Bien qu'on ne puisse conclure de cette analyse que la lourdeur règlementaire a effectivement augmenté ou diminué, elle démontre que le cout de la règlementation est devenu un enjeu de premier plan et spécifique à cette période.

Figure 7 : spécificités temporelles du corpus

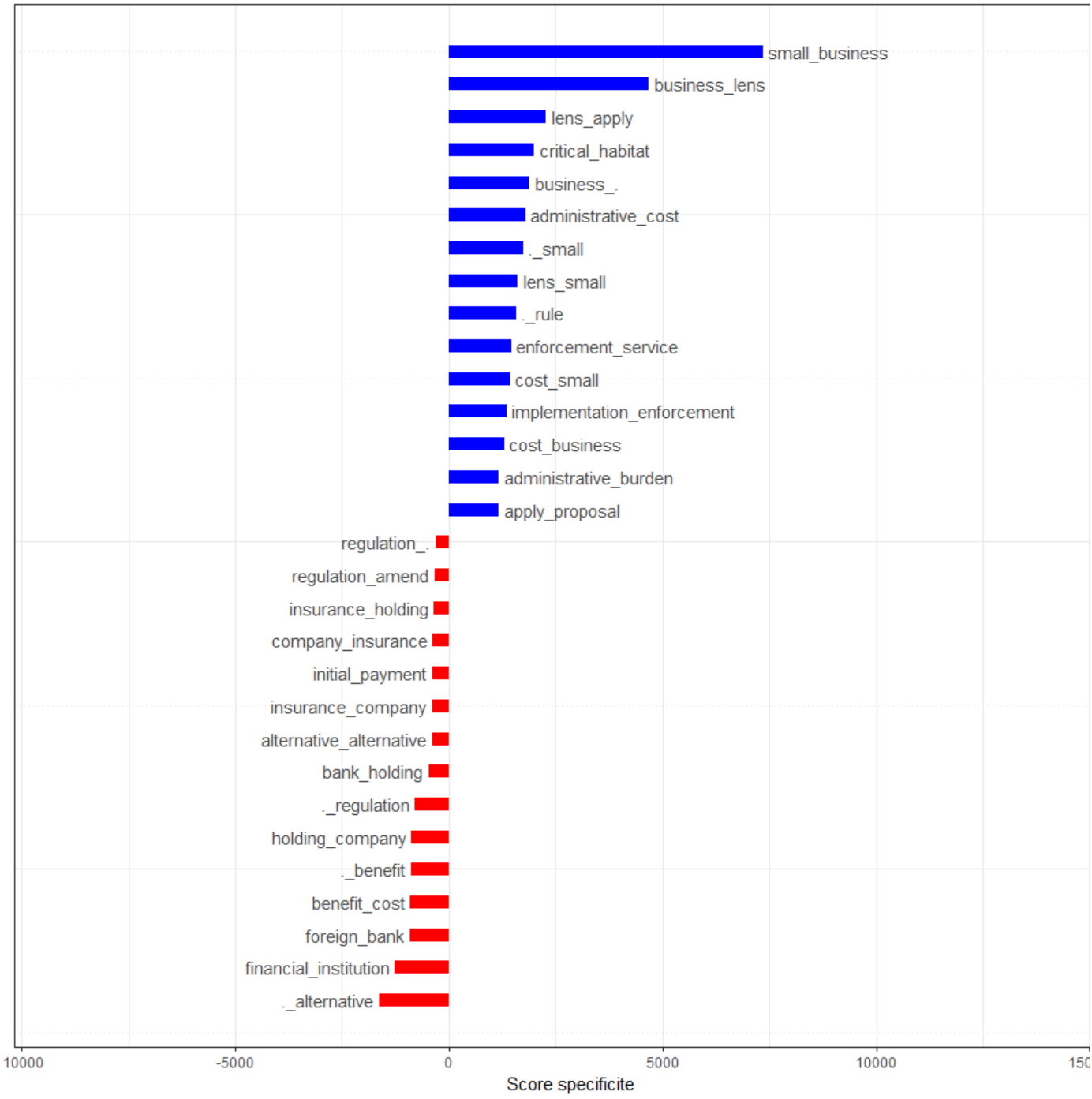
A. Spécificités des REIR de 2003 à 2007



B. Spécificités des REIR de 2008 à 2012



C. Spécificités des REIR de 2013 à 2017



Analyse de la diversité lexicale et de la lisibilité du corpus

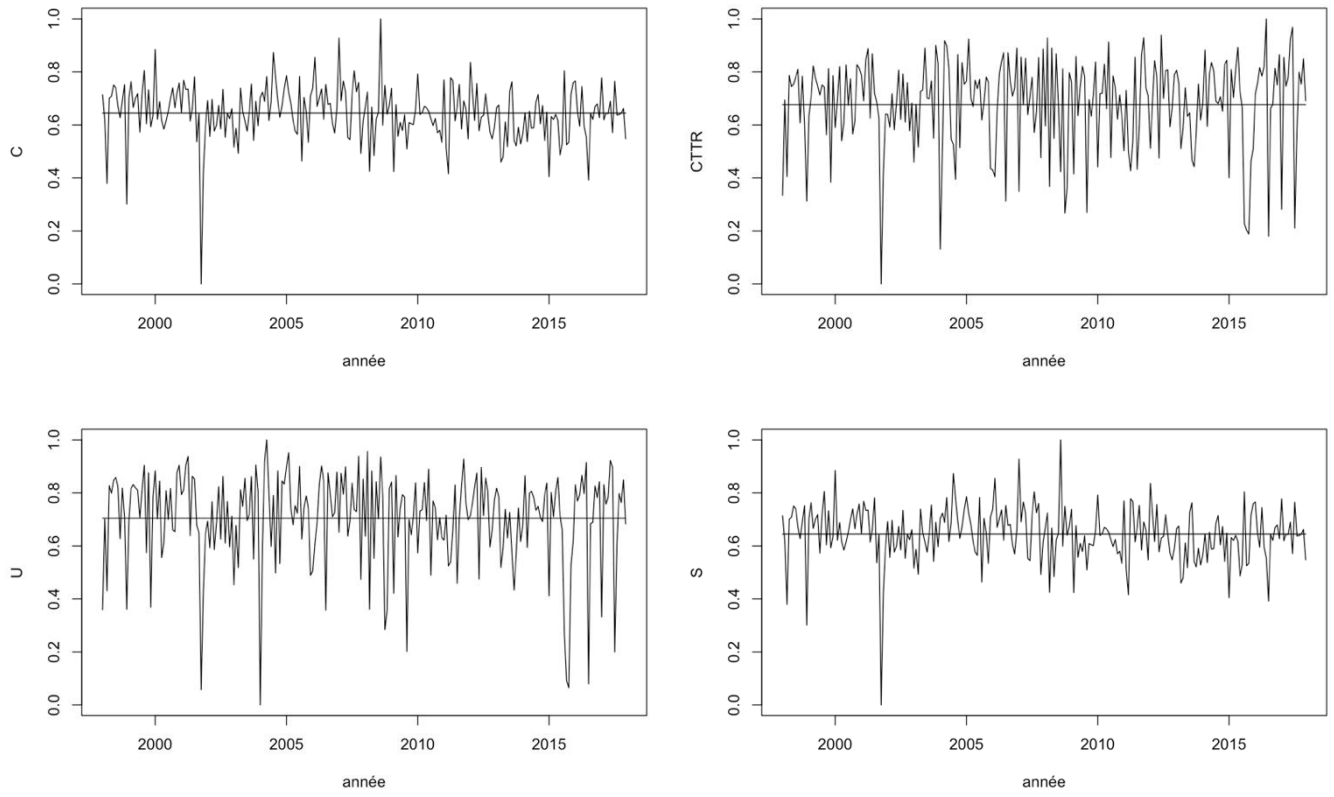
Jusqu'ici, nous avons montré qu'il existe des variations dans le temps et entre les ministères. Ces variations constituent une limite potentielle du corpus. Cependant, nous devons éviter de confondre les changements dans la façon de rédiger les REIR issus de directives administratives et les changements

dans les pratiques règlementaires que nous cherchons à observer. Pour vérifier l'évolution de la composition du texte comme tel, le nombre de mots contenus dans les REIR est non seulement insuffisant, il est trop sensible aux réformes de 2007 et de 2012. Nous mobilisons donc des indicateurs textuels complémentaires. La diversité lexicale et la lisibilité du texte des REIR figurent parmi ces indicateurs. Les indices de diversité lexicale utilisés sont des variantes du ratio de termes uniques par rapport au nombre total de mots, alors que les indices de lisibilité reposent sur la diversité du vocabulaire, le nombre de syllabes dans les mots utilisés et la longueur des phrases (Michalke et al. 2017). Pour plus de détails sur les indicateurs utilisés, on peut consulter le rapport méthodologique.

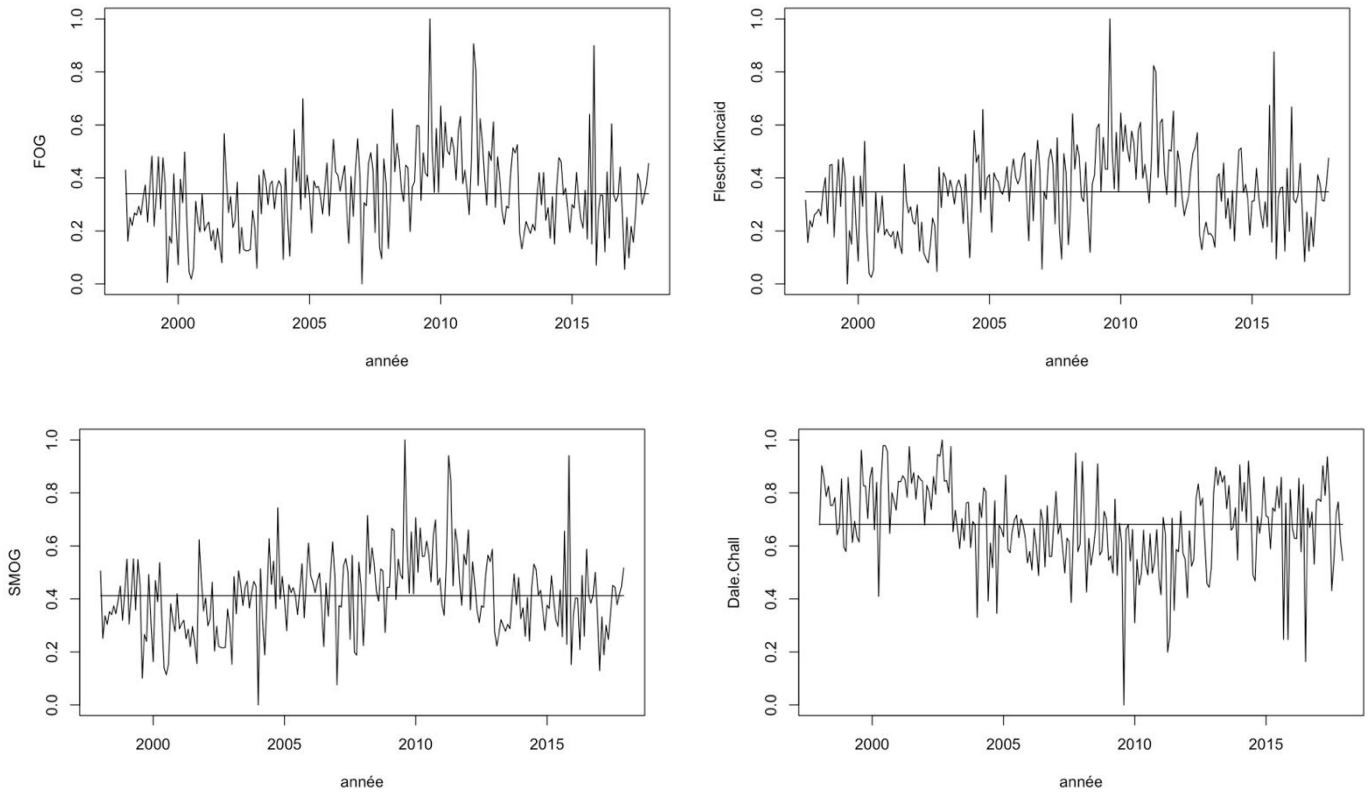
La figure 8 présente la série chronologique de huit indicateurs, quatre indicateurs de diversité lexicale et quatre indicateurs de lisibilité. Elle montre que, selon l'analyse de points de rupture, il n'y a eu aucun changement significatif de la moyenne pour les huit indicateurs. Autrement dit, les réformes administratives de 2007 et de 2012 n'ont eu aucun effet sur la diversité lexicale et la lisibilité des REIR. Cela nous rassure quant à la validité des REIR pour mesurer l'activité règlementaire sur l'ensemble de la période 1998-2017.

Figure 8 : diversité lexicale et lisibilité du corpus, 1998 et 2017

A. Diversité lexicale (indices C, CTTR, U et S)



B. Lisibilité (indices FOG, Flesch.Kincaid, SMOG et Dale.Chag)



En résumé

Dans cette section, nous avons voulu montrer la validité de notre corpus en analysant son évolution temporelle et les différences par ministère sur la base de plusieurs mesures. Bien que certains changements dans le nombre ou l'ampleur des REIR produits puissent être attribués aux réformes administratives de 2007 et de 2012, nous observons néanmoins une stabilité dans la diversité lexicale et la lisibilité du corpus. Grâce à l'analyse de spécificités, nous avons noté une tendance spécifique à la période 2013-2017, à savoir une attention portée à la question des couts de la règlementation. Au-delà de l'impact immédiat des réformes (l'ajout de nouvelles sections dans le REIR), nous croyons que cette nouvelle préoccupation n'est pas uniquement un artéfact des changements dans la façon de rédiger les REIR. Des entrevues réalisées avec des fonctionnaires fédéraux au printemps 2018 confirment son authenticité. Comment cette préoccupation s’est-elle traduite en termes de contrainte dans l’activité règlementaire ? La prochaine section se penche sur cette question.

4. Agenda et contrainte règlementaire

Les analyses qui suivent visent à vérifier dans quelle mesure les différents thèmes du corpus sont associés, à divers degrés, à un vocabulaire indiquant la présence de contraintes règlementaires. L'intérêt de circonscrire des thèmes plutôt que de baser notre analyse sur les ministères est de mesurer avec plus de précision l'agenda règlementaire. Un thème, les tarifs douaniers par exemple, peut être abordé par plusieurs ministères et agences, incluant les Affaires mondiales, Industrie Canada et l'Agence des services frontaliers. Examiner l'activité règlementaire du ministère le plus concerné par une thématique ne saurait donc suffire ; l'activité règlementaire de tous les ministères concernés doit être considérée pour bien rendre compte de l'agenda règlementaire.

On peut concevoir que l'activité règlementaire, selon sa thématique, peut engendrer différents niveaux de contrainte. Il est en effet commun de voir la règlementation comme une limite à la liberté des individus ou des entreprises. Si cette perception est fidèle à la réalité sur plusieurs thématiques, l'activité règlementaire peut également avoir pour but d'autoriser des activités ou même d'alléger des processus administratifs. Les règlements sur les produits toxiques, par exemple, contraignent souvent leur usage, comparativement aux règlements sur les semences, qui prévoient un mécanisme pour autoriser leur mise en marché. Bien entendu, on retrouve de la contrainte dans les deux types de règlements. Cependant, on peut s'attendre à un degré significativement moindre de contrainte si la thématique du règlement concerne la distribution des semences plutôt que les produits toxiques.

Pour identifier les thèmes plus susceptibles de produire de la contrainte règlementaire, nous présentons les résultats de trois analyses. D'abord, nous présentons les résultats d'une analyse thématique automatique du corpus. Cette analyse génère un modèle de 14 thèmes distincts qui permet d'observer avec un degré de finesse considérable l'agenda règlementaire canadien. Ensuite, nous utilisons un dictionnaire de mots pour mesurer le niveau de contrainte règlementaire associé à l'activité règlementaire entre 1998 et 2017 autour des thèmes générés par l'analyse thématique. Enfin, nous procédons à une analyse de correspondances dans le but de distinguer les différentes dimensions de la contrainte règlementaire, lesquelles sont confirmées par une analyse de *cluster*.

Analyse thématique du corpus

L'analyse thématique automatique est une technique d'apprentissage machine non supervisée qui vise à générer un modèle de thèmes constitués d'ensembles de mots reliés entre eux dans un corpus (Roberts et al. 2014). Chaque mot à une probabilité d'appartenir à un thème et chaque document du corpus, ici chaque REIR, peut être composé de plusieurs thèmes à la fois. Le modèle que nous avons généré comprend 14 thèmes distincts. On peut consulter la procédure détaillée de la création du modèle dans le rapport méthodologique.

Les figures 9 et 10 présentent les 14 thèmes du modèle ordonnés en fonction de leur proportion respective dans le modèle. Chaque thème est accompagné des huit mots les plus probables (figure 10)

et les plus exclusifs (figure 11). Les mots les plus probables sont ceux qui apparaissent le plus souvent dans le thème. Comme il s'agit fréquemment des mots courants, il est nécessaire d'examiner aussi l'exclusivité des mots, c'est-à-dire ces qui distinguent le mieux les thèmes les uns des autres. Par exemple, le mot *development*, un mot commun, apparait au thème 13 de la figure 10, mais disparait de la figure 11. Le mot *indian*, plus exclusif, est présent dans les deux figures. Notons qu'en plus de produire des listes de mots probables et exclusifs, nous pouvons isoler les REIR où la proportion de chaque thème est la plus forte. Afin de définir les thèmes, nous avons regroupé les dix REIR les plus associés à chacun des thèmes, dégagé les principales dimensions de chacun d'entre eux, puis croisé ces dimensions avec d'autres variables, telles que le ministère et l'année du règlement. Cette contextualisation de chaque thème, ainsi que la lecture attentive d'un échantillon de textes, nous a permis de définir les thématiques avec précision. La figure 11 présente les 14 thèmes ainsi définis selon leur proportion respective dans le modèle.

Figure 9 : les mots les plus probables du modèle thématique

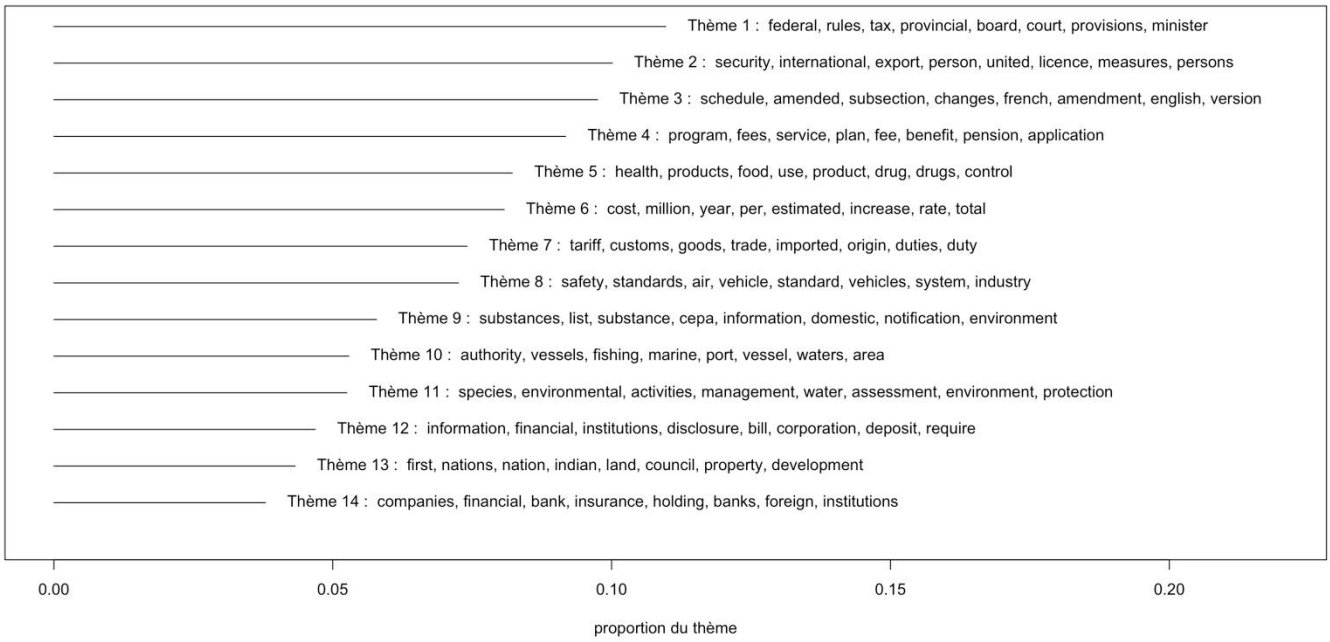
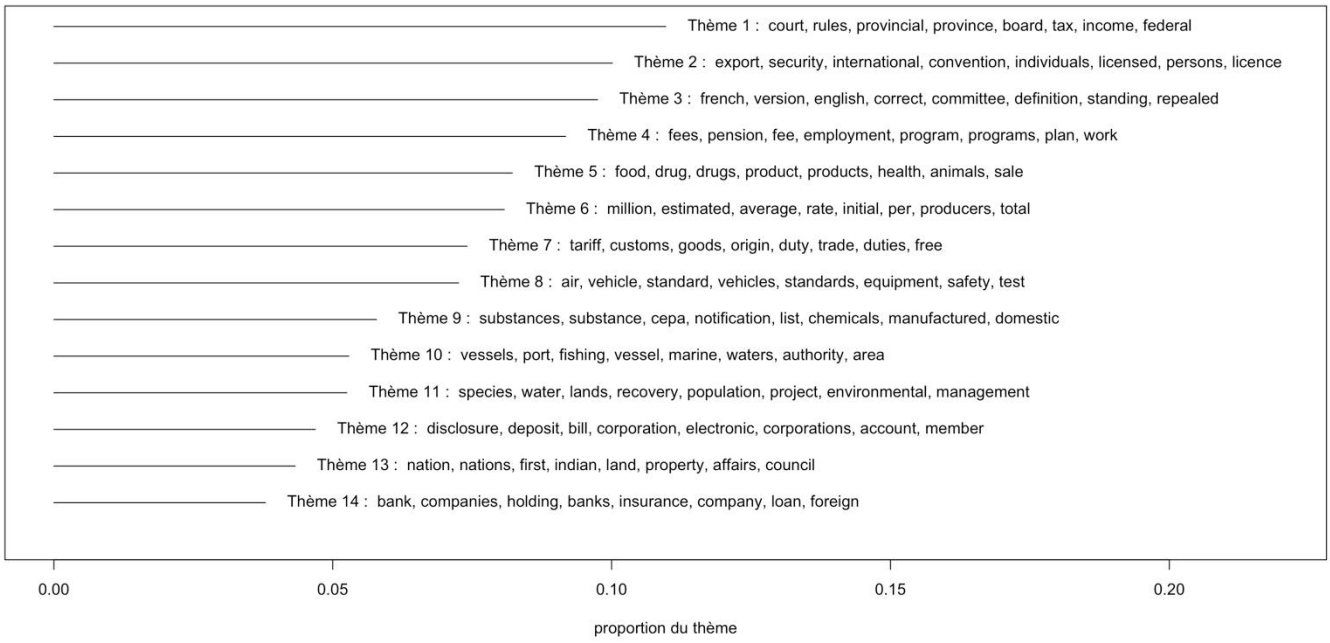


Figure 10 : les mots les plus exclusifs du modèle thématique

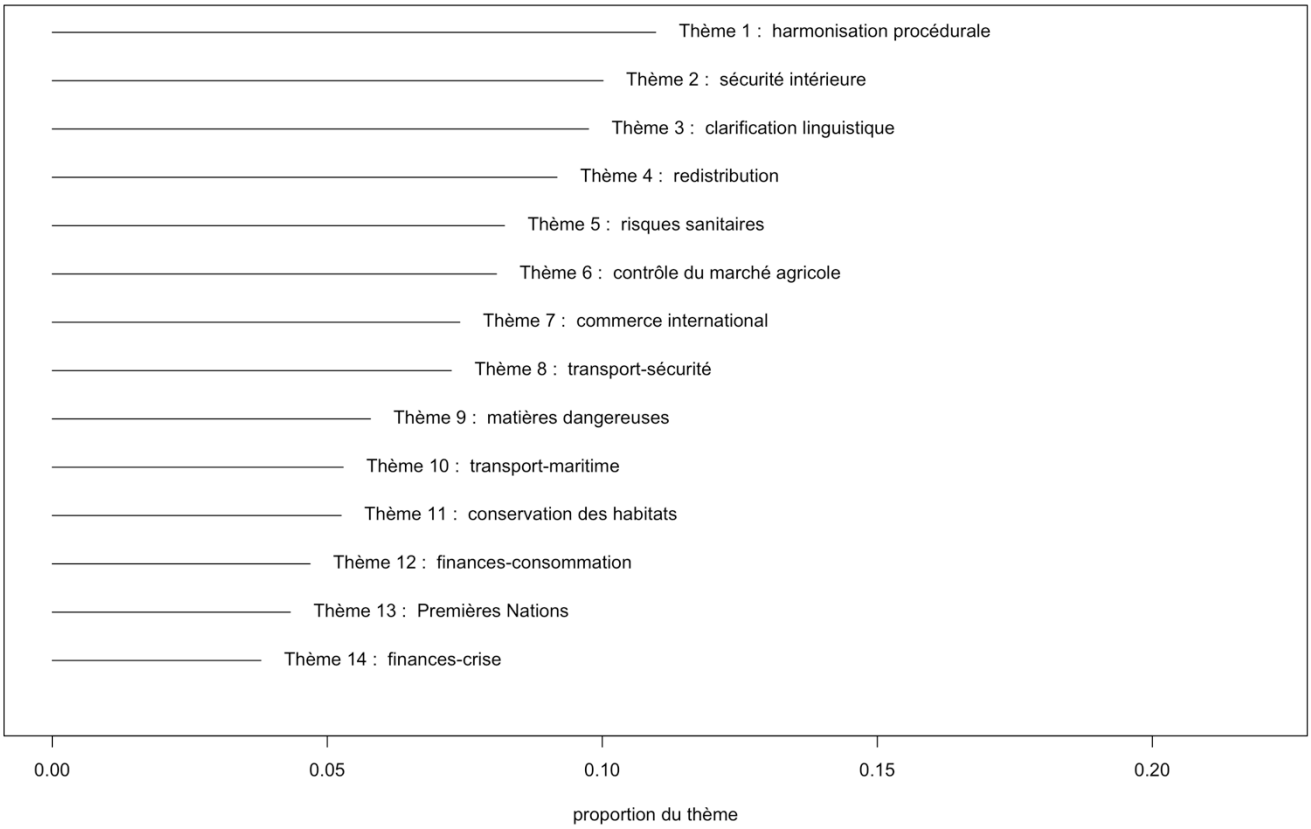


Le thème de l'harmonisation procédurale est le plus important du modèle. Les mots les plus probables de ce thème sont *federal*, *rules*, *tax*, *provincial*, *board*, *court*, *provision* et *minister* ; les mots les plus exclusifs sont *court*, *rules*, *provincial*, *province*, *board*, *tax*, *income* et *federal*. Plusieurs mots indiquent qu'il existe à l'intérieur de ce thème une dimension intergouvernementale. Effectivement, de nombreux REIR ont pour but d'harmoniser la règlementation des deux paliers de gouvernements. Des ententes entre le gouvernement fédéral et les gouvernements provinciaux ont en effet été conclues dans ce but, leur application exigeant des changements aux règlements. Les règlements adoptés en vertu de la Loi sur les contraventions et les ajustements aux transferts fédéraux (l'allocation canadienne pour enfants par exemple) fournissent des exemples. L'harmonisation, cependant, n'a pas toujours une dimension intergouvernementale. Bien souvent, elle reflète la volonté de résoudre des incongruités administratives qui ont été notamment identifiées dans les règles entourant le statut des réfugiés ou découlant du Code criminel, pour nommer deux exemples.

La sécurité intérieure est le deuxième thème et concerne des règlements du ministère des Affaires mondiales et, dans une moindre mesure, du ministère de la Justice. On y retrouve par exemple les règlements sur le port d'armes à feu, les interdictions de séjour pour des criminels internationaux et les règlements relatifs à la liste des marchandises et technologies d'exportations contrôlées (notamment les armes à feu). Le troisième thème concerne les clarifications linguistiques, qu'il s'agisse de révisions textuelles d'un règlement existant, de changements de terminologie, ou une précision concernant la traduction entre l'anglais et le français, ou vice-versa. Ces changements proviennent souvent de recommandations du Comité mixte permanent d'examen de la règlementation. Le quatrième thème est celui de la redistribution. On parle ici des règlements d'application des programmes sociaux, dont les

régimes de retraite, les prêts étudiants, l'assurance emploi, la péréquation, les compensations pour les employeurs de travailleurs étrangers et autres.

Figure 11 : les thèmes du modèle définis



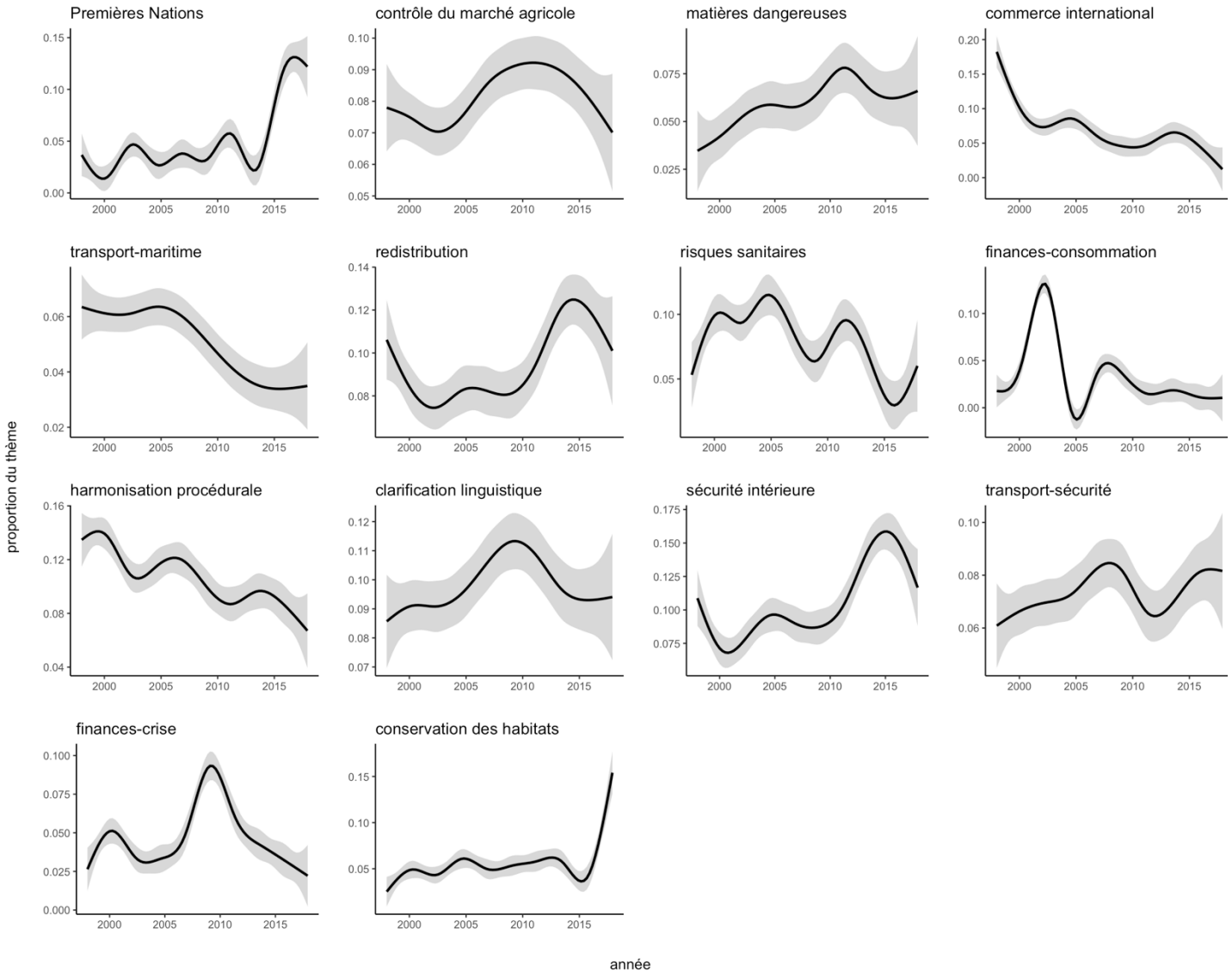
Les autres thèmes sont les risques sanitaires (thème 5), le contrôle du marché agricole (thème 6), le commerce international (thème 7), les standards de sécurité en matière de transport (thème 8), les matières dangereuses (thème 9), le transport-maritime (thème 10), la conservation des habitats (thème 11), la règlementation sur l'Agence de la consommation en matière financière du Canada (thème 12), les Premières Nations (thème 13) et la règlementation financière post-2007 (thème 14).

La figure 12 montre l'évolution des thèmes du modèle entre 1998 et 2017. Notons que l'échelle de l'axe diffère d'un thème à l'autre pour maximiser la visualisation de chacune des courbes. La figure 13 indique d'importantes variations de trajectoires d'un thème à l'autre. Le thème des Premières Nations, par exemple, est pratiquement absent de l'agenda règlementaire jusqu'en 2015, après quoi l'activité règlementaire augmente drastiquement. La promulgation de la Loi sur les élections au sein de Premières Nations explique sans aucun doute le niveau de l'activité règlementaire post-2015. Depuis 2000, le thème de la sécurité intérieure connait une hausse relativement constante, ce qui n'a rien d'étonnant ce thème étant devenu une préoccupation majeure de tous les gouvernements occidentaux depuis le 11 septembre 2001.

Plus étonnant, à première vue, est le déclin du thème de l'harmonisation procédurale. Cette tendance s'explique par le niveau très élevé d'activité règlementaire à la fin des années 1990, période à laquelle le gouvernement revoie le traitement des infractions de manière à distinguer les infractions criminelles exigeant une procédure sommaire de déclaration de culpabilité devant un tribunal (la conduite en état d'ébriété) des infractions moins graves qui peuvent être l'objet d'un envoi de contravention (pêcher sans permis). Chaque infraction classée dans cette seconde catégorie a fait l'objet d'un changement règlementaire distinct, d'où le volume du changement règlementaire sur le thème de l'harmonisation procédurale en début de période.

Le déclin du thème du commerce international s'explique de la même façon. De nombreux accords commerciaux internationaux ont été signés dans les années 1990 et leur mise en œuvre a nécessité des changements aux tarifs douaniers dans les années suivantes. Or, la réduction ou l'élimination d'un tarif douanier exige un changement règlementaire. Nous distinguons deux thèmes liés aux Finances, finances-consommation et finances-crise. Le premier fait référence à l'activité règlementaire touchant à la protection des consommateurs, un thème particulièrement important entre 2002 et 2005. À cette période, le gouvernement a modifié substantiellement sa règlementation pour que les institutions financières se conforment à de nouvelles exigences en matière de protection des consommateurs. Le second reflète plutôt la réaction gouvernementale à la crise financière de 2008.

Figure 12 : les thèmes du modèle dans le temps



La contrainte dans la règlementation fédérale canadienne

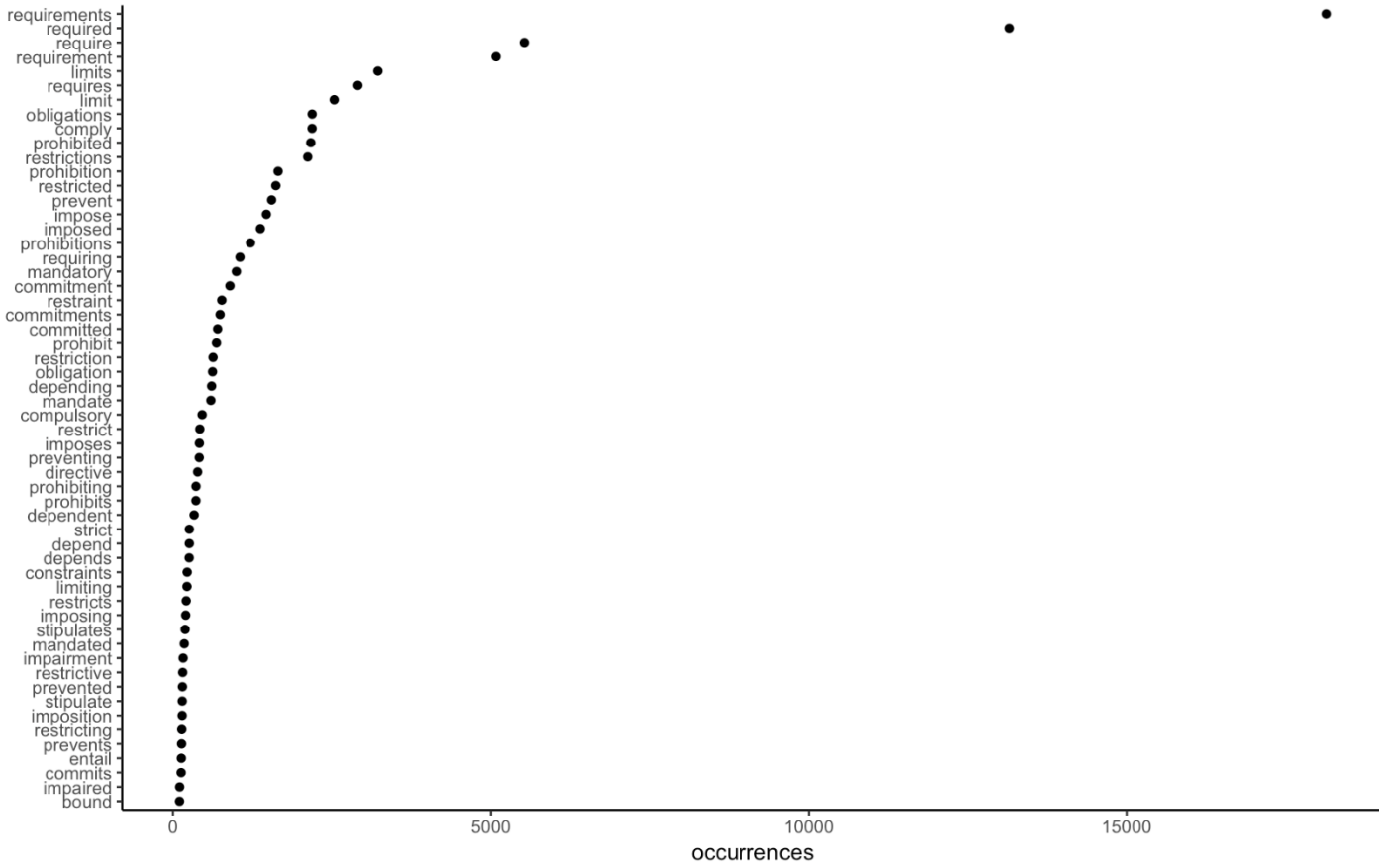
Dans cette section, nous cherchons à observer la contrainte dans l'activité règlementaire fédérale. Nous proposons un nouvel indicateur, bien qu'il en existe d'autres. L'un des projets les plus aboutis concernant la dimension contraignante de la règlementation est *RegData*, une banque de données qui quantifie le volume des restrictions règlementaires aux États-Unis (Al-Ubaydli et McLaughlin 2017). En plus de tenir compte du nombre de mots des règlements, les chercheurs de *RegData* mesurent le niveau de restriction règlementaire à partir d'un dictionnaire ou lexique de mots. Comptant cinq mots/expressions (*shall*, *must*, *may not*, *prohibited* et *required*), cet indicateur séduit par sa simplicité.

En 2012, le gouvernement du Canada a rendu public son Plan d'action pour la réduction du fardeau administratif. Depuis cette date, sous la supervision du Secrétariat du Conseil du Trésor, la plupart des ministères et agences responsables doivent dénombrer les exigences dans les règlements fédéraux qui imposent un fardeau administratif aux entreprises (Canada 2017). Ce dénombrement des exigences est similaire à celui utilisé par RegData. Il compte 16 expressions (*must*, *must not*, *shall*, *shall not*, *may*, *may not*, *is to be*, *is required*, *prohibited*, *is to have*, *is to ensure*, *is obliged*, *is under an obligation*, *is entitled to*, *has the right to*, *is required to*).

Bien qu'il s'agisse d'initiatives intéressantes, nous considérons ces deux indicateurs limités, surtout lorsqu'on les utilise sur un corpus plus riche lexicalement que le texte règlementaire lui-même. Aussi privilégions-nous une autre mesure. Nous utilisons un dictionnaire élaboré par Loughran et McDonald (2011 ; 2016) qui permet de mesurer la contrainte, et qui a surtout été appliqué dans l'étude du milieu financier. L'indice Loughran-McDonald propose six lexiques thématiques permettant de faire ressortir différents tons (« litigieux » ou « incertain » par exemple). Le lexique de l'un des tons, intitulé « contraignant », nous est tout particulièrement utile dans ce rapport. Comptant 177 mots, dont ceux de *RegData* et de l'indice du Secrétariat du Conseil du Trésor, ce lexique de la contrainte est particulièrement performant sur notre corpus de REIR. La figure 13 présente les mots de ce lexique les plus fréquemment utilisés (au moins 100 fois) dans notre corpus[2]. Pour nous assurer que le lexique ne mesure que la contrainte, nous n'avons pas pris en compte les mots précédés par des négations (*no*, *non*, *not*, *never*, *neither* ou *nor*).

[2] La liste de mots complète est consignée dans le Software Repository for Accounting and Finance de l'Université Notre-Dame (Software Repository for Accounting and Finance 2018).

Figure 13 : les mots associés au lexique « contraignant »



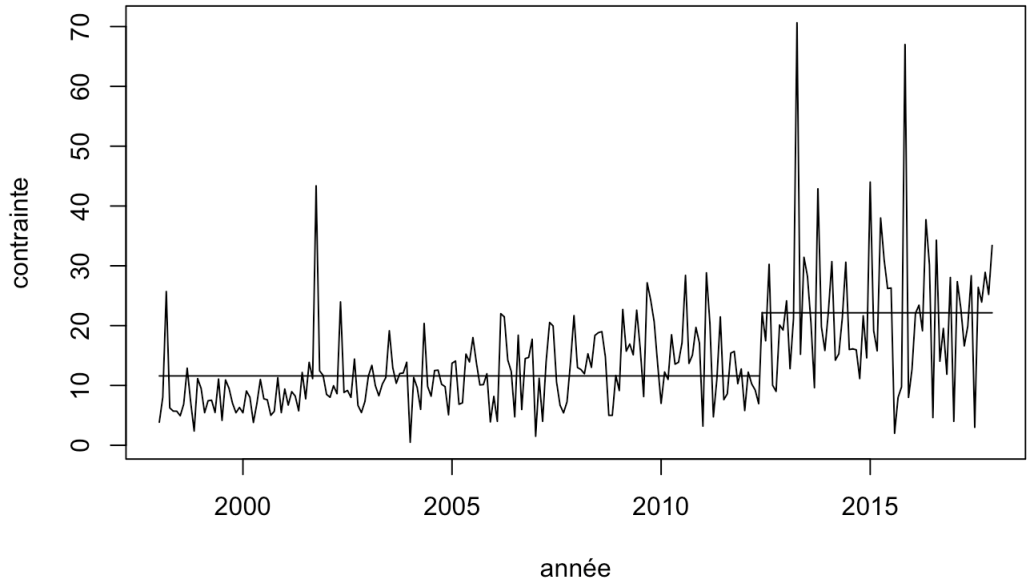
La figure 14 montre l'évolution temporelle de la contrainte règlementaire suivant deux mesures. La première est le nombre brut d'occurrences (14A). Sans surprise, on constate que le nombre d'occurrences augmente significativement à partir de 2012, année de la mise en oeuvre de la Directive sur la gestion de la règlementation, qui ajoutait aux REIR deux nouvelles sections (la règle du pour un et Lentilles des petites entreprises). La deuxième mesure divise les occurrences brutes par le nombre total de mots contenu dans chaque REIR, ce qui permet de contrôler l'effet de la réforme sur le nombre de mots contenus dans les REIR (14B). À partir de cet indicateur, nous concluons que la présence de la contrainte dans l'activité règlementaire est demeurée stable durant la période à l'étude. En effet, les deux séries chronologiques ont fait l'objet d'une analyse de points de rupture selon un test statistique. L'absence de rupture dans la ligne horizontale indique un niveau moyen constant de contrainte règlementaire.

Il faut comprendre que cette stabilité dans la présence de la contrainte dans l'activité règlementaire ne signifie pas que le gouvernement a évité d'ajouter de nouvelles contraintes au fil du temps. Cela ne signifie pas non plus que le gouvernement n'a pas, par son activité règlementaire, retiré des contraintes. En fait, si le lexique utilisé nous permet de mesurer la contrainte, il ne nous permet pas de mesurer ce que l'on pourrait appeler la dérèglementation, soit la diminution de la contrainte dans les règlements existants, par adoucissement des restrictions ou abrogation du règlement. Un lexique

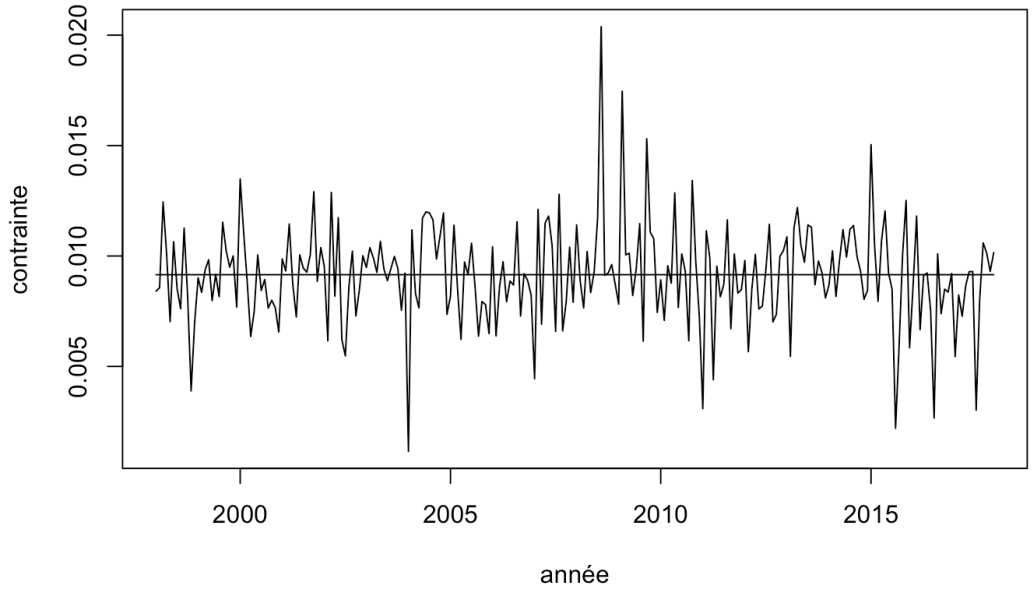
entièrement différent de celui que nous utilisons ici, et que nous développerons sous peu, serait toutefois nécessaire pour mesurer plus finement la dérèglementation. En d'autres termes, si la figure 14B indique que la contrainte produite par l'activité règlementaire fédérale de 1998 à 2017 est demeurée stable, nous ignorons si la contrainte accumulée durant cette période a été compensée par un effort de dérèglementation.

Figure 14 : contrainte règlementaire fédérale canadienne, 1998-2017

A. Occurrences

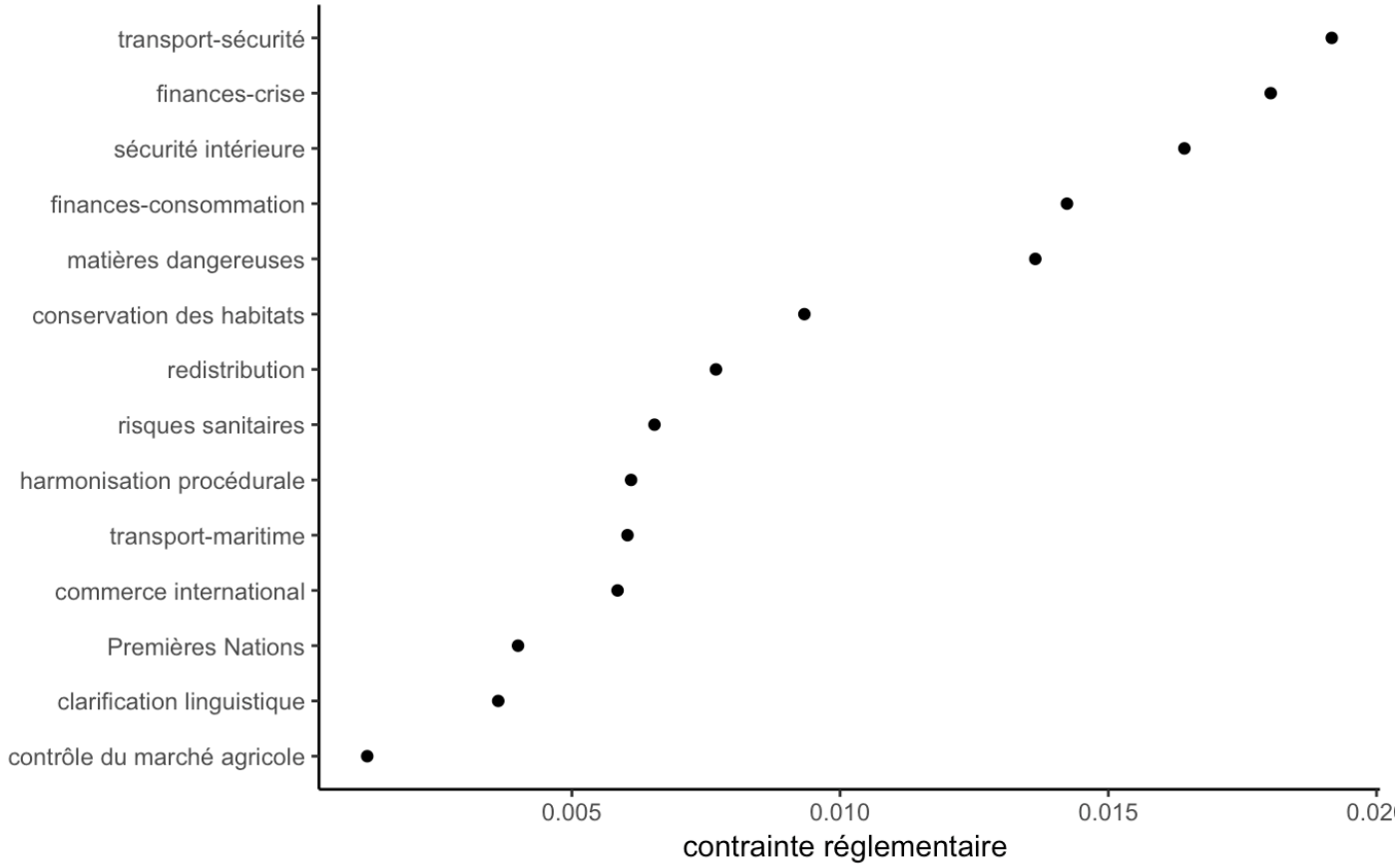


B. Occurrences/nombre de mots



Dans ce rapport, plutôt que d'examiner le ratio ajout/réduction de contrainte, nous cherchons à comprendre comment la contrainte se manifeste en fonction de thématiques précises. Nous cherchons notamment à voir si la contrainte, stable lorsque l'on examine le corpus en son entièreté, gagne ou perd de l'importance d'un thème à l'autre. Nous avons donc mesuré la contrainte pour chaque thème en incluant uniquement les REIR dans lesquels un thème est dominant à au moins 75%. Si cette opération réduit la taille du corpus (N=1537), elle permet une comparaison thématique plus nette. La Figure 15 montre la contrainte moyenne, toujours calculée à partir de l'indice Loughran-McDonald, exprimée par thème. Sur l'ensemble de la période, l'activité règlementaire des thèmes de la sécurité en matière de transport, de la règlementation financière, de la sécurité intérieure et du contrôle des matières dangereuses a produit le plus de contraintes. À l'inverse, l'activité règlementaire sur le contrôle du marché agricole et la gouvernance autochtone et l'activité règlementaire qui amène des clarifications linguistiques produisent moins de contraintes que les autres.

Figure 15 : niveau de contrainte règlementaire par thème



Pour approfondir notre compréhension de la contrainte dans chacun des thèmes, nous avons procédé à une analyse de correspondances du sous-corpus de la contrainte. L'analyse de correspondances est une forme d'analyse factorielle qui permet de faire ressortir les dimensions d'une problématique, ici la contrainte, et de visualiser les regroupements de mots que l'on retrouve sur différentes dimensions. On peut supposer que la contrainte associée à l'activité règlementaire sur les matières dangereuses,

qui vise sans doute à protéger des individus contre des risques, est de nature différente de la contrainte provenant de l'activité règlementaire sur la redistribution, qui régit les relations entre organisations et/ou individus. En revanche, la contrainte associée à l'activité règlementaire sur les risques sanitaires, qui vraisemblablement vise à protéger des individus contre des risques, peut s'apparenter à celle portant sur les matières dangereuses. L'analyse de correspondance regroupe les mots et les thèmes du texte en fonction de leur association relative les uns avec les autres. C'est la visualisation de ces regroupements qui nous permet de mieux comprendre la nature de la contrainte associée à chaque thème.

Les figures 16 et 17 offrent de telles visualisations sur les quatre premières dimensions du sous-corpus de la contrainte, soit deux par figure. Notons que ces quatre dimensions cumulent plus de 64% de la variance expliquée par l'analyse de correspondances. La couleur des mots qui apparaissent sur la figure est attribuée en fonction de leur contribution à la définition des dimensions. Plus les mots sont près les uns des autres, plus ils sont associés entre eux. Plus les mots s'éloignent de l'origine, plus ils sont exclusifs au thème auquel ils sont associés. En ce qui a trait aux thèmes, plus ils s'éloignent de l'origine, plus ils se démarquent au plan lexical des autres thèmes. La relation entre deux thèmes ou un thème et un mot s'évalue en fonction de l'angle formé en suivant leur position par rapport à l'origine. Plus l'angle est aigu, plus l'association est forte. Un angle de 90 degrés indique l'absence d'une association alors qu'un angle de 180 degrés, une opposition de thèmes.

Les figures 16 et 17 font ressortir trois formes de contrainte règlementaire : le contrôle des organisations, la gestion du risque et l'interdiction. Ces trois formes de contrainte se manifestent dans les six grappes de mots des figures 16 et 17 (trois dans la figure 16 et trois dans la figure 17). Il est important de noter que ces grappes ne sont pas mutuellement exclusives, dans la mesure où il s'agit d'un objet multidimensionnel projeté sur deux graphiques qui comptent chacun deux dimensions.

La figure 16 montre que la grappe dans le 3e cadran du graphique rassemble des mots liés à la règlementation financière. Les mots *disclose*, *account*, *statement*, *consumer*, pour n'en citer que quelques-uns, indiquent une contrainte qui s'exprime par de plus grandes exigences en matière de transparence de la part des institutions financières. La grappe de mots dans le 4e cadran n'est statistiquement pas liée avec la première (donc distincte) et renvoie plutôt à la règlementation des matières dangereuses. Les mots *substance*, *domestic*, *manufacture* et *assessment* indiquent en effet la gestion du risque entourant les règlements sur la Liste intérieure des substances, une liste publiée par Environnement Canada des substances fabriquées, importées ou utilisées au Canada de façon commerciale. La troisième grappe de mots, dans le 1er cadran, est moins distinctive. Elle regroupe néanmoins des mots associés à l'interdiction, notamment en ce qui concerne les titulaires d'une autorisation d'acquisition d'armes à feu.

La figure 17 montre aussi trois grappes, qui font aussi ressortir trois formes de contrainte, mais à partir de différents mots. Le contrôle des organisations se trouve dans la grappe de mots du haut du graphique. Les mots *obligation*, *tariff*, *preferential* montrent entre autres que, par la contrainte, le

gouvernement tente de contrôler des marchés par des tarifs préférentiels sur certains produits (cadran 1). La gestion du risque se retrouve dans la grappe de mots qui gravitent autour du thème de la sécurité dans le transport (cadran 4). Les mots *child*, *safety*, *manufacturer* et *vehicle* indiquent que les standards de sécurité des véhicules, notamment en ce qui concerne les sièges d'auto pour enfants, contribuent fortement à cette facette. Les thèmes de la sécurité intérieure et des matières dangereuses sont liés à la forme de contrainte la plus restrictive, soit l'interdiction (cadran 3). Soulignons qu'en incorporant les dimensions 3 et 4, le thème des matières dangereuses, associé à l'imposition de normes et de standards dans la figure 16, est également associé à des interdictions.

Afin de confirmer notre interprétation des différentes formes de la contrainte règlementaire, nous avons effectué une analyse de *clusters* (analyse d'apprentissage machine non paramétrique et non supervisée) sur les résultats de l'analyse de correspondance. L'analyse confirme nos précédents constats. Premièrement, les cinq *clusters* produits par l'analyse correspondent aux grappes identifiées dans l'analyse de correspondances. Tous peuvent être rattachés à l'une ou l'autre des formes de contraintes que nous avons identifiées. Deuxièmement, les mots les plus significatifs de ces *clusters* nous confirment que nous avons bien compris la nature des types de contraintes. Les résultats détaillés de cette analyse et leur interprétation sont disponibles dans le rapport méthodologique.

Figure 16 : analyse de correspondances (dimensions 1 et 2)

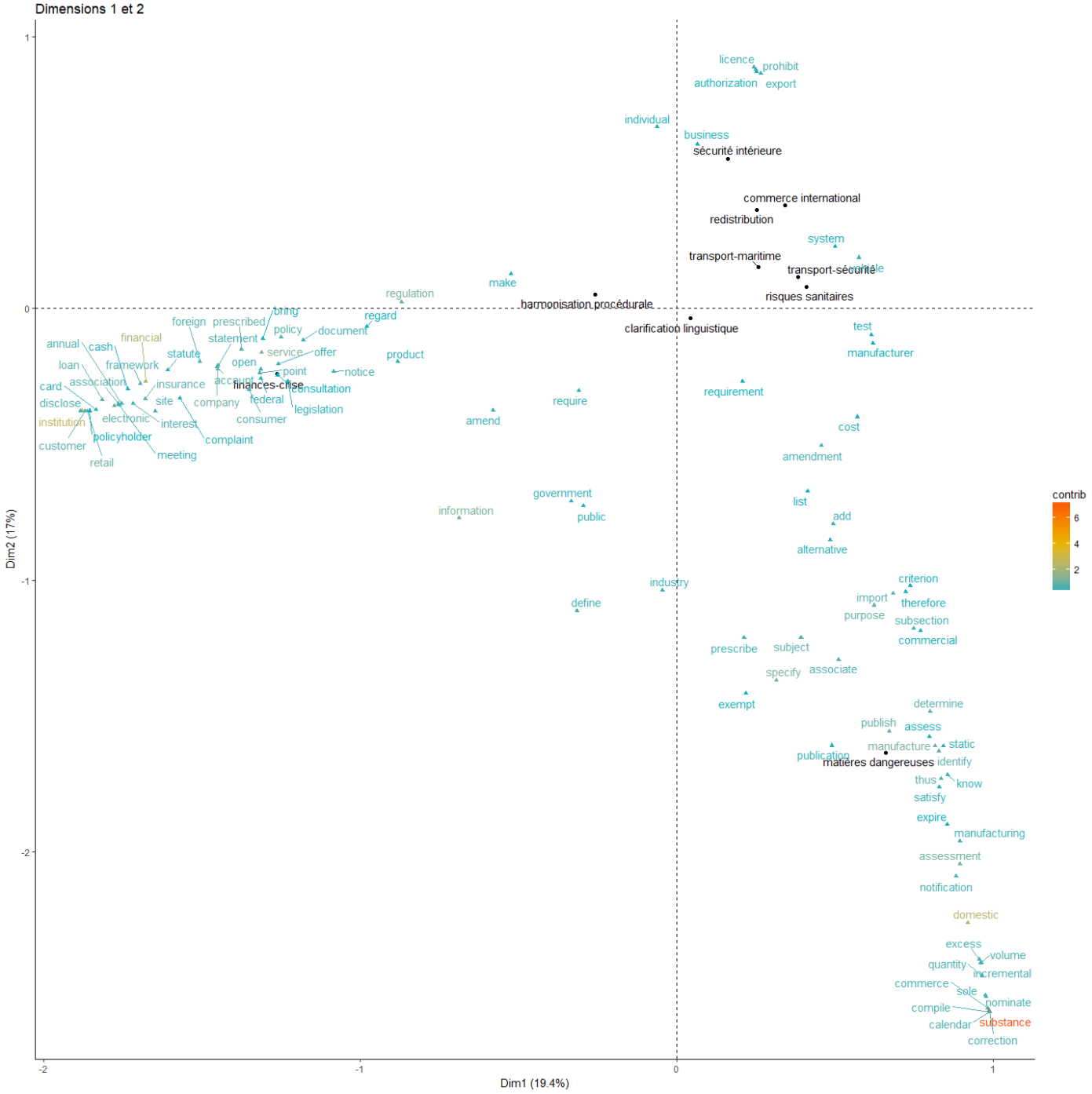
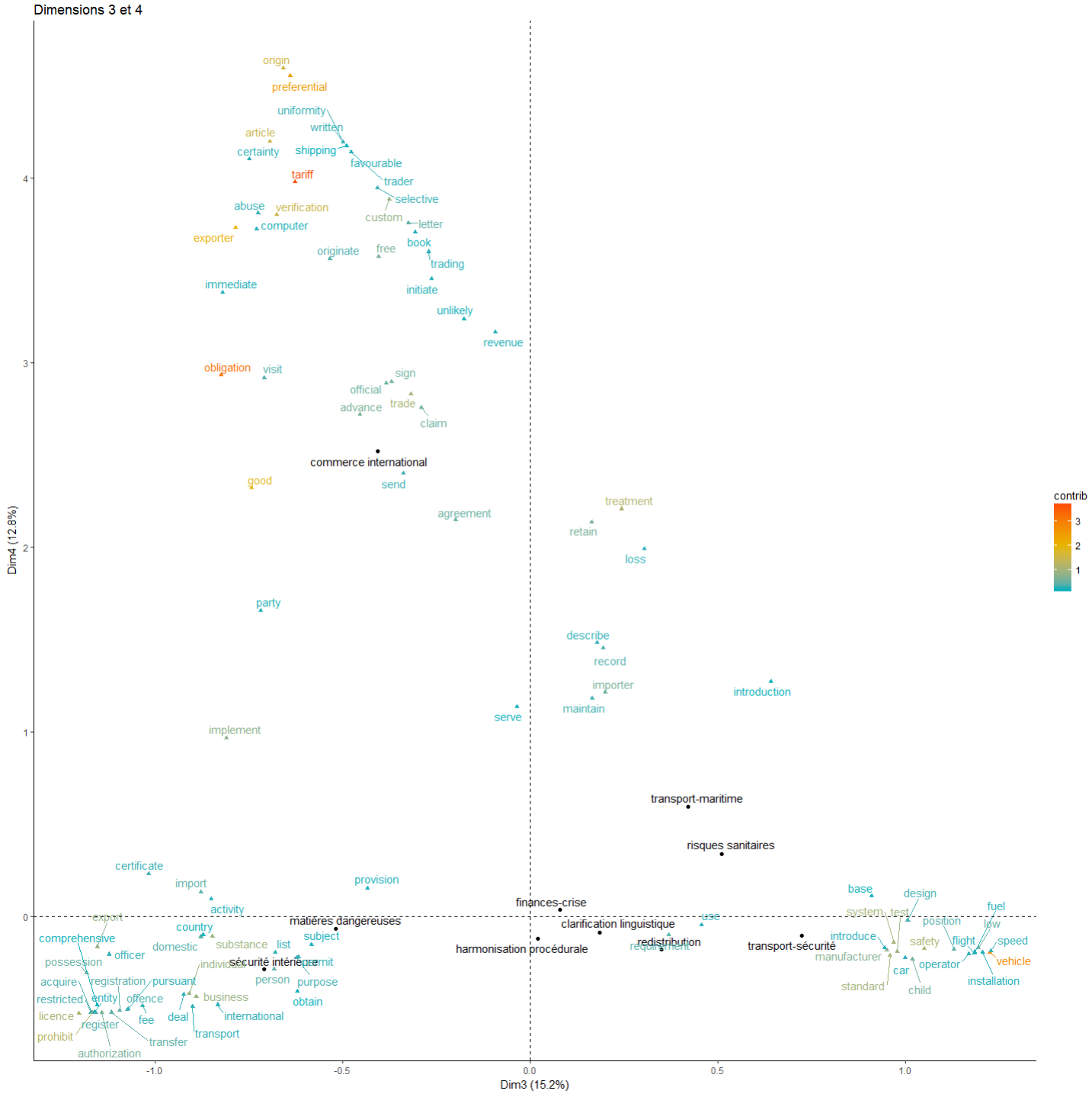


Figure 17 : analyse de correspondances (dimensions 3 et 4)



Conclusion et analyses futures

Dans ce rapport, nous avons montré les avantages d'utiliser les REIR plutôt que le texte de la règlementation pour analyser l'évolution de l'activité règlementaire canadienne au niveau fédéral. Si nos analyses indiquent une plus grande préoccupation pour la lourdeur règlementaire au cours des dernières années, des analyses poussées sont essentielles pour mieux comprendre ce que signifie cette préoccupation pour l'activité règlementaire. Ces analyses montrent que, globalement, la contrainte qui s'exprime dans l'activité règlementaire de 1998 à 2017 est demeurée stable. Cela ne signifie évidemment pas que le gouvernement a évité de mettre en place de nouvelles contraintes. Des analyses supplémentaires sont nécessaires pour mesurer de manière exacte l'accumulation de la contrainte règlementaire en tenant compte des règlements qui ont été abrogés et ceux visant à alléger la règlementation[3].

Nous avons par ailleurs montré qu'il existe de fortes variations d'une thématique à l'autre quant au niveau de contrainte induit par l'activité règlementaire. Si sur des thèmes comme les Premières Nations l'activité règlementaire exprime un faible niveau de contrainte, il en va autrement pour des thèmes comme la sécurité dans le transport. De manière encore plus intéressante, notre analyse de correspondances, confirmée par l'analyse de *clusters*, nous permet de conclure qu'il existe des différences de formes importantes dans la façon dont se manifeste la contrainte. En effet, la contrainte vise parfois à gérer des risques, parfois à contrôler des organisations, pour les contraindre à plus de transparence ou à payer des tarifs par exemple. Enfin, sur certaines thématiques, la contrainte se traduit par des interdictions. Il ne fait aucun doute que des analyses supplémentaires nous permettraient d'aller encore plus loin dans l'identification de formes de contraintes.

Dans le cadre de ce rapport, nous avons mobilisé les outils parmi les plus avancés de l'analyse textuelle, comme l'analyse thématique et l'analyse de *clusters*, qui relèvent de l'apprentissage machine non supervisée. À ce stade de notre recherche, une démarche non supervisée, plus inductive, était nécessaire pour circonscrire le contenu de notre corpus. Cette analyse complétée, nous avons maintenant des hypothèses nous permettant d'utiliser des techniques d'apprentissage machine supervisée pour faire des analyses prédictives.

Nous déploierons de tels outils. En ce qui concerne la question de la contrainte, par exemple, nous allons construire deux bases d'apprentissage du corpus. La première base reflètera la contrainte règlementaire et la deuxième son allègement. Nous allons ensuite utiliser ces bases afin d'entrainer des algorithmes à détecter l'expression de ces deux volets de la règlementation dans un segment de texte que nous aurions défini au préalable. L'intérêt de cette stratégie sera de pouvoir mesurer avec précision le ratio entre l'ajout et le retrait de contraintes règlementaires. D'autres analyses potentielles sont possibles. Une d'entre elles est d'isoler les types d'arguments présents dans le texte et de les

[3]Notons que le niveau de contrainte a sans doute diminué sur certaines thématiques, mais que le lexique utilisé dans ce rapport ne nous permet pas de le mesurer. Nous travaillons à la construction d'un lexique qui nous permettra mesurer la dérèglementation.

relier à des acteurs précis. Cela serait possible grâce à la détection d'entités nommées et à notre expérience de recherche sur les réseaux d'acteurs politiques. Par ailleurs, d'autres analyses de type non supervisé sont prévues. Nous voulons notamment analyser le degré de cohérence sémantique de certains concepts, notamment des termes juridiques, afin de détecter les incohérences potentielles présentes dans la règlementation fédérale canadienne. La présence de concepts ambigus en droit peut en effet poser problème lors de l'application des règlements.

Les possibilités sont multiples. Nous proposons donc au Secrétariat du Conseil du Trésor et à l'École de la fonction publique du Canada un partenariat. Comme nous l'avons mentionné au début du rapport, nous sommes tous membres du CIRST qui a récemment mis en place le Bureau des initiatives numériques dont le mandat est de constituer et d’entretenir des bases de données textuelles comme RegCan. Nous comptons mettre à profit cette base de données dans les mois et les années à venir par l'entremise du CIRST.

Références

Al-Ubaydli O. et P. A. McLaughlin. 2017. “RegData: A numerical database on industry-specific regulations for all United States industries and federal regulations, 1997–2012”. *Regulation & Governance*. 2017 11 (1): 109-23. https://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12107

Canada. 2007. *Directive du cabinet sur la rationalisation de la règlementation*. Secrétariat du Conseil du Trésor. http://publications.gc.ca/collections/Collection/BT22-110-2007F.pdf

Canada. 2011. Évaluation à mi-parcours de la mise en œuvre de la Directive du Cabinet sur la rationalisation de la règlementation. Secrétariat du Conseil du Trésor. https://www.tbs-sct.gc.ca/report/orp/2011/cdsr-dcrr01-fra.asp

Canada. 2012. *Directive du cabinet sur la gestion de la règlementation.* Secrétariat du Conseil du Trésor. https://www.canada.ca/fr/secretariat-conseil-tresor/services/gestion-reglementation-federale/lignes-directrices-outils/directive-cabinet-gestion-reglementation.html

Canada. 2014. *Au sujet de la Directive du cabinet sur la gestion de la règlementation*. Secrétariat du Conseil du Trésor. https://www.canada.ca/fr/secretariat-conseil-tresor/services/gestion-reglementation-federale/sujet-directive-cabinet-gestion-reglementation.html

Canada. 2016a. *La règle du "un pour un"*. Secrétariat du Conseil du Trésor. https://www.canada.ca/fr/secretariat-conseil-tresor/services/gestion-reglementation-federale/regle-un-pour-un.html

Canada. 2016b. *La Lentille des petites entreprises*. Secrétariat du Conseil du Trésor. https://www.canada.ca/fr/secretariat-conseil-tresor/services/gestion-reglementation-federale/lentille-petites-entreprises.html

Canada. 2017a. *Dénombrement des exigences réglementaires imposant un fardeau administratif*. Secrétariat du Conseil du Trésor. https://www.canada.ca/fr/secretariat-conseil-tresor/services/gestion-reglementation-federale/base-reference-fardeau-administratif/denombrement-exigences-reglementaires.html

Canada. 2017b. Les résultats de la base de référence du fardeau administratif pour l'ensemble du gouvernement. Secrétariat du Conseil du Trésor. https://www.canada.ca/en/treasury-board-secretariat/services/federal-regulatory-management/administrative-burden-baseline/government-wide-update-2016.html

Canada. 2018. *Acts and Regulations*. Department of Finance. https://www.fin.gc.ca/legislation/index-eng.asp

Hammit, J., Rogers, M., Sand, P. et Wiener, J.B. 2013. *The reality of precaution: Comparing risk regulation in the United States and Europe*. Routledge.

Killick, R. et I. Eckley. 2014. "changepoint: An R package for changepoint analysis". *Journal of statistical software*, 58 (3): 1-19. http://eprints.lancs.ac.uk/51975/

Loughran, T. et B. McDonald. 2011. « When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks ». *Journal of Finance*, 66 (1): 35–65. http://ssrn.com/abstract=1331573.

Loughran, T. et B. McDonald. 2016. « Textual Analysis in Accounting and Finance: A Survey ». *Journal of Accounting Research*. 54 (4) :1187–1230. http://ssrn.com/abstract=2504147

Michalke, M. 2017. *koRpus: An R Package for Text Analysis*. R package 0.10-2. https://cran.r-project.org/web/packages/koRpus/koRpus.pdf

Roberts, M. E., B. M. Stewart et D. Tingley. 2014. « stm: R package for structural topic models ». R package version 0.6.21. Software package. http://structuraltopicmodel.com

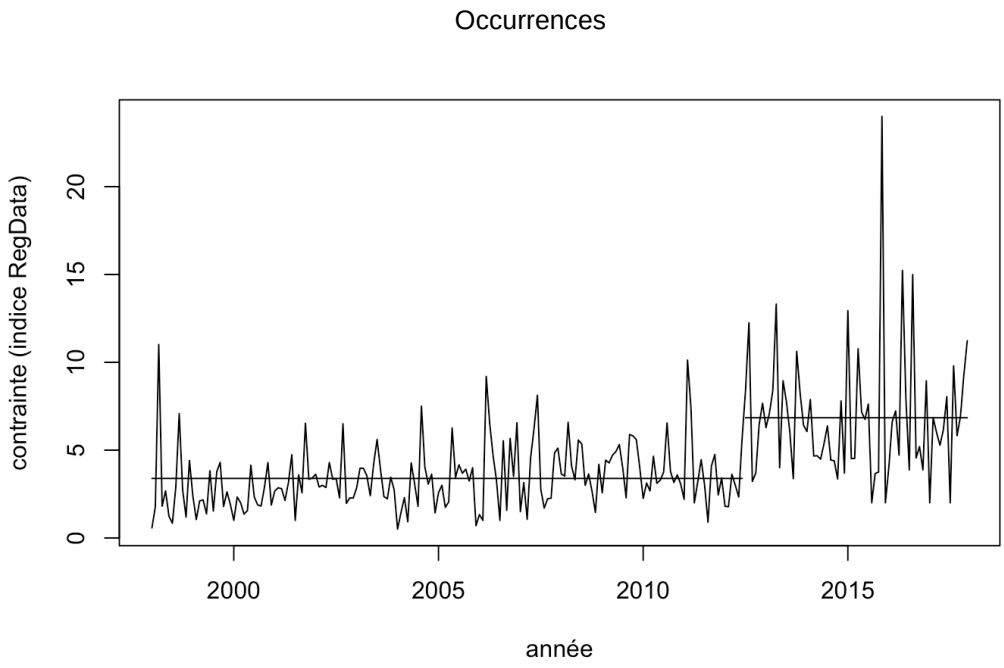
Software Repository for Accounting and Finance. 2018. “Software Repository for Accounting and Finance”. University of Notre-Dame. https://sraf.nd.edu/

Swedlow, B., Kall, D., Zhou, Z., Hammitt, J. K., et J.B. Wiener. 2009. “Theorizing and generalizing about risk assessment and regulation through comparative nested analysis of representative cases”. *Law & Policy*, 31 (2) : 236–269. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9930.2009.00296.x

Annexes

Annexe A : Indices de contrainte

Figure A1 : Regdata



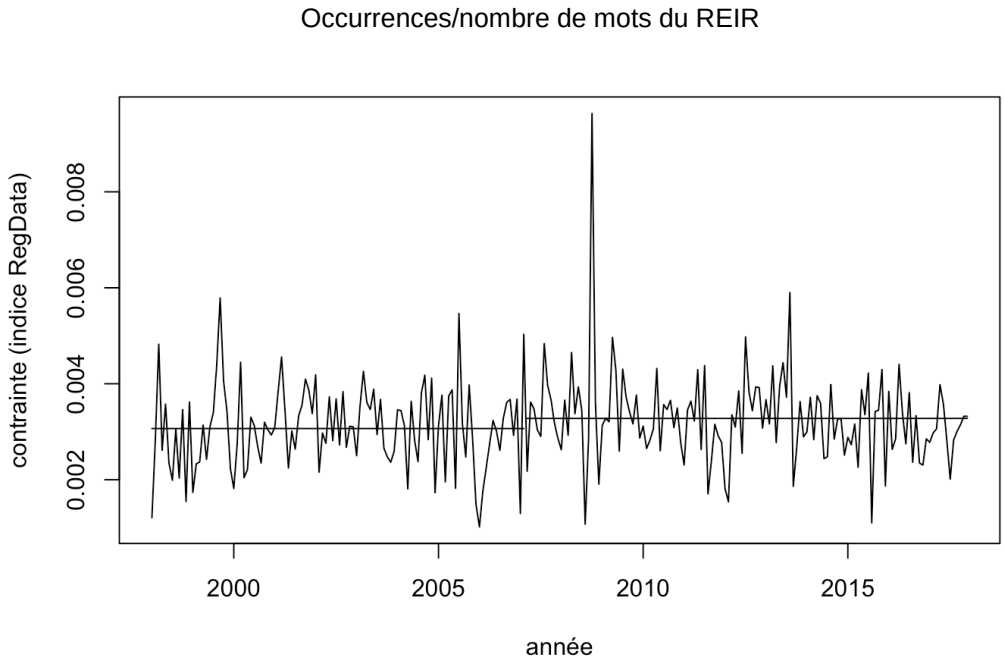
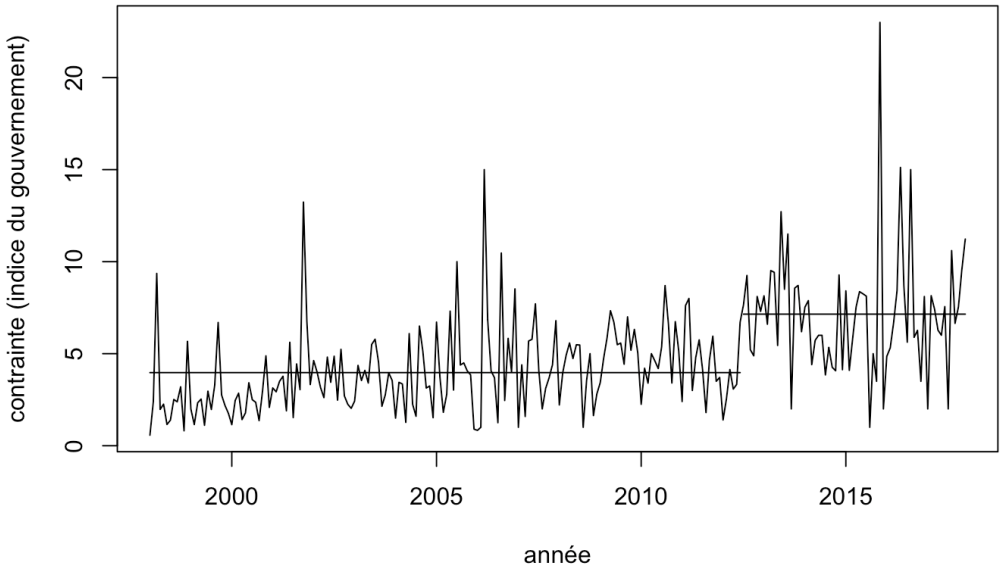
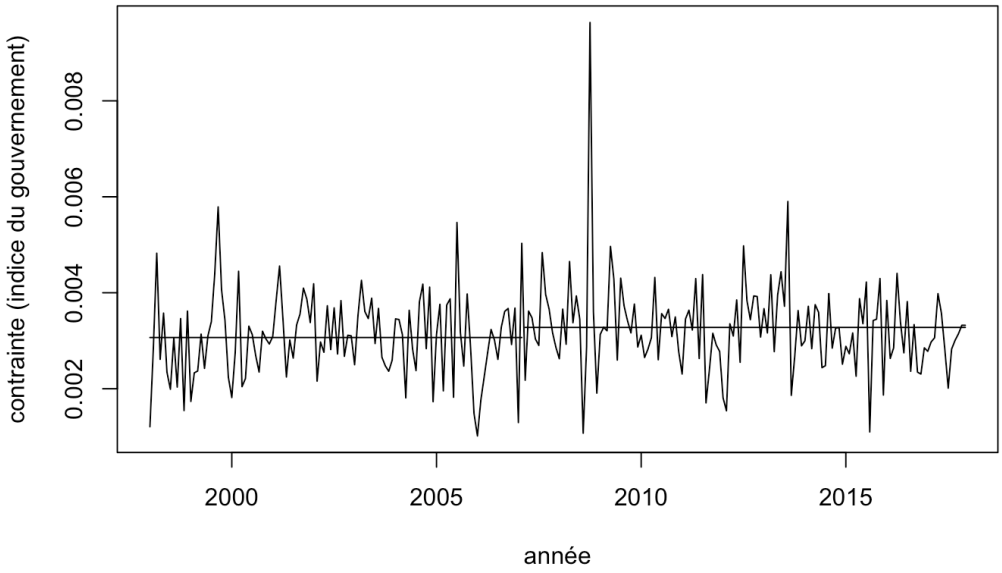


Figure A2 : Secrétariat du Conseil du Trésor

Occurrences



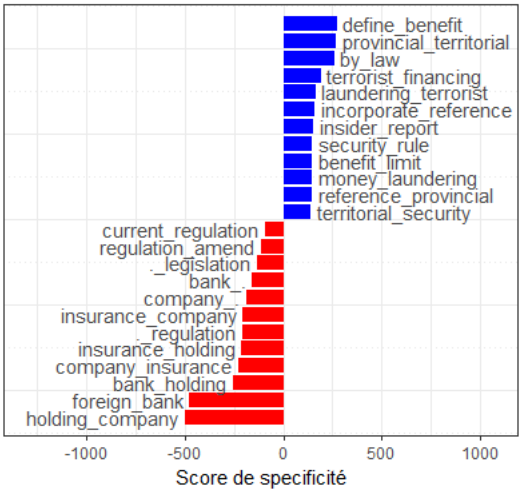
Occurrences/nombre de mots du REIR



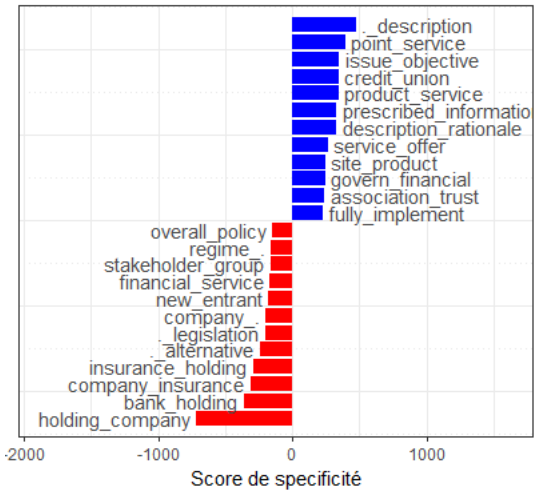
Annexe B: Spécificités lexicales par ministère

Ministère des finances

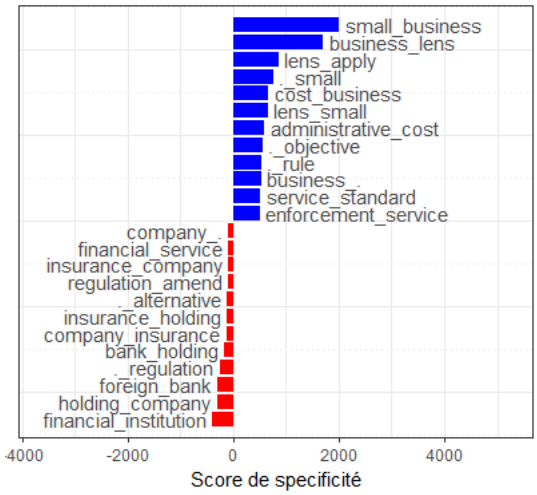
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

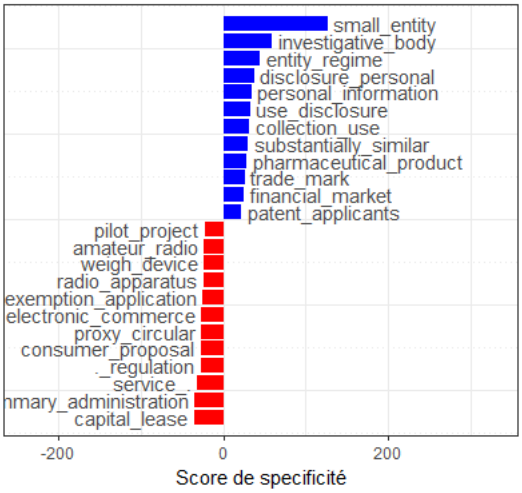


Spécificité de 2013 à 2017

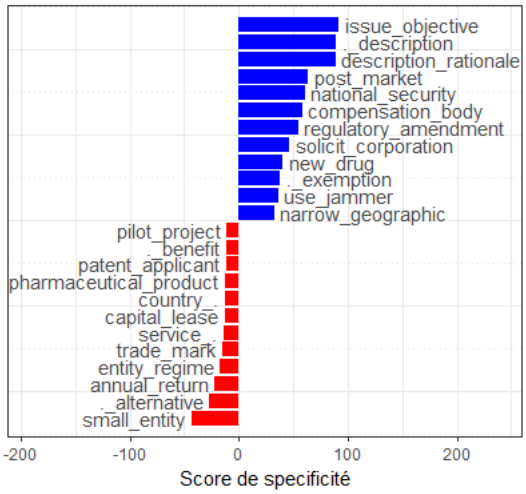


Ministère de l'industrie

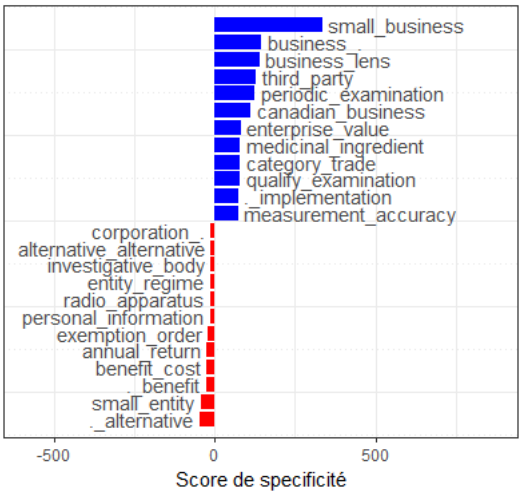
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

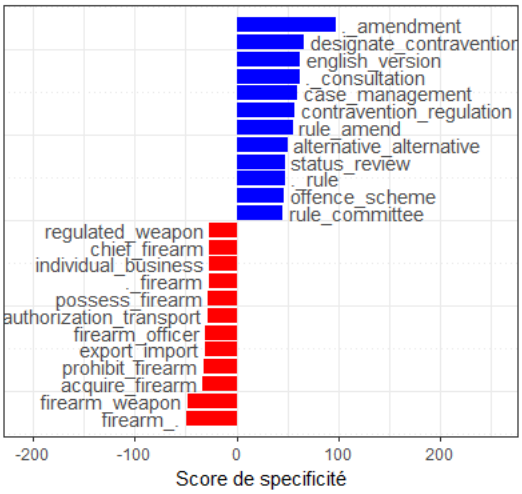


Spécificité de 2013 à 2017

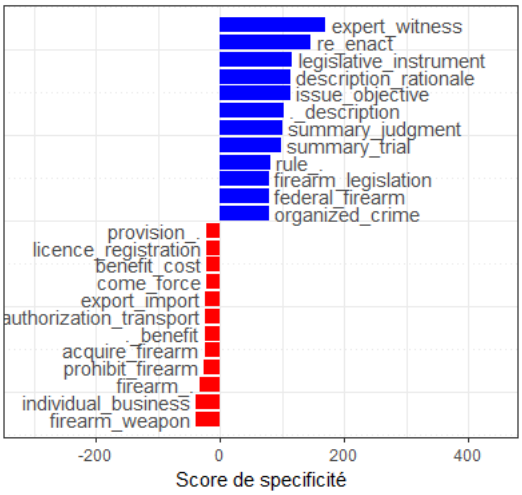


Ministère de la justice

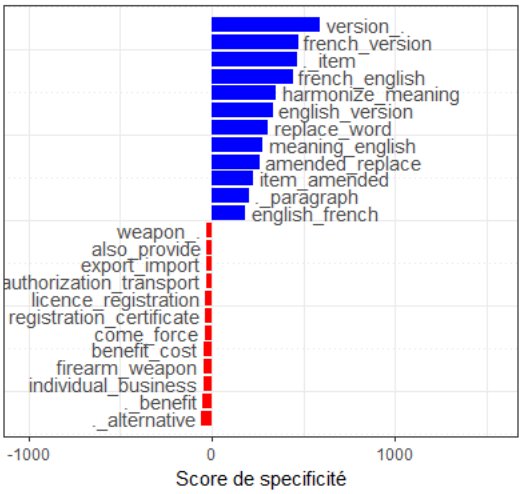
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

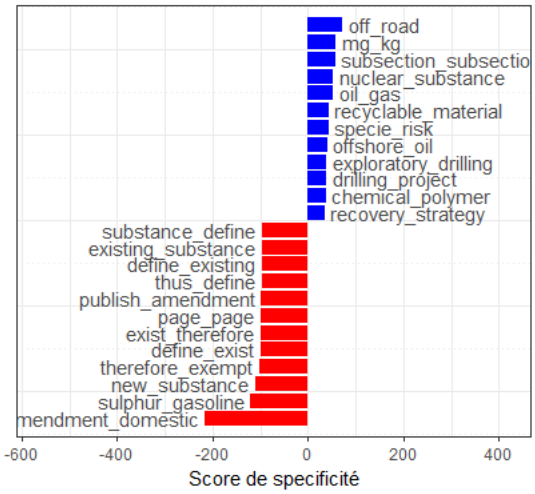


Spécificité de 2013 à 2017

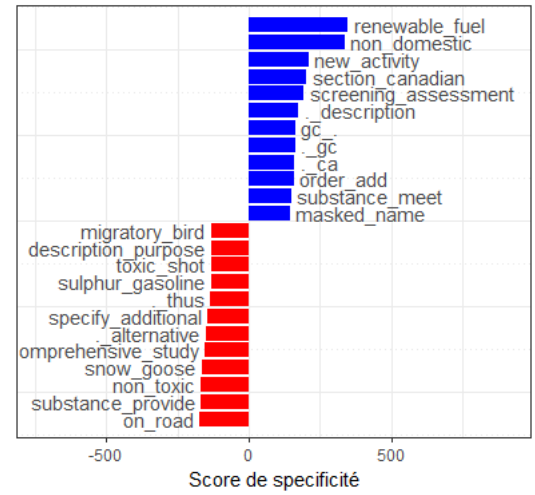


Ministère de l'environnement

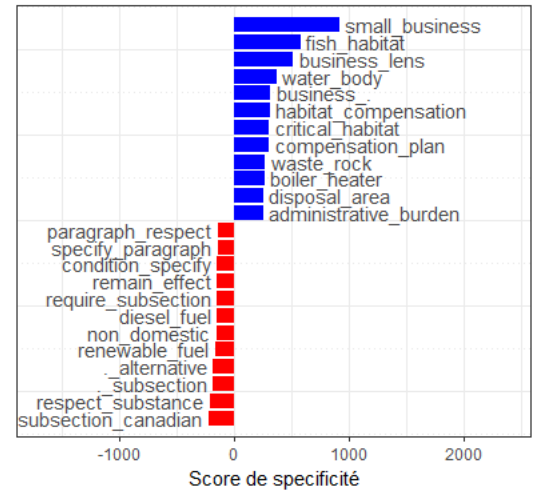
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

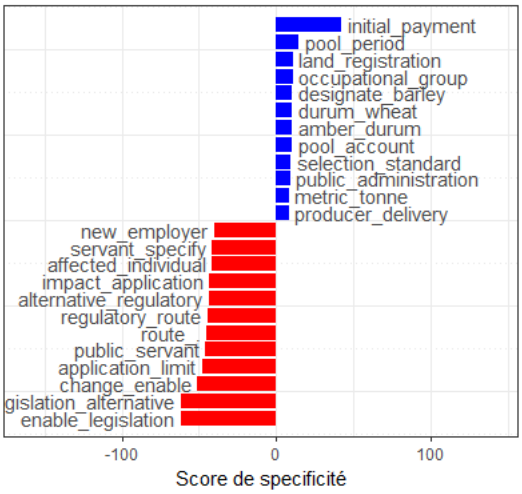


Spécificité de 2013 à 2017

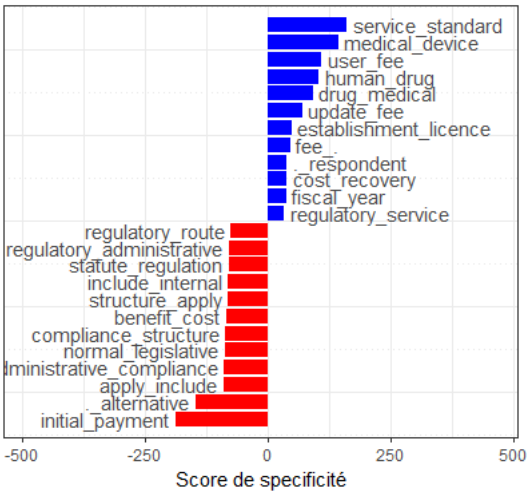


Conseil du Trésor

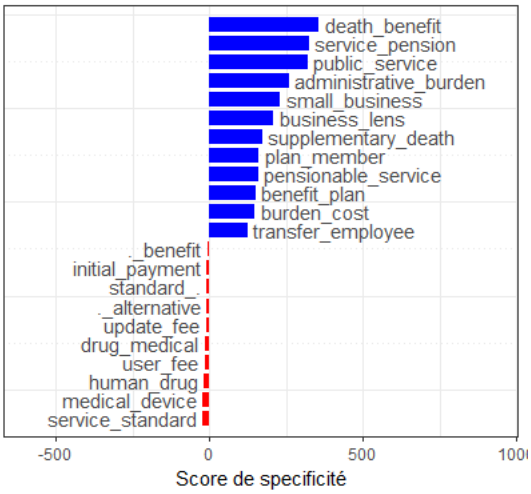
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

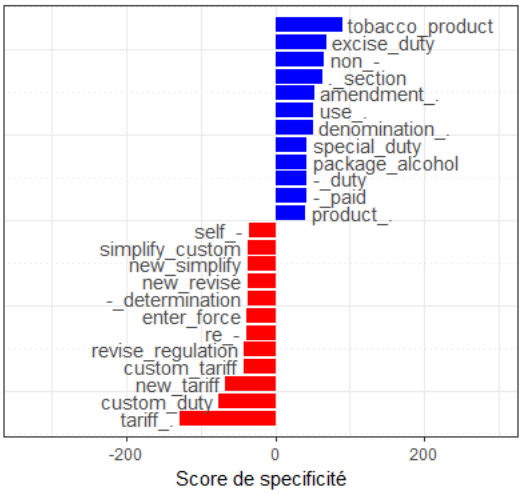


Spécificité de 2013 à 2017

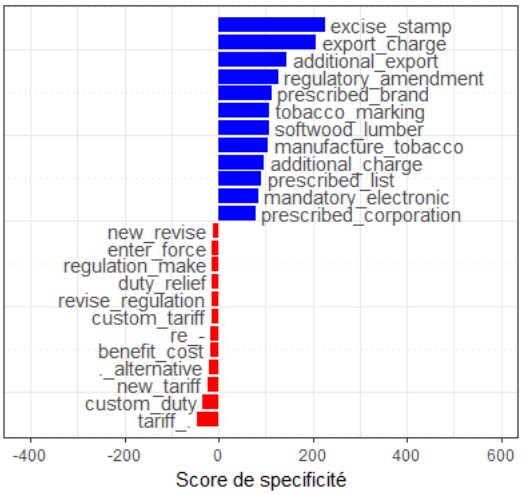


Ministère du revenu

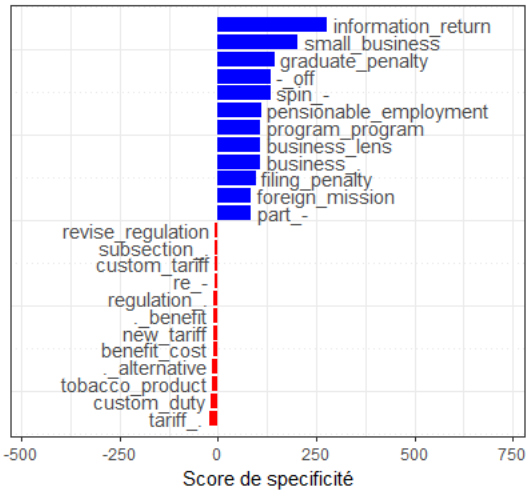
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

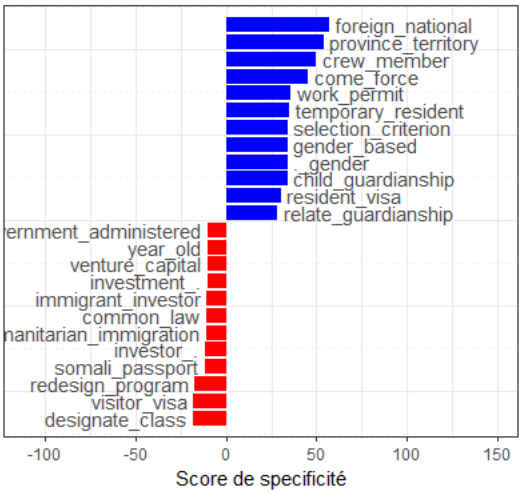


Spécificité de 2013 à 2017

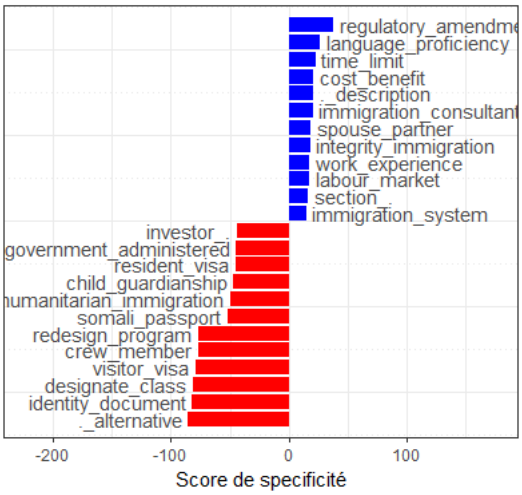


Ministère de l'immigration

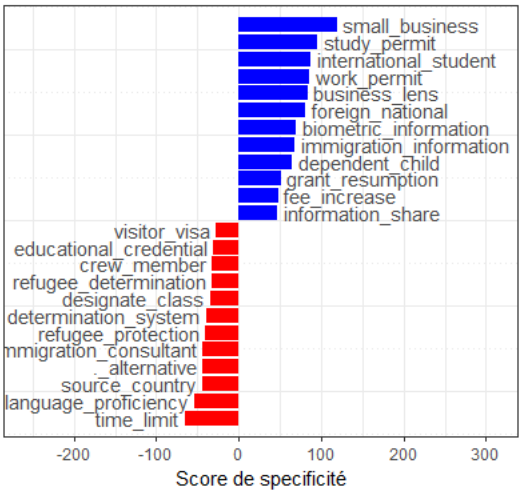
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

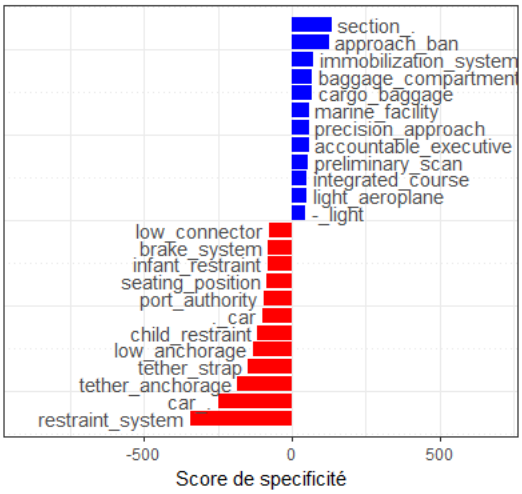


Spécificité de 2013 à 2017

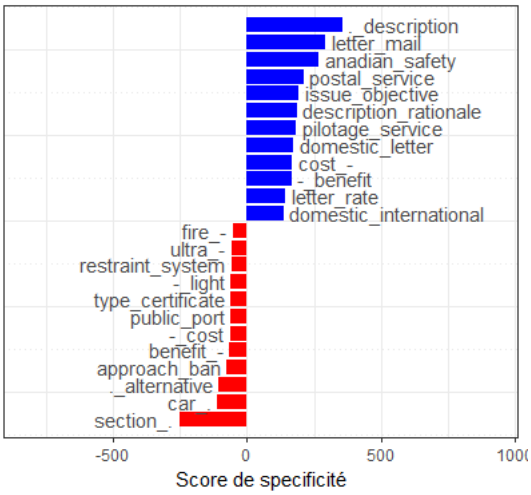


Ministère du transport

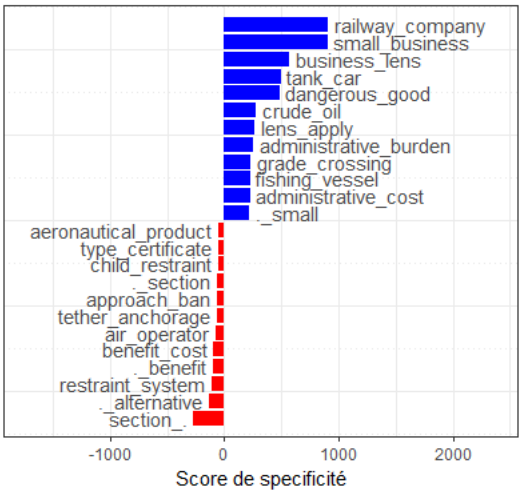
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

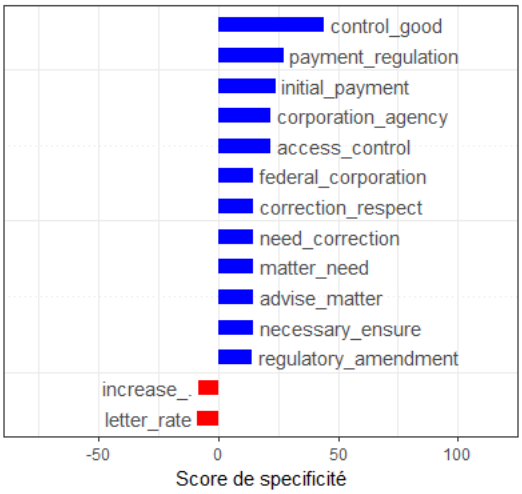


Spécificité de 2013 à 2017

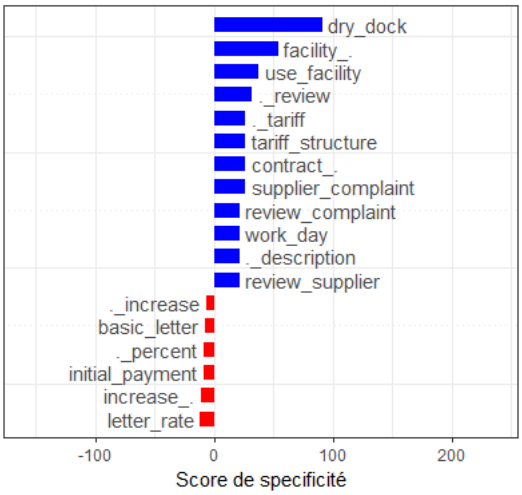


Ministère des travaux publiques

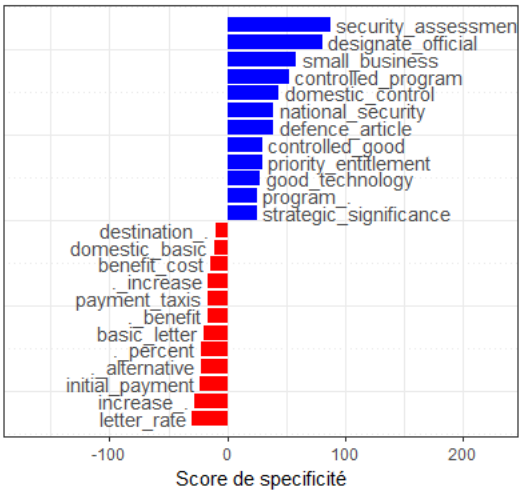
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

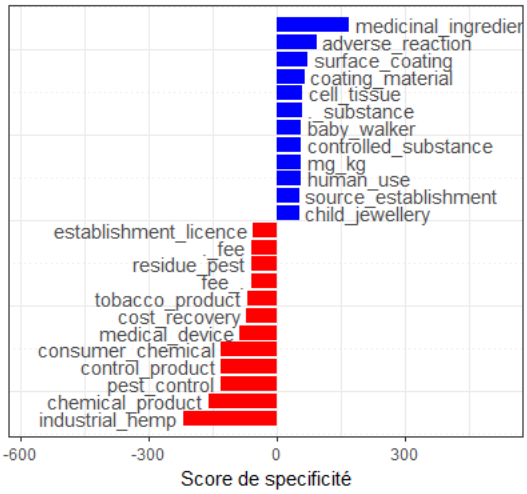


Spécificité de 2013 à 2017

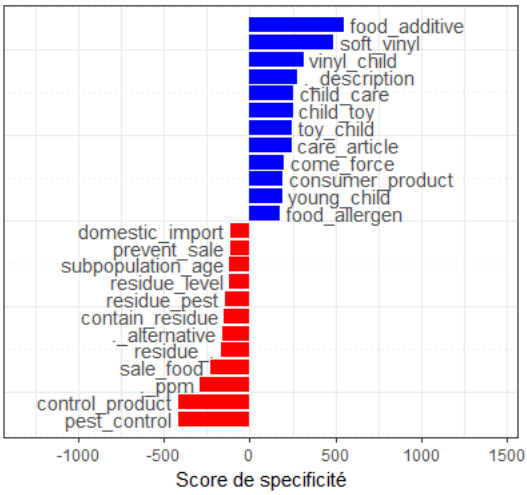


Ministère de la santé

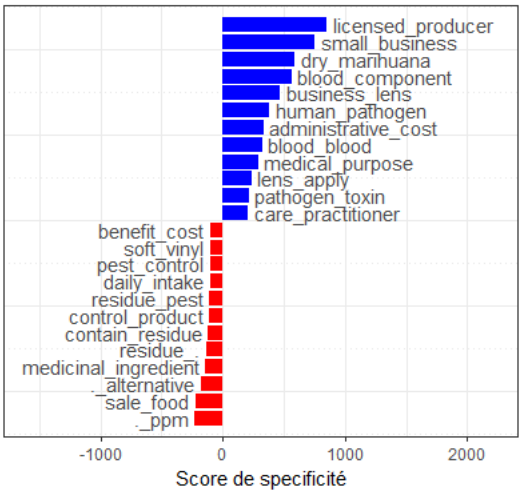
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

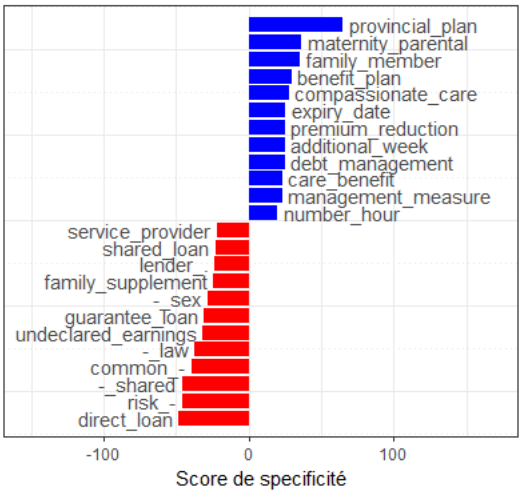


Spécificité de 2013 à 2017

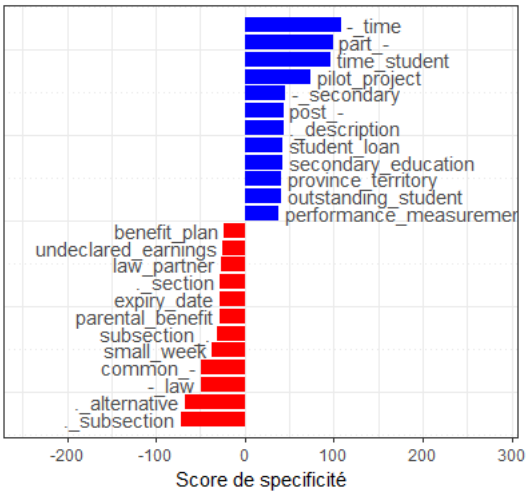


Ministère du développement social

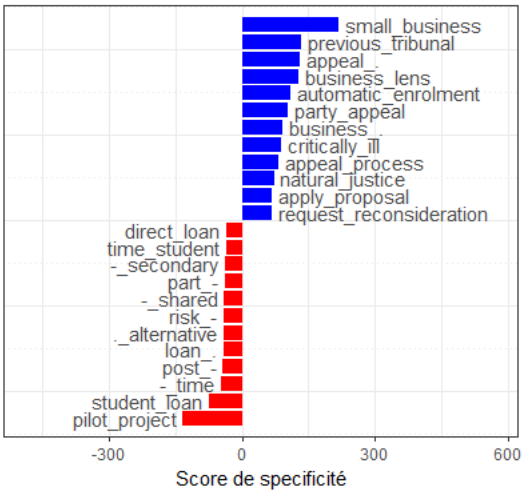
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

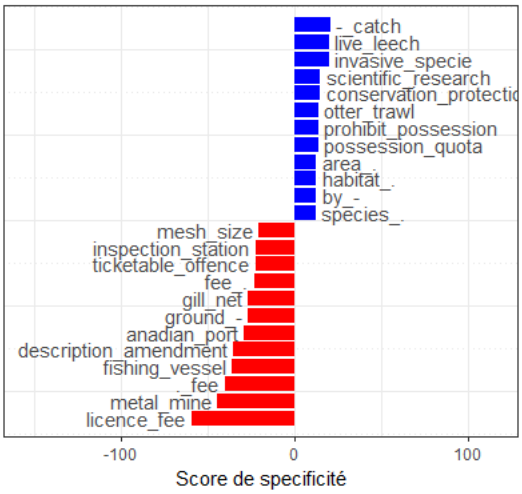


Spécificité de 2013 à 2017

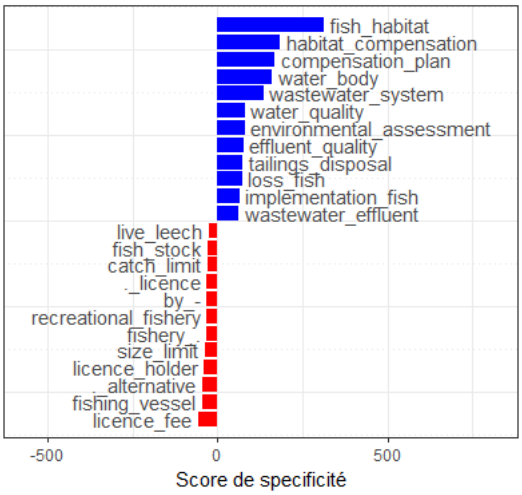


Ministère des pêches

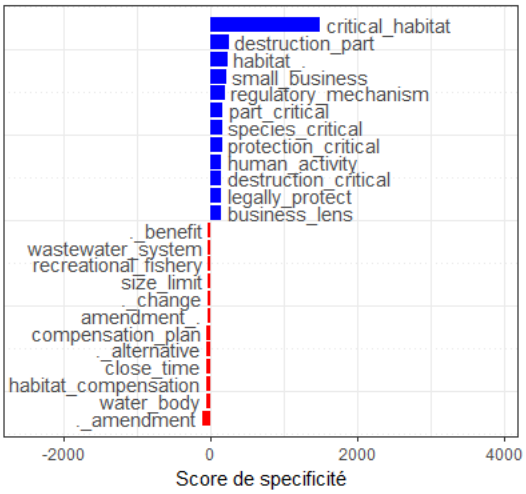
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

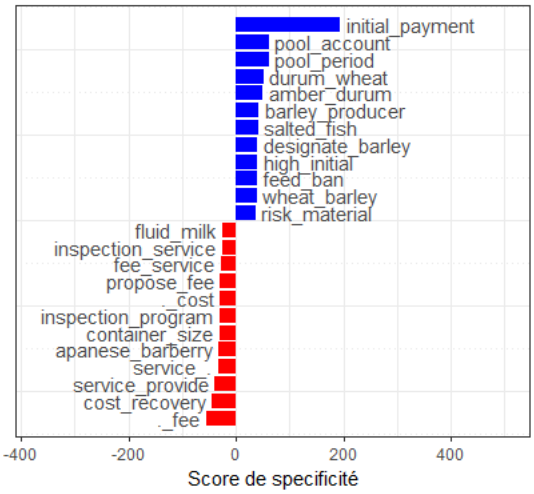


Spécificité de 2013 à 2017

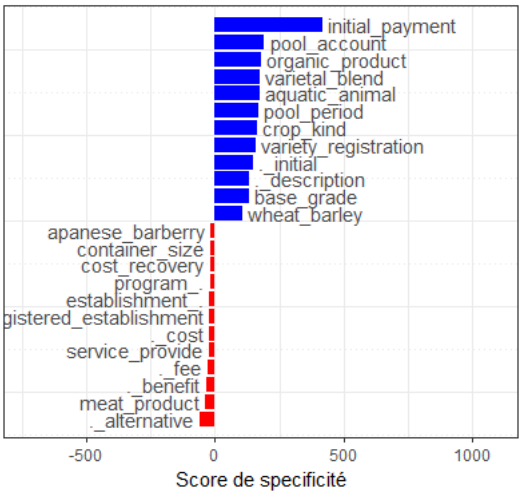


Ministère de l'agriculture

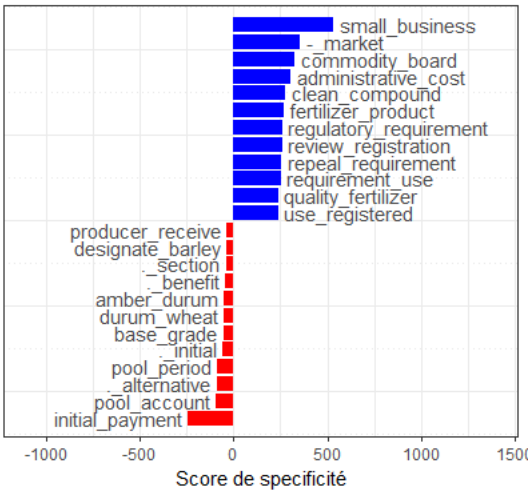
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

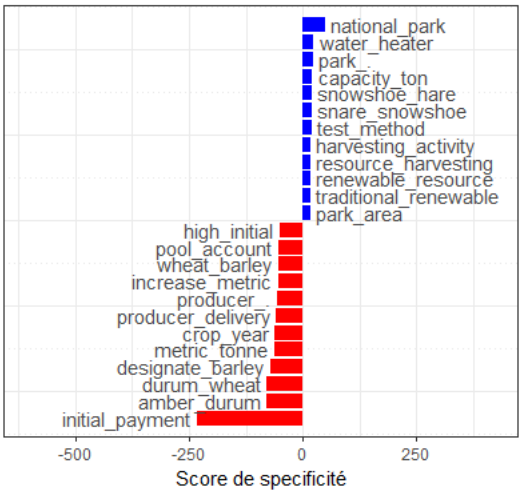


Spécificité de 2013 à 2017

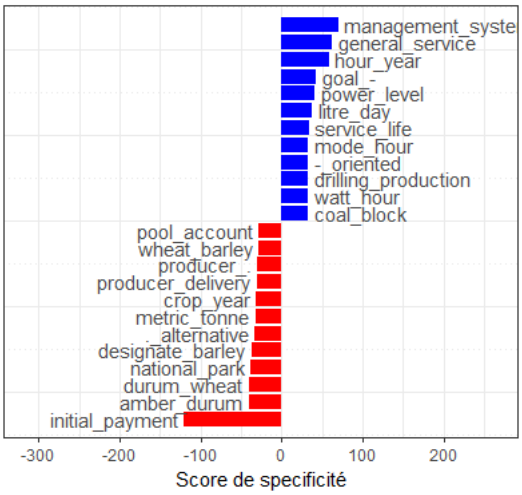


Ministère des ressources naturelles

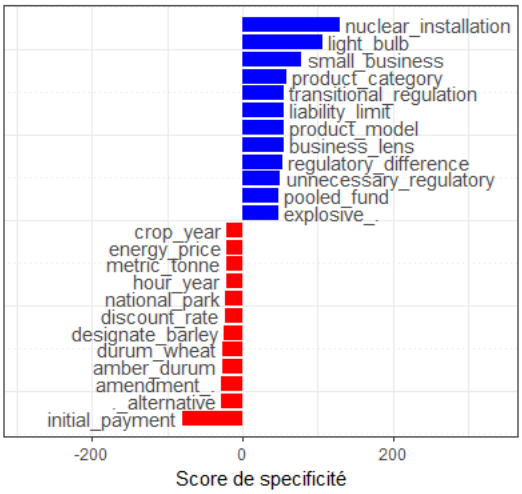
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

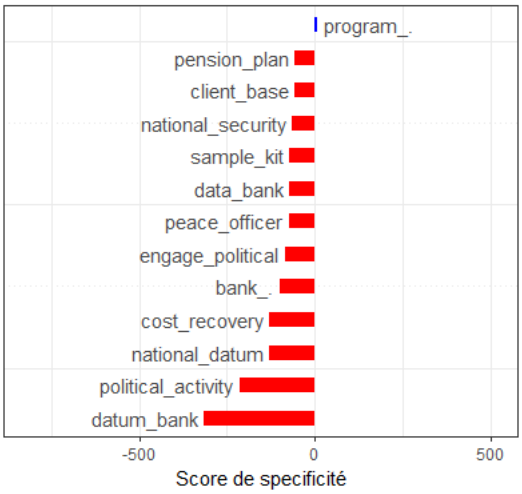


Spécificité de 2013 à 2017

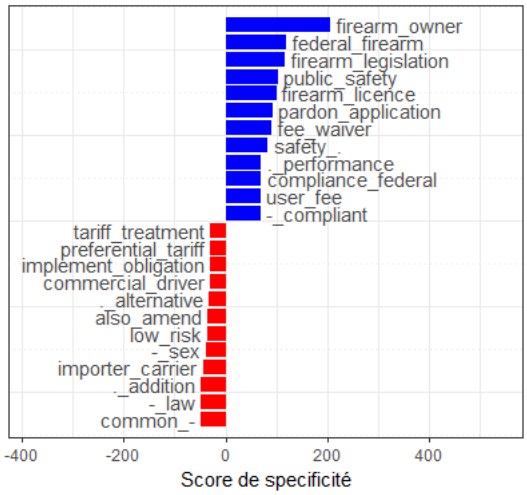


Ministère de la sécurité publique

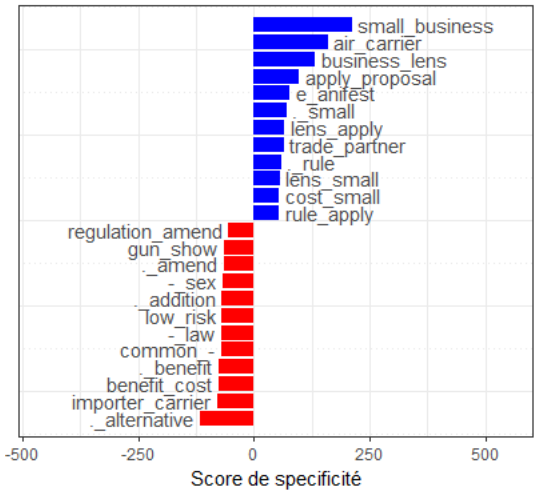
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

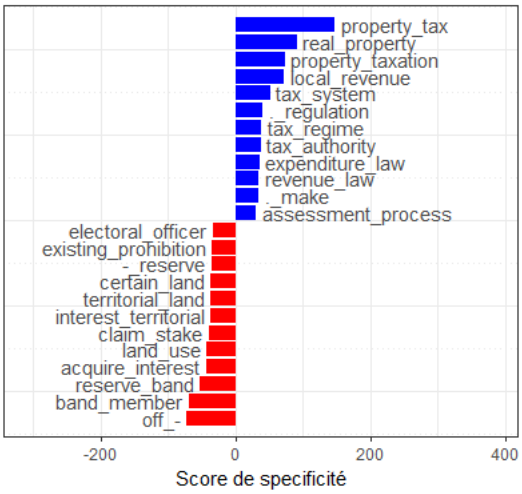


Spécificité de 2013 à 2017

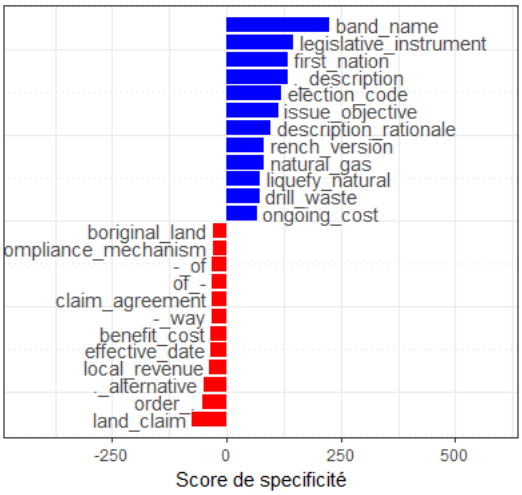


Ministère des affaires autochtone

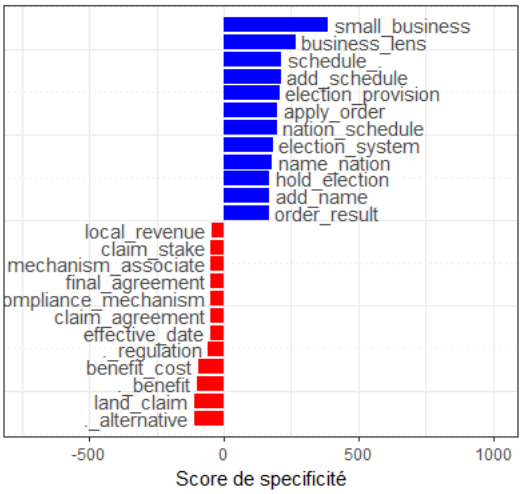
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

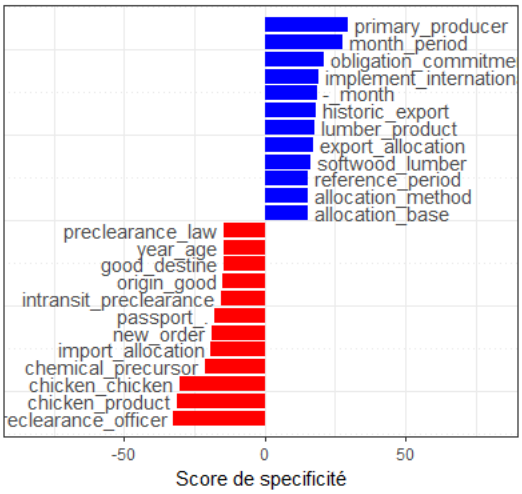


Spécificité de 2013 à 2017

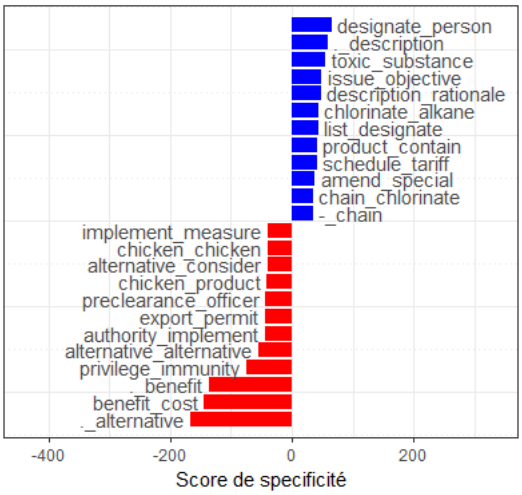


Ministère des affaires étrangères

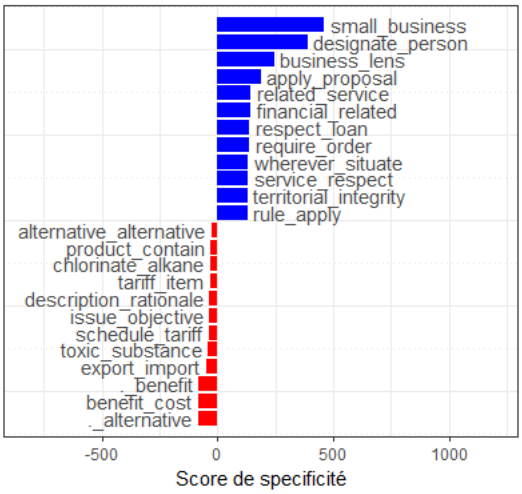
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

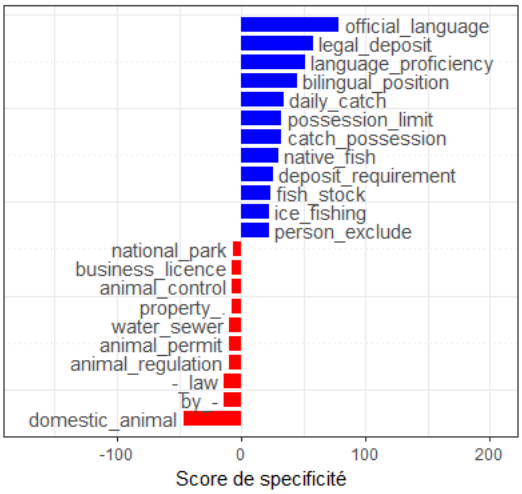


Spécificité de 2013 à 2017

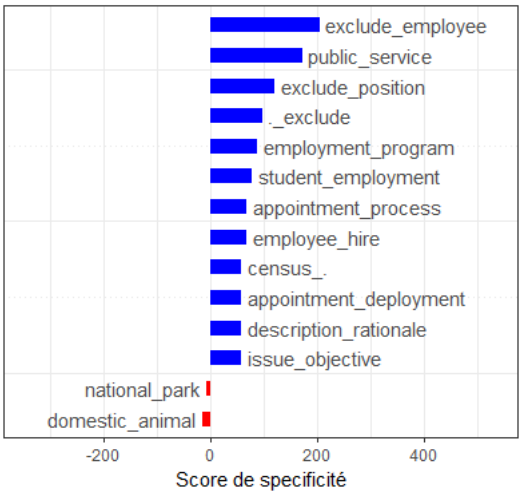


Ministère du patrimoine

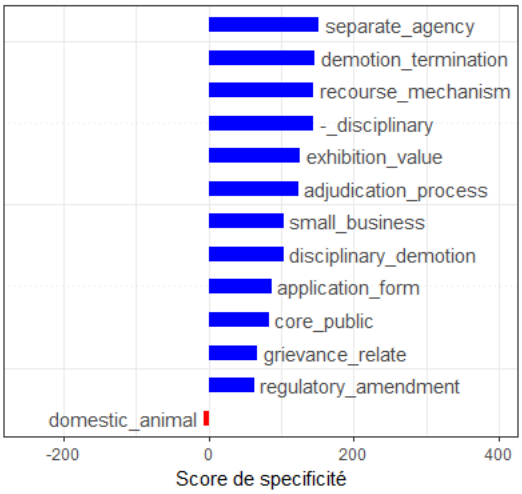
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

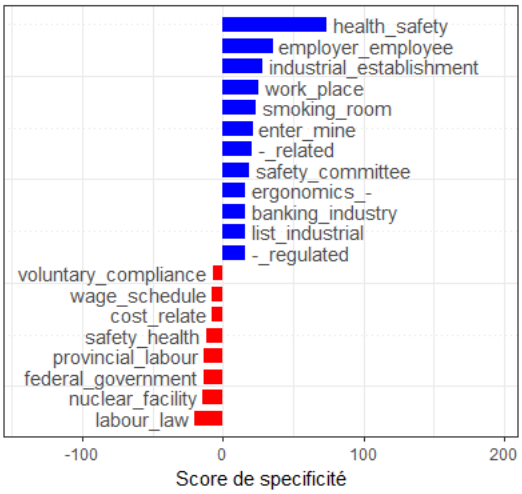


Spécificité de 2013 à 2017

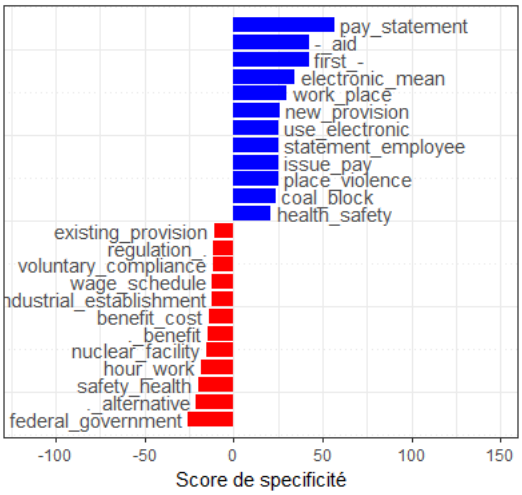


Ministère du travail

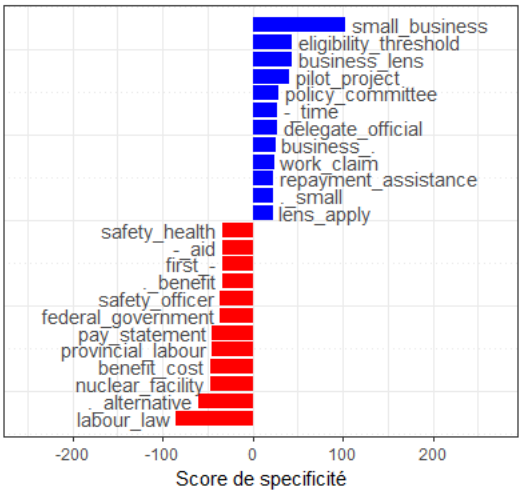
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012

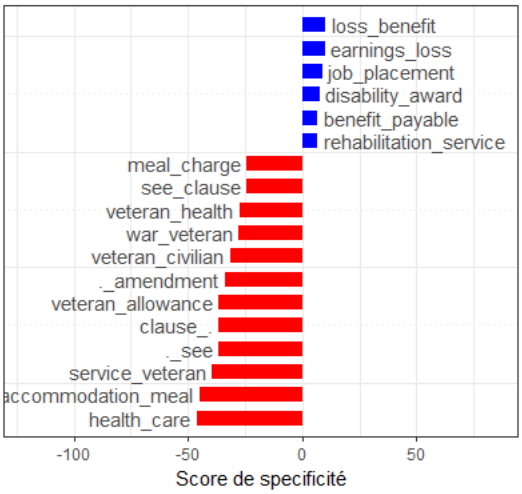


Spécificité de 2013 à 2017

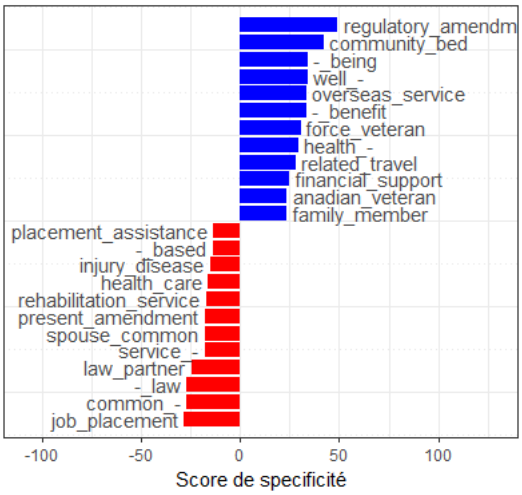


Ministère des vétérans

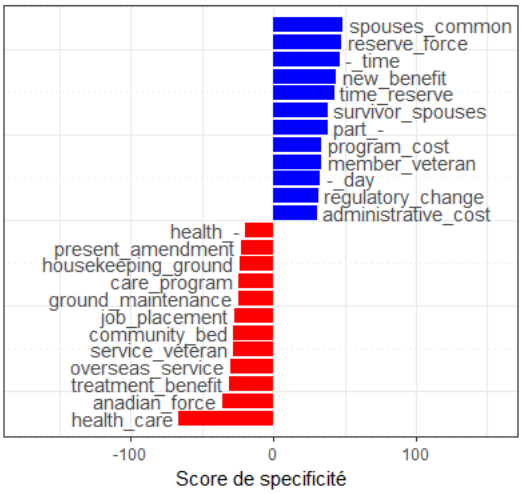
Spécificité de 2003 à 2007



Spécificité de 2008 à 2012



Spécificité de 2013 à 2017



Annexe C : Projet CRSH Éric Montpetit

Expert Information and Regulatory Adaptation **Éric Montpetit**

Objectives

It is overwhelmingly accepted that government regulation needs adaptation when technology evolves. Take genetic testing: the first genetic tests were pre-natal exams to detect genetic disorders in foetuses (Hudson et al. 2006). These non-intrusive tests were performed in a handful of clinics and laboratories to help couples decide whether they should consider testing that was more precise, but also more intrusive. While false-negative results were always a concern, the consequences of false-positives were minimal, mostly causing anxiety before more thorough tests were conducted. In most countries, governments opted for a light regulatory approach to oversee the validity of these genetic tests.

The information produced by the Human Genome Project accelerated the development of genetic tests, which are now conducted in a large number of clinics and laboratories for a wide array of medical decisions. For example, genetic testing has become a common tool in cancer treatment (Kolata 2016). While technological advances have expanded the field of application and the benefits of genetic testing, its risks have deeper implications than simple anxiety caused by false-positive pre-natal tests. Today, a false-positive test for a cancer patient can dangerously delay treatments or encourage unnecessary surgeries. The validity of the tests and their interpretation have thus become a serious concern, feeding into rational expectations that regulators will increase oversight of genetic testing.

As rational as regulatory adaptation might seem in this example, it is unlikely to occur straightforwardly, largely owing to information supply. When making decisions, regulators must rely on expert information, seemingly independent from political influence. Finding and selecting this expert information, however, is not a simple task. In the United States, information on genetic testing is eclectic, stemming from congressional committee hearings and various expert bodies of the executive branch. Therefore, regulators are often overwhelmed when trying to prioritize information and decide on adequate adaptations. The UK government tried to avoid this information overload trap in 1999 by creating the Human Genetics Commission (HGC), whose mandate was to draw from varied sources to provide unified advice for regulatory adaptation. The approach worked for a while, with HGC reports effectively informing regulatory adjustments. However, internal tensions over the credibility of various sources of information made it untenable in the longer term, and the HGC was disbanded in 2012. As seen in the American and British examples, informing regulatory adaptation is a difficult task; the main objective of this research proposal is to examine how Canadian regulators address this challenge.

While research on government regulations abounds in the United States and the United Kingdom, we know very little about regulation as a distinctive form of governmental activity in Canada. Canadian scholars have conducted informative case studies, but nothing comprehensive enough to tackle the big questions this project seeks to address. How do regulations change? Which expert information warrants these changes? Is this information likely to enable regulatory adaptations that are consistent with the evolution of scientific knowledge and technology? These questions are pressing, as technology and scientific knowledge are changing quickly, and not only in the field of genetic testing. With the arrival of self-driving cars, car safety regulations no longer suffice (Dwoskin and Fung 2016); regulations

that protect private information need revising to keep up with the progress of big data and machine learning (Keats Citron 2016); and GMO regulations can easily be bypassed by biotechnology developers using new gene-editing technologies (Ledford 2016).

This project seeks to overcome a major obstacle to the study of Canadian regulations, one that might explain why the topic has remained understudied. The Canadian federal government releases a large quantity of information about its regulatory activities, but in formats ill-suited for research (Roberts 2006, chapter 9). For example, since 1986, all new regulations, as well as amendments to existing ones, have to come with a Regulatory Impact Analysis Statement (RIAS). RIAS provide surprisingly rich and systematic information on the content and purpose of the regulatory changes, and on which experts ere consulted. However, RIAS are published in the Canada Gazette along with a large amount of other information (government appointments, prizes, etc.). The systematic extraction of RIAS from the Canada Gazette presents a challenge to researchers, one that this project will surmount. Beyond producing new knowledge about regulatory adaptation to technological change, my objective is to pull relevant information on the federal government's regulatory activities from the mass of government documents available on-line and compile it into a dataset that is accessible to other researchers and interested publics.

Context

One stream of literature that will be useful in this research prescribes "planned adaptive regulation"—that is, administrative procedures whereby regulations can be aligned with technological change within a reasonable timeframe (Eichler et al. 2015; McCray, Oye, and Petersen 2010). This project shares with this literature the definition of regulatory adaptation as decisions about government regulations, which can add to, discard or replace existing rules. This project, however, does not share the normative underpinning of the literature on planned adaptive regulation, which assumes that gaps between regulations and technological/scientific knowledge are anomalies that always need to be addressed through regulatory adaptation. Related complaints about the failure of regulators to adapt regulations to emerging technologies and scientific advances are reminiscent of a cybernetic ideal (e.g. Nature editorial 2016). Academic knowledge on cognitive limitations (Simon 1957) and on the nexus between values and facts (Douglas and Wildavsky 1982) raise serious doubts about the use of cybernetics in regulatory adaptation. Nevertheless, we learn from planned adaptive regulations (as well as from cybernetics) that expert information is key to regulatory change. A regulatory adaptation motivated by political pressure only—that is, an adaptation that ignores expert information—risks serious failures, especially in technical sectors. The selection of expert information sources by regulators, however, is problematic (Weible 2008). That is precisely the problem this project seeks to illuminate, and the literature on experts and public policy will be particularly useful in defining a working hypothesis.

Early scholarly literature on experts and public policy was concerned with technocracy and the use of expert reasoning for political decision-making (e.g. Akin 1977). Later work on experts and public policy focused on policy processes that allow for the exercise of power by a narrow elite of expert scientists (Jasanoff 1990), as well as on the difficulties arising from relying on experts to guide policy decisions (Harrison 1991; Brunner and Ascher 1992; Jasanoff 2003; Collingridge and Reeve 1986; Weible and Sabatier 2009; Montpetit 2011). These difficulties range from the lack of democratic accountability of

experts and their internal disagreements to the moving boundary between the political and scientific spheres. The literature further suggests that difficulties do not arise only from the experts themselves, but also from policy-makers' demands for scientific guidance. The timing of political decisions does not often leave scientists sufficient time to produce credible knowledge; policy-makers often prefer expert knowledge that confirms rather than challenges their prior beliefs; policy-makers demand clarity and simple messages while scientists prefer complexity and acknowledgement of uncertainty (Sarkki et al. 2014).

These observations coincide with an opening of policy processes to a wider diversity of experts, even to so-called "expert citizens" (Li and Marsh 2008). Scholars from both Europe and North America have witnessed a democratization of expertise and the implosion of policy monopolies once held by a narrow set of experts (Beck 1992; Radaelli 2002; Maasen and Weingart 2005; Montpetit and Rouillard 2008). This literature on experts and public policy suggests that over time, regulators have started to draw information from increasingly eclectic sources of expertise.

Baumgartner and Jones (2015) contend that eclectic information arises from what they call "entropic search." Entropic search, they argue, informs problem-relevance; they contrast this type of search with the narrower "expert search", which informs problem-understanding and assesses the solution's impact. Each type of search faces its own distinctive sets of problems. While expert-search facilitates the implementation of a solution, it also prioritizes a single dimension of an issue, potentially ignoring serious problems in other dimensions. In contrast, entropic search attends to several dimensions of an issue and it is therefore less likely to overlook problems. However, entropic search can overwhelm decision-makers, who have difficulty prioritizing and deciding on policy solutions. In the absence of decisions, policy stagnation prevails and eclectic information about problems in the policy environment accumulates until change becomes so compelling that policy-makers overreact. Entropic searches thus encourage both insufficient and excessive reactions to problems, instead of encouraging reactions that are proportional to the severity of the problems.

Transposing this reasoning to regulatory adaptation, I expect a lot of over- and under-adaptation in response to technological/scientific change, and few proportional adaptations. This expectation arises not only from Baumgartner and Jone's (2015) Politics of Information, but also from the difficulty in maintaining narrow expert searches in today's context of expertise democratization. My general working hypothesis in this project is therefore that the more entropic the information provided to regulators, the more likely they are to over- and under-adapt regulations.

The reasons for not relying more on the scholarly literature on government regulation to inform this project's working hypothesis need a brief explanation. Much of the existing literature was produced by economists looking into regulations as a means to overcome market failures (Atkinson and Tietenberg 1991) or as costs to industry (Ryan 2012; Helm 2006). But political scientists have also produced important work on regulations, examining the differences in national regulatory styles (Vogel 1981), studying long-term shifts in risk aversion reflected in regulatory approaches (Vogel 2002), analysing private regulation (Cashore et al. 2007), examining disjunctions in sectoral regulatory regimes within a single country (Hood, Rothstein, and Baldwin 2001), categorizing varieties of regulatory capitalism (Levi- Faur 2006), or looking into regulatory discourses in times of crisis (Lodge and Wegrich 2011). I rely on some of this literature in the methods section of this proposal, but it was of little help in defining a working hypothesis about short-term regulatory adaptation in a changing technological environment.

A seemingly relevant stream of literature is one that promotes the concept of "responsive regulations," which calls for the involvement of stakeholders and a flexible attitude on the part of regulators (Ayres and Braithwaite 1995; Grabosky 2013). While this project might contribute indirectly to scholarly debates on these matters, the main concern in this literature is compliance with rules (Nielsen and Parker 2009; Decker 2007; Baldwin and Black 2008), a somewhat distant concern from this proposal's focus on information feeding into regulatory adaptation.

Methods

Policy scholars have defined regulations in a variety of ways (Koop and Lodge 2015). Here I rely on the definition provided by the government of Canada:

> Regulations are a form of law, often referred to as delegated or subordinate legislation. Like Acts, they have binding legal effect and usually state rules that apply generally, rather than to specific persons or situations. However, regulations are not made by Parliament. Rather, they are made by persons or bodies to whom Parliament has delegated the authority to make them, such as the Governor in Council, a Minister or an administrative agency (Canada and Privy Council Office 2001, 176).

Although Acts of Parliament attract more attention, regulations account for a huge part of the federal government's activities. In 1998, it was estimated that the federal government had some 3000 regulations, together comprising over 30,000 pages of documents. These figures compare to about 450 federal statutes comprising some 13,000 pages. Every year, about 1000 regulatory proposals, including orders-in-council, are submitted to cabinet for consideration, versus 80 bills studied by Parliament (Canada and Department of Justice 1998, 3). Unsurprisingly, analysts have often suggested that to understand government activity, it is not enough to study budgets and Acts of Parliament; regulations must also be included as objects of analysis (Hood 1983). And yet Canadian scholars studying government activities tend to overlook regulatory activities. Most studies of regulations in Canada are case studies looking at sources of influence or regulatory impacts for particular groups or economic sectors (e.g. Montpetit 2005).

Centered on the issue of regulatory adaption, this project examines new regulations, as well as amendments to existing ones. It also includes the most important category of orders-in-council (Category A), which also contribute to regulatory adaptation (Canada, Privy Council Office, and Treasury Board Secretariat 2013, 3). These orders-in-council possess the attributes of a regulation, although they are adopted through a less stringent process, often in response to emergency situations. The government discusses new regulations, amendments to existing ones, and orders-in-council every week, making regulatory decisions in all policy sectors in cycles of just a few years. In fact, it is part of the government's regulation management strategy to encourage regular reviews of all regulations, notably to keep up with techno-scientific change (Canada 2012). Covering a period of 18 years, this project embraces the entire range of regulatory activity of the Government of Canada.

The Government of Canada is required by law to publish a wide range of information about new regulations, amendments to existing ones and orders-in-council. The process whereby regulations are made and amended in Canada is governed by the Statutory Instruments Act, which requires them to

be pre-published in Part I of the Canada Gazette before adoption. Moreover, since 1986, all new regulations and amendments to existing regulations have been required by law to be pre-published with a Regulatory Impact Assessment Statement (RIAS). Regulatory impact assessments have been promoted globally since the 1980s to make sure that regulators take into account the impact of their decisions on business (OECD 2009), but in Canada, RIAS have become far more significant. Although not formally part of the regulations, judges use them to inform their decisions in technical domains (Houle 2006). Owing to the fact that they must be written in plain language, RIAS provide a "layman's guide" to the regulation (Salembier and Bernhardt 2002, 15). More importantly for this project, RIAS systematically describe the intended change. They explain its objective and rationale, present the alternatives that have been considered, summarize a cost-benefit analysis, address issues of coordination with provincial governments and describe the stakeholder consultations that were conducted. In other words, RIAS inform both the nature of the regulatory adaption and the input that went into it. Category A orders-incouncils are not accompanied by a RIAS, but they do come with a plain-language explanation in Part I of the Canada Gazette. Note that there are far fewer orders-in-council than there are regulation proposals and amendments.

Beginning with Part I of the Canada Gazette, I will assemble a dataset that will inform regulatory adaptation. Archived in good quality electronic format since 1998, the Canada Gazette covers 18 years of regulatory activity by the federal government (more if we include the paper format), and yet to my knowledge, it has never been subjected to systematic analysis. Each line in the dataset will correspond to a new regulation, a regulatory amendment or a Category A order-in-council. For example, on February 15, 2014, the government pre-published an amendment to the Food and Drugs Regulations, authorized under the Food and Drugs Act. The corresponding RIAS presents the issue in the following terms: "Recent events of food-borne illness derived from mechanically tenderized beef (MTB) have underscored the need for consistent and enhanced food labelling." The RIAS further describes the amendment as "a requirement that all MTB sold in Canada be labelled clearly with information that it has been mechanically tenderized and safe cooking instructions." While the cost-benefit analysis indicates a net benefit, it also acknowledges "a one-time compliance cost, estimated to be $114,700," to industry. In other words, the RIAS clearly indicates that this amendment provides for more, not less rule (more below about how such information can be coded into variables). As with all RIAS, this one clearly presents the information that supports the regulatory amendment. It speaks of a health risk assessment produced by Health Canada that "showed a fivefold increase in risk from MTB products when compared to intact cuts of beef." It justifies the requirement to put cooking instructions on labels from "new scientific research specific to MTB that supports new cooking recommendations to achieve better protection." Lastly, the RIAS explains that prior to its announcement, the regulatory amendment was the object of consultations. It mentions four meetings with named stakeholders, including the Canadian Meat Council and the Retail Council of Canada. The opposition of some of these groups to the amendment is underlined. In addition to key stakeholders, the participation of "academic experts from the University of Manitoba, Alberta and British Columbia" were mentioned at one of the meetings, while "meat and microbiology experts" were mentioned in two others. In addition to the health risk assessment, input from various interest groups and experts obviously went into this amendment.

As illustrated by this example, the RIAS contain the information needed to test this project's hypothesis (including controls). To ensure the validity of this information, it will be crossed-checked using searches in Reports on Plans and Priorities that regulatory agencies are required to submit to parliament once a

year, and in communication reports that the Lobbying Act requires of any interest group communicating with public office holders, notably for the purpose of influencing a regulatory amendment or a new regulation. Part II of the Canada Gazette will also be used to verify that the cabinet finally approved the amendment or new regulation. After organising the RIAS (and the explanations of the orders-in-council) into string variables, a codebook will be prepared. While it is useful to know that a risk analysis informed the mandatory labelling of MTB products in the above example, this project is after more general patterns, such as the frequency of regulatory adjustments stemming from a risk estimate as opposed to other types of expert information. The codebook will enable to categorize and measure Canadian regulatory activity and to analyze adaptation as a function of expert information that went into it, while controlling for consulted groups and other variables.

Scholars such as Baumgartner and Jones (2015; 2004; Baumgartner et al. 2009) have stayed away from developing direct measures of insufficient, excessive and proportional policy reactions, comparing instead the frequency of government activities in different policy sectors and time periods, and identifying correlations with different patterns of information search. In their analyses, they use frequency of change as a proxy: excessive adaptations are most likely to occur in sectors where frequencies are highest; where they are lowest, insufficient adaption is the logical inference. I intend to use a similar strategy, using topic codes developed in Baumgartner and Jone's comparative agendas project to compare the frequency of regulatory adaptation by sector over time. The advantages to this strategy are twofold: firstly, topic codes will help identify sectors of regulation that are absent from the regulation agenda and that are therefore not being adapted at all; and secondly, they will cast light on variations that stem from known distinctions in sectoral logics (Levi-Faur 2006; Hood, Rothstein, and Baldwin 2001). Information search varies from sector to sector and if this project's hypothesis is correct, regulatory adaptation will vary as well. While this strategy will not allow us ascertain whether a regulatory adaptation is an over- or an under-reaction (although gaps with technological advances could be investigated on a case-by-case basis), it will show whether the sectors where changes are more and less frequent are also the sectors where information search is most entropic.

Beyond enabling sector comparisons, the information contained in an RIAS will allow for the development other proxies to estimate regulatory adaptation. Looking at regulatory change, Knill et al. (2012) provide for measurement of variations in regulatory density and intensity. Density is the difference between the number of instrument adoptions and the number of abolitions, while intensity measures increases and decreases in instrument strictness. While sector codes remain useful in making these calculations, they require coding at the level of regulatory change for instrument adoption and abolition, as well as for strictness. To measure strictness, Al-Ubaydli and McLaughlin (2015) suggest text analysis, a method that I plan to adapt to my own purpose. Perhaps even better than frequency of adaptation, measures of intensity, density and strictness will allow us to make prudent inferences about excessive and insufficient regulatory adaptation. In addition, Swedlow et al. (2009) have developed a method to study risk regulations organized around questions (e.g. How is the regulated risk defined? Who regulates the risk? What regulatory instrument is used? etc.). Some of the new regulations and

amendments pre-published in the Canada Gazette will not be risk regulations, but many will be, and the questions put forward by Swedlow et al. (2009) will be useful to reveal patterns of adaptation for this particular type of regulation. Together, these measures and categories help overcome the

“dependent variable problem,” which too frequently prevents the comparative analysis of regulatory change (Howlett and Cashore 2009). The codebook will provide for the faithful application of these measures, enabling their use in international comparisons. The dataset will be made public on a website.

In the past seven years, I have gained considerable experience in both manual and automatic content analysis of various kinds of documents (Montpetit and Foucault 2012; Montpetit, Lachapelle, and Harvey 2016; Montpetit 2016). I intend for this project to reach another level, contributing methodological advancements to content analysis. The project will begin with handwork: collecting the RIAS, coding their content and using other sources for validation. This handwork will not only produce data, it will also serve to automatize the production of these data in the future. RIAS have a systematically structured format and are available on a high-quality electronic support platform, enabling the development of a web crawling tool. A web crawling tool is a program that provides a computer with the necessary instructions to collect systematically relevant information from the vast amount of information available on the web. RIAS are buried among the mass of information contained in the Canada Gazette; a web crawling tool could automatically extract the text needed for this project. The information needed for the validation of the RIAS is just as systematically structured; web crawling can also be used for this purpose. Al-Ubaydli and McLaughlin (2015) measure regulatory strictness using

automatic text analysis. Not only will I adapt their algorithm, I plan on developing my own to measure other aspects of regulatory adaptation, such as intensity. Hand-coding will be instrumental in developing these tools, but by the end of this project, it will no longer be necessary, as the tools will be able to automatically update the dataset long after the project is over. In addition to my own experience, I will be able to count on the technical support of the Centre de recherche interuniversitaire sur la science et la technology (CIRST) for the development of the web crawling and automatic text analysis tools. My hope is that this dataset will become the most important source of comprehensive information on Canadian regulatory activities.

In countries with governments that value technological innovation as much as they value citizen safety and environmental protection, developing appropriate regulations ought to be a priority. However, satisfactory regulatory adaptation is difficult to achieve, if only because regulators are often torn about which expert information to rely on in making decisions. Very little knowledge of this problem exists in Canada; this project will fill the gap by drawing from the wealth of online information on the regulatory activities of the federal government.

RegCan : Rapport méthodologique

Marc Tremblay-Faulkner
Louis-Robert Beaulieu-Guay
Éric Montpetit
Maxime Sainte-Marie
Jean-François Chartier
Vincent Larivière

Document remis à l'École de la fonction publique du Canada dans le cadre de l'appel d'offres 24062-19-096 (Perspectives sur les règlements générés par l'intelligence artificielle)

Centre interuniversitaire de recherche sur la science et la technologie

14 septembre 2018

Rapport méthodologique

1. Extraction des REIR

L'extraction des REIR s'est faite à partir des documents PDF des publications de la Gazette du Canada. Ces documents ont été obtenus à l'aide d'un *web scraper* maison. La banque de PDF créer a été comparée avec celle fournie par TBS pour nous assurer d'être en possession de tous les documents.

Le texte des PDF à ensuit été extrait et converti en un format *data frame* sur le logiciel *R*. Une fois en format texte, le contenu des publications de la Gazette a été traité de manière à isoler les REIR et les textes règlementaires. Ce traitement s'est fait sur la base d'une série d'expressions régulières. En programmation, les expressions régulières sont des chaines de caractères qui mobilisent des syntaxes précises pour générer des opérations.

À l'aide d'une fonction maison, nous avons isolé le texte de langue française de celui en langue anglaise. Cette opération a été rendue possible en discriminant les colonnes en français et en anglais sur la base des espaces blancs les séparant et des retours sur charriots. Nous avons préféré cette approche à l'utilisation d'un algorithme de reconnaissante langagière, car elle s'est avérée beaucoup plus efficace et nous a permis d'obtenir un texte anglophone nettement plus propre. Cette technique n'est cependant pas infaillible et un nombre négligeable de mots en français est demeuré dans le texte.

2. Nettoyage et prétraitement du corpus

Une fois le texte anglophone isolé, nous avons procédé, à l'aide d'expressions régulières, à l'élimination des expressions françaises résiduelles, des caractères non alphanumériques, des entêtes et notes en bas de pages, des adresses URL et des numéros de page. Suivant cette même approche, nous avons isolé certaines sections des REIR de leur texte original (notamment les sections *Consultations* et *Contact*) et nous avons généré une panoplie de métadonnée (ex : date de publication, identifiant, ministère responsable de la rédaction, lois étant associé au REIR, type de règlementation, etc.).

Pour mener à bien nos analyses, nous avons étiqueté le texte des REIR à l'aide d'un étiqueteur morphosyntaxique basé sur l'arbre de décision développé par Schmid (1994). À partir des mots étiquetés, nous avons créé une version lemmatisée du texte des REIR et un antidictionnaire. Le résultat est un corpus de 5843 REIR publiés dans la partie II de la Gazette du Canada entre 1998 et 2017. Ce corpus est épuré des entités nommées (nom de lois, d'organisme, d'agence, etc.), car nous nous intéressons dans ce rapport aux mots utilisés dans les règlements et non aux noms propres ou aux acteurs en jeux.

Figure 1 : Un exemple des effets du nettoyage

A. Le texte PDF

Description

The *Income Tax Act* contains provisions that allow certain supplementary pension plans maintained by the federal and provincial governments to be treated as retirement compensation arrangements (RCAs). The effect of these provisions is to ensure that such plans are subject to the same tax treatment as funded supplemental pension plans established by private sector employers. In particular, amounts credited to the RCA account are subject to a special 50% tax (which is refunded as amounts are paid out of the plan) and amounts contributed by employees are tax deductible. These provisions apply only where the particular plan is prescribed by regulation. There are currently no prescribed plans.

Amendments in 1992 to the *Members of Parliament Retiring Allowances Act* and regulations made in 1994 and 1995 under the *Special Retirement Arrangements Act* established RCAs for Members of Parliament and federal public servants for pension benefits in excess of the registered pension plan limits.

Regulations are required to ensure that these supplemental pension plans are recognized as RCAs for tax purposes.

Accordingly, Part LXVIII of the *Income Tax Regulations* is amended to prescribe, for the purposes of paragraph 8(1)(*m*.2) and subsection 207.6(6) of the *Income Tax Act*, the following pension plans:

- the pension plan established as a consequence of the establishment, by section 27 of the *Members of Parliament Retiring Allowances Act*, of the Members of Parliament Retirement Compensation Arrangements Account, effective January 1, 1992;
- the pension plan established by the *Retirement Compensation Arrangements Regulations, No. 1*, effective December 15, 1994; and
- the pension plan established by the *Retirement Compensation Arrangements Regulations, No. 2*, effective April 1, 1995 (since this plan does not provide for employee contributions, it is only prescribed for the purposes of subsection 207.6(6)).

88

Description

Aux termes de la *Loi de l'impôt sur le revenu*, certains régimes de pension supplémentaires tenus par les administrations fédérale et provinciales sont considérés comme des conventions de retraite. Cette disposition a pour effet d'assurer que ces régimes font l'objet du même traitement fiscal que les régimes de pension supplémentaires par capitalisation établis par les employeurs du secteur privé. En effet, les montants portés au crédit du compte de convention de retraite sont assujettis à un impôt spécial de 50 p. cent (qui est remboursé à mesure que des sommes sont versées sur le régime) et les cotisations des employés sont déductibles d'impôt. Ces dispositions ne s'appliquent que si le régime est visé par règlement. À l'heure actuelle, aucun régime n'est ainsi visé.

Les modifications apportées en 1992 à la *Loi sur les allocations de retraite des parlementaires* et les règlements pris en 1994 et 1995 en application de la *Loi sur les régimes de retraite particuliers* ont eu pour effet d'établir, pour les parlementaires et les fonctionnaires fédéraux, des conventions de retraite prévoyant des prestations dépassant les plafonds des régimes de pension agréés.

Des dispositions réglementaires doivent être prises pour que ces régimes de pension supplémentaires soient reconnus à titre de conventions de retraite aux fins de l'impôt.

À cette fin, la partie LXVIII du *Règlement de l'impôt sur le revenu* est modifiée de façon que les régimes de pension suivants y soient énumérés pour l'application de l'alinéa 8(1)*m*.2) et du paragraphe 207.6(6) de la *Loi de l'impôt sur le revenu* :

- le régime de pension établi par suite de l'ouverture, par l'article 27 de la *Loi sur les allocations de retraite des parlementaires*, du compte de convention de retraite des parlementaires le 1er janvier 1992;
- le régime de pension établi par le *Règlement n° 1 sur le régime compensatoire* qui est applicable depuis le 15 décembre 1994;
- le régime de pension établi par le *Règlement n° 2 sur le régime compensatoire* qui est applicable depuis le 1er avril 1995 (étant donné que ce régime ne prévoit pas le versement de cotisations salariales, il n'est visé par règlement que pour l'application du paragraphe 207.6(6) de la *Loi de l'impôt sur le revenu*).

2000-01-05 *Canada Gazette Part II, Vol. 134, No. 1* *Gazette du Canada Partie II, Vol. 134, n° 1* **SOR/DORS/2000-12**

In addition to the above amendment, the income tax withholding rules in Part I of the Regulations are being amended so that there is no withholding requirement for the special 50% RCA tax at the time that amounts are credited under these prescribed plans.

Alternatives

These amendments are consequential on changes to the *Income Tax Act* and other legislation. Therefore, no alternatives were considered.

Benefits and Costs

It is not anticipated that these amendments will affect government revenues.

Consultation

These proposed amendments were developed in consultation with the Canada Customs and Revenue Agency and Treasury Board Secretariat. They were pre-published on October 16, 1999 in Volume 133, No. 42, of the *Canada Gazette*, Part I, pp. 3058-3060. No comments were received as a result of pre-publication.

Compliance and Enforcement

The Act provides the necessary compliance mechanisms for these Regulations. It allows the Minister of National Revenue to conduct audits and to assess and reassess tax payable, interest and penalties.

En outre, les règles sur la retenue d'impôt énoncées à la partie I du *Règlement de l'impôt sur le revenu* sont modifiées de sorte qu'aucune retenue au titre de l'impôt de 50 % sur les conventions de retraite ne soit à opérer au moment où des montants sont portés au crédit des régimes visés.

Solutions envisagées

Les modifications font suite à des changements apportés à la *Loi de l'impôt sur le revenu* et à d'autres lois. Par conséquent, aucune autre solution n'a été envisagée.

Avantages et coûts

Il est peu probable que les modifications influent sur les recettes de l'État.

Consultations

Les modifications proposées ont été mises au point en consultation avec l'Agence des douanes et du revenu du Canada et le Secrétariat du Conseil du Trésor. Elles ont fait l'objet d'une publication préalable le 16 octobre 1999 dans la *Gazette du Canada* Partie I, vol. 133, n° 42, p. 3058 à 3060. Aucun commentaire n'a été reçu à la suite de cette publication.

Respect et exécution

Les modalités nécessaires sont prévues par la *Loi de l'impôt sur le revenu*. Elles permettent au ministre du Revenu national de faire des vérifications et d'établir des cotisations et des nouvelles cotisations concernant l'impôt payable, les intérêts et les pénalités.

B. Le texte tel qu'extrait

Description The Income Tax Act contains provisions that allow certain supplementary pension plans maintained by the federal and provincial governments to be treated as retirement compensation arrangements (RCAs). The effect of these provisions is to ensure that such plans are subject to the same tax treatment as funded supplemental pension plans established by private sector employers. In particular, amounts credited to the RCA account are subject to a special 50% tax (which is refunded as amounts are paid out of the plan) and amounts contributed by employees are tax deductible. These provisions apply only where the particular plan is prescribed by regulation. There are currently no prescribed plans. Amendments in 1992 to the Members of Parliament Retiring Allowances Act and regulations made in 1994 and 1995 under the Special Retirement Arrangements Act established RCAs for Members of Parliament and federal public servants for pension benefits in excess of the registered pension plan limits. Regulations are required to ensure that these supplemental pension plans are recognized as RCAs for tax purposes. Accordingly, Part LXVIII of the Income Tax Regulations is amended to prescribe, for the purposes of paragraph 8(1)(m.2) and subsection 207.6(6) of the Income Tax Act, the following pension plans: •= the pension plan established as a consequence of the establishment, by section 27 of the Members of Parliament Retiring Allowances Act, of the Members of Parliament Retirement Compensation Arrangements Account, effective January 1, 1992; •= the pension plan established by the Retirement Compensation Arrangements Regulations, No. 1, effective December 15, 1994; and •= the pension plan established by the Retirement Compensation Arrangements Regulations, No. 2, effective April 1, 1995 (since this plan does not provide for employee contributions, it is only prescribed for the purposes of subsection 207.6(6)). 88 2000-01-05 In addition to the above amendment, the income tax withholding rules in Part I of the Regulations are being amended so that there is no withholding requirement for the special 50% RCA tax at the time that amounts are credited under these prescribed plans. Alternatives These amendments are consequential on changes to the Income Tax Act and other legislation. Therefore, no alternatives were considered. Benefits and Costs It is not anticipated that these amendments will affect government revenues. Consultation These proposed amendments were developed in consultation with the Canada Customs and Revenue Agency and Treasury Board Secretariat. They were pre-published on October 16, 1999 in Volume 133, No. 42, of the Canada Gazette, Part I, pp. 30583060. No comments were received as a result of pre-publication. Compliance and Enforcement The Act provides the necessary compliance mechanisms for these Regulations. It allows the Minister of National Revenue to conduct audits and to assess and reassess tax payable, interest and penalties.

C. Le texte filtré et lemmatisé

description contain provision allow certain supplementary pension plan maintain federal provincial government treat retirement compensation arrangement . effect provision ensure plan subject tax treatment funded supplemental pension plan establish private sector employer . particular amount credit account subject special tax refund amount pay plan amount contribute employee tax deductible . provision apply particular plan prescribe regulation . currently prescribe plan . amendment member regulation make special establish federal public servant pension benefit excess registered pension plan limit . regulation require ensure supplemental pension plan recognize tax purpose . accordingly amend prescribe purpose paragraph m. subsection . follow pension plan pension plan establish consequence establishment section member member effective pension plan establish effective pension plan establish effective plan provide employee contribution prescribe purpose subsection . . addition amendment income tax withhold rule part amend withhold requirement special tax time amount credit prescribed plan . alternative amendment consequential change legislation . therefore alternative consider . benefit cost anticipate amendment affect government revenue act provide necessary compliance mechanism . allow conduct audits assess reassess tax payable interest penalty .

3. Analyse de points de rupture

Les figures 2, 5, 7, 15a et 15b ont fait l'objet d'une analyse de points de rupture (*changepoint analysis*). Lorsque les lignes horizontales de ces graphiques sont rompues, cela signifie qu'il y a eu un

changement de moyenne significatif dans une série chronologique. Les changements mesurés sur le nombre de REIR produits, le nombre de mots des REIR et le nombre d'occurrences de la contrainte règlementaire, dans les figures 2, 5 et 15a sont mesurés à partir d'un test des sommes cumulées, qui ne postule pas une distribution normale des données (Csorgo et Horvath 1997). Puisque les distributions des indicateurs de diversité lexicale, de lisibilité et de l'indice de contrainte règlementaire contrôlé pour le nombre de mots épousent une forme normale, les changements mesurés dans la figure 7 et 15b sont mesurés à partir d'un test de rapport de vraisemblance, une méthode d'abord développée par Hinkley (1970). Cette méthode consiste à dériver la distribution asymptotique de la statistique du test du rapport de vraisemblance à partir des changements à la moyenne d'une distribution normale. Toutes les analyses de points de rupture dans ce rapport utilisent la méthode AMOC (*at most one change*), qui cherche à mesurer un seul changement significatif dans une distribution. Pour une présentation complète de la méthode, on peut consulter Killick et Eckley (2014).

4. Diversité lexicale et lisibilité

La diversité lexicale et la lisibilité ont été mesurées à partir du texte brut. Les indicateurs de diversité lexicale de la figure 15, qui sont standardisés sur une échelle de 0 à 1 pour des fins de comparaison, sont tous des indicateurs adaptés de la mesure classique de diversité lexicale TTR (*type-token ratio*), soit le ratio de termes uniques (V) par rapport au nombre total d'occurrences (N). Les formules sont les suivantes :

L'indice C de Herdan : $\frac{logV}{logN}$

L'indice CTTR, le TTR corrigé de Carroll : $\frac{V}{\sqrt{2N}}$

L'index Uber de Dugast : $\frac{(logN)^2}{(logN - logV)}$

L'index de Summer : $\frac{log\ logV}{log\ log\ N}$

Les indicateurs de lisibilité utilisés sont les suivants. L'indice FOG mesure le pourcentage de syllabes dans les mots utilisés et la longueur moyenne des phrases. L'indice Flesch-Kincaid mesure le nombre moyen de mots par phrase multiplé par le nombre de syllabes par mots. L'indice SMOG mesure la racine carrée du nombre de mots multisyllabiques de 30 phrases dans le document. L'indice Dale-Chall diffère des autres et repose plutôt sur le nombre de mots difficiles et compliqués dans le texte (on peut consulter la liste dans Dale et Chall (1995).

Toutes les formules des différents indicateurs de diversité lexicale et de lisibilité ont été implémentées via le package *koRpus* (Mikalke 2017).

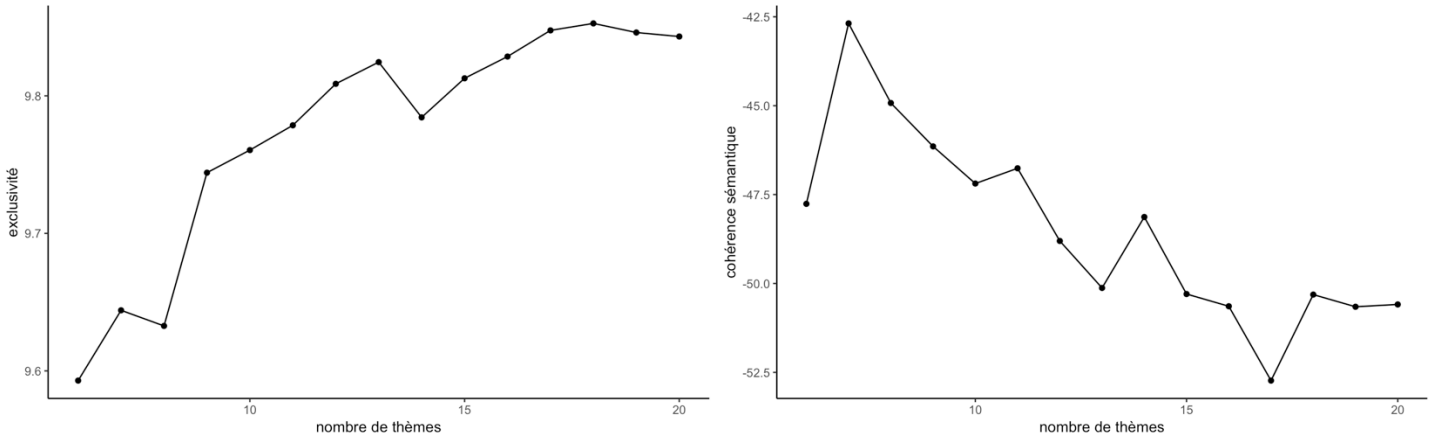
5. Analyse thématique automatique : prétraitement et création du modèle

L'analyse thématique automatisée présentée dans la section 3 du rapport est exécutée sur le texte brut des 5843 REIR du corpus. Le modèle thématique est un modèle probabiliste généré automatiquement dans lequel chaque mot à une probabilité d'appartenir à un thème et chaque document du corpus, ici chaque REIR, peut être composé de plusieurs thèmes à la fois (Roberts et al. 2014).

Le prétraitement du texte suit deux étapes. La première étape est de filtrer le texte brut en supprimant les "mots courants" (*stop words*) de la langue anglaise, puis par l'entremise d'une tokénisation (supprimer la ponctuation, les chiffres et les symboles). La liste de mots courants est celle du package *tm* (Feinerer et Hornik 2012). La deuxième étape consiste à filtrer les mots selon un double seuil de fréquences. Nous avons choisi un double seuil de 5%/50%. Le seuil inférieur supprime tous les mots qui ne se retrouvent que dans moins de 5 % du corpus, tandis que le seuil supérieur supprime tous les mots qui apparaissent dans plus de 50 % du corpus. Après ce traitement, le texte des 5843 documents du corpus est réduit à 1381 termes uniques et 1 042 105 occurrences.

Le modèle peut dès lors être généré sur la base de trois paramètres. Le premier paramètre est l'initialisation. Le type d'initialisation que nous avons choisie est déterministe, c'est-dire qu'elle utilise une factorisation par matrices non négatives (ou décomposition spectrale) des cooccurrences. Cela signifie qu'avec les mêmes paramètres le résultat sera toujours le même (contrairement à une stratégie d'initialisation aléatoire par semence). Le deuxième paramètre est le choix des covariables. Notre modèle a été généré avec trois covariables : le ministère responsable du REIR, l'année-mois où le REIR a été enregistré dans la Gazette du Canada et le nombre de mots du REIR. Le troisième paramètre est le choix du nombre de thèmes. Il n'existe pas de nombre « optimal » de thèmes dans un corpus. Nous avons choisi le nombre de 14 thèmes sur la base de deux indicateurs. Le premier indicateur est la cohérence sémantique respective des thèmes, maximisée lorsque les mots les plus probables d'un thème cooccurrent (Mimno et al. 2011). Cette mesure est insuffisante, dans la mesure où l'on peut se retrouver avec peu de thèmes dominés par des mots très communs. Elle doit donc être complétée avec l'exclusivité, calculée par le taux d'un mot de ce thème divisé par la somme du taux de ce mot dans tous les autres thèmes (Airoldi et Bischof 2016). La combinaison de ces deux indicateurs nous permet de générer un modèle de manière éclairée. La figure 2 montre la courbe des deux indicateurs en fonction du nombre de thèmes du modèle, que nous avons limité à un maximum de 20. Nous avons choisi le modèle de 14 thèmes, car il maximise à la fois la cohérence sémantique et l'exclusivité des thèmes.

Figure 2 : Exclusivité et cohérence sémantique



6. Analyse de spécificités

L'analyse de spécificités a été conduite à partir du texte lemmatisé des REIR. Le texte a été formaté en objet corpus à l'aide du package *quanteda* (Benoît et Nulty 2016). Il a ensuite subi un nettoyage supplémentaire par l'entremise d'une tokénisation (retrait des nombres, de la ponctuation, des symboles) et à été converti en matrice document-mot. Les spécificités ont ensuite été calculées sur les collocations de mots à l'aide du test Chi2 à $p < 0.01$.

7. Analyse de correspondances

Pour mener l'analyse de correspondance, le package *FactoMineR* a été utilisé (Lê et Husson 2008). Comme pour l'analyse de spécificités, le texte lemmatisé des REIR a été utilisé. De plus, un sous corpus a été constitué. Ce sous-corpus est défini en deux temps. D'une part, il discrimine en fonction de l'appartenance d'un REIR à un thème précis, avec un seuil de 75% d'association a un thème comme critère de rétention. D'autre part, seuls les documents faisant partie du 85e percentile des documents les plus contraignants selon l'indice Loughran-McDonald ont été retenus. Ainsi, seuls les REIR dominés par un thème et contraignants ont été retenus, assurant ainsi la validité de nos interprétations. Des 5843 documents du corpus de base, 1537 ont été sélectionnés pour leur association à un thème, et de ce nombre, 231 ont été identifiés comme étant particulièrement contraignants.

Les textes sélectionnés ont ensuite été regroupés de sorte que leur vocabulaire soit amalgamé sous un seul thème, puis une matrice de contingence a été créée de cette association mot(colonne)/thème(rang). Ainsi, l'analyse de correspondances que nous avons conduite compare deux vecteurs-thèmes sur la base de leurs propriétés-mots. Lors de cette opération, un seuil minimal de fréquence de mot parmi les thèmes a été fixé à 15%, de sorte que les dimensions de l'analyse de correspondance ne soient pas qu'entièrement définies par un ou deux documents traitant d'un sujet particulièrement spécifique à un nombre limité de REIR et, par conséquent, peu représentatif de l'activité règlementaire canadienne.

Le tableau 1 présente les valeurs propres de chaque dimension et la variance expliquées par celle-ci, tandis que le tableau 2 affiche le cosinus carré des mots, nous permettant ainsi d'éviter les erreurs d'interprétation. Ces résultats plus détaillés ont guidé, avec la représentation graphique, notre interprétation de l'analyse de correspondance.

Tableau 1 : La variance expliquée par les dimensions

	Valeur propre	% variance	% cumulatif de la variance
dimension 1	0.4405	19.441	19.441
dimension 2	0.3847	16.9768	36.4177
dimension 3	0.345	15.2226	51.6403
dimension 4	0.29	12.7984	64.4387
dimension 5	0.233	10.2833	74.722
dimension 6	0.2114	9.329	84.051
dimension 7	0.1519	6.7017	90.7528
dimension 8	0.1447	6.3853	97.1381
dimension 9	0.0649	2.8619	100

Tableau 2 : Définition des axes selon les mots du corpus

	Dimension 1		Dimension 2		Dimension 3		Dimension 4	
	Mot	Chi2	Mot	Chi2	Mot	Chi2	Mot	Chi2
Les 15 mots au plus fort cosinus^2	notices	0.983365	industry	0.965961	clarify	0.888529	trading	0.939713
	prescribed	0.982382	specify	0.921026	activity	0.829951	selective	0.938775
	foreign	0.981321	exempt	0.919103	relative	0.770561	maintain	0.937274
	bring	0.981156	prescribe	0.873794	note	0.769141	origin	0.932142
	participant	0.981072	publication	0.87125	introduce	0.767562	custom	0.930201
	regard	0.976975	define	0.836946	half	0.764244	unlikely	0.928582
	orally	0.975793	publish	0.815527	everyone	0.760653	inspection	0.925843
	consent	0.973543	construction	0.799785	harm	0.760653	preferential	0.925587
	consistency	0.973239	volume	0.799014	stage	0.756602	uniformity	0.92474
	communication	0.972802	pdf	0.796919	harmonize	0.755452	written	0.92474
	financial	0.972136	manufacturing	0.795188	asp	0.752838	trader	0.924477
	framework	0.971885	compile	0.794984	inventory	0.75137	favourable	0.924477
	phase	0.971062	calendar	0.794965	final	0.749734	advance	0.923178
	statement	0.967964	substance	0.79488	person	0.741833	letter	0.921139
	company	0.967255	correction	0.794784	competition	0.741174	book	0.920526
Les 15 mots au plus faible cosinus^2	withhold	5.15E-09	ultra	1.25E-06	prosperity	2.42E-06	hour	1.95E-08
	employer	1.13E-06	rectify	1.25E-06	inspector	3.13E-06	label	4.11E-08
	tax	1.87E-06	employer	3.43E-06	hazardous	5.22E-06	calculate	3.99E-07
	employee	3.42E-06	inform	4.44E-06	formulation	5.22E-06	acting	4.76E-07
	authority	1.14E-05	average	6.63E-06	promptly	5.22E-06	share	1.99E-06
	declare	1.98E-05	exemption	1.04E-05	locally	5.22E-06	regardless	3.23E-06
	salary	3.13E-05	telephone	2.02E-05	already	5.93E-06	point	3.87E-06
	rate	4.17E-05	liability	2.55E-05	link	9.07E-06	experience	4.31E-06
	submit	4.79E-05	abuse	2.62E-05	fail	1.16E-05	amount	5.09E-06
	ultra	0.000116	streamline	2.73E-05	facilitate	1.40E-05	alone	5.62E-06
	rectify	0.000116	identification	4.31E-05	temporary	1.85E-05	privately	7.49E-06
	streamline	0.000156	skin	4.56E-05	regulatory	1.91E-05	traffic	8.38E-06
	prevent	0.000184	growth	4.81E-05	statute	2.24E-05	calculation	8.96E-06
	documentation	0.000388	salary	6.72E-05	identification	2.40E-05	adjustment	9.20E-06
	operate	0.000482	avoid	7.43E-05	provincial	2.75E-05	open	9.35E-06

8. Analyse de *clusters*

L'analyse de *clusters* est opérationnalisée à partir de la réduction dimensionnelle issue de l'analyse de correspondances. Elle suit la méthode de Ward, basée sur la mesure d'inertie, afin d'établir une règle de classification hiérarchique (Husson et al. 2010). Le nombre de *clusters* pour l'analyse a été défini de manière à maximiser la perte relative d'inertie. Cela nous donne cinq *clusters* distincts.

Le tableau 3 présente les mots les plus près du centroïde de chaque *cluster*. Ceci est mesuré à l'aide de la distance, dans l'espace dimensionnel euclidien, entre le mot et le centre du *cluster*. Ces mots nous permettent d'identifier la nature des *clusters* et de les associer aux types de contraintes précédemment identifiés.

Tableau 3 : Les 15 mots les plus près du centroïde de chaque cluster

Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
Mot	Distance	Mot	Distance	Mot	Distance	Mot	Distance	Mot	Distance
disclose	2.46226	licence	2.21625	origin	4.919220	prescription	4.96049	correction	3.25543
institution	2.45948	scheme	2.19680	preferential	4.889654	treat	4.95710	calendar	3.25184
customer	2.45644	freeze	2.19291	article	4.533419	supervision	4.92476	compile	3.25090
retail	2.45127	acquire	2.19227	uniformity	4.529034	food	4.86290	substance	3.24804
policyholder	2.44365	prohibit	2.19029	written	4.529034	skin	4.84959	commerce	3.24339
card	2.42586	entity	2.18527	shipping	4.495603	guidance	4.61365	nominate	3.19232
closing	2.40642	restricted	2.18416	trader	4.475043	characterize	4.54187	sole	3.19073
interact	2.39776	register	2.17875	favourable	4.475043	essential	4.44984	pdf	3.15083
efficiently	2.39776	authorization	2.17624	certainty	4.441964	scientific	4.32763	quantity	3.11089
timing	2.39321	house	2.17557	tariff	4.308533	distributor	4.32763	incremental	3.05883
electronic	2.37572	nuclear	2.17516	custom	4.218293	severe	4.25018	volume	3.05558
online	2.37462	handling	2.16866	selective	4.192958	medical	4.22479	excess	3.05274
loan	2.37462	dispose	2.16526	abuse	4.161289	growth	4.17416	fair	2.95106
cessation	2.37462	licensing	2.16339	verification	4.130925	advice	4.15876	domestic	2.92741
meeting	2.36214	exclusively	2.15924	letter	4.092925	critical	4.158756	construction	2.83920

Les tableaux 4 à 8 décrivent l'appartenance des thèmes à chaque *cluster*. Une valeur-test plus élevée que 1.96 signifie une appartenance plus grande au *cluster* par rapport à la moyenne des autres thèmes à $p<0.05$.

En observant parallèlement les tableaux 3 et les tableaux de 4 à 8, il est possible de corroborer nos hypothèses précédentes. En effet, le *cluster* 1 est propre au thème de la finance (tableau 4) et, additionnellement, le tableau 3 nous indique qu'il est caractérisé par un vocabulaire financier. À la lumière de nos précédentes analyses, il est possible d'affirmer que ce vocabulaire fait appel à l'imposition de l'État aux institutions financières de pratiques plus transparentes. Ce champ lexical peut lui-même être rattaché à l'une des facettes de la contrainte, soit celle du contrôle des organisations.

Tableau 4 : L'appartenance thématique du cluster 1

Thème	Intern %	glob %	Intern freq	Glob freq	p.value	v.test
finance-crise	**60.0965**	**20.8789**	**15824**	**21180**	**0**	**Inf**
transport-maritime	0.2962	0.4081	78	414	8.00E-04	-3.354
redistribution	1.8685	3.4611	492	3511	0	-17.4578
risques sanitaires	1.6103	3.1318	424	3177	0	-17.588
commerce international	2.0926	4.0516	551	4110	0	-19.9901
transport-sécurité	16.1065	30.1335	4241	30568	0	-Inf
sécurité intérieure	13.4936	27.5014	3553	27898	0	-Inf
matières dangereuses	4.2194	10.2492	1111	10397	0	-Inf

En observant le tableau 3, on peut reconnaitre un vocabulaire associé à l'interdiction dans le *cluster* 2. En croisant cette observation avec le tableau 5, il est possible d'aller plus loin et de lier ce vocabulaire à la sécurité intérieure. Si l'analyse de correspondances nous indique que ce type de contrainte est partagé par les thèmes de la sécurité intérieure et des matières dangereuses, l'analyse par *cluster*, sans nous permettre de définir l'interdiction comme étant exclusive à la sécurité intérieure, nous démontre que ce dernier thème reste dominant dans ce type de contrainte (comme nous l'indiquait d'ailleurs la figure 17).

Tableau 5 : L'appartenance thématique du cluster 2

Thème	Intern %	glob %	Intern freq	Glob freq	p.value	v.test
sécurité intérieure	**56.4659**	**27.5014**	**19841.0**	**27898.0**	**0**	**Inf**
redistribution	3.1305	3.4611	1100.0	3511.0	0	-4.2056
transport-maritime	0.2504	0.4081	88.0	414.0	0	-5.9062
commerce international	3.4578	4.0516	1215.0	4110.0	0	-7.0498
risques sanitaires	1.8413	3.1318	647.0	3177.0	0	-17.9036
transport-sécurité	21.7542	30.1335	7644.0	30568.0	0	-Inf
matières-dangereuses	3.9302	10.2492	1381.0	10397.0	0	-Inf
finance-crise	9.0187	20.8789	3169.0	21180.0	0	-Inf

D'après le tableau 3, le troisième *cluster* est caractérisé par des termes associés aux tarifs et au commerce. Sans surprise, ce *cluster* est propre à la thématique du commerce international. Deux autres thèmes sont aussi plus associés à ce cluster que la moyenne, mais leur composition, à l'intérieur de celui-ci, reste très petite. Il est donc possible d'affirmer que leur contribution à ce *cluster* demeure marginale. Comme pour la thématique du *cluster* 1, celle du *cluster* 3 se rapporte au contrôle des organisations (dans ce cas ci, des marchés).

Tableau 6 : L'appartenance thématique du cluster 3

Thème	Intern %	glob %	Intern freq	Glob freq	p.value	v.test
commerce international	**59.3271**	**4.0516**	**1587**	**4110**	**0**	**Inf**
transport-maritime	**1.1589**	**0.4081**	**31**	**414**	**0**	**5.001**
risques sanitaires	**3.8505**	**3.1318**	**103**	**3177**	**0.0399**	**2.0547**
redistribution	1.7944	3.4611	48	3511	0	-5.158
sécurité intérieure	14.9907	27.5014	401	27898	0	-15.5604
matières dangereuses	1.4953	10.2492	40	10397	0	-18.4399
transport-sécurité	14.5047	30.1335	388	30568	0	-19.0602
finances crise	2.8411	20.8789	76	21180	0	-27.8556

Les *clusters* 4 et 5 peuvent être analysés en bloc. En effet, les deux se rapportent, selon le tableau 3 et nos analyses précédentes, à la notion de gestion du risque par l'émission de normes et de standards. Le vocabulaire du *cluster* 4 reste assez général, mais il peut être lié en grande partie au thème du transport (tableau 7). Le vocabulaire du cluster 5, à la lumière du tableau 8, peut être quant à lui associé au thème des matières dangereuses.

Tableau 7 : L'appartenance thématique du cluster 4

Thème	Intern %	glob %	Intern freq	Glob freq	p.value	v.test
transport-sécurité	**65.3925**	**30.1335**	**15602**	**30568**	**0**	**Inf**
risques sanitaires	**7.1671**	**3.1318**	**1710**	**3177**	**0**	**37.4522**
redistribution	**6.836**	**3.4611**	**1631**	**3511**	**0**	**30.2646**
transport-maritime	**0.7125**	**0.4081**	**170**	**414**	**0**	**7.8685**
commerce international	1.6891	4.0516	403	4110	0	-23.1186
matières dangereuses	4.2416	10.2492	1012	10397	0	-38.0449
sécurité intérieure	9.225	27.5014	2201	27898	0	-Inf
finance-crise	4.5308	20.8789	1081	21180	0	-Inf

Tableau 8 : L'appartenance thématique du cluster 5

Thème	Intern %	glob %	Intern freq	Glob freq	p.value	v.test
matières dangereuses	**50.9934**	**10.2492**	**6853**	**10397**	**0**	**Inf**
risques sanitaires	2.1802	3.1318	293	3177	0	-7.0927
commerce international	2.6341	4.0516	354	4110	0	-9.4219
redistribution	1.7858	3.4611	240	3511	0	-12.3479
transport-sécurité	20.0387	30.1335	2693	30568	0	-28.2996
sécurité intérieure	14.1528	27.5014	1902	27898	0	-Inf
finances-crise	7.6643	20.8789	1030	21180	0	-Inf

Références

Airoldi E.M. et J.M. Bischof. 2016. “Improving and Evaluating Topic Models and Other Models of Text.” *Journal of the American Statistical Association*, 111(516), 1381–1403. https://www.tandfonline.com/doi/full/10.1080/01621459.2015.1051182

Benoit, Kenneth et P. P. Nulty. 2016. "Quanteda: Quantitative Analysis of Textual Data". R package version 0.9. 6-9. https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html

Csorgo, M. et L. Horvath. 1997. *Limit Theorems in Change-Point Analysis*. John Wiley & Sons Inc.

Feinerer, I. et K. Hornik, 2012. *tm: Text mining package. R package version 0.5-7.1*, *1* (8). https://cran.r-project.org/web/packages/tm/tm.pdf

Gunning, R. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Husson, Francois, Julie Josse, et Jerome Pages. 2010. « Principal Component Methods - Hierarchical Clustering - Partitional Clustering: Why Would We Need to Choose for Visualizing Data? », Agrocampus.

Killick, R. et I. Eckley. 2014. "changepoint: An R package for changepoint analysis". *Journal of statistical software*, 58 (3): 1-19. http://eprints.lancs.ac.uk/51975/

Lê, S., Josse, J. et F. Husson. 2008. "FactoMineR: an R package for multivariate analysis". *Journal of statistical software*, 25(1): 1-18. http://factominer.free.fr/more/article_FactoMineR.pdf

Michalke, M. 2017. *koRpus: An R Package for Text Analysis*. R package 0.10-2. https://cran.r-project.org/web/packages/koRpus/koRpus.pdf

Roberts, M. E., B. M. Stewart et D. Tingley. 2014. « stm: R package for structural topic models ». R package version 0.6.21. Software package. http://structuraltopicmodel.com

Mimno D., H.M. Wallach, E. Talley, M. Leenders et M. McCallum. 2011. “Optimizing Semantic Coherence in Topic Models.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’11, pp. 262–272. Association for Computational Linguistics, Stroudsburg. http://dirichlet.net/pdf/mimno11optimizing.pdf