Bonjour tout le monde. Je suis heureux de pouvoir présenter devant vous aujourd’hui.

Il y aura deux parties à ma présentation. D’abord, la première partie, extraire le sens, où je discuterai ma vision de l’analyse computationnelle dans les sciences sociales. Dans la deuxième partie, je vais vous présenter une recherche scientifique en cours liée à ce programme de recherche.

La première partie …

**Extraire le sens**

Le thème central de ma présentation aujourd’hui sur ma vision de la recherche, c’est une réflexion sur les éléments qui viennent distinguer l’analyse computationnelle en science sociale de l’analyse quantitative traditionnelle. Une des difficultés de l’analyse computationnelle des données de grande taille c’est que cette donnée n’est souvent pas crée, n’est souvent pas produite avec l’objectif de l’analyser pour extraire un sens social scientifique. C’est souvent une donnée non structurée, une donnée complexe à analyser. Je vais noter trois points.

Premierement, peut importe la donnée utilisée, on cherche à observer ou collecter, comprendre et expliquer. Par comprendre et expliquer, je peux référer à une analyse descriptive ou causale ou interprétative. Il y a bien sur des nuances et des distinctions à apporter mais il y a plus de similarités que de différences. Pourquoi? On pose une question une question de recherche : par exemple comment est-ce que la consomation médiatique tend à affecter la susceptibilité à la mésinformation, comment les partis politiques utilisent une politique de brèche, par exemple la COVID-19. Il faut commencer par la question de recherche. Où c’est différent avec l’analyse computationnelle, c’est sur, c’est qu’on peut regarder très loin et très proche. Vous voyez un télescope et un microscope. Le téléscope c’est regarder au loin. C’est collecter des données en temps réel dans d’autres pays. C’est faire de l’analyse textuelle sur des textes du 17e siècle. Mais j’ai mis un microscope aussi. Quand on collecte quelque centaines de milliers de données, on peut regarder étudier les sous-groupes. Si on a un groupe social avec 1%, 1% de 200000 c’est 2000, on peut peut-être même etudier l’age à l’intérieur de ce sous-groupe. J’ai participer à des efforts comme ça dans le passé, à Vox Pop Labs où collectaient des données comme ça.

Deuxièmement, j’aimerais faire un point sur la proposition de valeur d'un scientifique des données, d’un analyse des données, d’un scientifique social computationnel, peu-importe comment on l’appel - à mon humble avis - est d'amener les gens à réfléchir à leurs problèmes de manière différente. C’est deux choses. Une manière technique, une manière scientifique. Technique : qu’est ce qui est possible. Bases de données, est-ce que les données existent. Est-ce qu’on peut joindre des données d’une source externe? Scientifique : si on expérimente pas. Est-ce que ça fait du sens au niveau causal ça? Notre échantillon, il représente quelle population?

Troisièmement : on est formés en science sociale. Moi je suis formé en science politique. J’aime inverser des matrices de temps en temps, apprendre sur la gestion des version, github, et je crois qu’il y a beaucoup de valeur à comprendre les équations, comment les base de données dans le nuage fonctionnent, AWS, Redshift, Colab pour l’apprentisage machine quand les modèles sont trop complexes pour etre executés localement. Mais ultimement on est pas informaticiens, on est pas mathématiciens, il est impératif de collaborer avec les gens qui sont des experts dans ça. Nous on est les experts dans la question de la formulation des questions de recherche. J’inclus reproductivilite et transparence la dedans. Ce que je veux dire c’est que pour collaborer efficacement, il faut apprendre les flux de travail (workflow). Il faut comprendre la façon dont ces gens-là, les informaticiens, travaillent. Aussi plus généralement, avoir un bon flux de travail avec les données c’est être transparent avec son équipe, avec le public aussi, si possible. Une science sociale forte c’est une science ouverte, reproductible.

**Plan de recherche quinquennal**

Ça, c’est la vue d’ensemble. Le tableau général. Maintenant, je veux en dire un peu plus sur les trois axes qui vont structurer ma recherche sur les cinq prochaines années.

Le premier axe c’est la science du comportement, la recherche d’opinion publique. Comportements d’un côté, attitudes, préférences de l’autre. Comment l’un affecte l’autre. Une recherche que j’ai en cours avec des collègue au Bureau du conseil privé, on a présenté il y a deux semaines à la conférence BIG DIFFERENCE BC une conférence sur les connaisances comportementales organisée en collaboration avec UBC Saunders School of Business et plusieurs groupes gouvernementaux qui font de la science du comportement, s’intéresse à l'effet des perceptions sociales erronées. Ici on voit le pourcentage de la population qui croient certains énoncés faux mesuré avec un échantillon recruté de façon probabiliste, versus la perception. En gros, quand on leur demande 6% des Canadiens Canadiennes croient que les composantes à basses fréquences des éolienne peuvent causer le cancer. Mais quand on demande aux gens combien de gens pensez vous croient ça, ils disent 35%. 9% des gens croient que la COVID peut causer le cancer, l’estimé de consensus moyen c’est 36%. Similaire pour d’autres questions. Ensuite, on a remarqué qu’il y avait une corrélation claire entre les gens qui surestimaient ces valeur de consensus, et ceux qui croyaient les énoncés faux. L’hypothèse c’est que le recalibrage des perceptions du consensus social des individus peut créer un changement évolutif grâce à l'influence des normes sociales. À tester expérimentalement.

Dans une autre recherche, on a trouvé une association assez forte en l’utilisation de Reddit (déclarée par sondage) et la résistance à la mésinformation. La question, évidemment, c’est est ce que ça pourrait être causal. Nous ne pouvons pas l'exclure. On a tester des dizaines de spécifications différentes, incluant avec l'appariement des scores de propension. Il faudrait probablement un design expérimental pour conclure définitement. Mais même si c’est juste une association, il y a quelquechose de fascinant à observer que beaucoup de gens sur Reddit cherche une expérience différente. Reddit c’est un ecosystème plus partificatpatif, potentiellement toxique aussi, mais certainement avec un caractère communautaire et décentralisé.

Ça c’était science du comportement. Maintenant, deux mots sur l’opinion publique. Une des valeurs ajoutées c’est de trouver des façons novatrices de visualiser, de communiquer. Ici avec des collègues de McMaster, j’étudie le transfert du vote en 2019 et 2021 entre les différents partis. On était particulièrement intéressé à comprendre qui sont ces conservateurs avant transféré vers le Parti Populaire. Dans un modèle de régression logistique multinomiale, on montrait que parmi ceux qui étaient contre la vaccination obligatoire l’effet spécifique était immense sur le vote pour le Parti Populaire. Pas tant sur les autres. Enfin, dans un article récemment publié avec des collègues, on utilisaient l’analyse expérimentale conjointe pour examiner les préférences des Britanniques en lien avec le partenariat Royaume-Uni Europe suite au Brexit. Un des points de l’article c’est que les citoyens britanniques sont favorables à une relation plus souple avec l’Europe. Ici on utilisait une visualisation novatrice pour communiquer ces préférences. L’analyse computationnelle c’est tout le flux de la collecte à la visualisation.

Développements d’outils méthodologiques

Une des frustrations qu’on a au gouvernement c’est la question de la pondération des sondages. Toutes les firmes de sondages pondèrent avec des variables différentes. D’un point de vue scientifique c’est un peu choquant. Il y du travail qui a été fait ici par William Poirier et Anne-Sophie Charest. C’est un travail qui mériterait d’être publicisé davantage et étendu à la question de l’inférence basée sur les modèles. C’est une question technique qu’on peut discuter à un autre moment mais avec la montée en force des échantillons non-probabiliste et semi-probabiliste, c’est une question qui mérite d’être posée. Extraire le sens des données c’est aussi corriger les biais dans les échantillons.

Enfin, le troisième axe sur lequel je veux travailler c’est l’analyse computationnelle pour le texte.

Je vais donner trois exemples. Une question classique en analyse des données en politique canadienne c’est à quelle point faite vous confiance à tel ou tel institution. Pas du tout confiance, un peu confiance … Totalement confiance. Une question que j’explore présentement avec des collègues chercheurs c’est la question de la confiance mesurée avec les questions ouvertes de suivi. Vous avez répondu exemple « Pas du tout confiance » dites nous pourquoi. On utilise un modèle BERT-phrase (sentence-BERT) un modèle transformeur, comme les Grand modèle de langage. Le modèle vient encoder/positionner chaque phrase dans un espace à plusieurs dimensions. Ensuite on peut extraire de façon non supervisée certain thèmes, qui est proche.

Les modèles de sujets fonctionnent en groupant les mots qui apparaissent fréquemment ensemble, supposant que ces groupes représentent des sujets cachés au sein des textes. J’ai deux tels projets en ce moment. Un lié à la présentation scientifique qui va suivre et un lié au discours journalistique au Québec.

Une autre avenue super intéressante en analyse computationnelle du texte c’est le changement sémantique. En gros, le sens des mots change, et par extension le sens des concepts politiques change aussi. Dans la deuxième partie aujourd’hui je vais vous présenter un exemple.

Je vais terminer la première partie de ma présentation sur cette diapositive résumé. Deux photos. Le point est simple, avant si on faisait du sondage c’était par téléphone. Les taux de réponses étaient assez haut. On pouvait prétendre que l’échantillon était probabiliste. Aujourd’hui, beaucoup des collectes de données se font par ordinateur. Il peut y avoir une étape de plus pour corriger le biais. Aussi, l’information textuelle (discours des politiciens, contributions, données spatiales) est plus facilement accessible. Sur l’ordinateur.

Dans un sens l’analyse computationnelle c’est, comme je l’ai dit, observer, mesure, expérimenter. Dans un sens c’est vraiment la continuité des approches qui ont précédé l’analyse computationnelle. Mais certaines questions deviennent plus importantes. Je pense à une conférence récente de collègues à Montréal, Erick Lachappelle, Yoshua Bengio et trois étudiants. Ils ont générer des images personnalisées visant à aider les citoyens à visualiser les effets possibles du changement climatique là où ils vivent; démontrant des niveaux plus élevés d'inquiétude, de plus grandes intentions comportementales. On vit dans une époque où les choses bougent vite. On peut récolter des données à grande échelle, génrer des images personalisées, par exemples les deux images ici ont été générees avec stable-diffusion-v1-5, un modèle code source ouvert, un modèle latent de diffusion texte-image qu’on peut exécuter dans le nuage parceque ça prend des Processeur graphique.

La proposition de valeur de l’analyse computationelle c’est de faire le pont entre ces disciplines là. Collaborer, réfléchir aux enjeux éthique des biais, poser les bonne questions.

FIN

Ma conférence scientifique s’intitule l’évolution du concept de nation au Canada

Avant de vous présenter le plan, l’approche théorique, les données la méthodologie les résultats, j’aimerais simplement résumer l’argument. Je pense que ça peut être utile pour voir un peu la vue d’ensemble. Cinq points. D’abord qu’est ce que je fais c’est une étude du changement sémantique du concept de « nation » au Canada. On va discuter qu’est-ce que la nation. Nation-État-Canada. État plurinationaux. Québec. Canada. Premières nations. 2 Pourquoi j’étudie ça. La thèse implicite c’est que on peut en apprendre beaucoup sur la construction nationale en étudiant la façon dont le discours sur la nation a changé. 3. Par le passé, on pouvait étudier certains textes de façon qualitative. L’idée ici c’est qu’on peut étudier de façon systématique tous les discours des politiciens à la chambre des communes qui traitent du concept de la nation. Qu’est ce que je trouve? **REVENIR**

**Plan**

**Concept de Nation 1**

Le Canada se caractérise par une mosaïque culturelle exceptionnelle, reflétant la coexistence de diverses communautés, allant des groupes autochtones aux immigrants de diverses origines. Cette diversité se manifeste dans la pluralité linguistique, avec deux langues officielles, l'anglais et le français, ainsi que dans une multitude de langues parlées par les différentes communautés ethniques. Les régions canadiennes, chacune avec ses propres traditions et son histoire, contribuent à une richesse régionale distincte, renforçant le caractère unique du paysage culturel et social du Canada.

Bien que le Canada embrasse une grande diversité, il existe un sentiment d’unité nationale, soutenu par des symboles communs, des valeurs partagées telles que la liberté, la démocratie et le bilinguisme, ainsi que par des institutions nationales. Cette unité est parfois mise à l'épreuve par des tensions régionales ou linguistiques, comme celles entre le Québec francophone et d'autres provinces majoritairement anglophones, mais elle reste un pilier central de l'identité canadienne. Le gouvernement canadien a activement œuvré à promouvoir l'unité nationale à travers des politiques de bilinguisme et de multiculturalisme, cherchant à intégrer harmonieusement la diversité dans un cadre national cohérent.

L’identité collective canadienne est fortement influencée par le concept de multiculturalisme, qui reconnaît et célèbre la diversité des cultures et des peuples qui composent la nation. L'identité collective au Canada est également marquée par l'engagement envers des principes de paix, d’égalité et de respect de l'environnement, reflétant une vision commune du pays et de son rôle sur la scène internationale.

**Concept de Nation 2**

Dans le contexte canadien, le discours et les actions des leaders politiques jouent un rôle crucial dans la formation et la définition de l'identité collective. Les déclarations et politiques des élites politiques peuvent soit renforcer l'unité nationale, soit exacerber les divisions. Par exemple, les discours sur le bilinguisme et le multiculturalisme, fréquemment abordés par les politiciens canadiens, contribuent non seulement à façonner les politiques publiques, mais aussi à définir l'identité nationale comme inclusive et diverse. Les réactions des élites politiques face à des enjeux majeurs, tels que les relations avec les peuples autochtones ou les débats sur l'immigration, ont un impact significatif sur la perception publique de ce qui constitue l'essence de l'identité canadienne et sur la manière dont les citoyens s'identifient au concept de la nation.

**Méthode**

La théorie de la signification fondée sur l’usage, développée par Ludwig Wittgenstein, postule que la signification des mots est déterminée par leur usage dans le langage et non par une correspondance avec des objets ou des idées spécifiques. Selon cette approche, comprendre un mot implique de connaître les contextes et les pratiques linguistiques dans lesquels il est employé, plutôt que de se référer à une définition fixe ou à une essence intrinsèque.

les mots qui apparaissent fréquemment dans des contextes similaires ont tendance à avoir des significations proches ou liées. Cette approche met l'accent sur le fait que la signification et la fonction d'un mot peuvent être déduites de ses modèles de cooccurrence avec d'autres mots dans le langage naturel.

La diapositive présente la méthode de vectorisation des mots, également connue sous le terme anglais "word embeddings", où les mots sont représentés par des vecteurs dans un espace multidimensionnel. Sur le graphique, on observe comment des mots distincts comme "pays", "nation", "chat", "chien", "bmw" et "toyota" sont positionnés dans cet espace, suggérant des relations sémantiques basées sur leur proximité vectorielle.

Les propriétés arithmétiques de ces vecteurs, illustrées par un graphique qui montre des mots comme "roi", "reine", "homme", "femme", "marcher" et "marche", démontrant comment des relations sémantiques et syntaxiques peuvent être modélisées mathématiquement dans cet espace vectoriel.

On obtient ces vecteur par optimisation. Par exemple disons que nation et voted apparaisent 78 fois ensemble en prenant le produit scalaire de ces vecteurs on viendra approximer ce chiffre. On répète ça pour toutes les cooccurences, pour toutes les paires de mot.

L’important, disons, c’est que chaque mot est positioné dans l’espace. Il est quelquepart. La faiblesse de cette méthode, évidemment, que chaque mot a une position. « Nation » dans le contexte de nation Québécois, ou Canadiene, ou Nation en général. C’est toute la même position.

Plusieurs solutions existent. Si on veut les position à chaque décennie. Nous pouvons prendre un sous-ensemble de données pour chaque décennie. Exécuter le modèle. Il y a quelques problèmes avec ça. 1 c’est long, mais 2 et c’est plus important on peut manque de données. Rodriguez, Spirling & Stewart proposent des vectorisation à la Carte. C’est l’adaption d’un article de 2018 de science informatique. L’idée est super simple. On prend la moyenne des vectorisations des mots autour du mot qu’on étudie. Il y a une transformation additionnelle, en fait on prend le vecteur et on le multiplie par une matrice de pondération qui vient diminuer l’influence des mots très fréquent et augmenter l’influence des mots peu fréquent, ce qui est souvent fait. Donc, l’idée c’est qu’au lieu d’avoir une position dans l’espace pour nation, on a une position pour chaque instance de nation. On peut maintenant prendre la moyenne des position de nation sur la période 1910-1920, 1920-1930, 2000-2010, par parti, etc. Tout ça est descriptif. Dans le même article, les auteurs proposent une façon de tester statistiquement si la signification a changé. L’intuition est la suivante. D’habitude on modèle une variable en fonction d’une autre : disons la satisfaction envers la démocratie en fonction de, chez pas, l’idéologie, ou si vous avez gagné vos élections. Ici on n’a pas une variable dépendante. On en a 300. On utilise une régression multivariée. Chaque coefficient, au lieu d’en avoir un, on en a 300. On vient prendre la distance euclidienne. On appelle ça la norme. La norme comme telle n’est pas interpretable. Par contre on peut les comparer entre elles. C’est ça qu’on fait. Quand je vais montrer les résultats ce sera plus clair. Mais de façon informelle, on peut dire : le concept de nation mesurer semantiquement a change 3 fois plus sur 1900-2020 que sur 1900-1950.

Les données proviennent du Canadian Hansard Dataset. Ils ont été collectés par un groupe de politiologuqes, informaticiens et historiens. L’initiative s’appelle LIPDAD (Linked Parliamentary Data Project). 4.8 M textes. 51000 avec la mention nation. Certains en ont plus qu’une mention, donc 86,000 textes. Dans un exercice comme celui-ci il y a des choix à faire. Par exemple, je voulais mettre l’accent sur la « nation » donc « nations » avec un s est exclu. Nations est souvent en lien avec les nations du monde, donc pas la nation au Canada. Évidemment, plusieurs textes mentionnent les nations autochtones. En pratique souvent « nous discutons aujourd’hui des nations …listes … » ces cas sont exclus. Par contre si le texte mentionne « La Nation anishnabe » là il est inclus. Dans un sens c’est un choix théorique. L’idée est de considerer l’évolution du discours sur la nation. Ça peut être la nation canadienne, québécois, ça peut être une nation autochtone, mais c’est une nation au singulier. Étudier « nations » c’est un autre concept, qui mérite aussi de l’attention, mais c’est un autre concept. J’y reviendrai à la fin. Les mots sont mis en minuscule, les mots vides (stop words en anglais) sont retirés. On a mis ensemble les ngrames les plus importants. La méthodes est automatisées pour les identifier (bigrames et trigrames) mais avec mon codage manuel. Le tout est mis en corpus. Le corpus est accompagné de quelques variables, que je vais maintenant décrire.

Données

D’abord la première méta donné associée aux textes est l’année. C’est assez clair qu’il y a plus de textes récents que au début du XXe siècle. Il y a une année où il y en a plus, 1964, mais c’est pas une donnée aberrante. C’est seulement environ 25% plus haut que les grosses années.

Celui-ci est assez frappant. Il n'y a pratiquement pas de discours de femmes, et ils sont tous récents. C'est logique. Mais bien sûr, c'est quelque chose qu'il faut garder à l'esprit en termes d'analyse de genre. Nous ne disposons tout simplement pas de beaucoup de données.

La distribution est plus équilibrée pour les partis. Certains regroupements sont faits. En particulier 7 partis sont groupés sous la bannière conservatrice. Le parti conservateur actuel. Les progressistes conservateurs, l’ancien parti conservateur (1942), l’Alliance, le parti réformiste, le parti unioniste, gouvernement national. On a testé différentes catégorisation et en général c’est stable. J’y reviendrai.

La validation en analyse textuelle automatisée est cruciale pour quelques raisons. La validation permet de s'assurer que les résultats obtenus par l'analyse automatisée sont fiables et reflètent correctement les données. Dans certains contextes c’est plus facile. Par exemple, lorsqu’on fait une classification supervisée ou non-supervisée de thèmes on peut coder les thèmes à la main puis ensuite comparer les codages manuels. Ici, par contre, on vient situer les mots concepts dans un espace multidimensionnels donc la comparaison est différente. Je propose deux validation une interne et une externe. Celle interne fonctionne très bien. On vient comparer la proximité de la vectorisation GloVe sur tout le corpus avec la vectorisation à la carte moyenne. 0.966 il n’y a pas de problème. La second validation est une validation externe. On prend des mots (ici nation, immigration, et quebec) et on compare les mots les plus proches sur le modèles entrainé par les chercheurs américains créateur du Glove Standard et nos vectorisations à la carte. À première vue, c’est clair que les mots les plus proches du Glove Standard sont plus générique, nation est proche the american, continent, largest, states. En utilisat la vectorisation à la carte les mots les plus proches dans notre corpus parlementaire snt country (aussi) world, must, canada, mais aussi become future past strong. C’est le genre de concepts que les politiciens utilisent. Notre nation dans le future sera forte, toujours … Pour immigration et quebec. Les mots les plus proches sont bons aussi dans le cas du Glove Standard. C’est juste moins spécifiques. Deportation, citizenship, illegal, asylum. On parle ici d’un context américain. Dans le cadre des discours de politiciens, politiciennes au Canada, les mots les plus proches sont immigrants departement (ministères), citoyenneete, loi, application, ministre. Le contexte est celui d’un discours politique, plustot que d’un discours général, américain. Avec le Glove Standard les mots les plus proche de quebec sont ontario, canada, montreal, manitoba, etc. Ça fait du sens. C’est descriptif. Avec la vectorisation à la carte c’est int´ressant les mots les plus proches sont province, majorité, minorité, ontario, gouvernment fédéral, p[rovincial, federal, cosnittution. Le sens de ces trois mots dans nos vectoridations à la carte est situé dans un réseau de référents plus politiques alors que le sens des mots est plus générique dans le modèle original du gant. C’est une validation plus subjective, mais pour l’instant on juge cela suffisant.

Ici, on a des résultats descriptifs. Par décénie, quels sont les mots les plus proches sémantiquement du concept de nation. Sur 1900-1910, empire, to-day (une analyse etymologique révèle que le mot est écrit en deux mots jusqu'au 16e siècle, après quoi il s'écrivait généralement avec un trait d’union jusqu'au début du 20e siècle; donc ce mot là est sémantiquement proche mais on peu croire que c’est un artifice de changement d’orthographe), british, england, great britain, empire est le mot sémantiquement le plus proche pour les trois première décenies ensuite il glisse hors du top 10. On observe que certains mots, comme "empire" au début du 20e siècle et "canadians" vers la fin, apparaissent fréquemment près de "nation". Guerre est dans le top 10 deux fois, entre 1910-1920 et entre 1940-1950. Au milieu 1940-1960, on voit que nation est proche sémantiquement de great greater greatest. Croire. World. À partir des années 1980, il y a un deuxieme tournant. C’est en 1980-1990 qu’on voit recognize/reconnaitre pour la première fois. Reconaitre reste dans les plus proches voisin entre 1980-2020 pour les 4 dernières décennies. D’autre plus proches voisins : société, future, quebec, fort, fière, communaturé, quebec.

C’est un exercice nécessairement qualitatif, et je tiens à le dire, nécessairement superficiel vis-à-vis ce que la théoricienne ou le théoricien politique peut produire. En même temps, c’est assez clair. 1900-1940, la nation c’est associé, dans les discours des politiciens, et ici je dirai pas politicienne parceque y’en avait pas, ou une je crois Agnes MacPhail. Puis on a un tournant vers le monde, avec des concepts universalistes. « Great » « Grand » le monde. Puis à partir des années 1980, on a un tournant pluraliste. Reconnaitre, dans le sens de reconaitre la nation. On parle de culture. Quebec. Superficiel dans un sens, mais fascinant tout de même.

Donc, chaque mot a une position à chaque décenie. On peut regarder les 10 mots qui se raprochent le plus de nation. Là encore, c'est éclairant et assez clair. Autochtones, Bloc (Bloc Quebecois).

Est-ce que ces différences sont statistiquement significative. Oui. Si on prend la période 1900-19010 comme période de référence on voit que le glissement sémantique est graduel. Chaque effet est par rapport à la période de référence 1900-1910 (c’est pour ça qu’il n’y a pas d’intervalle de confiance et que le point est à 0. En gros, on peut dire que le concept a dérivé autant entre 1900 et 1950 qu'entre 1950 et aujourd'hui. Encore une fois, la magnitude des coefficients ne peut être interprété. Ce qui peut être interprété c’est la distance relative.

La méthode permet une grande flexibilité. Ici, on explore la différence par parti. En fait, ici on regarde l’effet par parti, par décennie. Puisque le changement sémantique dans le temps est très important, c’est innaproprié de mettre ensemble les discours libéraux du début du xxe avec ceux de la dernière décennie. Ici, ça prend un parti de référence : ce parti est le parti libéral. On peut noter 3 points. Premièrement, la norme du coeffcient sur les discours conservateurs est stable à travers les années et assez faible, autour de 1925. En comparaison, entre 1900-1910 et 1910-1920, la distance est plus grande. Deuxìemement ça saute à l’œil, le parti qui saute à l’œil, pas de surprise, c’est le Bloc. Pourquoi? Parcequ’il parle de la nation Québécoise. Troisiement, plus d’hétérogénéité avec les autres partis et Credit (en vert).

On peut faire le même exercice avec la différence sémantique homme femmes. Toujours plus bas que 0.5, et il faut garder en tête que quand la taille de l’échantillon augmentes (dans les dernière années) la différence diminue. Je ne le montre pas ici mais j’ai également réalisé une analyse des termes les plus proches de nation dans les discours des hommes et des femmes. Certain termes liées à la culture et à la société sont plus proche, mais c’est vraiment pas évident. C’est en autre pourquoi je ne m’y attarde pas plus ici.

Enfin, l’analyse par régression proposée permet de mesure l’influence relative des différentes variables. Par exemple, si tout le changement sémantique récent était du aux discours nationalistes du Bloc, l’effet observé de la temporalité serait plus petit. Résumons ainsi : l’effet observé où le sens du terme nation perdure dans le temps demeure. Comme attendu, l’effet du Bloc (vs les libéraux en référence) est important. Ils ne parlent simplement pas de la même nation.

La dernière analyse que je présente est une analyse en composantes principales. Cette analyse permet au statisticien de résumer l'information en réduisant le nombre de variables. Les nouvelles variables qu’on identifient sont les variables les plus décorélées, les plus différentes. Ces cinq composantes (dimensions si on veut) sont les plus importantes. Dans les données, ce sont les dimensions sur lesquelles on trouve le plus de variation

Résumé

Le changement sémantique est clair. Au début du siècle la nation est sémantiquement proche d’Angleterre, Grande Breatgne, Empire. Dans les 40 dernières années, on parle du futur (implicitement des enfants) de reconaissance, de communauté. Au milieu du siècle, de grandeur, du tout, de croire et de réalizer. Statistiquement, ces différences sont statistiquement significatives et demeure après l’inclusion du genre des candidats et du parti.

Limites

C’est une approche nécessairement plus superficielle qu’une analyse qualitative profonde. Mais elle a d’autres avantages. On peut étudier le corpus dans son enesemble. On peut trouver des tendances statisiques.

Directions futures