

Un service Web pour l'analyse de la cooccurrence

William Martinez ¹, François Daoust ², Jules Duchastel ²

¹ ILTEC – Instituto de Linguística Teórica e Computacional – Lisbonne – Portugal

² UQAM – Centre ATO – Québec – Canada

Résumé

Les analyses de cooccurrences font partie des outils classiques de la statistique textuelle. Un calcul de cooccurrence constitue aussi un objet documentaire pouvant être repris et consulté de façon relativement autonome par rapport au corpus sur lequel il s'appuie. Mais, cela suppose que les données nécessaires et suffisantes à la production du calcul prennent la forme d'un document bien établi. Nous présentons un modèle de document de cooccurrence qui utilise une syntaxe XML-TEI. Dans ce document, les contextes se présentent sous forme d'empans textuels dont les frontières sont définies par des pointeurs sur des mots dans le document source. Nous présentons aussi une implantation du programme de calcul qui emprunte la forme d'un service Web opérant sur ces documents. On utilisera notre modèle pour comparer deux algorithmes de cooccurrence appliqués à un même corpus et dont les résultats seront cumulés dans le document de cooccurrence.

Abstract

Cooccurrences' analysis is one of the standard statistical textual tools. Cooccurrences's computation may also be considered as a documentary object, which can be referred to independently from the corpus itself. But, this requires that the data needed for computation be properly translated into a well-established document. We submit a model for establishing this kind of document, using an XML/TEI protocol. In the document, the context of a word will be presented as a textual segment delimited by markers set on tokens in the original document. We are also proposing an implementation of the computation program in the form of a Web service applied on the XML/TEI document. Finally, we use our model to compare two different algorithms on one corpus for which the resulting data will cumulate in the cooccurrences' document.

Keywords : co-occurrences, web service, TEI

1. Introduction

Les analyses de cooccurrences font partie des outils classiques de la statistique textuelle. Plusieurs stratégies statistiques ont été proposées pour évaluer les résultats des calculs de cooccurrence (voir, entre autres, la recension dans Lebart et Salem, 1994). Peu de logiciels, cependant, donnent accès à plusieurs méthodes permettant des comparaisons sur corpus. De nouveaux paradigmes de programmation et de formalisation de formats de documents d'annotation permettent de reposer ce problème d'accès dans une perspective très ouverte basée sur un principe d'interopérabilité des traitements appuyé sur un modèle documentaire cohérent.

L'objectif de cette communication est donc triple ¹. Il s'agit d'abord de faire état de ce nouveau contexte informatique et documentaire et de formuler une proposition concrète appliquant ces

¹ En raison des contraintes d'espace imposées pour les Actes des JADT, il n'est pas possible d'exposer en détail

nouveaux paradigmes au problème du calcul des cooccurrences. Cette proposition a mené à une implantation informatique permettant de comparer deux méthodes statistiques sur un même corpus, l'une faisant appel à la loi binomiale et l'autre aux *spécificités* dépistées par la loi hypergéométrique. Après avoir exposé les principes de ces deux méthodes, nous montrons les résultats de leur application sur un corpus politique sur le thème du discours constitutionnel canadien.

2. Contexte informatique, documentaire et discursif

Dans son sens le plus général, on peut voir le service Web comme l'implémentation logicielle d'une ressource, identifiée par un *URI (Universal Resource Identifier)*, accessible en utilisant les protocoles Internet :

«Web services provide a standard means of interoperating between different software applications, running on a variety of platforms and/or frameworks. Web services are characterized by their great interoperability and extensibility, as well as their machine-processable descriptions thanks to the use of XML. They can be combined in a loosely coupled way in order to achieve complex operations. Programs providing simple services can interact with each other in order to deliver sophisticated added-value services» (<http://www.w3.org/2002/ws/Activity>).

Cette définition générale des services Web par le W3C met en évidence la notion d'interopérabilité d'applications logicielles indépendamment du langage utilisé pour les programmer, de l'architecture des calculateurs et des systèmes d'exploitation gérant ces calculateurs. Les requêtes, avec leurs données et leurs résultats, peuvent s'exprimer en utilisant diverses syntaxes concrètes, mais elles ont en commun de circuler à travers un réseau utilisant les standards du Web, en particulier l'URI (*Universal Resource Identifier*) qui correspond à l'idée courante d'*adresse Internet*. Un des modèles d'implantation de services Web emprunte l'architecture REST (*Representational State Transfer*), d'après le terme inventé par Roy Fielding en 2000. Cette architecture fait largement appel au protocole HTTP couramment utilisé par les internautes dans leur navigation quotidienne. Elle a notamment pour particularité que chaque requête est *sans états*, dans le sens qu'elle ne dépend pas d'un état antérieur de l'interaction et qu'elle repose uniquement sur les données transmises, lesquelles peuvent faire référence à des ressources existantes accessibles par URI. C'est ce modèle que nous avons employé dans le prototype fonctionnel que nous présentons ici.

La programmation d'applications sous forme de services Web n'implique pas qu'il faille toujours utiliser le Web pour accéder au service. Par exemple, au sein d'une grappe de calculateurs, un réseau interne à haut débit pourrait être utilisé pour fédérer des applications sans dépendre, au-delà du protocole de communication, d'environnements logiciels et matériels particuliers. On peut aussi faire tourner des services Web sur un même ordinateur qui s'autoréférence par son adresse IP locale. Donc, le service Web est d'abord une architecture de développement qui mise sur l'ouverture et qui peut se déployer à diverses échelles, y compris en mode autonome sur le poste de travail de l'utilisateur.

Les analyses statistiques de cooccurrence se prêtent bien à ce type d'architecture. D'abord, le modèle de calcul peut être défini de façon formelle et autonome en conformité avec l'architecture REST. On a souvent l'habitude de présenter la sémantique de la cooccurrence statistique à partir de modèles probabilistes illustrés par des exemples de tirages aléatoires de boules de différentes couleurs ou portant divers numéros. Dans le cadre de l'analyse textuelle, c'est souvent le mot

les aspects techniques entourant chacun des trois volets du projet. On en trouvera cependant l'exposé dans une version longue de l'article consultable en ligne à l'adresse « <http://www.atonet.net/publications> ».

qui sera pris comme l'équivalent de la *boule* et la *pige* du tirage prendra la forme d'empans textuels, par exemple la phrase, rassemblant un ensemble de mots. Pour rester dans des termes généraux, nous avons convenu de définir notre modèle de données comme étant constitué de deux ensembles : un ensemble d'objets, avec leur description, pouvant se retrouver dans un ensemble de contextes. Ce qui nous intéresse, c'est de savoir quels sont les objets qui, d'après un certain modèle probabiliste, occurrent ensemble dans les contextes avec une fréquence difficilement explicable par le hasard. Cette fréquence de coapparition peut être beaucoup plus ou beaucoup moins élevée que prévue par le modèle probabiliste. Sur ce double ensemble de données, on peut donc procéder à plusieurs tests en faisant varier le modèle probabiliste, les paramètres du modèle, de même que le mot pôle par rapport auquel seront identifiés les objets dont la cooccurrence positive ou négative sera jugée significative en fonction du modèle statistique choisi.

Sur la base de ce modèle abstrait, on doit déterminer une syntaxe concrète pour représenter les deux ensembles de données, les paramètres de la requête et les résultats obtenus. En accord avec la proposition d'ATONET de format XML-TEI pour l'échange de corpus annotés (Daoust et Marcoux 2006), nous utiliserons les recommandations de la *Text Encoding Initiative* (TEI) pour représenter en XML les données de départ et l'enrichissement amené par les résultats de l'analyse de cooccurrence. Même si les données manipulées par le service Web sont principalement de nature numérique, le choix de considérer ces données comme faisant partie d'un texte est congruent avec notre proposition de modèle documentaire de dépôt de données adapté à la constitution de corpus de recherche (Daoust et al., 2008). Ce modèle propose de publier l'annotation analytique sous la forme de documents numériques portant sur d'autres documents considérés comme primaires par rapport au document d'annotation. Dans son aspect documentaire, chacun des documents est une ressource possédant son URI. Les documents peuvent être décrits par des fiches de métadonnées pouvant être récoltées par des moteurs de métadonnées gérant le descriptif du document, à partir d'un noyau *Dublin Core*. On peut aussi décrire et publier les relations entre documents sous la forme de relations RDF (*Resource Description Framework*, W3C 2000) pouvant également faire l'objet de requêtes.

Cette mise en relation à l'échelle du document se superpose à une mise en relation beaucoup plus fine entre le document d'annotation et des éléments dans le document faisant l'objet de l'annotation. Pour établir ces relations, nous faisons appel aux structures de pointage recommandées par la TEI. L'utilisation d'éléments syntaxiques communs entre les documents *primaires*, sujets de l'annotation, et les documents *secondaires*, annotant les documents primaires, permet de rendre compte du mouvement réel de l'intertextualité dans lequel un texte commentant un autre texte pourra lui-même devenir objet de commentaires, d'analyses et d'annotations. Sous cet aspect, le document qui sera soumis à l'analyse de la cooccurrence demeure un texte sur un texte, même si son contenu emprunte davantage une forme numérique que prosaïque.

Le document de cooccurrence a donc, en quelque sorte, un double statut. Du point de vue du service Web calculant la cooccurrence, il est autonome et autosuffisant. Pour l'interprétation statistique de la cooccurrence, le document *secondaire* est nécessaire et suffisant. Il peut faire l'objet d'analyses successives, avec des algorithmes différents. Il peut être comparé à d'autres documents de cooccurrence établissant des relations entre *objets* sur la base de *contextes* différents pouvant même référer à d'autres textes. Ainsi, pourra-t-on constater les différences entre réseaux de cooccurrents construits à partir de corpus différents.

D'un autre côté, la validation interprétative des résultats de la cooccurrence, du point de vue de leur portée discursive, exigera probablement un retour aux sources textuelles dans lesquelles les objets comptés prennent leur sens à l'intérieur de contextes inscrits dans une textualité

concrète déployant de multiples réseaux de relations. Voilà pourquoi il est essentiel de prévoir, dans le document de cooccurrence, tous les mécanismes permettant de référer aux éléments du document primaire qui ont servi à construire le document de cooccurrence.

3. Syntaxe XML-TEI du document de cooccurrence

Pour concrétiser notre proposition, nous présentons les diverses composantes d'un document de cooccurrence en format TEI ².

- 1) **teiheader** ³. Cette section correspond à l'entête standard de tout document TEI. On y retrouve une description du contenu du document avec ses références bibliographiques. Le contenu de l'élément *encodingDesc* fait référence au fichier de déclaration de structures de traits utilisé pour l'analyse de cooccurrence.
- 2) **div type="Statistiques"**. Après l'entête TEI, on retrouve trois divisions dans le corpus du texte. La division de type *Statistiques* contient deux blocs de données sous les balises *ab* (*arbitrary bloc*).
Le bloc **ab type="Sommaire"** donne des informations quantitatives générales comme valeurs des attributs *quantity* des éléments *measure*. L'attribut *type* indique la nature de la mesure : taille du corpus source en nombre d'occurrences (*Corpus_OccNbr*); nombre d'occurrences dans les contextes considérés (*Contexte_OccNbr*); nombre de contextes considérés (*Contexte_Nbr*) et nombre d'objets qui sont dénombrés dans ces contextes (*Objets_Nbr*). On y trouve aussi le nombre de requêtes exécutées sur le fichier (*Req_Nbr*).
Le bloc **ab type="Cooccurrence"** contient les données relatives à chacune des requêtes de cooccurrence. On aura autant de ce type de bloc qu'on aura soumis de requêtes au service Web de cooccurrence. Chaque bloc possède un identifiant unique (*xml:id="req1"*). Les éléments *rs* (*referring string*) et *measure* décrivent les paramètres de la requête.
- 3) **div type="fs"**. Cette division fait appel au formalisme des structures de traits pour décrire les propriétés (éléments *f* pour *feature*) de chacun des objets à dénombrer dans les contextes. Chaque objet est numéroté (*f name="Numéro"*) et l'ensemble de la structure *fs* reçoit un identifiant unique (attribut *xml:id*). Un autre trait (*f name="id"*) renvoie, si pertinent, à un identifiant dans le corpus primaire sur lequel a été construit le document de cooccurrence.
- 4) **div type="Contexte"**. Cette section du document est composée d'un ensemble d'éléments *span* qui définissent des empan textuels dans le document source *discourscc.xml*. Les attributs *from* et *to* du *span* pointent sur des identificateurs *xml:id* dans *discourscc.xml*. Ces identificateurs sont associés à des balise *w* qui découpent le corpus source en mots. L'attribut *n* donne la longueur de l'empan textuel dans le corpus de référence. Cette longueur se calcule en nombre d'occurrences. Le contenu du *span* est composé d'une suite de paires formées d'une balise vide *<cb/>* suivie d'un nombre. La balise *cb* marque une frontière de colonne dans une ligne de texte. On utilise l'attribut *n* pour indiquer le numéro de la colonne, ce qui permet d'omettre les frontières de colonnes pour lesquelles il n'y a pas de contenu. Cette idée de colonne est une traduction directe de la représentation des données sous forme de tableau

² On peut consulter, à l'adresse <http://www.atonet.net/publications>, un exemple de document de cooccurrence généré par le logiciel SATO (Daoust, 2009), appliqué au *corpus constitutionnel canadien 1941-1987* (Bourque et Duchastel, 1996) pour la période 1941-1950.

³ Le service Web pourra, au besoin, ajouter une ligne **xml-styleheet** en début de fichier. Cette ligne provoque l'exécution d'une feuille de style XSLT qui produira une représentation HTML des résultats contenus dans le document. Ainsi, si on ouvre cette ressource dans un navigateur Web à partir de son URI, on aura un affichage convivial. L'ensemble du fichier XML est cependant disponible et pourra être conservé dans son intégralité.

en format tabulaire. Dans ce format, chaque ligne représente un contexte et chaque colonne le nombre d'occurrences de l'objet compté dans l'ordre séquentiel des objets identifiés dans la ligne d'entête. Ces tableaux sont des matrices creuses composées d'un très grand nombre de zéros. Dans le format XML, on assume qu'une colonne qui n'est pas décrite contient zéro comme valeur implicite. On peut donc omettre de la représentation XML la grande majorité des colonnes. On aurait aussi pu choisir de représenter le décompte des objets dans les contextes comme des séries de mesures balisées par des éléments *measure*. On aurait alors utilisé un attribut *ana* pour pointer sur la structure de traits de l'objet dénombré et un attribut *quantity* pour donner le résultat du comptage.

4. Présentation des mesures de cooccurrence fournies par le service Web

Les choix méthodologiques qui président à la conception de notre application visent essentiellement à :

- circonscrire un type de phénomène d'attraction au sein de la nébuleuse cooccurentielle;
- définir un instrument de statistique textuelle à même de distinguer les attractions lexicales les plus pertinentes dans un corpus.

Nous avons opté pour une définition ouverte de la cooccurrence lexicale en cherchant à préciser le calcul statistique grâce à deux mesures différentes mais complémentaires.

4.1. Les cooccurrences lexicales empiriques

Bien que nombreuses dans toute langue, concrètes dans tout énoncé et centrales dans toute analyse linguistique, les unités lexicales complexes sont des entités qui n'ont pas un statut univoque. À peine recensées dans les dictionnaires, les phraséologies éludent les théories et règles de formation, tout en multipliant leurs réalisations en contexte. Depuis les expressions figées et arbitraires jusqu'aux combinaisons discontinues, mais tributaires d'une logique syntactico-sémantique, toutes ces cooccurrences constituent un fonds de vocabulaire incontournable qu'il est nécessaire d'étudier si l'on souhaite dépasser le stade analytique de la segmentation lexicométrique et parvenir à synthétiser les mots et le sens qu'ils portent.

Les nombreux travaux réalisés sur la *cooccurrence lexicale* ont mis en évidence une multitude de critères de formation dont le nombre d'éléments en jeu, leur orientation, leur contiguïté et leur autonomie (versus leur dynamique) sémantique. Toute étude entreprise sur ces entités s'avère soumise à la théorie que l'on souhaite privilégier sur la nature de leurs relations, et cette dépendance lexicale donne lieu à autant de définitions qu'il existe d'objectifs en amont de l'analyse des attractions entre mots. Ici nous avons choisi d'adopter une acception ouverte du phénomène cooccurentiel en tant que « groupe de mots apparaissant fréquemment ensemble dans une fenêtre contextuelle donnée ». Cette conception toute empirique de la cooccurrence présente l'avantage de considérer l'attraction lexicale au sens large en faisant abstraction de toute contrainte linguistique et permet de saisir la réalité cooccurentielle de manière exhaustive. On vérifiera à l'issue des expériences menées dans quelle mesure les systèmes statistiques dégagés par notre application font immédiatement sens car ils coïncident avec des systèmes sémantiques.

Afin de déterminer les conditions formelles de réalisation d'une cooccurrence, les deux méthodes statistiques proposées par le service Web ont en commun d'opérer une sélection par

sanction statistique des associations significatives en contexte par le biais d'un calcul contrastif qui tire profit du volume de la masse textuelle au lieu de s'y noyer.

4.2. Méthodes statistiques

Ces dernières années plusieurs modèles statistiques ont été appliqués avec succès à l'analyse des cooccurrences lexicales. Parmi les méthodes éprouvées, on citera l'*Information Mutuelle* (Church et Hanks, 1990), les *Méthodes des Cooccurrences* de P. Lafon (Lafon, 1984) et le *Test de Significativité des Cooccurrences* (Beauchemin et Cucumel, 1995). Ces deux dernières nous ont inspirés pour la conception d'un module double qui réalise un premier calcul basé sur le *modèle hypergéométrique* et un second sur le *modèle binomial*.

Si les deux algorithmes explorent l'unité naturelle de la phrase afin d'intégrer à leur calcul les caractéristiques d'organisation sémantique et syntagmatique du texte, ils se distinguent en favorisant, pour le premier, les volumes contextuels explorés (en nombre d'occurrences), pour le second le nombre de contextes de rencontre entre pôle et cooccurrent. C'est cette distinction et ses conséquences dans l'identification de cooccurrences remarquables que nous nous proposons de mesurer.

4.2.1. Calcul sur la base de la Loi hypergéométrique

Quand il applique le modèle hypergéométrique à l'analyse des cooccurrences, P. Lafon (1984) cherche à cerner les échanges réciproques entre le discours et les mots qui le constituent. Il s'agit, en comparant l'échantillon contextuel formé par les phrases où se retrouve une cooccurrence donnée avec la totalité du corpus ainsi que les fréquences globales et locales des mots associés, de sanctionner par le biais d'un indice statistique, les cooccurrents les plus caractéristiques du pôle suivant qu'ils apparaissent dans son voisinage plus ou moins souvent que prévu par la loi de probabilité.

La simplicité du paramétrage du modèle hypergéométrique (voir le modèle présenté ci-après) et la lecture aisée de ses résultats autorisent, outre son usage initial dans le calcul des spécificités par partie (Lebart et Salem, 1994), de nombreuses adaptations à l'analyse des cooccurrences. En particulier le calcul des cooccurrences relatives aux sous-parties (chrono-) logiques d'un corpus et la détection de réseaux de cooccurrence (Martinez, 2003).

Le Modèle Hypergéométrique détermine la valeur la plus probable d'après les paramètres suivants :

T : le nombre d'occurrences dans le corpus

t : le nombre d'occurrences dans les contextes du pôle

F : la fréquence du cooccurrent dans le corpus

f : la fréquence du cooccurrent dans les contextes du pôle

$$P[X = f] = \frac{\binom{F}{f} \binom{T-F}{t-f}}{\binom{T}{t}}$$

Le Modèle hypergéométrique est fondé sur la distribution en probabilité du nombre de rencontres de toutes les permutations possibles des formes étudiées dans l'hypothèse d'équiprobabilité. A partir de la valeur probable estimée, on calcule un diagnostic de spécificité signalant l'écart par rapport à la valeur attendue - un écart qui peut être positif, négatif ou nul. Si la fréquence réelle est supérieure à la fréquence attendue, alors la forme est spécifique positive et nous l'indiquons par le code $+x$. Si la fréquence réelle est inférieure à la fréquence attendue, la forme est spécifique négative et nous l'indiquons par le code $-x$. Enfin, si la fréquence réelle est égale à la fréquence attendue, alors la forme est banale. La valeur numérique de l'indice mesure quant

à elle le degré de probabilité de l'événement : un indice de 3 signalera une probabilité de 1 sur 1000, 4 une probabilité de 1 sur 10 000, etc.

4.2.2. Calcul sur la base de la Loi binomiale

Pour Beauchemin et Cucumel (1995), l'importance d'une association lexicale est relative aux emplois isolés des mots et à leurs emplois conjoints, et elle doit donc être estimée sur la base de leurs contextes d'apparition et de coïncidence. Ils proposent alors, à partir d'un test construit sur la loi binomiale⁴, une méthode qui permet de mesurer la significativité d'une cooccurrence à la fois par rapport à l'abondance des rencontres et par rapport à leur rareté.

Le Modèle binomial opère avec les paramètres suivants :

- n : le nombre de phrases du corpus
- f_p : le nombre de phrases contenant le pôle
- f_c : le nombre de phrases contenant le cooccurrent
- f_{pc} : le nombre de phrases contenant à la fois le pôle et le cooccurrent

La fréquence espérée des phrases contenant les deux formes peut s'estimer par :

$$e = (f_p \times f_c) / n$$

Si p représente la proportion de phrases dans le corpus contenant le pôle et le cooccurrent, alors on peut tester l'hypothèse nulle :

$$H_0 : p = e / n$$

Si $f_{pc} \geq e$, l'hypothèse alternative est alors :

$$H_1 : p > e / n$$

Pour une cooccurrence, la variable aléatoire étudiée, que nous appellerons X , est le nombre de phrases du corpus contenant cette cooccurrence. On peut alors calculer la probabilité que X soit supérieur à f_{pc} sous H_0 :

$$P [X \geq f_{pc} / p = e / n]$$

Si cette probabilité est petite, on conclut que la situation observée est peu probable et qu'il est peu vraisemblable que e/n soit la proportion de phrases dans le corpus contenant à la fois le pôle et la forme cooccurrente. On rejette l'hypothèse nulle. Il suffit de fixer un seuil en dessous duquel on considère la probabilité comme étant trop faible, par exemple 0.05. Ce seuil correspond au risque de rejeter H_0 alors que cette hypothèse est vraie. Le rejet de l'hypothèse nulle signifie que la cooccurrence est abondante dans le corpus.

Si $f_{pc} \leq e$, on effectue un test unilatéral à gauche en posant l'hypothèse alternative :

$$H_1 : p < e / n$$

et en calculant :

$$P [X \leq f_{pc} / p = e / n]$$

Si cette probabilité est petite, on en conclut qu'il est peu probable que e/n soit la proportion de phrases contenant le mot pôle et le mot cooccurrent. On peut alors rejeter l'hypothèse nulle. Dans ce deuxième cas, c'est la rareté de la cooccurrence qui est significative.

On notera que cette méthode ne comptabilise, ni les fréquences individuelles des formes, ni le volume de l'échantillon analysé, mais qu'elle compare le nombre de contextes où apparaissent ces formes avec le nombre de contextes où s'opère une cooccurrence. Ici encore, la simplicité

⁴ Un calcul binomial établit la probabilité d'une réalisation par rapport à une réalisation alternative unique soit $P(x)$ par rapport à $P(y)$.

du paramétrage du modèle statistique permet d'imaginer des variantes intéressantes pour ce calcul telles que les *cooccurrences étendues* (au-delà des associations binaires) qui ont été implémentées par Beauchemin et al., 2000.

4.3. L'implantation informatique

Chacun des deux algorithmes de cooccurrence a fait l'objet d'une implantation informatique indépendante, illustrant ainsi la flexibilité du mode de développement par service Web. Le modèle binomial a été implanté en Perl alors que le modèle hypergéométrique a fait l'objet d'un programme en PHP. Le programme en Perl peut être déployé sous la forme d'un appel classique de type *CGI (common gateway interface)* ou il peut être appelé par un lanceur, telle l'interface Web de SATO, qui fournit les paramètres au programme au moyen d'un fichier sur le serveur.

Par leur nature même, ces services Web pourraient être appelés depuis un programme quelconque qui préparera les données et récupérera les résultats sans que l'utilisateur ait à se soucier de la syntaxe concrète du fichier XML. SATO et son interface Web utilisent les services Web de cette façon. Mais, toute autre application pourrait utiliser ces services qui sont en accès public sur le site d'ATONET. Une application cliente pourrait ainsi fournir des interfaces très flexibles sans qu'il soit nécessaire de refaire les programmes de calcul de la cooccurrence. Les services Web peuvent aussi évoluer de leur côté pour ajouter de nouvelles fonctionnalités et des améliorations en terme d'efficacité.

Aux fins d'illustration du système, nous pouvons aussi appeler l'un ou l'autre des deux programmes au moyen d'un simple formulaire Web accessible depuis le site d'ATONET⁵. Pour la présentation des résultats, on peut sélectionner une des feuilles de style XSLT développées pour le projet ou appliquer sa propre feuille de style pour la mise en forme du fichier XML-TEI. D'autres services Web pourraient être mis à contribution pour fournir d'autres calculs, par exemple des réseaux de cooccurrents. En effet, le format XML-TEI que nous privilégions est très facilement extensible, notamment en raison du formalisme des structures de traits qui a aussi fait aussi l'objet d'une norme ISO. On peut donc enrichir de façon cumulative le document de cooccurrence ou, si l'on préfère, créer de nouveaux documents d'annotation en utilisant le même formalisme. La navigation à travers le document principal et les divers documents d'annotation ne relève pas des services Web, mais du logiciel d'analyse textuelle qui agit comme poste de travail intégré.

5. Une expérimentation sur corpus

L'expérimentation sur corpus vise trois objectifs : appliquer le service Web pour l'analyse de la cooccurrence à un corpus de recherche qui a déjà fait l'objet d'analyses; comparer les deux méthodes statistiques d'analyse des cooccurrences, implantées dans le service Web; confirmer ou infirmer les résultats d'analyse de la cooccurrence (Bourque et Duchastel, 1996), obtenus à l'aide SATO et du test de la loi binomiale (Cucumel et Beauchemin, 1995). Le corpus comprend les allocutions des Premiers ministres du Canada et des Provinces, à l'occasion des conférences constitutionnelles de 1941 et 1987. Nous avons observé que la notion d'*unité* n'apparaît plus comme valeur significative à partir de la première séance des conférences sur les peuples autochtones en 1983. Cet effacement de la problématique de l'unité s'est poursuivi dans les conférences qui ont mené à l'entente de Charlottetown en 1992, qui aspiraient à régler

⁵ L'adresse du formulaire pour la binomiale est « <http://www.atonet.net/bino.html> ». Une page d'entrée permet d'avoir accès aux services actuels : « <http://www.atonet.net/stats.html> ».

l'imbroglia créé par l'échec de l'Accord du Lac Meech (1987) sur les conditions d'adhésion du Québec à la Constitution de 1982. Cette disparition de la notion d'*unité* a été compensée par l'apparition de la notion d'*identité* qui semble exercer la même fonction discursive que la notion d'*unité* avant elle. Les deux notions partagent, en effet, dans leur voisinage un certain nombre de cooccurrents communs, tout en se distinguant par des vocabulaires spécifiques. Ce déplacement des valeurs nous apparaissait comme le passage d'une problématique de l'unité nationale à une problématique de la diversité des identités au sein du Canada.

L'application du service Web pour l'analyse de la cooccurrence sur le corpus 1941-1987 donne les résultats attendus. Nous trouvons (dans Bourque et Duchastel, 1996) un tableau (Tab. 1) des cooccurrences significatives de la notion d'*unité* durant la période 1941-1987. Si nous le comparons à la liste des cooccurrents significatifs à partir de la loi binomiale générée par le service Web (voir Tab. 2), nous obtenons des résultats très proches ⁶.

Nationale (160, 38); Canada (1581, 35); pays (801, 31); canadienne (324, 24); canadiens (569, 19); conférence (744, 15); citoyens (194, 7); divergences (43, 7); prospérité (51, 7); constitutionnel (158, 6); diversité (49, 6); gouvernement (169, 6); harmonie (25, 5); ouest (127, 5); coopération (45, 4); droits (77, 4); concessions (17,3); espoir (42, 3); espoirs (25, 3); loyauté (18, 3).

Note : Covoisinage dans la phrase : seuil de probabilité à 99%; seuil de cofréquence de 3 : la fréquence totale et la cofréquence du cooccurrent.

Tableau 1: Cooccurrences significatives de la notion d'unité durant la période 1941-1987

Le second objectif consiste à comparer les résultats obtenus par l'application de deux méthodes d'analyse de cooccurrence implantées dans le service Web. En fixant un seuil de probabilité de .05 pour l'application de l'algorithme fondé sur la loi binomiale (Tab. 2), on obtient une liste de 29 cooccurrents très proches des 30 cooccurrents apparaissant dans le lexique sous l'application utilisant la loi hypergéométrique (Tab. 3). La différence entre les deux lexiques s'explique par la présence de sept mots dans le premier lexique qui n'apparaissent pas dans le second, soient les mots : Provinces au pluriel, nation, débat, intérêts au pluriel, nationales au pluriel, gouvernement et espoir au singulier. Elle s'explique également par l'absence, dans la première liste, de sept mots associés négativement au pôle, mais dont cinq réapparaissent si on accroît le seuil de probabilité à .09. On peut conclure que les deux méthodes convergent fortement, surtout pour les mots dont l'écart à la fréquence attendue ou la spécificité sont plus élevés.

Le troisième objectif vise à valider l'analyse produite en 1996 sur la comparaison entre le vocabulaire des cooccurrents des mots *unité* et *identité* (Tab. 4 et 5). Quelle que soit la méthode de cooccurrence retenue, nous obtenons des résultats très semblables à ceux de notre première analyse. Les deux méthodes associent aux deux mots pôle un vocabulaire commun : *nationale, traditions, canadienne*, et montrent une association négative avec les mots *autochtones*. La méthode appuyée sur la loi binomiale identifie un vocabulaire commun additionnel : *canadiens, droits, langues, nation*. C'est dire que *unité* et *identité* partagent un voisinage semblable qui nous laisse croire qu'ils exercent une même fonction discursive.

⁶ Cependant, les nombres pour chaque cooccurrent semblent légèrement supérieurs dans les résultats de 1996. Cela peut s'expliquer par de légères modifications dans la définition du corpus et par le choix des délimiteurs de phrase.

Contextes qui contiennent...

	Nombre	Avec pôle	Attendu	Écart	Prob.
<i>nationale</i>	154	37	1	+36	0.0000
<i>canadienne</i>	312	22	3	+19	0.0000
<i>pays</i>	732	27	6	+21	0.0000
<i>prospérité</i>	51	7	0	+7	0.0000
<i>divergences</i>	42	6	0	+6	0.0000
<i>diversité</i>	47	6	0	+6	0.0000
<i>harmonie</i>	23	4	0	+4	0.00006
<i>Canada</i>	1418	28	12	+16	0.00009
<i>coopération</i>	44	4	0	+4	0.00067
<i>autochtones</i>	748	0	7	-7	0.00145
<i>Canadiens</i>	499	12	4	+8	0.00186
<i>citoyens</i>	180	6	2	+4	0.00557
<i>intérêt</i>	90	4	1	+3	0.00857
<i>droits</i>	91	4	1	+3	0.00889
<i>juridiction</i>	19	2	0	+2	0.01235
<i>provinces</i>	1314	20	11	+9	0.01407
<i>effort</i>	61	3	1	+2	0.01701
<i>langues</i>	62	3	1	+2	0.01775
<i>espoirs</i>	24	2	0	+2	0.01914
<i>ouest</i>	117	4	1	+3	0.02039
<i>opinion</i>	26	2	0	+2	0.02221
<i>minorité</i>	27	2	0	+2	0.02382
<i>nation</i>	184	5	2	+3	0.02411
<i>inégalités</i>	28	2	0	+2	0.02547
<i>débat</i>	34	2	0	+2	0.03629
<i>intérêts</i>	83	3	1	+2	0.03728
<i>nationales</i>	38	2	0	+2	0.04432
<i>gouvernement</i>	153	4	1	+3	0.04686
<i>espoir</i>	40	2	0	+2	0.04855

Tableau 2: Liste des cooccurrents significatifs de *Unité* sur la base de la loi binomiale (seuil de prob. de .05)

	Globale	Co-Fréq	Spéc.
<i>nationale</i>	157	38	38
<i>canadienne</i>	318	23	12
<i>divergences</i>	43	7	7
<i>pays</i>	767	27	7
<i>prospérité</i>	51	7	6
<i>diversité</i>	49	6	5
<i>harmonie</i>	25	5	5
<i>Canada</i>	1505	29	3
<i>Canadiens</i>	524	14	3
<i>coopération</i>	44	4	3
<i>droits</i>	95	5	3
<i>citoyens</i>	181	6	2
<i>constitutionnel</i>	158	5	2
<i>effort</i>	61	3	2
<i>espoirs</i>	24	2	2
<i>intérêt</i>	91	4	2
<i>inégalités</i>	28	2	2
<i>juridiction</i>	20	2	2
<i>langues</i>	63	3	2
<i>minorité</i>	27	2	2
<i>opinion</i>	28	2	2
<i>ouest</i>	124	5	2
<i>traditions</i>	49	3	2
<i>accord</i>	267	0	-2
<i>aujourd'</i>	333	0	-2
<i>autonomie</i>	284	0	-2
<i>fédéral</i>	977	4	-2
<i>parlement</i>	263	0	-2
<i>province</i>	272	0	-2
<i>autochtones</i>	800	0	-5

Tableau 3: Cooccurrents spécifiques du pôle *unité* sur la base de la loi hypergéométrique

Par contre, chacun de ces mots pôle comporte une liste de cooccurrents qui lui est propre et qui constitue ainsi un champ sémantique spécifique. *Unité* s'accompagne de *divergence*, *harmonie*, *Canada*, *Canadiens*, *coopération*, *droits*, *citoyens*, *constitutionnel*, *effort*, *espoirs*, etc., qui renvoient tous à une problématique de «*nation building*». Par contraste, *Identité* se caractérise par la coprésence de mots comme *culture*, *reconnaissance*, *patrie*, *société*, *culturelle*, *continent*, *droit*, *nation*, *nationaux*, *patrimoine*, etc. Cela semble confirmer l'interprétation de 1996 qui détectait la présence à la fois d'un fonctionnement discursif commun et la démarcation de deux champs sémantiques indépendants. L'*unité* était invoquée comme finalité à atteindre pour la nation canadienne, alors qu'*identité* ouvrait le vaste espace de la diversité et des politiques du multiculturalisme.

On peut donc affirmer que l'expérience appliquée au cas du discours constitutionnel montre la pertinence du modèle de service Web, indique la convergence des méthodes bien qu'il subsiste des différences mineures entre elles, ce qui ne semble pas pour autant remettre en cause les interprétations possibles des résultats.

Contextes qui contiennent...

	Nombre	Avec pôle	Attendu	Écart	Prob.
<i>culture</i>	83	7	0	+7	0.00000
<i>traditions</i>	47	5	0	+5	0.00000
<i>nationale</i>	154	7	0	+7	0.00000
<i>société</i>	125	6	0	+6	0.00000
<i>reconnaissance</i>	55	4	0	+4	0.00003
<i>patrie</i>	19	3	0	+3	0.00003
<i>culturelle</i>	36	3	0	+3	0.00020
<i>nation</i>	184	5	1	+4	0.00028
<i>peuples</i>	289	6	1	+5	0.00029
<i>canadienne</i>	312	6	1	+5	0.00043
<i>autochtones</i>	748	9	2	+7	0.00056
<i>patrimoine</i>	22	2	0	+2	0.00210
<i>épanouissement</i>	22	2	0	+2	0.00210
<i>continent</i>	27	2	0	+2	0.00314
<i>nationaux</i>	29	2	0	+2	0.00360
<i>langue</i>	106	3	0	+3	0.00428
<i>droit</i>	42	2	0	+2	0.00736
<i>participants</i>	44	2	0	+2	0.00805
<i>droit</i>	146	3	0	+3	0.01024
<i>régionales</i>	50	2	0	+2	0.01027
<i>Québec</i>	318	4	1	+3	0.01653
<i>Canadiens</i>	499	5	2	+3	0.01874
<i>provinces</i>	1314	0	4	-4	0.01906
<i>Autochtones</i>	109	2	0	+2	0.04346

Tableau 4: Liste des cooccurrents significatifs de Identité sur la base de la loi binomiale (seuil de prob. de .05)

	Globale	Co-Fréq	Spéc.
<i>culture</i>	85	7	7
<i>reconnaissance</i>	58	6	7
<i>nationale</i>	157	7	6
<i>traditions</i>	49	5	6
<i>patrie</i>	20	3	5
<i>société</i>	127	6	5
<i>culturelle</i>	36	3	4
<i>autochtones</i>	800	9	3
<i>canadienne</i>	318	6	3
<i>continent</i>	27	2	3
<i>droit</i>	151	4	3
<i>nation</i>	190	5	3
<i>nationaux</i>	29	2	3
<i>patrimoine</i>	22	2	3
<i>peuples</i>	300	6	3
<i>épanouissement</i>	22	2	3
<i>droit</i>	42	2	2
<i>langue</i>	115	3	2
<i>participants</i>	44	2	2
<i>régionales</i>	50	2	2
<i>éducation</i>	69	2	2
<i>conférence</i>	741	0	-2
<i>provinces</i>	1424	0	-3

Tableau 5: Cooccurrents spécifiques du pôle identité sur la base de la loi hypergéométrique

6. Conclusions et perspectives

L'utilisation de plusieurs méthodes de calcul, correspondant à différentes façons de voir les données, permet de conforter l'interprétation. Mais, encore faut-il que l'accès à ces diverses mesures n'oblige pas le chercheur à des manipulations complexes de conversion de format et d'apprentissage de logiciels différents. Dans le cas de l'analyse de cooccurrence, l'utilisation de services Web offre une solution alléchante qui n'impose aucun langage de programmation, type d'ordinateur ou système d'exploitation. La seule exigence est l'adhésion à des normes largement répandues de communication Web, sans pour cela exiger de transiter par l'Internet.

L'utilisation d'une forme XML de balisage des données facilite l'accès à une variété de services Web dans un contexte d'interopérabilité. Le format XML-TEI d'annotation externe est particulièrement adapté à cette tâche en ce qu'il permet d'établir le lien entre les données textuelles du corpus et les données numériques de la cooccurrence. De plus, ce format permet d'enrichir de façon incrémentielle le document de cooccurrence par l'ajout des résultats des requêtes successives au service de cooccurrence. L'intégration du document de cooccurrence à un système de dépôt de données permettrait de situer l'analyse de la cooccurrence dans son contexte de production et de la réintégrer au besoin dans de nouvelles chaînes de traitement sur corpus. Nous estimons que ce type d'architecture pourrait être étendu à d'autres types d'analyse textométrique, contribuant ainsi à faciliter l'accès à nos méthodes d'analyse.

Références

- ATONET. Réseau pour l'échange de ressources et de méthodologies en analyse de texte assistée par ordinateur. <http://www.atonet.net>.
- Beauchemin J., Cucumel G. and Gendreau V. (2000). Analyse de stratégies argumentatives dans le cadre méthodologique de la cooccurrence étendue, In M. Rajman et J.-C. Chappelier, éd., *Actes des 5^{èmes} Journées Internationales d'Analyse Statistique des Données Textuelles*, École Polytechnique Fédérale de Lausanne, Lausanne, pp. 331-338.
- Bourque G. and Duchastel J. (avec la collaboration de Armony V.) (1996). *L'identité fragmentée: nation et citoyenneté dans les débats constitutionnels canadiens*. Fidès [Description du corpus : <http://www.chaire-mcd.uqam.ca/ato-mcd/>].
- Brunet E. (2007). *Séquences et fréquences. Mises en œuvre dans HyperBase*, Lexicometrica, Topographie et topologie textuelles.
- Church K. W. & Hanks P. (1990) Word association norms, mutual information, and lexicography, *Computational Linguistics*, n. 16, MIT Press, Boston, pp 22-29.
- Cucumel G. and Beauchemin J. (1995). Stratégies discursives et test de significativité des cooccurrences lexicales. In S. Bolasco, L. Lebart, A. Salem, éd., *Actes des JADT-1995, Analisi statistica dei dati testuali*. Vol. 2, pp. 13-20.
- Daoust F. (2009). Système d'analyse de texte par ordinateur, SATO, Manuel de référence, version 4.3. Centre d'analyse de texte par ordinateur, UQAM, 2007 [modifié en 2009 ; <http://www.ling.uqam.ca/sato/satoman-fr.html>].
- Daoust F., Duchastel J., Marcoux Y. and Rizkallah E. (2008) Pour un modèle de dépôt de données adapté à la constitution de corpus de recherche. In S. Heiden, B. Pincemin, L. Vosghanian, éd., *Actes des JADT-2008*, vol. 1, pp. 355-367 [<http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/daoust-duchastel-marcoux-rizkallah.pdf>].
- Daoust F. et Marcoux Y. (2006). Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés. In *Les Cahiers de la MSH Ledoux n° 3, Actes des JADT-2006*, vol. 1, pp 327-340, Presses universitaires de Franche-Comté, 2006. <http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/029.pdf>
- Dublin Core Metadata Initiative. Site Web : <http://dublincore.org/>.
- Fielding, T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Thèse de doctorat, University of California [<http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>].
- Harris Z.S. (1968). *Mathematical Structures of Language*. Editions John Wiley and Sons.
- Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.
- Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.
- Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de Doctorat en Sciences du Langage, Université de la Sorbonne nouvelle – Paris 3, sous la direction d'André Salem, Paris.
- TEI Consortium (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium [<http://www.tei-c.org/Guidelines/P5/>].
- Tournier M. (1985). Texte propagandiste et cooccurrences. Hypothèses et méthodes pour l'étude de la sloganisation. *Mots*, 11 : 155-187.
- W3C (2000). *Harvesting RDF Statements from XLinks*. [Editors: Ron Daniel Jr. (Metacode Technologies Inc.) <http://www.w3.org/TR/xlink/rdf>].