Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex

Serge Heiden

ICAR UMR5191 – ENS-LSH – 69342 Lyon BP7000 Cedex 07 – France
slh@ens-lsh.fr

Abstract

There are a lot of different statistical tests that can help us to judge if two different words do co-occur or not in a corpus. Instead of trying to give a significance level to a particular test on linguistic data, we try to use it to explore the co-occurrence space through the use of an hypertext interface. We use a co-occurrence model conceived in our laboratory. We have implemented it in our tool called *Weblex*. Through the example of a corpus of speeches from different speakers of French "Assemblée constituante", we illustrate the to and fro movement available. This is done through different levels of synthesis with the ply between concordances, lexicograms and recursive lexicograms. Finally we show that it is the dynamic nature of skimming that permits to build the interpretation. Future work will concern better semi-automatic parameter tuning and on the fly annotation of information.

Résumé

Il existe de nombreux indices statistiques permettant de juger de l'attirance entre deux mots dans un corpus de textes. Plutôt que d'accorder une valeur de vérité à un indice particulier dans le domaine linguistique, nous cherchons à l'utiliser au mieux afin d'explorer l'ensemble de l'espace de cooccurrence au moyen d'une interface hypertextuelle. Nous exploitons un modèle de calcul direct de cooccurrence propre à notre laboratoire. Nous avons implémenté ce modèle dans notre outil *Weblex*. Sur la base d'un corpus exemple de discours d'orateurs de l'Assemblée constituante, nous illustrons le va-et-vient offert dans l'espace de cooccurrence. Ceci est réalisé à travers différents niveaux de synthèse possibles dans un parcours successif à travers des concordances, des lexicogrammes et enfin des lexicogrammes récursifs. Finalement nous montrons que c'est la dynamique du parcours qui permet de construire l'interprétation. Les développements futurs concernent de nouveaux asservissements de réglage de paramètres ainsi que des possibilités d'annotation d'informations en cours de parcours.

Mots-clés : cooccurrence, concordance, lexicogramme, lexicogramme récursif, hypertexte, Weblex.

1. Introduction

Il existe une vaste littérature sur les différents modèles de cooccurrences disponibles et sur leurs multiples applications : en lexicographie, analyse de discours, génération de texte, analyse syntaxique, etc. (Manning *et al.*, 2002 ; Oakes, 1998). L'indice *t* permet, par exemple, d'évaluer si la rencontre entre deux mots est due au hasard ou non. Il permet encore d'obtenir les cooccurrents les plus discriminants entre deux mots. L'information mutuelle permet de comparer la probabilité que deux mots cooccurrent (se rencontrent) avec celle de l'apparition indépendante de chaque mot. La log-vraisemblance, l'écart réduit Z, le test du Khi^2, l'Entropie, etc. sont tous susceptibles à leur manière de fournir des informations utilisables dans ce domaine. Dans cet article nous présentons l'usage que nous faisons d'un test de calcul direct de probabilité de cooccurrence conçu dans notre laboratoire (Lafon, 1984), dans la même lignée que le test de Fisher. Nous ne cherchons pas à établir si tel test ou tel autre est particulièrement « significatif » sur des données du domaine linguistique (Manning *et al.*, 2002 :

166), ou à leur accorder une interprétation linguistique directe comme certains cherchent à le faire (Krenn, 2001), voire pire à leur donner une valeur de vérité dans le domaine de l'attribution d'auteur (Labbé, 2001). Plus modestement, nous cherchons à exploiter un indicateur nous permettant de repérer des attirances entre mots de toutes natures : lexicales quand la distance moyenne et l'écart-type entre mots sont faibles, syntaxiques pour les distances plus grandes, thématiques au delà, etc. Comme la quantité d'informations sur les cooccurrences dans les grands corpus est potentiellement énorme, nous allons construire différentes vues sur cet espace de cooccurrence au moyen d'un hypertexte. De nombreux paramètres vont par ailleurs nous permettre de régler le type d'attirance mis en évidence : tris supplémentaires sur la distance moyenne, choix d'élagage systématique dans le vocabulaire considéré, choix du contexte de cooccurrence, des lexies prises en compte (lexème et/ou locutions), etc. L'objectif de cet article est donc de présenter les enjeux de l'exploitation d'un indice de cooccurrence particulier dans le cadre d'une interface informatique de type hypertexte.

Après avoir présenté le modèle utilisé dans la section 2, nous présenterons à la section suivante l'usage hypertextuel en couches de synthèses que nous faisons de la probabilité de cooccurrence dans l'outil *Weblex*. Enfin, en conclusion nous ferons la synthèse des enjeux d'interfaçage de ce parcours et nous présenterons quelques développements futurs.

2. Un modèle de cooccurrences reposant sur un calcul direct

Le modèle présenté dans cet article est implémenté dans l'outil *Weblex* développé au laboratoire ICAR de l'Ecole Normale Supérieure de Lettres et sciences humaines (Heiden, 2002). La mise en œuvre est réalisée en deux phases : tout d'abord une phase de préparation du corpus permet d'expliciter au fil du corpus un certain nombre d'informations qui seront exploitées dans la phase suivante (les lexies et les contextes) ; ensuite une phase d'analyse met en œuvre le modèle sur la base de la préparation. Pour des raisons de place, nous n'exposerons que le modèle lui-même suivi de la deuxième phase de sa mise en œuvre dans l'outil.

Le modèle de cooccurrences de Weblex

Nous nous situons dans le cadre où deux phénomènes linguistiques se rencontrent dans le même contexte sans être obligatoirement juxtaposés[1]. Le phénomène considéré dans cet exposé est le lexème, c'est à dire un mot simple, alors que n'importe quel phénomène, en fonction de la phase de préparation, peut être supportée par le modèle (locution, propriété d'occurrences : comme les parties du discours, lemme, etc.). Nous ne considérerons par ailleurs que la succession (orientée donc) de ces phénomènes que nous désignerons par « couple de cooccurrents », alors que les paires de mots (non orientées) font l'objet d'un autre modèle. Enfin le contexte de rencontre sera la phrase orthographique (essentiellement délimitée par la ponctuation forte : ., !, ?, …), alors que d'autres contextes sont envisageables comme le syntagme, la proposition, le paragraphe, etc.

Le modèle utilisé est celui de Lafon (1984). Étant donnés : les fréquences totales f_A et f_B de deux mots A et B du texte et V son vocabulaire, le calcul direct de la probabilité que ces deux mots se rencontrent exactement r fois dans les P phrases du texte est donnée par (la flèche dénotant la succession dans une même phrase) :

$$P(card\{A \in Vet B \in V / A \rightarrow B\} = r) = \frac{C_r^{fA} \times C_{fB-r}^{P+fB}}{C_{fB}^{fA+fB+P}}$$

[1] Ceci justifie l'appellation de cooccurrences en opposition aux collocations classiquement juxtaposées.

L'indice utilisé dans *Weblex* est finalement celui de la probabilité que ces deux mots se rencontrent le nombre de fois qu'on le constate effectivement dans le texte et plus encore, à concurrence de la fréquence minimum des deux mots. Cette probabilité est obtenue en sommant les valeurs de la probabilité discrète sur l'ensemble de la queue de la distribution.

L'implémentation efficace en langage C de cet indice repose sur une double série de fonctions rémanentes[2], ainsi que sur un agencement précis de l'ordre des opérations arithmétiques afin d'éviter tout phénomène d'underflow.

3. Mise en œuvre hypertextuelle du modèle

Dans cet article nous illustrerons un parcours descriptif basé sur un corpus historique fermé constitué de discours des orateurs de l'Assemblée constituante. N'étant pas historien nous-même, nous ne chercherons pas à confirmer ou à infirmer une hypothèse de travail mais plutôt à décrire l'analyse en cooccurrences de divers fonctionnements discursifs. Donc plutôt que l'aspect probatoire de la méthode ce sera son aspect heuristique — de découverte par la lecture transversale du corpus — que nous illustrerons.

Comme nous utiliserons comme indice quantitatif fondamental la fréquence de chaque forme du vocabulaire de chaque orateur dans nos analyses, présentons d'abord les limites quantitatives globales du corpus à travers le tableau des dimensions lexicométriques pour chaque locuteur (voir figure 1).

Corpus	**Occurrences**	**Formes**	**Phrases**
Mirabeau	96404	8865	4120
Sieyès	13875	2233	677
Total	110279		4797

Figure 1.Dimensions lexicométriques du corpus des orateurs.

La colonne *corpus* indique l'orateur concerné, *occurrences* le nombre de mots, *formes* la taille du vocabulaire de l'orateur et *phrases* le nombre de phrases orthographiques.

3.1. *Réalisations de la base « constitution »*

Prenons comme objet d'étude le champ sémantique de *constitution* tel que la surface graphique du discours de chaque locuteur nous permet de l'appréhender. Le choix de ce champ est fortuit, on procède selon la même méthode pour n'importe quel champ notionnel.

Indépendamment des différentes formes que pourra prendre la base « constitution » dans ce corpus, l'interprétation de chacune de ses réalisations ne peut, au final, s'effectuer qu'en contexte, c'est à dire en situant chaque mot dans le discours, par exemple la phrase où il apparaît. Le calcul classique des *Contextes* d'apparition d'une forme permet d'obtenir des concordances classiques où le mot est mis en évidence et inséré dans ses différents contextes, les uns à la suite des autres.

[2] Nous utilisons une fonction rémanente de calcul du logarithme népérien du coefficient du binôme, mémorisant les appels de (1,1) à (100,100), reposant sur une fonction rémanente de calcul du logarithme népérien de la factorielle, mémorisant les appels de 1 à 100, utilisant elle-même la fonction gammaln de la librairie NRC (Press W.H. *et al.*, 1992).

3.2. *Concordances KWIC*

Systématisons plus avant la lecture des contextes. Partant des constats que :

– la lecture des contextes d'apparition d'une notion procède le plus souvent d'une lecture allant de la notion vers le début ou vers la fin du contexte ;
– des contextes d'apparition similaires ont tendance à provoquer une catégorisation similaire de la notion analysée (nous assimilons ici le travail de dépouillement des contextes d'apparition d'une base à celui de la catégorisation des notions qu'elle sous-tend).

Une variante de l'outil *Contexte*, appelé *Concordances*, nous permet de synthétiser de manière plus efficace encore les contextes d'apparition. L'outil Concordances ne se distingue des Contextes que par seulement deux points, mais essentiels :

– chaque contexte est affiché sur une seule ligne. Ceci permet d'aligner les apparitions de la notion les unes au dessus des autres ;
– les lignes sont triées selon le contexte gauche ou droit en fonction de la notion étudiée et notamment des propriétés morphosyntaxiques des mots qui la réalisent.

Dans ces conditions, un usage approprié de tris multiples permet d'obtenir une liste de contextes qui *rapproche* les apparitions situées dans des contextes similaires. Ce rapprochement est alors utilisé comme une heuristique de lecture. La figure 2 présente un extrait de la concordance KWIC de Constitution chez Mirabeau triée à droite.

MIR26, p812	peuples . on dénonce de toute part la	**Constitution**	civile du clergé , décrétée par vos
MIR25, p810	que l' exposition des principes de la	**Constitution**	civile du clergé , récemment publiée
MIR25, p798	à ce que vous avez statué sur la	**Constitution**	civile du clergé ; mais que vous
MIR26, p829	à sceller de votre serment la nouvelle	**Constitution**	civile du clergé que par l'
MIR30, p850	serait donc pas une , surtout dans une	**Constitution**	comme la nôtre , dont le premier
MIR26, p829	entraîner dans sa chute la liberté et la	**Constitution**	de l' empire . l' une n' aspire à voir
MIR25, p805	oubli des principes élémentaires de la	**Constitution**	de l' Église . sans rechercher en quoi
MIR20, p747	n' avions pas eu le droit de changer la	**Constitution**	de l' État , ou que l' exercice d
MIR25, p803	sacerdotales . on cherche à paralyser la	**Constitution**	de l' État , pour faire revivre l'

Figure 2.
Extrait de la Concordance KWIC de « Constitution » chez Mirabeau triée à droite. Les contextes sont volontairement réduits pour des raisons de place, la référence renvoie à la page intégrale de l'édition pour la lecture élargie.

Au début de chaque contexte, la zone soulignée en bleu forme la *référence* qui situe l'apparition dans l'œuvre et donc dans le corpus. Elle est composée d'une réduction du nom de l'orateur (MIR), du n° de son discours (01, 03…), puis du numéro de page. La référence renvoie, de plus, par le biais d'un lien hypertextuel, directement à la page correspondante de l'édition en ligne du corpus des orateurs. A l'aide d'un simple lien, on accède donc aisément au contexte élargi de Constitution à travers la page qui contient l'occurrence du mot, tout en pouvant revenir aussi facilement aux Contextes d'où on est arrivé. La figure 3 présente un exemple de page d'édition en ligne, la page 805 de l'œuvre. La page d'édition est conçue de sorte à représenter le plus fidèlement possible le fac-similé de l'œuvre d'origine à l'aide de la mise en page, de la typographie, etc. Elle est elle-même bien sûr reliée aux autres pages de

l'œuvre par des liens hypertextuels. Ce réseau de pages constitue de fait le contexte définitif de la réalisation du champ sémantique, c'est-à-dire la lecture de l'ensemble de l'œuvre. On notera que le lien hypertextuel de la référence relativise la gêne occasionnée par la taille limitée du contexte affiché (même si la taille de ce contexte est paramétrable et volontairement limité ici) : le rôle du contexte est, en quelque sorte, de permettre une présélection focalisée sur un champ, pour approfondir, éventuellement, vers les pages de lecture complètes. On peut, en ce sens, parler d'un premier niveau de lecture « dynamique » focalisée, dont le support est un hypertexte.

Cette concordance est triée selon le contexte droit. L'ordre lexicographique des contextes droits à permis « d'empiler » les apparitions de Constitution participant aux mêmes locutions comme « Constitution civile du clergé » ou « Constitution de l'État ». Ici, le tri de concordances nous a permis de regrouper ensemble pour l'analyse les locutions dans lesquelles la base participe, locutions que l'outil d'analyse du vocabulaire n'avait pas construit initialement. Dans ce cas, on peut alors focaliser l'étude de la notion à partir de la locution elle-même, voire mettre à jour le corpus en y forçant cette locution comme unité lexicale, de sorte à ce qu'elle fasse partie de son vocabulaire de base.

Le réglage des divers tris ainsi que le parcours hypertextuel de l'édition du corpus nous font ici entrer dans un usage dynamique de l'outil lexicométrique où on ne peut plus se contenter de dépouiller des listings imprimés sur le papier[3]. La paramétrisation de l'instrument donne accès au corpus selon divers prismes et rend la lecture de ce dernier dynamique : ce qu'on y voit ou trouve dépend des réglages des parcours transversaux que l'on y effectue. En opposition au dépouillement d'un listing, ceci permet d'engager une sorte d' « interaction » avec le corpus.

En consultant la colonne des références de cette concordance (la première colonne), on peut par ailleurs constater que l'ordre de présentation des apparitions de Constitution ne correspond plus à l'ordre naturel du texte du corpus. De fait, cette délinéarisation du texte, provoquée par le tri, entraîne une lecture paradigmatique du matériau textuel. Et c'est ce type de lecture que nous allons continuer à suivre dans la suite de cet article. Bien sûr le lien associé à la référence (soulignée en bleu) donne immédiatement accès à la page où se trouve l'occurrence, ce qui permet toujours de revenir à la linéarité « naturelle » du texte.

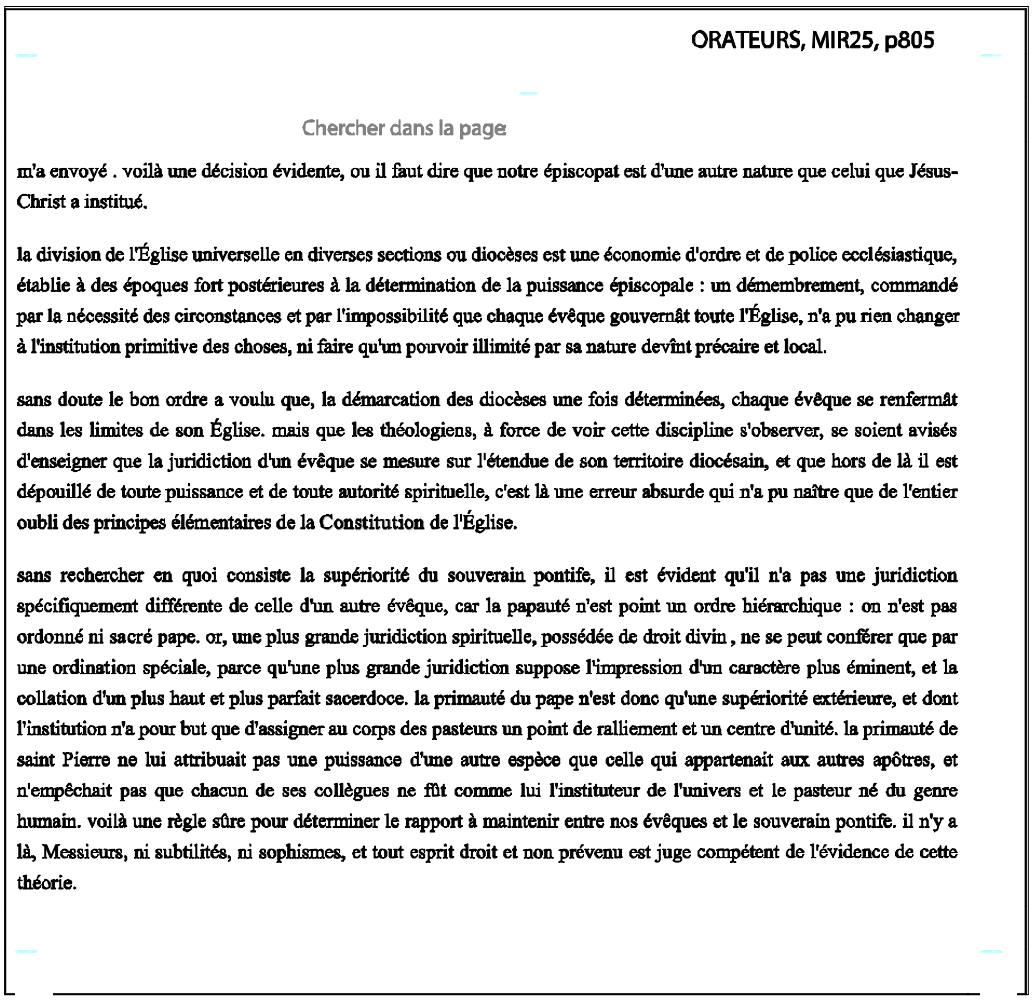
La concordance KWIC triée forme le deuxième niveau de synthèse paradigmatique de *Weblex* après le premier qui est lui constitué par les contextes simples d'apparition.

3.3. *Le lexicogramme de « Constitution » chez Mirabeau*

Un des problèmes des concordances KWIC triées est que seuls certains rapprochements de fonctionnements discursifs contigus, comme dans les locutions, sont offerts immédiatement à la lecture. Or, de nombreux liens de cooccurrence « à distance » sont susceptibles d'intéresser un dépouillement notionnel. L'outil *Lexicogramme* va être un moyen de palier à ce problème à l'aide de l'indice quantitatif de cooccurrence présenté à la section 2. Le lexicogramme d'un mot s'interprète comme une synthèse des cooccurrents gauches et droits d'un mot, à l'intérieur de toutes les phrases où il apparaît. Il peut aussi s'interpréter approximativement comme les listes hiérarchiques du vocabulaire des contextes gauches et droits d'une concordance KWIC. Le mot faisant l'objet du lexicogramme est appelé pivot du lexicogramme. Afin

[3] Ceci n'enlève rien au confort naturel de la lecture sur le papier, qui reste nécessaire à certains moments du dépouillement quand les données restent en quantités importantes.

d'illustrer l'usage de cet outil, et dans la continuité de l'effort de synthèse de sa concordance KWIC triée, la figure 4 présente le lexicogramme de Constitution chez Mirabeau. Les colonnes de gauche renseignent sur les cooccurrents situés à gauche de Convention dans le texte (en probabilité), les colonnes de droite sur les cooccurrents situés à droite.


ORATEURS, MIR25, p805

Chercher dans la page

m'a envoyé . voilà une décision évidente, ou il faut dire que notre épiscopat est d'une autre nature que celui que Jésus-Christ a institué.

la division de l'Église universelle en diverses sections ou diocèses est une économie d'ordre et de police ecclésiastique, établie à des époques fort postérieures à la détermination de la puissance épiscopale : un démembrement, commandé par la nécessité des circonstances et par l'impossibilité que chaque évêque gouvernât toute l'Église, n'a pu rien changer à l'institution primitive des choses, ni faire qu'un pouvoir illimité par sa nature devînt précaire et local.

sans doute le bon ordre a voulu que, la démarcation des diocèses une fois déterminées, chaque évêque se renfermât dans les limites de son Église. mais que les théologiens, à force de voir cette discipline s'observer, se soient avisés d'enseigner que la juridiction d'un évêque se mesure sur l'étendue de son territoire diocésain, et que hors de là il est dépouillé de toute puissance et de toute autorité spirituelle, c'est là une erreur absurde qui n'a pu naître que de l'entier oubli des principes élémentaires de la Constitution de l'Église.

sans rechercher en quoi consiste la supériorité du souverain pontife, il est évident qu'il n'a pas une juridiction spécifiquement différente de celle d'un autre évêque, car la papauté n'est point un ordre hiérarchique : on n'est pas ordonné ni sacré pape. or, une plus grande juridiction spirituelle, possédée de droit divin , ne se peut conférer que par une ordination spéciale, parce qu'une plus grande juridiction suppose l'impression d'un caractère plus éminent, et la collation d'un plus haut et plus parfait sacerdoce. la primauté du pape n'est donc qu'une supériorité extérieure, et dont l'institution n'a pour but que d'assigner au corps des pasteurs un point de ralliement et un centre d'unité. la primauté de saint Pierre ne lui attribuait pas une puissance d'une autre espèce que celle qui appartenait aux autres apôtres, et n'empêchait pas que chacun de ses collègues ne fût comme lui l'instituteur de l'univers et le pasteur né du genre humain. voilà une règle sûre pour déterminer le rapport à maintenir entre nos évêques et le souverain pontife. il n'y a là, Messieurs, ni subtilités, ni sophismes, et tout esprit droit et non prévenu est juge compétent de l'évidence de cette théorie.

Figure 3. Édition en ligne de la page 805 du corpus des orateurs. Cette page se situe dans le discours n°25 et fait partie d'un discours de Mirabeau comme l'indique son en-tête. Le mot Constitution y est mis en évidence en couleur rouge car la page a été accédée à partir d'un lien hypertextuel issu d'une concordance de ce mot. Les flèches situées aux quatre coins de la page sont des liens hypertextuels vers les pages précédentes et suivantes.

En fait la liste des cooccurrents gauches, par exemple, d'un pivot est potentiellement l'ensemble de tout le vocabulaire se trouvant à sa gauche dans les phrases, qu'ils lui soient contigus ou non. Chez Mirabeau il s'agit de 747 mots différents situés à gauche de Constitution dans ses discours. Afin d'obtenir une liste exploitable (ou lisible), c'est-à-dire plus limitée et constituée des seuls mots « les plus cooccurrents avec » ou « les plus attirés par » Constitution, nous utilisons le modèle probabiliste de cooccurrence. Ce modèle nous permet de trier la liste des mots cooccurrents afin de pouvoir n'afficher que ses premiers éléments, et

c'est sa seule vocation[4]. Un paramétrage de seuils permet alors de faire varier le nombre maximum de mots cooccurrents que l'on désire afficher. Dans l'usage de ce modèle, le chercheur a donc un rôle actif de réglages de l'instrument d'analyse. En aucun cas il s'agit d'essayer d'interpréter une « réalité » sous-jacente calculée par la machine, mais plutôt d'opérer un parcours interprétatif en « filtrant » à la demande la richesse de l'espace de cooccurrence du corpus utilisé.

Les lexicogrammes peuvent être triés selon leurs différentes colonnes afin d'orienter la lecture. Les tris et les seuils les plus utilisés sont ceux en probabilité de cooccurrence et en distance moyenne (dont le calcul est tout à fait indépendant de celui de la probabilité de cooccurrence). Dans la lecture du lexicogramme ces deux dimensions sont utilisées conjointement pour interpréter le lien de cooccurrence : les attirances fortes de mots rapprochés, en moyenne, correspondent aux figements lexicaux, aux locutions, voire aux syntagmes, les attirances fortes de mots plus éloignés, correspondent plus aux fonctionnements discursifs, voire thématiques des cooccurrents. Enfin, comme la lecture des lexicogrammes, qui forment une sorte de synthèse de la contextualisation de leur pivot — soit une synthèse de concordance KWIC — emmène souvent la lecture des propres lexicogrammes des cooccurrents du pivot courant, *Weblex* fournit un lien hypertextuel direct vers le calcul du lexicogramme de chaque cooccurrent à travers le clic sur sa forme.

Les lexicogrammes forment le troisième niveau de synthèse paradigmatique de *Weblex*.

Constitution (153)									
cooccurrents gauches					**cooccurrents droits**				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
comité	32	8	1e-05	1.0	consacrés	7	4	5e-05	6.8
Déclaration	16	6	1e-05	8.8	gouvernement	45	7	9e-04	15.6
principes	124	14	8e-05	7.3	française	23	5	1e-03	1.6
royal	5	3	4e-04	14.0	principes	124	11	4e-03	15.2
nouvelle	29	6	4e-04	0.0	résistance	19	4	4e-03	7.0
concilier	9	3	3e-03	4.0	civile	21	4	6e-03	0.0
changer	18	4	3e-03	10.2	voeux	11	3	6e-03	8.0
rapport	30	5	4e-03	6.2	délégués	12	3	8e-03	6.0
rédaction	11	3	6e-03	13.0	désormais	12	3	8e-03	23.3
organisation	13	3	1e-02	14.3	maintenir	13	3	1e-02	7.7
esprit	39	5	1e-02	4.4	exécution	15	3	1e-02	17.7
ancienne	14	3	1e-02	7.7	matière	17	3	2e-02	32.0
droits	109	9	1e-02	6.7	entièrement	19	3	3e-02	2.7
veto	41	5	1e-02	14.2	égalité	19	3	3e-02	5.3
voir	28	4	2e-02	32.0	État	67	6	3e-02	11.7
doit	133	10	2e-02	12.5	rendre	34	4	3e-02	5.2
lui-même	32	4	3e-02	11.2	jour	34	4	3e-02	15.5
arrêter	20	3	3e-02	13.0	part	21	3	4e-02	29.7
travail	21	3	4e-02	17.0	social	22	3	4e-02	9.3

Figure 4. Lexicogramme de « Constitution » chez Mirabeau.

*A droite des colonnes de formes de cooccurrents gauches et droits, la colonne **f** donne la fréquence du cooccurrent, **cf** la co-fréquence ou nombre de rencontres avec le pivot dans les phrases du corpus, **p** la probabilité de cooccurrence calculée et **d_m** la distance moyenne, en*

[4] Un modèle de cooccurrences en discours « complet » devrait au moins aussi tenir compte d'attirances distributionnelles en langue, ce qui n'est pas le cas ici.

*nombre de mots, séparant le cooccurrent du pivot dans le corpus. Pour afficher ce lexicogramme, les seuils : **f** ≥ 3, **cf** ≥ 3, **p** ≤ 5.0E-2, d_m ≤ 1000.0, ont été utilisés.*

3.4. *Comparaison entre lexicogrammes*

Ces synthèses de cooccurrents peuvent bien sûr se lire en comparaison les unes avec les autres. La figure 5 présente ainsi le lexicogramme de Constitution chez Sieyès. On accède alors de manière très synthétique à la réalisation de ce champ chez ces deux orateurs.

Constitution (21)

cooccurrents gauches					cooccurrents droits				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
raisonnée	4	3	9e-05	25.0	constituant	10	2	3e-02	13.0
exposition	4	3	9e-05	26.0	appartient	11	2	4e-02	2.5
réformer	5	3	2e-04	11.7	présenter	11	2	4e-02	16.5
bonne	5	2	8e-03	0.0	donner	13	2	5e-02	5.5
française	6	2	1e-02	17.0	objet	15	2	7e-02	22.0
parties	12	2	5e-02	2.0	publics	16	2	8e-02	8.5
partie	17	2	9e-02	14.5	peuple	25	2	2e-01	7.5
droits	47	3	1e-01	26.3	pouvoirs	26	2	2e-01	7.5
nation	48	3	1e-01	16.0	pouvoir	61	3	2e-01	9.7
moyens	24	2	2e-01	8.5					
citoyen	24	2	2e-01	27.0					
peuple	25	2	2e-01	9.0					
homme	30	2	2e-01	30.0					
doit	44	2	4e-01	7.0					

Figure 5. Lexicogramme du pôle « Constitution » dans les discours de Sieyès.
*Seuils : **f** 3, **cf** 2, **p** 5.0E-1, d_m 1000.0*

3.5. *Descente de contrôle vers les concordances de couples*

L'interprétation complète (ou fine) d'un couple de cooccurrents donné, dépend de la lecture précise de leurs contextes de rencontre. L'outil *Weblex* fournit donc un lien hypertextuel (associé à la fréquence **cf** de leur rencontre, soulignée en bleu, dans les lexicogrammes) provoquant le calcul de la concordance KWIC de l'apparition effective du couple dans le corpus. La figure 6 illustre, en exemple, la concordance obtenue en cliquant sur la co-fréquence « 7 » de la deuxième ligne des cooccurrents droits du lexicogramme de Constitution (voir la figure 4), c'est à dire le lien vers la concordance du couple (Constitution – gouvernement).

L'accès à ces concordances permet de « descendre » d'un niveau de synthèse paradigmatique, les concordances donnant elles-mêmes accès au niveau de lecture totale du corpus qui forme lui-même le niveau de base. Un principe fondamental du parcours hypertextuel offert par l'outil *Weblex* est donc le suivant : à chaque **montée en synthèse** paradigmatique doit correspondre une possibilité de **descente de contrôle** vers le niveau de synthèse inférieur.

1	MIR11, p664	politique a le droit inaliéna-ble d' établir , de modifier ou de changer la	**Constitution , c' est-à-dire la forme de son gouverne-ment**	, la distribution et les bornes des différents pouvoirs qui le composent .
2	MIR11, p666	des grands États , et surtout de l' empire français , que chaque progrès dans leur	**Constitution , dans leurs lois , dans leur gouverne-ment**	, agrandit la raison et la per-fectibilité humaine . elle vous sera due , cette
3	MIR20, p754	ême mon profond regret , que l' homme qui a posé les bases de la	**Constitution , et qui a le plus contribué à votre grand ouvrage , que l' homme qui a révélé au monde les véritables prin-cipes du gouvernement**	représentatif , se condamne lui-même à un silence que je déplore , que je trouve c

Figure 6.

Trois premières lignes de Concordance des sept rencontres de « Constitution » suivi de « gouvernement » dans les discours de Mirabeau. Comme pour les concordances précédentes, la référence de la ligne de concordance permet d'accéder à la page d'apparition de l'occurrence du couple pour une lecture plus approfondie dans l'archive elle-même.

4. Conclusion : le lexicogramme récursif de « Constitution » chez Mirabeau et Sieyès

Le parcours successif, de cooccurrents de cooccurrents, etc, à travers le réseau de lexicogrammes, permet d'accéder à une certaine image synthétique de plus en plus raffinée de la contextualisation d'un pivot initial. *Weblex*, en synthèse supérieure, permet d'afficher l'image de la totalité de ce parcours sous la forme d'un graphe appelé lexicogramme récursif. Pour construire ce graphe, l'outil parcours lui-même l'ensemble des lexicogrammes jusqu'à saturation du vocabulaire, puis dessine le graphe correspondant au parcours. Bien sûr aux seuils de calcul des lexicogrammes calculés précédemment, le graphe de parcours serait trop grand pour être représenté sur une page. L'outil cherche donc automatiquement un seuil en probabilité de sorte à obtenir un graphe ayant un nombre maximum prédéfini de mots cooccurrents. De plus, un seuil supplémentaire **pl** (pour palier) est utilisé afin de limiter la profondeur du parcours à partir du pivot. Dans un lexicogramme récursif, chaque nœud représente une forme du vocabulaire (présente une seule fois dans le graphe par définition), et chaque arc un lien de cooccurrence entre les nœuds où l'étiquette indique la force de la cooccurrence[5]. La figure 7 présente le lexicogramme récursif de Constitution chez Mirabeau, puis la figure 8 celui de Constitution chez Sieyès. Dans le même esprit de contrôle du graphe de cooccurrence obtenu, *Weblex* associe un lien hypertexte à chaque nœud du graphe vers le calcul de son lexicogramme (plus détaillé), donnant lui-même accès aux concordances de couples de cooccurrents, elles mêmes donnant accès aux pages d'édition où ces couples apparaissent.

L'implémentation de la méthode lexicométrique dans l'outil *Weblex* utilise donc la métaphore de l'hypertexte pour favoriser le va-et-vient nécessaire entre la montée en synthèse assistée par des indices quantitatifs et les descentes de contrôle dans la colonne paradigmatique d'un corpus donné. C'est la dynamique du parcours qui construit l'interprétation et non la lecture de résultats statiques donnés une fois pour toute. Cette dynamique est contrôlée par le réglage de paramètres variés : le choix de tri des cooccurrents, les seuils d'élagage quantitatifs (sur le

[5] Précisément : l'étiquette correspond au logarithme de la probabilité de cooccurrence entre les nœuds, plus le nombre est grand plus la probabilité de rencontre est faible, et donc plus l'étonnement est grand et le couple cooccurrent.

modèle comme la valeur de probabilité maximale à considérer, et sur l'interface comme le nombre maximum d'éléments à lister, le nombre maximum de nœuds d'un graphe, son algorithme de placement), les moyens d'élagage qualitatifs (prise en compte ou non de mots particuliers, de mots outils, numéraires, etc.), le choix du contexte de rencontre, les choix initiaux de segmentation des lexies (lexèmes et/ou locutions). Nos développements futurs concernent d'une part une meilleure prise en charge de réglages semi-automatiques de parcours, notamment au passage entre les différents niveaux de synthèse où les seuils d'élagage (par exemple) peuvent être adaptés semi-automatiquement en fonction de la quantité d'informations à afficher, et d'autre part la possibilité de déposer des annotations au fil d'un parcours afin de pouvoir mémoriser la progression d'une interprétation et éventuellement de la transmettre pour évaluation.

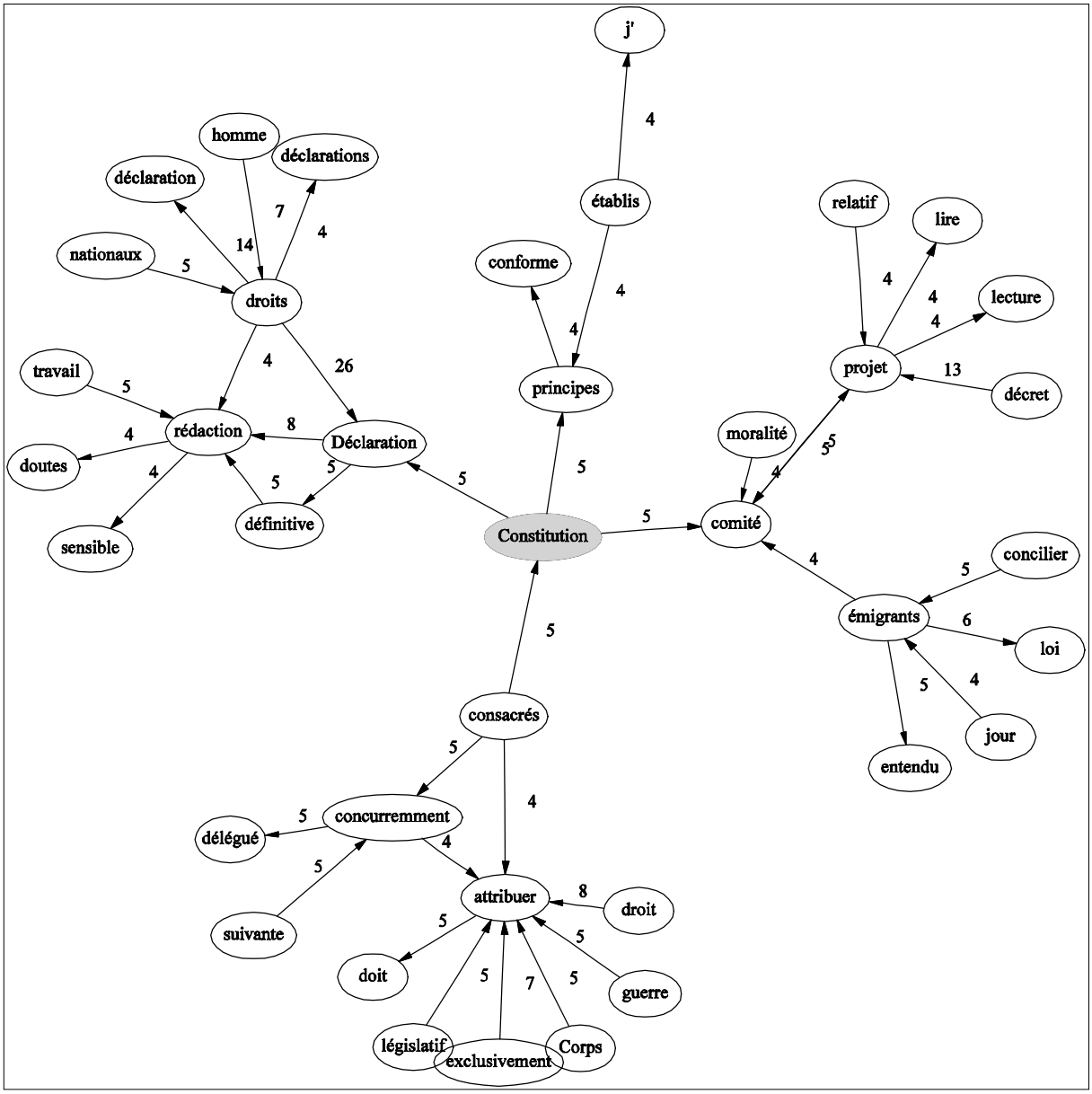


*Figure 7. Lexicogramme récursif du pôle « Constitution » dans les discours de Mirabeau. Seuils : **p** 4e-04, **r** 2, **f** 3, $\boldsymbol{d_m}$ 1000.0, **pl** 3*

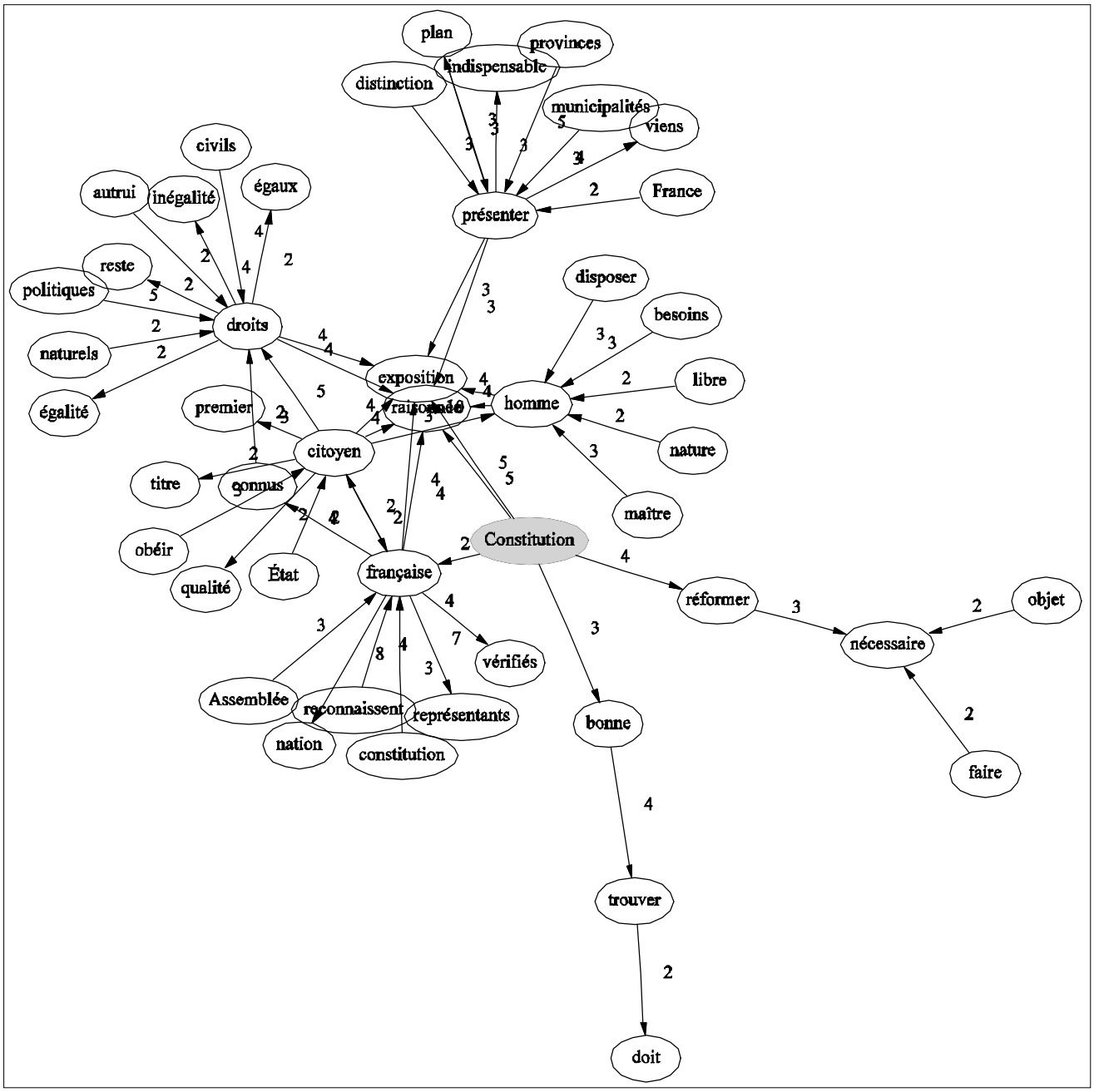


Figure 8. Lexicogramme récursif du pôle « Constitution » dans les discours de Sieyès.
*Seuils : **p** 2e-02, **r** 2, **f** 3, $\mathbf{d_m}$ 1000.0, **pl** 3*

Références

Heiden S. (2002). *Manuel Utilisateur de Weblex. Version 4.1*. Janvier 2002, ICAR CNRS/ENS-LSH, <http://weblex.ens-lsh.fr/doc/weblex.pdf>.

Krenn B. et Evert S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of ACL Workshop on Collocations*, Toulouse.

Labbé D. et Labbé C. (2001). Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Journal of Quantitative Linguistics*, vol. (8) : 213-231.

Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*. Slatkine-Champion.

Manning C.D. et Schütze H. (2002). *Foundations of statistical natural language processing*. MIT Press : 151-189.

Oakes M.P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Press W.H. *et al.* (1992). *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press.