Illustration d'une méthode lexicométrique des cooccurrences sur un corpus historique

Serge Heiden

Illustration d'une méthode lexicométrique des cooccurrences sur un corpus historique

Serge Heiden*

La méthode lexicométrique procède d'une démarche d'analyse à la fois descriptive et interprétative de divers corpus textuels à l'aide de l'ordinateur sur la base d'indices quantitatifs de divers fonctionnements discursifs. En *explicitant* et en *figeant* le témoin textuel d'une certaine réalité discursive, et en *systématisant* son analyse on espère, au delà de capacités illustratives, obtenir une méthode offrant une certaine valeur probatoire à l'analyse.
En rendant l'objet analysé *explicite* et transmissible entre les acteurs d'une discipline, on espère pouvoir systématiser la confrontation des points de vue et ainsi obtenir une analyse consensuelle. Enfin, en inscrivant les outils informatisés instrumentant l'objet dans le cadre strict d'une méthodologie *reproductible*, on espère obtenir une certaine garantie de généralité des analyses effectuées.
Dans cet article nous illustrerons un parcours descriptif implémenté par l'outil Weblex, développé au laboratoire ICAR de l'Ecole Normale Supérieure de Lettres et sciences humaines, basé sur un modèle de collocations ou « cooccurrences » appliqué à 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 collocations de divers fonctionnements discursifs. Donc plutôt que son aspect probatoire, ce sera l'aspect heuristique – de découverte par la lecture transversale du corpus – de la méthode que nous illustrerons.

Dimensions du corpus orateurs, partition par noms

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).
Dans ce tableau, la dernière ligne de la colonne « Occurrences » nous indique que ce corpus est constitué de 443949 mots en tout. De son côté, la colonne « Formes » indique la taille du vocabulaire de chaque orateur, sachant qu'il faut la relativiser par rapport au nombre total de mots prononcés par l'orateur. Par exemple, l'orateur Cazalès a prononcé 7312 mots, son vocabulaire étant constitué de 1692 formes différentes.

* slh@ens-lsh.fr

Corpus	Occurrences	Formes	Phrases
Aiguillon	1042	444	44
Antraigues	2618	676	104
Barnave	33431	4439	1414
Bergasse	12896	2145	466
Boisgelin	25813	3406	1208
Cazales	7312	1692	327
ClermontTonnerre	16662	2926	755
Duport	21736	3275	1074
DuvaldEprEmesnil	5160	1225	327
LallyTollendal	16510	3012	767
LeChapelier	17670	3103	755
Malouet	26598	4053	1109
Maury	40594	5955	1884
Mirabeau	96404	8865	4120
Mounier	55007	5879	2560
Sieyes	13875	2233	677
Talleyrand	19188	3056	816
Thouret	31433	4210	1252
Total	443949		19659

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. Les noms de corpus correspondent aux noms d'orateurs normalisés par le logiciel pour les diacritiques et les espaces. La correspondance est la suivante : Aiguillon - D'Aiguillon, Antraigues - D'Antraigues, Cazales - Cazalès, ClermontTonnerre - Clermont-Tonnerre, DuvaldEprEmesnil - Duval D'Èprémesnil, LallyTollendal - Lally-Tollendal, LeChapelier - Le Chapelier, Sieyes - Sieyès.

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. Afin d'apprécier la diversité des réalisations de la base lexicographique de constitution dans l'ensemble du corpus, un moteur de recherche générique va d'abord nous permettre d'extraire du corpus l'ensemble du vocabulaire construit autour de cette base. Pour cela nous construisons l'expression de recherche générique[†] `".*constitution.*"%cd` qui désigne dans le vocabulaire l'ensemble des formes ayant constitution comme base, quels que soient

[†] Une expression de recherche générique permet d'exprimer la recherche dans un vocabulaire de l'ensemble des formes (ou chaînes de caractères) ayant une morphologie obéissant à un ensemble de contraintes plus ou moins lâches sur les caractères.

les préfixes et les suffixes possibles. L'expression est de plus assortie d'un modifieur en fin d'expression « `%cd` » exprimant le fait que la casse (distinction minuscule/majuscule) et la présence de diacritiques particuliers (accents, cédilles) sont indifférent. Cette dernière précaution permet d'obtenir des formes dont l'orthographe est non sinon peu normalisée dans le texte. Le calcul de l'*index* de cette expression permet alors d'obtenir la liste triée par fréquence décroissante de l'ensemble des formes correspondantes (voir la figure 2).

	f	forme
1	835	Constitution
2	163	constitution
3	40	constitutionnels
4	32	constitutionnelle
5	22	constitutionnel
6	20	constitutionnelles
7	15	constitutionnellement
8	9	constitutions
9	5	inconstitutionnel
10	2	Constitutions
11	2	anticonstitutionnels
12	2	inconstitutionnelle
13	2	inconstitutionnelles
14	1	CONSTITUTION
15	1	constitutionnaire
16	1	inconstitutionnels
17	1	reconstitution
	1153	au Total

Figure 2.
Index des occurrences de `".*constitution.*"%cd` dans le corpus orateurs.
La colonne **f** indique la fréquence de la forme située sur la même ligne dans la colonne **forme**.

Réalisations de la base « constitution » orateur par orateur

La partition du corpus par orateur nous permet maintenant de réaliser la même extraction orateur par orateur afin de contraster les diverses réalisations de la base et de ce fait du champ lui-même pour chaque orateur (voir la figure 3)..

Auteur	Index	Total
Aiguillon	1 constitution	1
Antraigues	9 Constitution, 1 constitution, 1 constitutions	11
Barnave	74 Constitution, 7 constitutionnelle, 5 constitutionnels, 2 constitution, 2 constitutionnellement, 1 Constitutions, 1 constitutionnel, 1 constitutionnelles, 1 constitutions	94
Bergasse	20 constitution, 19 Constitution, 1 CONSTITUTION, 1 constitutions	41
Boisgelin		0
Cazales	10 Constitution, 1 constitutionnelle, 1 constitutionnelles	12
ClermontTonnerre	17 Constitution, 6 constitution, 4 constitutionnels, 2 constitutionnel, 2 constitutionnellement, 1 constitutionnelles	32
Duport	39 Constitution, 3 constitution, 3 constitutionnelle, 1 constitutionnels	46
DuvaldEprEmesnil	23 constitution	23
LallyTollendal	34 Constitution, 2 constitution, 2 constitutionnelles, 2 constitutionnels, 1 Constitutions, 1 constitutionnelle, 1	43

	constitutions	
LeChapelier	51 Constitution, 10 constitution, 3 constitutionnelles, 2 constitutionnels, 2 inconstitutionnelles, 1 anticonstitutionnels, 1 constitutionnelle, 1 constitutionnellement, 1 constitutions	72
Malouet	71 Constitution, 30 constitution, 5 constitutionnels, 2 constitutionnelle, 1 constitutionnel, 1 constitutionnellement	110
Maury	45 Constitution, 10 constitution, 9 constitutionnels, 6 constitutionnel, 2 constitutionnelles, 1 constitutionnelle, 1 constitutionnellement	74
Mirabeau	153 Constitution, 10 constitution, 5 constitutionnelle, 4 constitutionnelles, 3 constitutionnel, 3 constitutionnellement, 3 constitutionnels, 2 inconstitutionnelle, 1 anticonstitutionnels, 1 constitutionnaire, 1 constitutions, 1 inconstitutionnel	187
Mounier	137 Constitution, 12 constitution, 5 constitutionnelle, 3 constitutionnels, 1 constitutionnel, 1 constitutionnelles	159
Sieyes	21 Constitution, 13 constitution, 1 constitutionnel, 1 constitutions	36
Talleyrand	17 Constitution, 2 constitution, 2 constitutionnels, 1 constitutionnellement, 1 constitutions	23
Thouret	94 Constitution, 10 constitution, 7 constitutionnel, 4 constitutionnelle, 4 constitutionnelles, 4 constitutionnels, 3 constitutionnellement, 3 inconstitutionnel, 1 constitutions, 1 reconstitution	131

Figure 3.
Index ventilé des occurrences de `".*constitution.*"%cd` pour chaque orateur.

La base « constitution » en contexte

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 des *Contextes* d'apparition d'une forme permet d'obtenir des concordances classiques où le mot est mis en évidence et inséré dans son contexte. Pour des raisons de place, dans la figure 4, nous ne montrons qu'une sélection des contextes d'apparition du mot Constitution pour l'orateur Mirabeau. L'analyse effective du champ nécessiterait la lecture de tous les contextes d'apparition pour chaque locuteur intéressant l'étude. Nous présenterons à la section suivante un outil permettant de s'affranchir, en partie, de la lecture linéaire de *tous* ces contextes.

Par ailleurs, encore pour des raisons de place, et alors que c'est bien à l'analyse contrastive entre tous les orateurs qu'il faut effectivement procéder dans ce corpus, nous limiterons, dans cet article, la suite de l'exposé au seul corpus des discours de Mirabeau.

MIR01, p617 s' adjugent eux-mêmes leurs prétentions ? laissez-les faire , Messieurs , ils vont nous donner une **Constitution** , régler l' État , arranger les finances ; et l' on vous apportera solennellement l' extrait de

MIR01, p619 point l' Assemblée nationale comme un bureau de subdélégués ; nous qui croyons que travailler à la **Constitution** est le premier de nos devoirs , et la plus sainte de nos missions ; nous qui savons qu' il est

MIR03, p625 a senti qu' il fallait donner à la France une manière fixe d' être gouvernée , c' est-à-dire une **Constitution** , on oppose à ses volontés , et aux voeux de son peuple les vieux préjugés , les gothiques

MIR03, p626 par cent cinquante et un individus pourrait arrêter le roi et vingt-quatre millions d' hommes ; une **Constitution** où deux ordres qui ne sont ni le peuple , ni le prince , se serviront du second pour pressurer le

Figure 4.
Les quatre premiers contextes (sur 187 en tout) de `".*constitution.*"%cd` chez Mirabeau.

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 5 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 (comme le lieu du saut de page, la disposition des paragraphes…), 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'oeuvre. 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.

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 5

Edition 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.

La base « constitution » en contexte KWIC[3]

Systématisons maintenant 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

[3] pour KeyWord In Context.

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 6 présente un extrait de la concordance KWIC de Constitution chez Mirabeau. 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[4]. 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 comme pour les Contextes, 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.

[4] 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.

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'
MIR26, p815	les mêmes qui nous disaient que cette	**Constitution**	devait perdre l' État et déshonorer la
MIR20, p748	lui aura donné , sera l' ennemi de cette	**Constitution**	dont il doit être le garant et le
MIR19, p723	corps , de troubler l' harmonie d' une	**Constitution**	dont l' égalité politique , c'
MIR25, p803	l' État , pour faire revivre l' ancienne	**Constitution**	du clergé ; on aspire à faire évanouir
MIR25, p799	vous ne vous hâtez de recommencer la	**Constitution**	du clergé sur les principes exposés par
MIR16, p699	les rapports qui la lient à la nouvelle	**Constitution**	du royaume , aux principes de la morale

Figure 6
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 toujours à la page intégrale de l'édition pour la lecture élargie.

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 collocation « à distance » sont susceptibles d'intéresser un dépouillement notionnel. L'outil *Lexicogramme* va être un moyen de palier à ce problème à l'aide d'un indice quantitatif de cooccurrence. Le lexicogramme d'un mot s'interprète comme une synthèse des collocations gauches et droites d'un mot, à l'intérieur de toutes les phrases où il apparaît[5]. 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 d'illustrer cette notion, et dans la continuité de l'effort de synthèse de sa concordance KWIC triée, la figure 7 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.
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 un modèle probabiliste de cooccurrence[6].

[5] Dans la suite de cet article, le contexte de cooccurrence sera celui de la phrase. En fait, tout contexte peut être utilisé : syntagme, proposition, phrase, paragraphe, section, partie…
[6] Sur la base de quatre paramètres (**F** la fréquence du pivot, **f** la fréquence du cooccurrent considéré, **cf** la co-fréquence entre le pivot et le cooccurrent dans les phrases et **P** le nombre total de phrases du corpus),

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[7]. 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.
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 7
Lexicogramme de « Constitution » chez Mirabeau.

le modèle calcule la probabilité que le pivot et le cooccurrent se rencontrent effectivement le nombre de fois que l'on constate dans le corpus (soit cf) et plus encore (à concurrence de la fréquence minimale des deux mots : ils ne peuvent se rencontrer plus de fois qu'ils apparaissent eux-mêmes).

[7] 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.

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 $\mathbf{d_m}$ la distance moyenne, en nombre de mots, séparant le cooccurrent du pivot dans le corpus. Pour afficher ce lexicogramme, les seuils : **f** >=[8] 3, **cf** >= 3, **p** <=[9] 5.0E-2, $\mathbf{d_m}$ <= 1000.0, ont été utilisés.

Comparaison entre lexicogrammes

Ces synthèses de cooccurrents peuvent bien sûr se lire en comparaison les unes avec les autres. La figure 8 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**	$\mathbf{d_m}$		**f**	**cf**	**p**	$\mathbf{d_m}$
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 8
Lexicogramme du pôle « Constitution » dans les discours de Sieyès.
Seuils : **f** 3, **cf** 2, **p** 5.0E-1, $\mathbf{d_m}$ 1000.0

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

[8] Plus grand ou égal à
[9] Plus petit ou égal à

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 9 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 7), c'est à dire le lien vers la concordance du couple (Constitution – gouvernement).
L'accès à ces concordances de contrôle permet de « descendre » d'un niveau de synthèse paradigmatique, les concordances donnant elles-mêmes accès au niveau de lecture totale du corpus.

1 MIR11, p664	. toute association politique a le droit inaliénable d' établir , de modifier ou de changer la	**Constitution , c' est-à-dire la forme de son gouvernement**	, la distribution et les bornes des différents pouvoirs qui le composent . " Article 4 . le bien
2 MIR11, p666	est l' influence des grands États , et surtout de l' empire français , que chaque progrès dans leur	**Constitution , dans leurs lois , dans leur gouvernement**	, agrandit la raison et la perfectibilité humaine . elle vous sera due , cette époque fortunée , où
3 MIR20, p754	us l' apporte ; je ne cacherai pas même 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 principes du gouvernement**	représentatif , se condamne lui-même à un silence que je déplore , que je trouve coupable , à
4 MIR29, p847	que reposent sur les lois , et les lois sur le respect qu' on leur porte , le chef-d ' oeuvre d' une	**Constitution , le chef-d ' oeuvre d' un gouvernement**	est de pouvoir échapper au malheur d' un mauvais roi , même d' un mauvais administrateur . or ,
5 MIR20, p746	ce danger par rapport à notre Constitution , à nous-mêmes , et au roi . par rapport à notre	**Constitution , pouvons - nous espérer de la maintenir , si nous composons notre gouvernement**	de différentes formes opposées entre elles ? j' ai soutenu moi-même qu' il n' existe qu' un seul pr
6 MIR16, p699	si je voulais envisager une aussi grande question sous tous les rapports qui la lient à la nouvelle	**Constitution du royaume , aux principes de la morale , à ceux de l' économie politique , j' examinerais d' abord s' il convient au nouvel ordre de choses que nous venons d' établir que le gouvernement**	, distributeur de toutes les richesses ecclésiastiques par la nomination des titulaires , conserve
7 MIR20, p738	? pour moi , j' établis le contrepoids des dangers qui peuvent	**Constitution même , dans le balancement des pouvoirs , dans le concours des deux délégués de la nation , dans les forces**	représentatif , contre une armée placée aux frontières : et félicitez - vous , Messieurs , de cette

naître du pouvoir royal dans la	**intérieures que vous donnera cette garde nationale , seul équilibre propre au gouvernement**	

Figure 9
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.

Parcours de lexicogrammes successifs

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 cooccurrent 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 figures 10 à 12 illustrent ainsi l'enchaînement en profondeur du calcul direct des lexicogrammes de « gouvernement », puis « Église » et enfin « France » à partir du lexicogramme initial de « Constitution » par un simple parcours hypertextuel.

gouvernement
(45)

cooccurrents gauches					**cooccurrents droits**				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
Constitution	153	7	9e-04	15.6	monarchique	7	5	2e-09	1.6
forme	35	3	6e-03	1.3	représentatif	6	4	2e-07	1.5
principes	124	4	4e-02	11.8	régime	15	3	5e-04	10.3
j'	214	3	4e-01	24.3	pouvoirs	72	3	4e-02	16.0
nation	252	3	5e-01	9.0	ordre	87	3	6e-02	5.3
je	703	7	5e-01	27.0	Église	88	3	7e-02	2.0
					roi	251	3	5e-01	4.3

Figure 10
Lexicogramme du pôle « gouvernement » dans les discours de Mirabeau.
Seuils : **f** 3, **cf** 3, **p** 5.0E-1, **d_m** 1000.0

Église
(88)

cooccurrents gauches					cooccurrents droits				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
biens	110	29	6e-26	4.5	universelle	16	5	1e-05	0.0
donnés	13	6	1e-07	10.5	biens	110	10	7e-05	9.4
évêque	14	5	6e-06	16.4	France	49	5	3e-03	1.0
clergé	132	11	6e-05	17.0	fondations	24	3	1e-02	15.0
pasteurs	37	6	9e-05	20.2	ecclésiastique	29	3	2e-02	17.0
pasteur	11	3	1e-03	5.7	culte	32	3	3e-02	7.0
côté	25	4	2e-03	8.2	pu	48	3	8e-02	4.7
premiers	31	4	4e-03	15.2	propriétés	53	3	1e-01	4.3
évêques	32	4	4e-03	11.8	propriété	64	3	1e-01	14.3
juridiction	17	3	5e-03	20.0	société	71	3	2e-01	19.3
fondations	24	3	1e-02	13.3	droits	109	4	2e-01	23.5
propriétaire	28	3	2e-02	7.0	nation	252	7	2e-01	12.6
propriétés	53	4	2e-02	6.8	peuple	217	6	3e-01	11.5
premier	53	4	2e-02	14.8	temps	88	3	3e-01	13.7
contraire	34	3	3e-02	11.3	clergé	132	4	3e-01	18.0
lieu	36	3	4e-02	23.0	religion	96	3	3e-01	17.0
gouvernement	45	3	7e-02	2.0					
autorité	47	3	7e-02	12.3					
pu	48	3	8e-02	8.0					
doivent	48	3	8e-02	27.7					
dit	95	4	1e-01	6.5					
public	65	3	1e-01	19.0					
ministres	69	3	2e-01	10.0					
doute	70	3	2e-01	15.3					
nation	252	7	2e-01	10.6					
j'	214	6	3e-01	14.0					
temps	88	3	3e-01	25.0					
doit	133	4	3e-01	23.2					
corps	100	3	3e-01	10.3					
principes	124	3	5e-01	13.0					

Figure 11
Lexicogramme du pôle « Église» dans les discours de Mirabeau.
Seuils : **f** 3, **cf** 3, **p** 5.0E-1, **d_m** 1000.0

France

(49)

cooccurrents gauches					cooccurrents droits				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
résolu	13	4	1e-05	39.2	oppresseurs	6	3	3e-05	6.3
Assemblée	230	11	3e-05	11.7	siècles	15	3	6e-04	25.3
pasteurs	37	5	6e-05	23.0	bonheur	23	3	2e-03	20.0
privilégiées	9	3	1e-04	39.3	peuple	217	7	1e-02	18.1
classes	15	3	6e-04	40.3	manière	39	3	1e-02	10.0
nationale	148	7	1e-03	11.6	publique	104	3	1e-01	22.7
général	24	3	3e-03	8.3	nation	252	5	1e-01	21.0
Église	88	5	3e-03	1.0	liberté	146	3	2e-01	14.7
aujourd'	34	3	7e-03	3.3	roi	251	4	3e-01	15.5
jamais	108	5	7e-03	26.4	peut	228	3	5e-01	16.3
hui	35	3	7e-03	2.3	Assemblée	230	3	5e-01	16.0
peuple	217	7	1e-02	1.7					
nécessaire	39	3	1e-02	7.0					
est-à-dire	40	3	1e-02	27.3					
veto	41	3	1e-02	28.0					
nombre	44	3	1e-02	29.3					
représentants	88	4	2e-02	3.0					
députés	50	3	2e-02	34.7					
Messieurs	185	5	6e-02	39.4					
roi	251	6	6e-02	8.2					
clergé	132	4	6e-02	21.0					
religion	96	3	1e-01	14.7					
doit	133	3	2e-01	27.7					
faire	223	4	2e-01	21.8					

Figure 12
Lexicogramme du pôle « France» dans les discours de Mirabeau.
Seuils : **f** 3, **cf** 3, **p** 5.0E-1, **d_m** 1000.0

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, 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[10].
La figure 13 présente le lexicogramme récursif de Constitution chez Mirabeau, puis la figure 14 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.

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

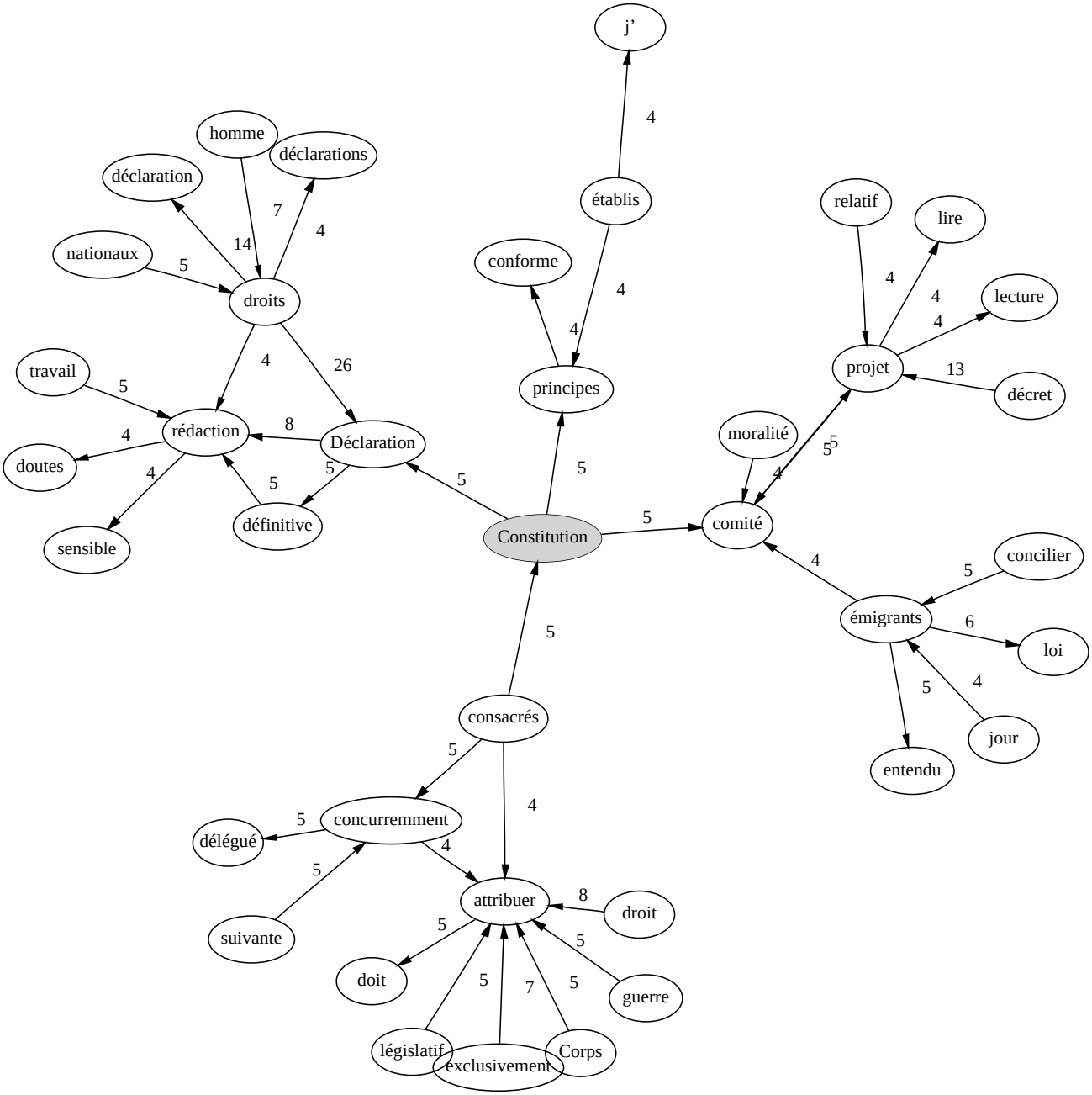

Figure 13
Lexicogramme récursif du pôle « Constitution » dans les discours de Mirabeau.
Seuils : **p** 4e-04, **r** 2, **f** 3, $\mathbf{d_m}$ 1000.0, **pl** 3

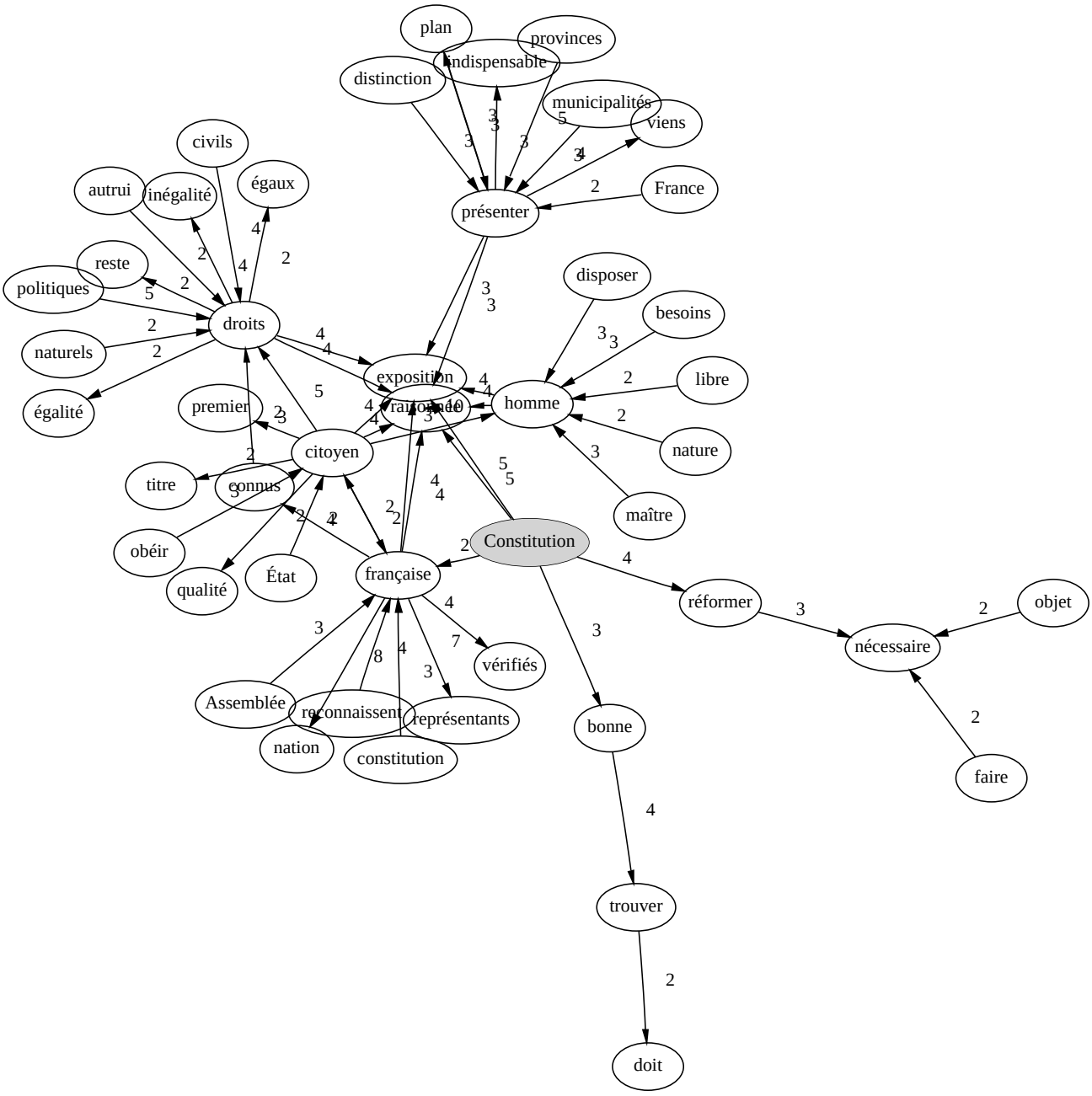

Figure 14
Lexicogramme récursif du pôle « Constitution » dans les discours de Sieyès.
Seuils : **p** 2e-02, **r** 2, **f** 3, $\mathbf{d_m}$ 1000.0, **pl** 3