Devoir de statistique textuelle

Martin André, Chahinaze Kandi, Antoine Legendre

Mai 2025

1 Introduction

L'objectif de cette étude est d'analyser le corpus numéro 2, qui contient les discours des présidents F. Mitterrand, J. Chirac et F. Hollande prononcés à l'occasion de la journée internationale des droits de la femme, le 8 mars 1977. L'analyse est réalisée à l'aide du logiciel IRaMuTeQ. Une première étape consiste à enrichir les dictionnaires en ajoutant des expressions telles que «Partie civile», «tenir compte», ou encore «nations-unies». Le corpus est ensuite chargé pour effectuer une analyse statistique, notamment la détection et la gestion des formes inconnues. L'étude s'intéresse ensuite à la richesse lexicale des trois locuteurs, puis aux différences de vocable à l'aide de tests du chi-2 et de nuages de mots. Une classification sur segments de textes permettra d'identifier des thématiques ou styles distincts. Enfin, des graphes de similitude viendront compléter l'analyse.

2 Préparation du corpus

Avant de lancer l'analyse du corpus `femmes.txt` sous IRaMuTeQ, nous avons enrichi le dictionnaire en y ajoutant plusieurs expressions composées fréquemment rencontrées dans les discours, afin d'améliorer la qualité de l'analyse textuelle. Parmi celles-ci figurent : *Partie civile*, *de sorte que*, *tenir compte*, *prendre en compte*, *à juste titre*, *bien sûr*, *par rapport* et *nations-unies*. Cette étape vise à éviter que ces expressions soient analysées comme des mots isolés sans signification propre.
Après chargement du corpus dans IRaMuTeQ et lancement de l'analyse statistique, nous avons consulté la liste des formes inconnues (notées `nr`). Celles-ci étaient exclusivement constituées de noms propres, comme *Europe*, et ne nécessitaient donc pas d'ajouts supplémentaires au dictionnaire. Nous avons également vérifié que les expressions ajoutées figuraient bien dans les formes actives : par exemple, *bien sûr* apparaît bien dans la liste des formes actives, ce qui confirme sa prise en compte dans l'analyse. Aucune autre modification du dictionnaire n'a été jugée nécessaire à ce stade.

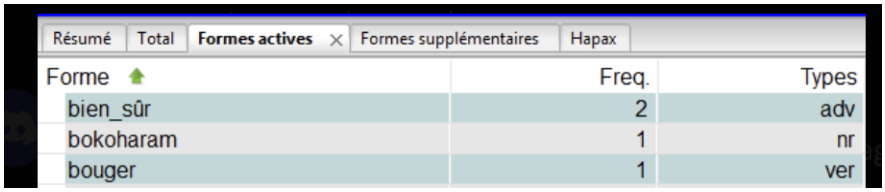


Figure 1: Apparition de *bien sûr* parmi les formes actives dans l'analyse du corpus

3 Etude de la richesse du vocabulaire et des spécificités

3 : L'analyse des discours de François Mitterrand, Jacques Chirac et François Hollande à l'aide du logiciel IRaMuTeQ permet de comparer la richesse lexicale de chacun à travers plusieurs indicateurs : le nombre total d'occurrences, le nombre de formes différentes, et le nombre d'hapax (formes apparaissant une seule fois).

Discours de François Mitterrand :

- Nombre d'occurrences : 4827
- Nombre de formes : 1042
- Nombre d'hapax : 563 (11,66% des occurrences ; 54,03% des formes)

Discours de Jacques Chirac :

- Nombre d'occurrences : 774
- Nombre de formes : 291
- Nombre d'hapax : 171 (22,09% des occurrences ; 58,76% des formes)

Discours de François Hollande :

- Nombre d'occurrences : 2604
- Nombre de formes : 655
- Nombre d'hapax : 403 (15,48% des occurrences ; 61,53% des formes)

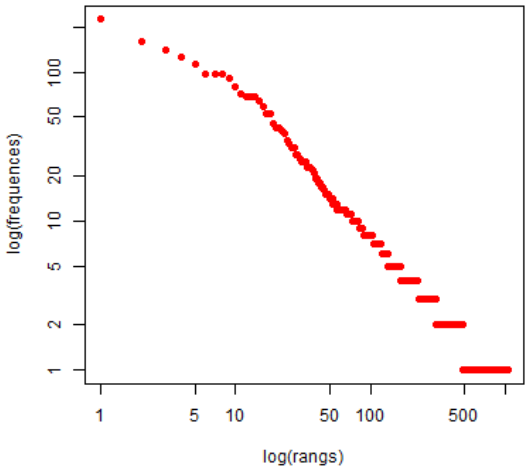
Comparaison des discours :
Pour comparer la richesse du vocabulaire, on peut regarder deux choses : le nombre de mots différents par rapport au nombre total de mots (diversité lexicale), et la part des mots qui n'apparaissent qu'une seule fois (hapax), ce qui donne une idée de l'originalité du vocabulaire.

- Le discours de François Mitterrand est le plus long (4827 mots) et contient aussi le plus grand nombre de mots différents (1042). Mais la proportion de mots rares reste assez modérée (54,03% des formes).
- Celui de Jacques Chirac est beaucoup plus court (774 mots), mais il utilise proportionnellement plus de mots rares, avec un taux d'hapax élevé (58,76%). Cela montre une certaine densité dans le vocabulaire.
- Le discours de François Hollande se situe entre les deux : il est plus long que celui de Chirac mais plus court que celui de Mitterrand (2604 mots), et c'est lui qui utilise le plus de mots rares en proportion (61,53%), ce qui traduit une plus grande originalité lexicale.

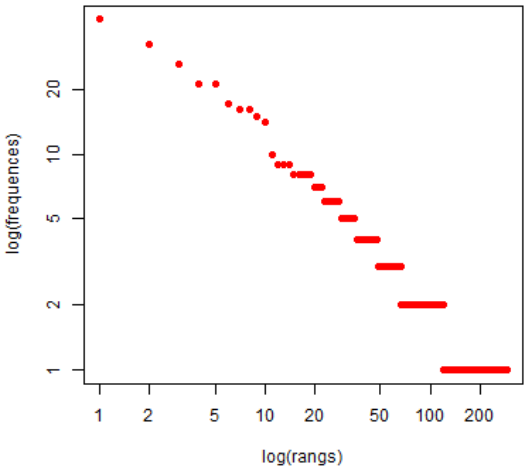
Données globales (ensemble des 3 discours) :

- Nombre de textes : 3
- Nombre total d'occurrences : 8205
- Nombre total de formes : 1448
- Nombre d'hapax : 761 (9,27% des occurrences ; 52,56% des formes)
- Moyenne d'occurrences par texte : 2735,00

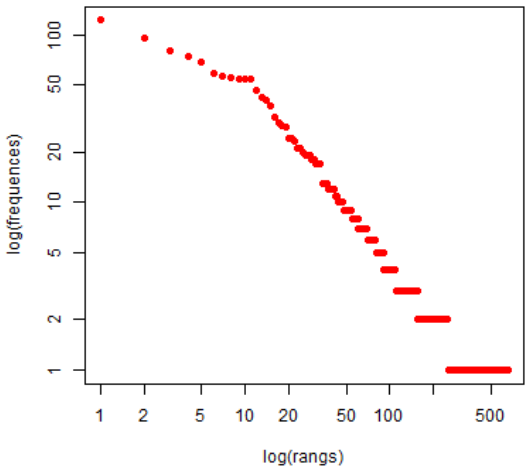
En résumé, si François Mitterrand a un discours plus long, c'est François Hollande qui semble faire preuve de la plus grande richesse lexicale relative, suivi de Jacques Chirac. Le discours de Mitterrand, plus long, est plus répétitif, ce qui atténue la proportion de formes rares.



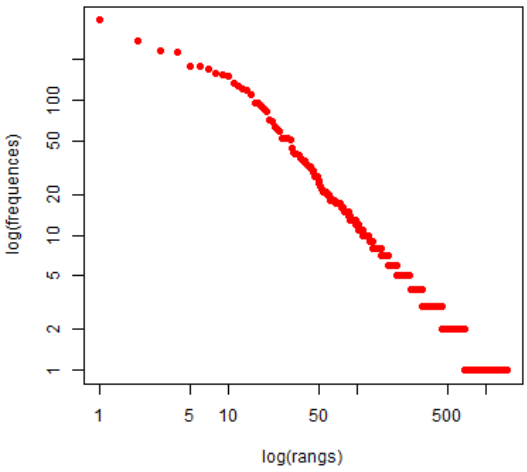
(a) Discours de Mitterand



(b) Discours de Chirac



(c) Discours d'Hollande



(d) Ensemble des 3 discours

Figure 2: Graphique des logarithmes des fréquences des formes en fonction des logarithmes des rangs

Ces graphiques viennent confirmer ce que l'on a pu observer précédemment. En effet, on observe que la pente pour le discours d'Hollande est moins abrupte ce qui témoigne d'une plus grande richesse de vocabulaire. On constate également bien que le discours de Chirac est bien plus court et est donc moins riche lexicalement.

4 Les différences de vocable entre les trois présidents

4.1 Nuages de mots

Afin de comparer les différences de vocable entre les présidents, nous allons réaliser des nuages de mots dans lesquels la taille des formes est proportionnelle à leur fréquence. Pour réaliser ces nuages nous avons choisi d'afficher uniquement les formes actives de fréquences supérieures ou égales à 10. On obtient alors les nuages suivants :

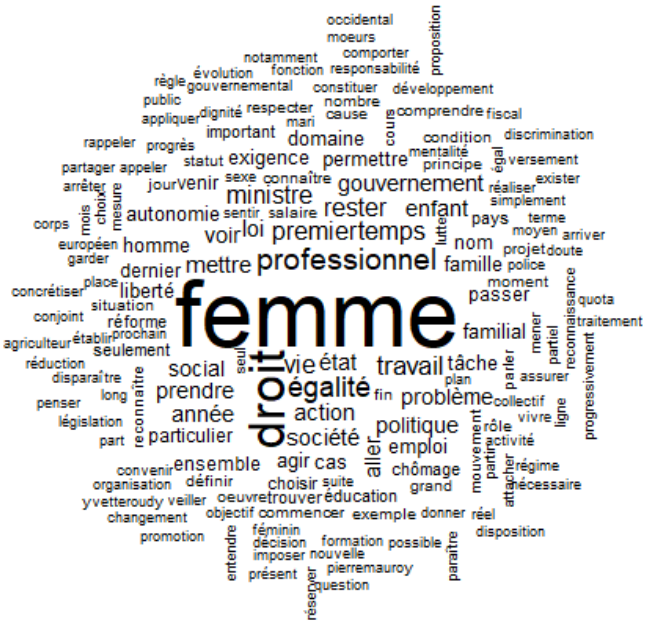


Figure 3: Nuage du discours de Mitterand

Figure 4: Nuage du discours de Chirac

Figure 5: Nuage du discours de Hollande

L'analyse des fréquences montre que le mot femme occupe une place centrale dans les discours de François Mitterrand et François Hollande, où il apparaît bien plus fréquemment que tout autre terme. En revanche, dans le discours de Jacques Chirac, ce mot est relégué derrière d'autres termes. Le mot droit est également très présent chez Mitterrand et Hollande, tandis que Chirac l'emploie de manière plus modérée. À l'inverse, ce dernier fait un usage marqué des mots combat et violence, deux thématiques également présentes chez Hollande, mais quasiment absentes chez Mitterrand. Les discours de Mitterrand, Hollande et Chirac révèlent des focalisations lexicales distinctes : Mitterrand et Hollande insistent davantage sur les enjeux liés aux droits des femmes, tandis que Chirac met l'accent sur les violences faites aux femmes.

4.2 Test du chi-deux

Les données étant sous forme de tableau de contingence, les observations étant indépendantes, les effectifs attendus suffisament grands (80% des fréquences attendues ≥ 5 et aucune fréquence < 1), les catégories ne se recouvrant pas et la taille du corpus étant suffisament grande, on peut appliquer le test du chi-2. Nous avons effectué ce test du chi-deux afin d'évaluer l'indépendance entre les mots employés et les présidents. Le tableau de contingence obtenu présente les fréquences d'apparition de chaque mot dans les discours de Chirac, Hollande et Mitterrand.
Le test donne une statistique de $\chi^2 = 257{,}81$, avec 86 degrés de liberté, et une *p-valeur* inférieure à $2{,}2 \times 10^{-16}$. Cela indique une dépendance significative entre les mots utilisés et les présidents.

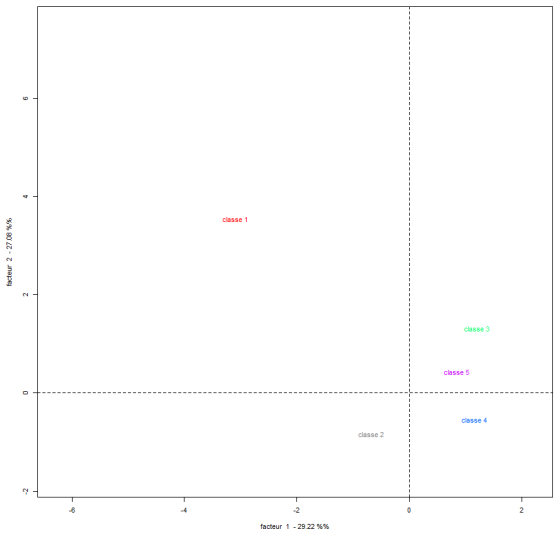
Remarque : les conditions classiques d'application du test du khi-deux (fréquences théoriques suffisamment élevées dans chaque case du tableau) peuvent ne pas être totalement respectées dans ce cas. En particulier, certaines cases ont des effectifs faibles, ce qui peut affecter la validité de l'approximation asymptotique. Les résultats doivent donc être interprétés avec prudence.

4.3 Tests d'adéquation par président (test du χ^2 simulé).

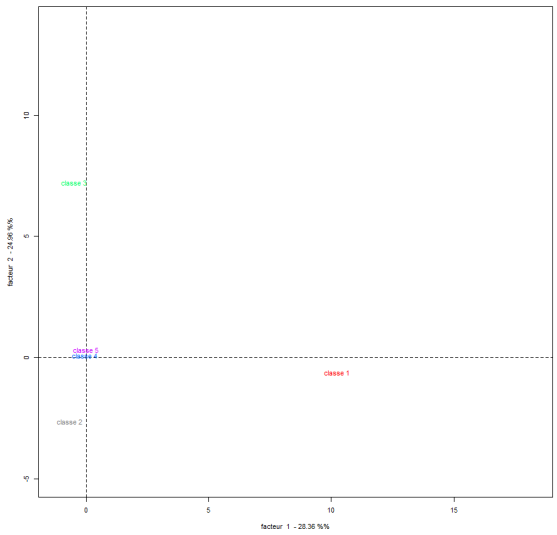
Nous avons effectué un test du χ^2 d'adéquation, avec estimation de la *p-value* par simulation de Monte Carlo (10 000 répétitions), pour comparer la distribution des formes pleines utilisées par chaque président à celle des deux autres.
Les résultats montrent que pour **Chirac**, la statistique du test est de 6,76 avec une *p-value* de 0,2135, ce qui ne permet pas de rejeter l'hypothèse nulle : sa distribution n'est pas significativement différente de celle des deux autres présidents.
En revanche, pour **Hollande**, la statistique de test est de 16,29 avec une *p-value* de 0,0066, ce qui est significatif au seuil de 5% : cela indique que la distribution des formes pleines qu'il emploie est significativement différente de la distribution moyenne des deux autres présidents.
Enfin, pour **Mitterrand**, la statistique est de 10,69 avec une *p-value* de 0,0621, ce qui correspond à un résultat marginalement significatif : on ne peut pas rejeter l'hypothèse d'une distribution similaire à celle des autres présidents, mais une tendance à la différence est tout de même observable.
Ces résultats suggèrent que Hollande se démarque nettement par son usage des formes pleines, tandis que les usages de Chirac et de Mitterrand sont plus proches de ceux des autres.

5 Classification sur segments de textes

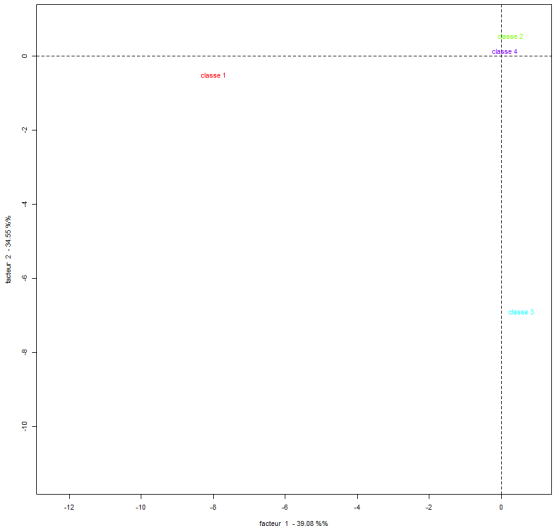
Nous avons ensuite réalisé des classifications (avec ou sans les noms supplémentaires, avec ou sans les verbes supplémentaires dans les formes actives) sur des segments de texte de taille 40. On a alors choisi, au vu notamment de la séparation des classes sur le graphique de l'AFC (voir Figure 6), la classification avec les noms et les verbes supplémentaires.



(a) Avec noms et verbes supplémentaire



(b) Avec noms mais sans verbes supplémentaires



(c) Avec verbes mais sans noms supplémentaires

Figure 6: Répartition des classes

En choisissant donc la classification avec noms et verbes supplémentaires, on obtient alors :

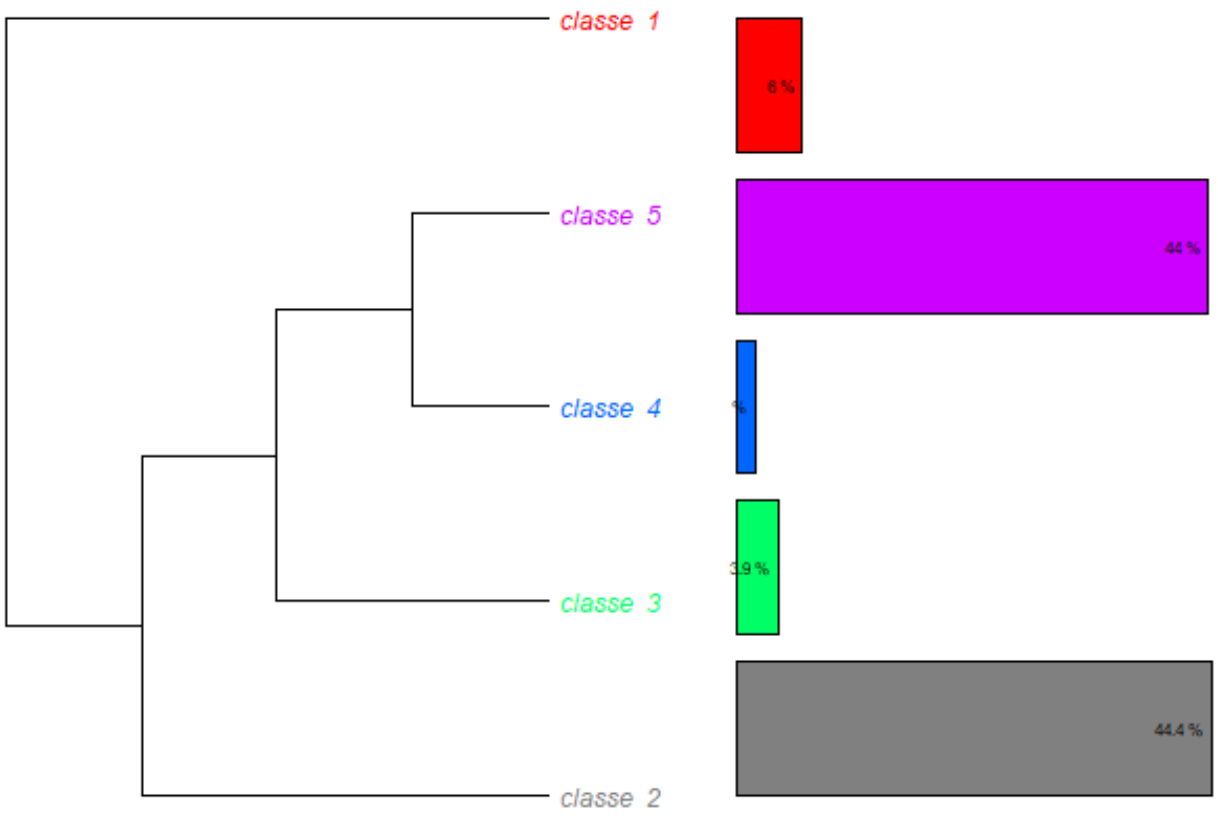


Figure 7: Dendrogramme de la classification

- Nombre de textes: 3
- Nombre de segments de texte: 233
- Nombre de formes: 1916
- Nombre d'occurrences: 8205
- Nombre de lemmes: 1448
- Nombre de formes actives: 1200
- Nombre de formes supplémentaires: 69
- Nombre de formes actives avec une fréquence >= 3: 292
- Moyenne de formes par segment: 35.214592
- Nombre de classes: 5
- 232 segments classés sur 233 (99.57%)

On remarque notamment que la classification est composée de 2 grandes classes regroupant à elles deux plus de 88% des segments et que tous les segments sauf 1 ont été classés et qu'on a donc 99.57% de segments classés ce qui est un très bon résultat.
Cette classification donne alors la classification suivante pour les formes actives et supplémentaires :

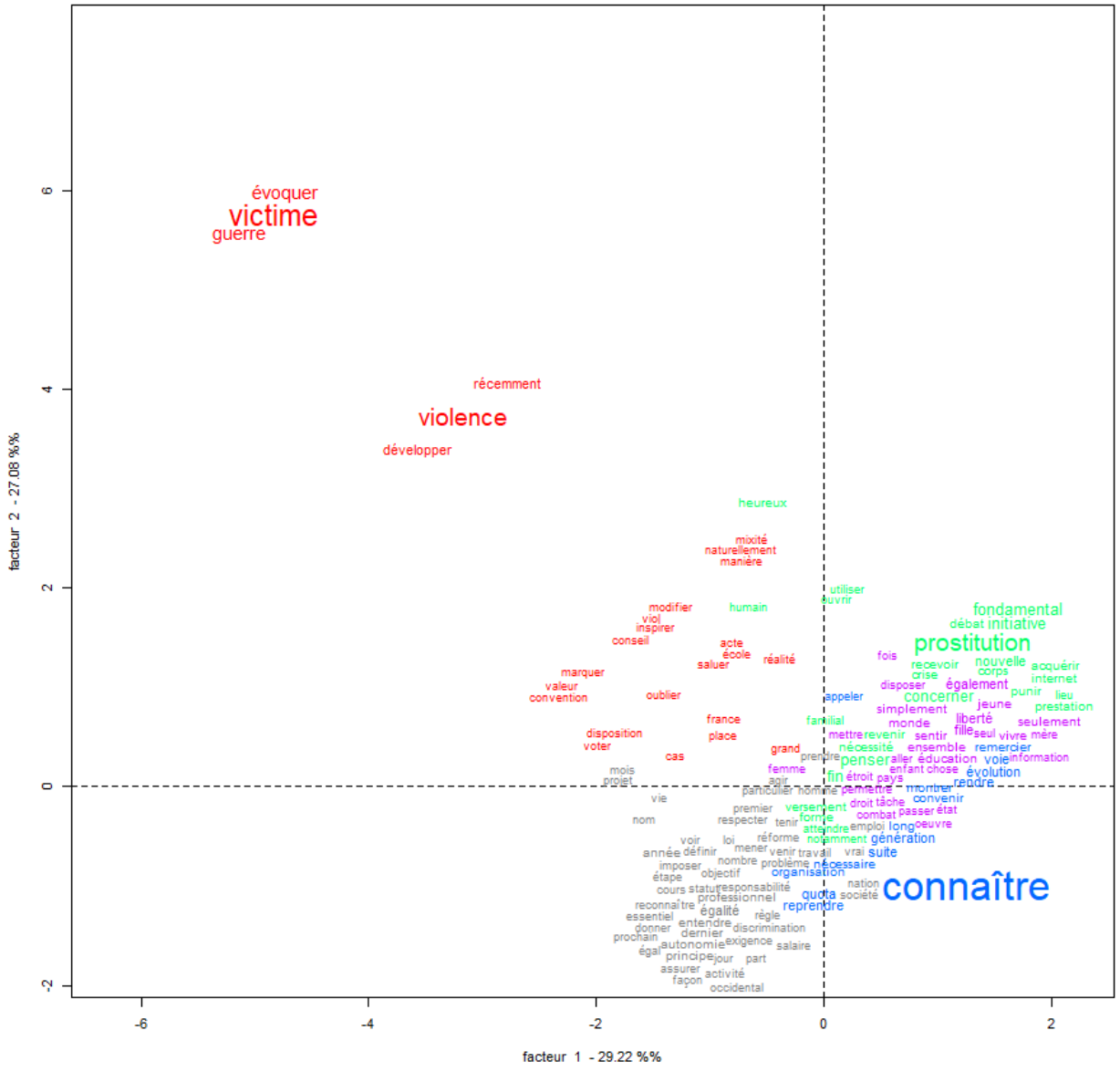


Figure 8: Classification des formes actives

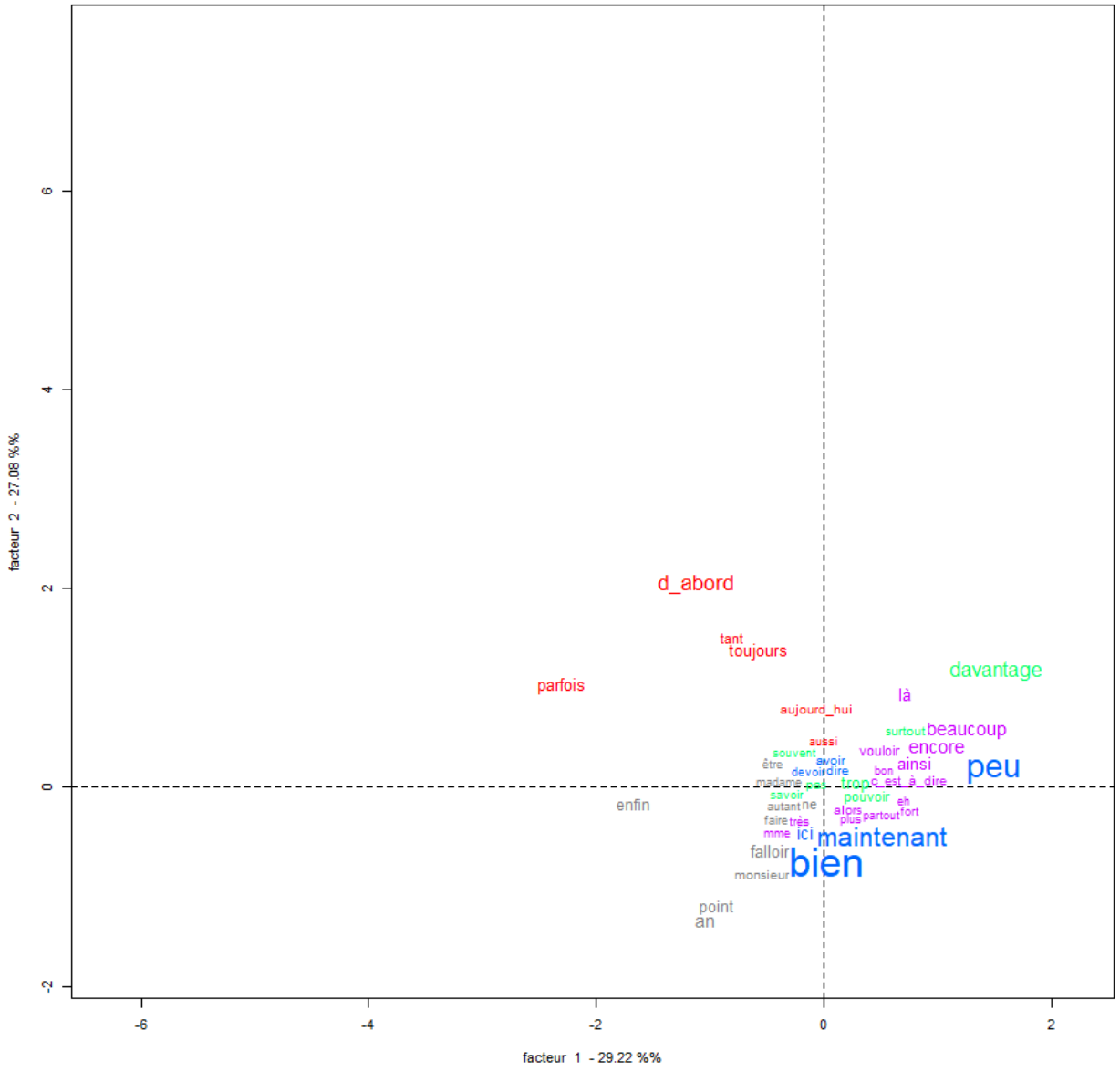


Figure 9: Classification des formes supplémentaires

En observant ces graphiques ainsi que les chi2 d'association, les concordanciers relatifs aux formes caractéristiques de chaque classe, les segments caractéristiques, les segments répétés, les profils et anti-profils de chaque classe, on obtient :

Classe 1 (6.03% des segments) : "Violences et victimes"
Principaux mots clés significatifs :

- victime ($\chi^2 = 112.39$, p = 2.93e-26)
- violence ($\chi^2 = 82.04$, p = 1.33e-19)
- évoquer ($\chi^2 = 47.33$, p = 6.01e-12)
- guerre ($\chi^2 = 47.33$, p = 6.01e-12)

- récemment (χ^2 = 19.70, p = 9.04e-06)
- développer (χ^2 = 19.70, p = 9.04e-06)

Interprétation :
Cette classe évoque des violences physiques et leurs victimes. Elle correspond à des parties des discours sur :

- Les violences faites aux femmes("victime", "violence", "viol").
- Des récits émotionnels ou mémoriels ("évoquer", "récemment", "développer").

Classe 2 (44.4% **des segments) : "Égalité professionnelle et droits sociaux"**
Mots clés :

- égalité (χ^2 = 19.80, p = 8.60e-06)
- professionnel (χ^2 = 14.28, p = 1.57e-04)

Interprétation :
Thématique centrée sur :

- Les luttes contre les discriminations ("égalité", "discrimination").
- Les réformes législatives ("loi").
- Les droits sociaux ("professionnel", "statut", "travail").

Classe 3 (3.88% **des segments) : "Prostitution et enjeux sociétaux controversés"**
Mots clés :

- prostitution (χ^2 = 79.41, p = 5.05e-19)
- penser (χ^2 = 36.57, p = 1.48e-09)
- initiative (χ^2 = 32.13, p = 1.44e-08)
- fondamental (χ^2 = 32.13, p = 1.44e-08)
- fin (χ^2 = 25.12, p = 5.40e-07)
- concerner (χ^2 = 23.22, p = 1.45e-06)
- nouvelle (χ^2 = 17.88, p = 2.35e-05)
- corps (χ^2 = 11.80, p = 5.92e-04)

Interprétation :
Sujets polémiques ou moraux :

- Régulation de la prostitution ("prostitution", "fin", "corps", "versement", "punir", "prestation").
- Références à des crises ("fondamental", "nécessité", "crise").

Classe 4 (1.72% **des segments) : "Évolution sociétale et futur"**
Mots clés :

- connaître (χ^2 = 184.79, p = 4.37e-42)
- génération (χ^2 = 17.92, p = 2.30e-05)
- évolution (χ^2 = 17.92, p = 2.30e-05)
- voie (χ^2 = 17.92, p = 2.30e-05)

- suite (χ^2 = 17.92, p = 2.30e-05)
- reprendre (χ^2 = 17.92, p = 2.30e-05)
- rendre (χ^2 = 17.92, p = 2.30e-05)
- quota (χ^2 = 17.92, p = 2.30e-05)
- remercier (χ^2 = 13.01, p = 3.09e-04)
- nécessaire (χ^2 = 13.01, p = 3.09e-04)

Interprétation :

- Thèmes de changement social ("génération", "évolution", "voie").
- Mesures ("quota").
- Lexique de transition ("suite", "reprendre", "rendre").

Classe 5 (43.97% **des segments) : "Libertés individuelles et éducation"**
Mots clés :

- liberté (χ^2 = 16.13, p = 5.92e-05)
- fille (χ^2 = 14.72, p = 1.25e-04)
- également (χ^2 = 14.72, p = 1.25e-04)

Interprétation :

- Droits reproductifs ("IVG", "contraception").
- Émancipation et éducation des jeunes ("fille", "jeune", "éducation", "enfant", "mère", "garçon").
- Libertés individuelles ("liberté", "protection", "vivre", "ensemble").

On remarque que 2 classes sont majoritairement évoquées, les classes 2 et 5 ce qui montre que les sujets les plus abordés sont les enjeux sociaux et les questions de libertés et d'éducation.

Nous allons maintenant essayer de déterminer à quels discours chaque classe peut le plus se rattacher afin de pouvoir déterminer les thématiques sur lesquelles chaque président s'est focalisé lors de son discours. En utilisant les différentes données à disposition et notamment la table de contingence et les valeurs du Chi-2 des tables de profils et d'anti-profils, on obtient que :

- Mitterand évoque principalement la classe 2 (59% de ses segments) et dans une moindre mesure la classe 5 (36%). Il est également le seul à parler de la classe 4, même s'il n'en parle que 3% des segments. Les classes 1 et 3 sont elles très peu évoquées dans son discours.
- Chirac évoque majoritairement la classe 5 mais son discours étant très court cela représente 4 fois moins de segments que Mitterand ou Hollande pour cette même classe. En revanche, proportionnellement à la taille de son discours, il représente un grand nombre des segments de la classe 1(23% de son discours contre 11% pour Hollande et moins d'1% pour Mitterand). Les classes 3 et 4 ne sont pas évoquées et la classe 2 est évoquée de manière négligeable par rapport à Hollande et surtout à Mitterand.
- Enfin, Hollande évoque principalement la classe 5 (57% de ses segments) (autant de segments que Mitterand) ainsi que la classe 3 (en comparaison des 2 autres discours) où il représente 78% des segments de la classe. Il évoque aussi les classes 1 et 2 même s'il évoque bien moins la classe 2 que Mitterand. Seule la classe 4 n'est pas évoquée dans son discours.

6 Analyse des similitudes

Nous allons maintenant réaliser des graphiques de similitude afin de voir quels termes sont les plus associés dans chaque discours.

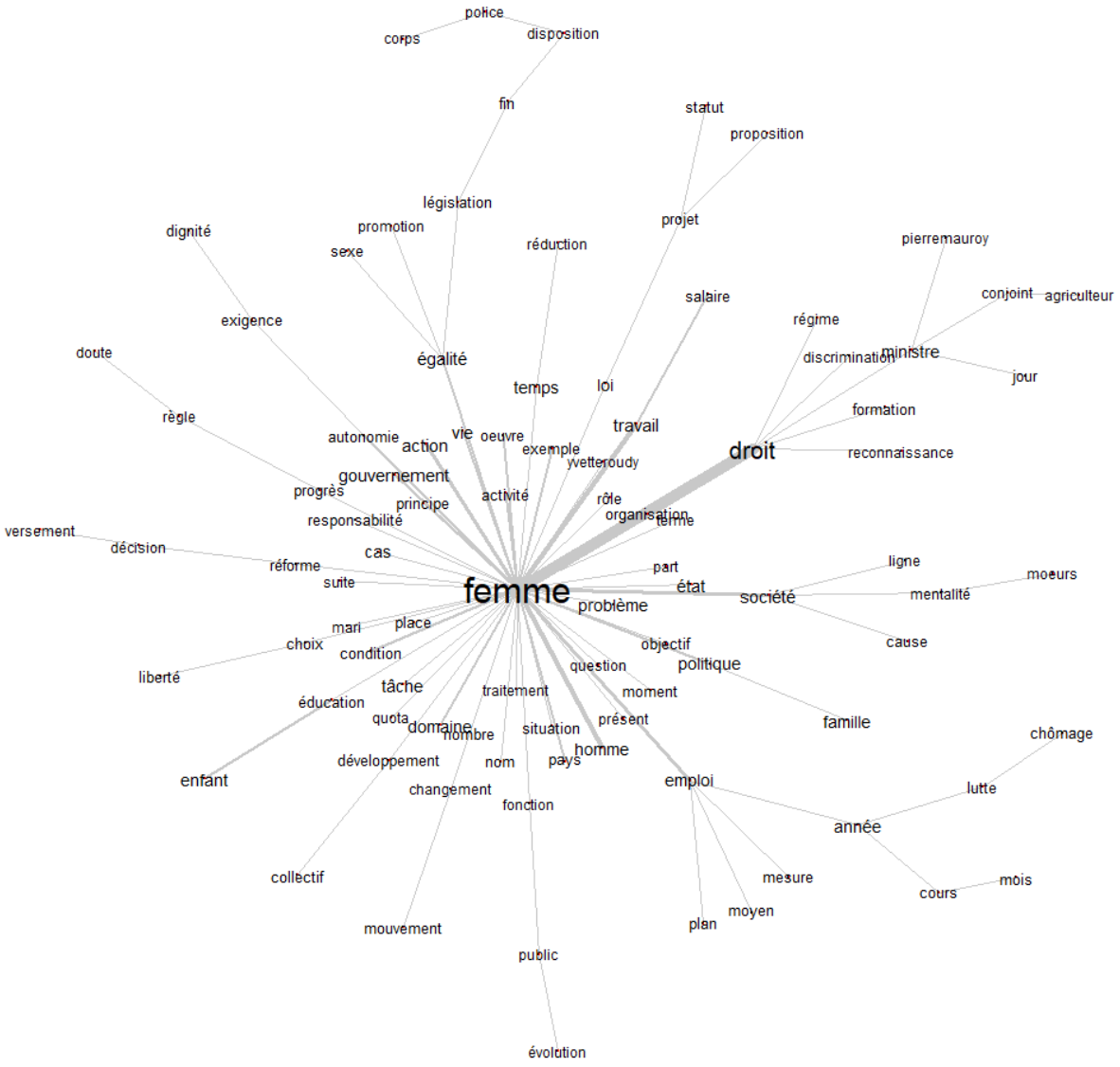


Figure 10: Graphe des similitudes pour le discours de François Mitterand avec uniquement les formes actives

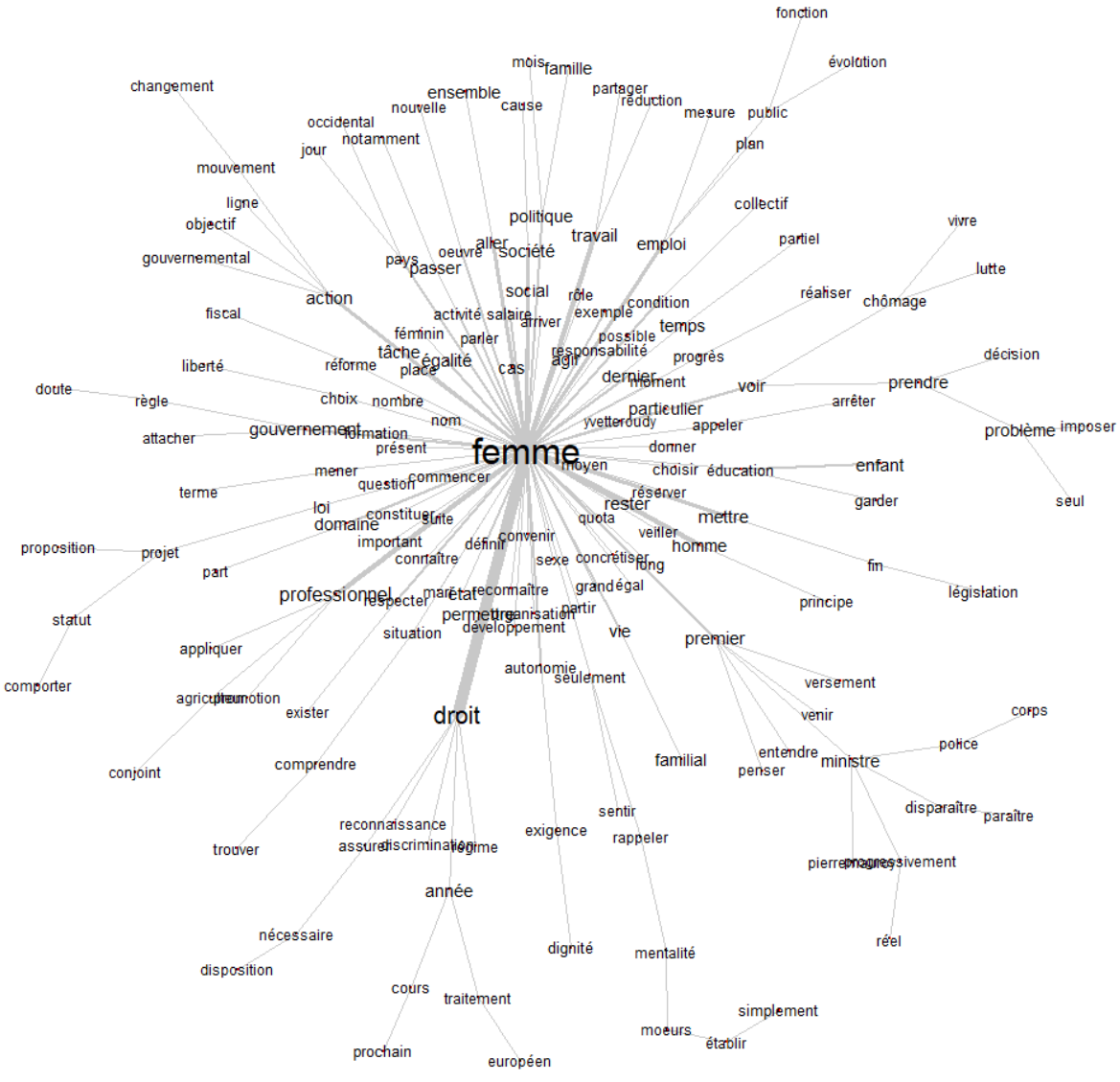


Figure 11: Graphe des similitudes pour le discours de François Mitterand avec les formes actives et les verbes supplémentaires

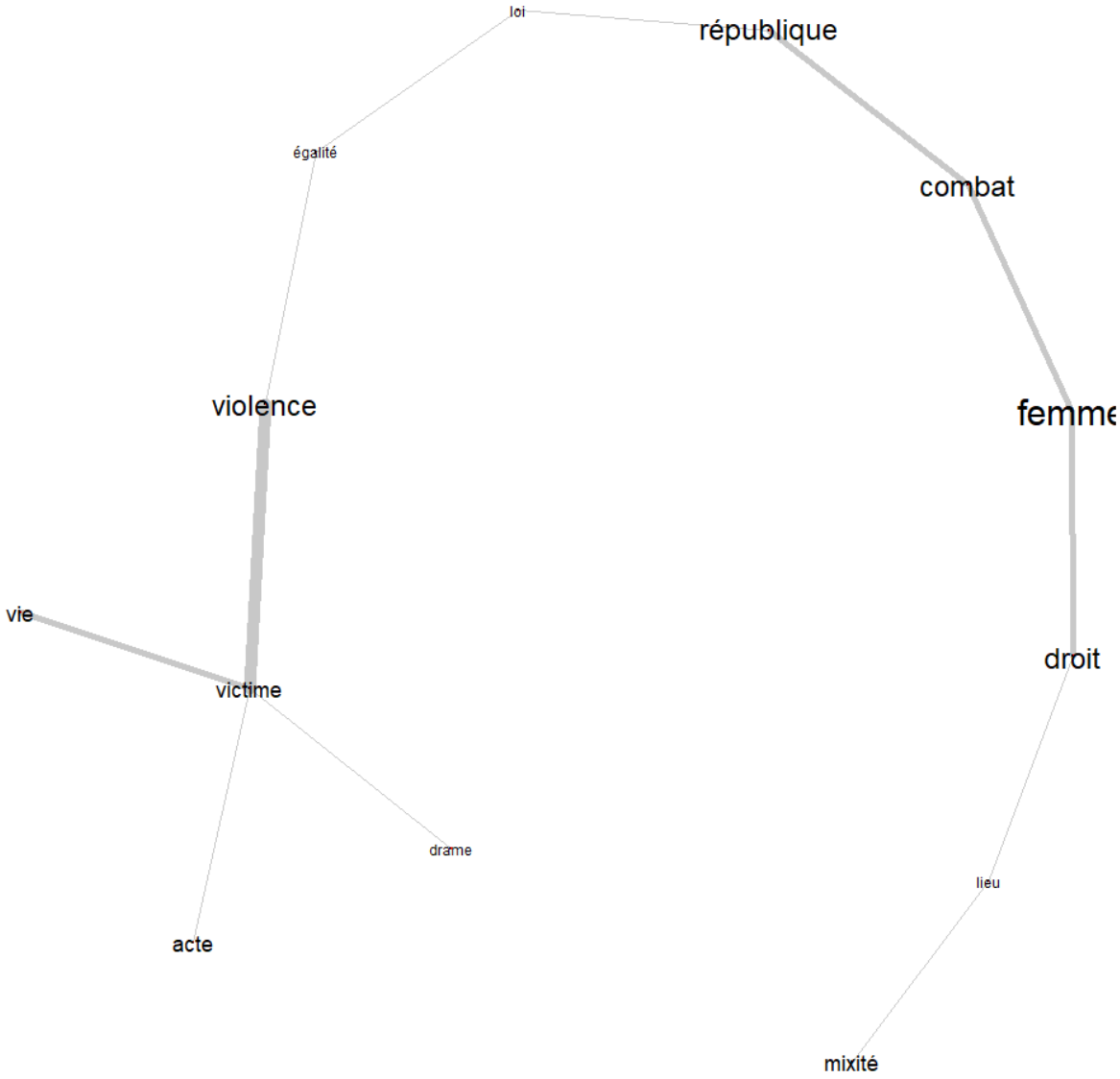


Figure 12: Graphe des similitudes pour le discours de Jacques Chirac avec uniquement les formes actives

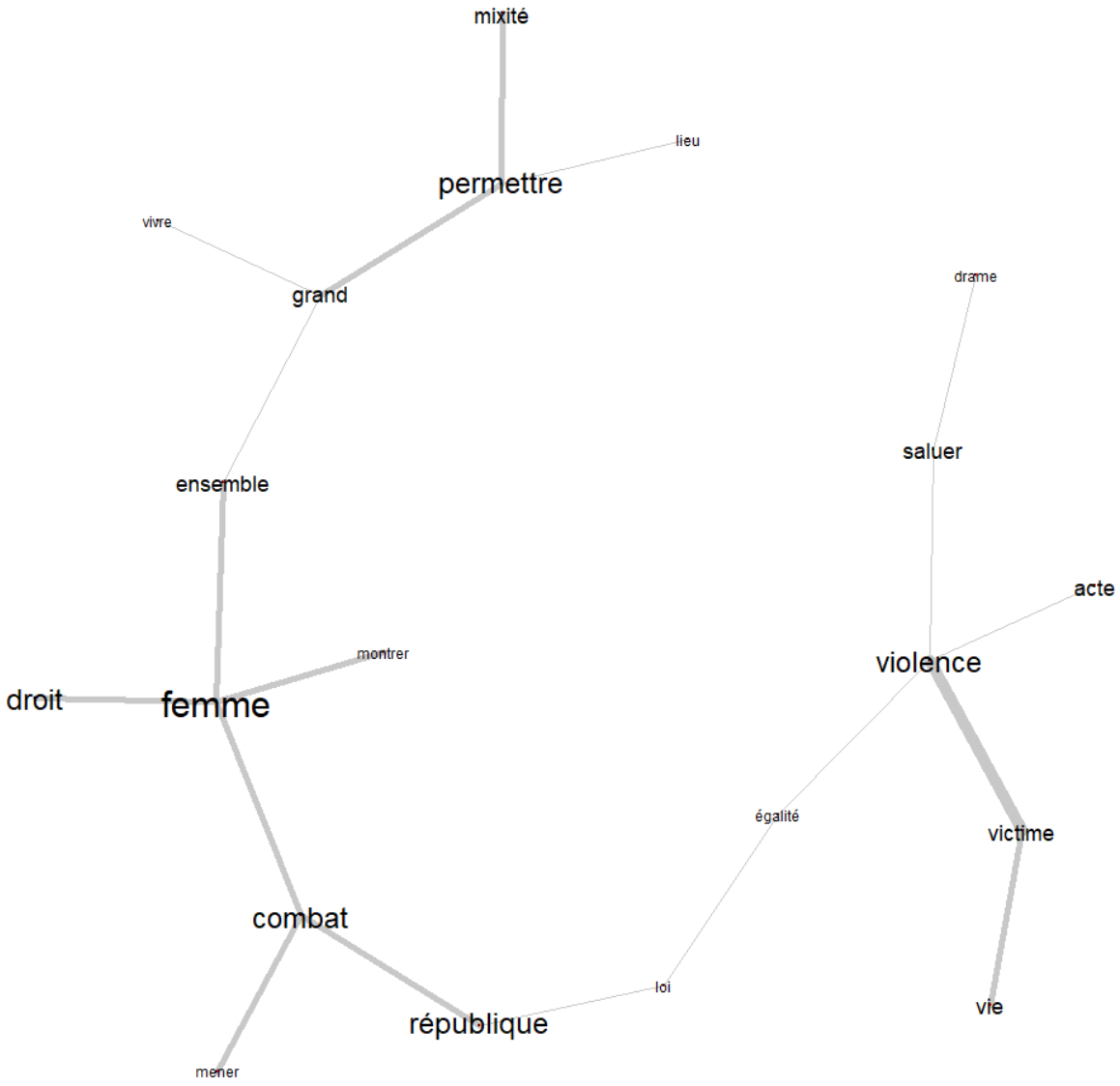


Figure 13: Graphe des similitudes pour le discours de Jacques Chirac avec les formes actives et les verbes supplémentaires

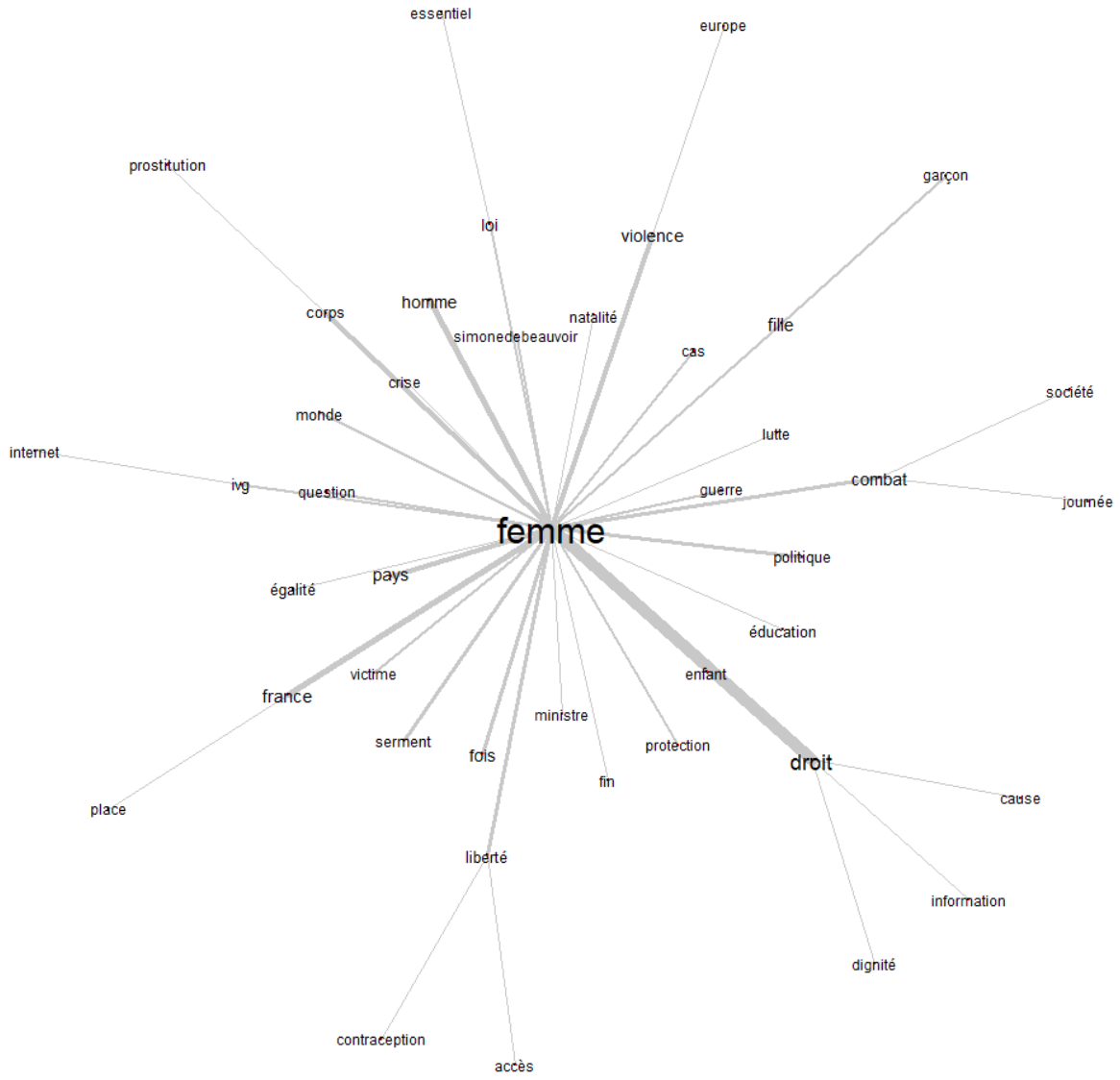


Figure 14: Graphe des similitudes pour le discours de François Hollande avec uniquement les formes actives

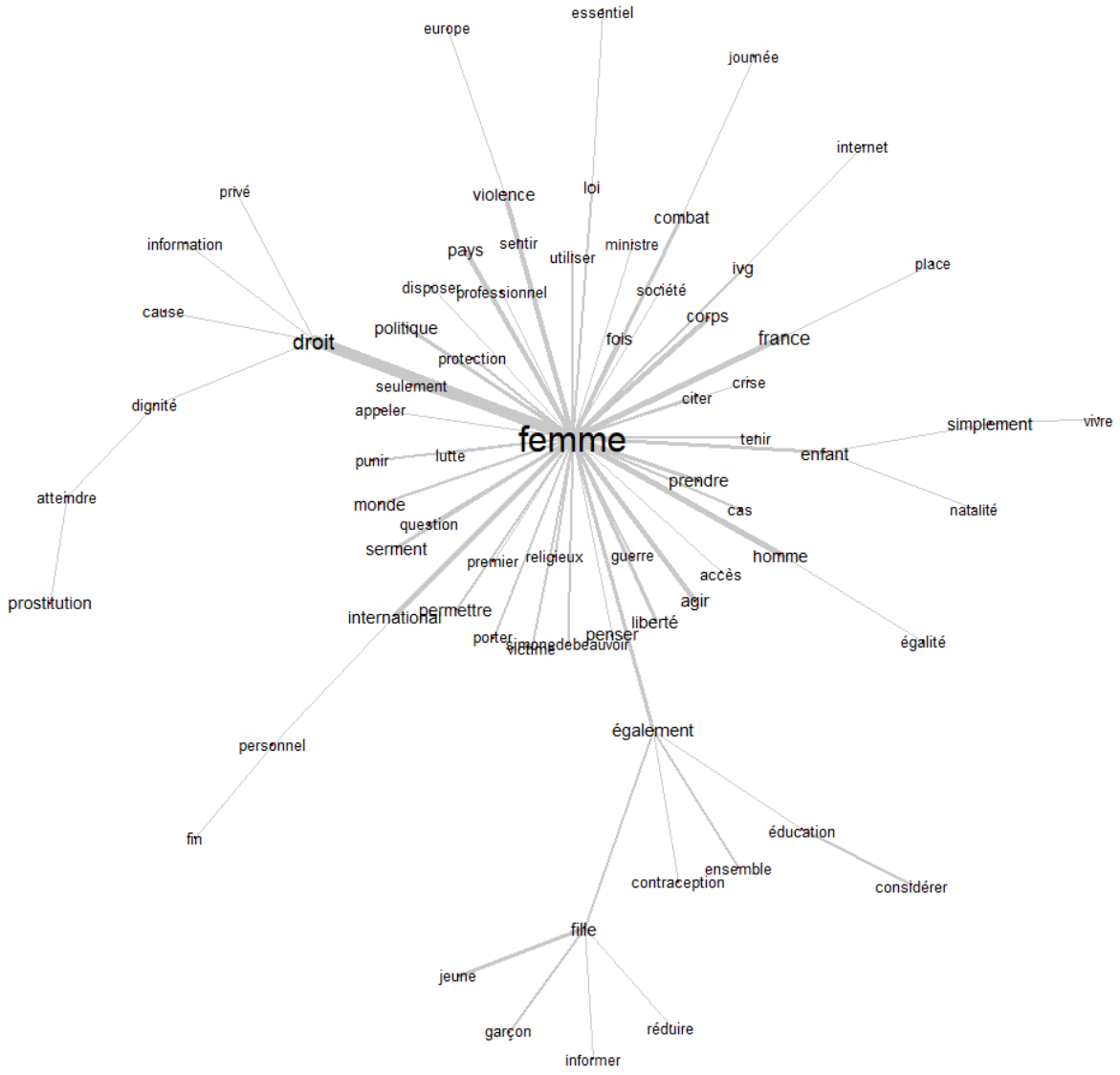


Figure 15: Graphe des similitudes pour le discours de François Hollande avec les formes actives et les verbes supplémentaires

Comme le laissaient présager les analyses précédentes, le discours de Jacques Chirac comporte nettement moins de formes apparaissant au moins cinq fois, contrairement à ceux de François Mitterrand et François Hollande, qui se distinguent par une plus grande richesse lexicale et une présence marquée de formes actives. Ces deux derniers discours présentent d'ailleurs une forte association entre les termes femme et droit, ce qui confirme, comme l'avaient déjà suggéré les nuages de mots, qu'ils sont particulièrement centrés sur les droits des femmes.
Dans le discours de François Mitterand, on retrouve clairement les thématiques de l'emploi et de l'égalité, très présentes dans la Classe 2, ainsi que la notion de futur, caractéristique de la Classe 4.
Du côté de Jacques Chirac, les principaux axes portent sur les violences et les victimes, thématiques de la Classe 1, ce qui corrobore les observations faites à partir du nuage de mots et de la classification. La question de la mixité y est également abordée.

Enfin, le discours de François Hollande met surtout l'accent sur les thématiques de la Classe 5 — l'éducation des jeunes, la contraception et l'IVG — ainsi que celles de la Classe 1 liées aux violences.

7 Conclusion

En conclusion, cette étude révèle des différences marquées dans le traitement des enjeux féministes par les trois présidents lors de la Journée internationale des femmes. Le discours de Mitterrand (1982) se concentre massivement sur l'égalité professionnelle (82 segments sur 138 en classe 2), reflétant la préparation de la loi Roudy sur l'égalité professionnelle (votée en 1983) et son orientation politique de gauche. Il évoque également la classe 5 (49 segments) sur les libertés et l'éducation tandis que les autres thèmes (violences, prostitution) sont quasiment absents, probablement en raison des tabous et de leur moindre visibilité politique à l'époque. Chirac (2006) adopte une approche minimaliste (seulement 22 segments au total) et sélective, évoquant, en comparaison avec les deux autres discours, principalement les violences (classe 1, 5 segments) sans aborder les questions sociales (0 segment en classe 2), ce qui peut s'expliquer par son orientation politique de droite. Il évoque majoritairement dans son discours les thématiques de la classe 5 sur les libertés et l'éducation mais il en parle tout de même 4 fois moins que Mitterand ou Hollande.
Hollande (2017) présente le discours le plus diversifié : il combine les violences (classe 1, 8 segments), l'égalité professionnelle (classe 2, 16 segments, moins que Mitterrand mais bien plus que Chirac), la prostitution (classe 3, 7 segments), sûrement grâce à la loi de 2016 sur la prostitution, et surtout les libertés individuelles (classe 5, 41 segments), illustrant l'élargissement des revendications féministes aux questions corporelles et sociétales. Ces écarts reflètent des priorités politiques différenciées ainsi qu'une évolution des revendications féministes entre les années 1980 et 2010.