

# Topic Modeling sur deux versions latines du Lévitique

## Description du travail effectué

Alice Leflaëc

Avril 2023

### 1 Présentation du projet

Dans le cadre de notre travail postdoctoral, nous préparons l'édition commentée et la traduction française du livre du Lévitique dans l'*Heptateuque*, un poème latin du V<sup>e</sup> siècle. L'*Heptateuque* est une réécriture en vers des sept premiers livres de l'Ancien Testament.

Le procédé de la paraphrase appelle naturellement une comparaison entre le texte source et la réécriture. Cette comparaison concerne aussi bien la forme que le fond. Dans le cas de l'*Heptateuque*, le changement de forme est particulièrement visible puisque l'on passe de la prose aux vers. On observe également, pour le livre du Lévitique en particulier, un phénomène de contraction : les 27 chapitres du Lévitique biblique sont condensés dans 309 vers. Cette condensation oblige le poète à faire des choix au niveau du contenu et à écarter certains passages. Les choix opérés par le poète et les procédés de réécriture seront étudiés attentivement dans notre édition commentée. Il nous a semblé qu'une première approche par *topic modeling* pouvait nourrir de manière intéressante cette réflexion. Nous avons donc décidé de faire du *topic modeling* sur le Lévitique dans la Vulgate et dans l'*Heptateuque* puis de comparer les résultats obtenus pour voir si les principaux *topics* étaient les mêmes dans les deux textes.

Il convient de préciser que l'auteur de l'*Heptateuque* n'utilisait non pas la Vulgate mais une *Vetus Latina*. Malheureusement, nous ignorons sur quelle version il s'appuyait et nous avons fait le choix de comparer le texte de l'*Heptateuque* à celui de la Vulgate. Comme il s'agit de s'intéresser avant tout aux thèmes les plus traités dans chacun de ces textes et que ces derniers ne diffèrent pas entre les *Veteres Latinae* et la Vulgate, travailler sur le texte de la Vulgate n'était pas trop gênant. Cela exclut une étude comparative trop précise des termes employés dans le texte biblique et dans le poème de l'*Heptateuque*, mais, de toute façon, le phénomène même de paraphrase appelle la *variatio*, notamment synonymique. Il pourrait, toutefois, être intéressant, pour compléter notre étude, de faire du *topic modeling* sur une *Vetus Latina*, par exemple la *Pentateuchi versio latina antiquissima e codice Lugdunensi* éditée par Ulysse Robert (le texte est disponible [en ligne](#)).

## 2 Description des données

Nous avons travaillé sur deux versions latines du livre du Lévitique.

**Lévitique dans la Vulgate** : Nous avons récupéré le texte de la Vulgate sur le site [Bible-Gateway](#). Il a fallu copier le texte dans un éditeur de texte (SublimeText) avant de l'enregistrer au format .txt.

**Lévitique dans l'*Heptateuque*** : Le texte de l'*Heptateuque* a directement été récupéré en format .txt sur la bibliothèque numérique [Perseus Scaife Viewer](#)<sup>1</sup>.

## 3 Description de la méthode

Pour faire du *Topic Modeling* sur les deux textes sélectionnés, nous avons utilisé le logiciel d'analyses statistiques R.

Nous ne commenterons pas toujours en détail dans ce document le code utilisé, car cela a été fait abondamment dans le script dans R. Nous nous contenterons d'exposer les principales étapes de la méthode appliquée.

### 3.1 Préparation des données

Avant de pouvoir analyser la fréquence des termes dans nos deux textes et de faire du *topic modeling*, il a fallu préparer les données.

1. *Définition de la session de travail* : Après avoir défini le chemin vers le notebook, il a fallu récupérer les textes (cf. section 2) et associer chaque texte à une variable créée dans le script R grâce à la fonction `readLines`.
2. *Premier nettoyage du texte : suppression de la ponctuation et lemmatisation* : Nous avons supprimé la ponctuation à l'aide de la fonction `removePunctuation` qui a nécessité l'installation de la *library* `tm`, puis nous avons rangé les textes nettoyés dans un dossier baptisé « Clearer ». Nous avons ensuite lemmatisé chacun des textes au moyen de l'interface Pyrrha. Les textes ont été exportés depuis cette interface en format .tsv et convertis en format .csv avec un convertisseur en ligne.
3. *Second nettoyage : réduction à la minuscule et retrait des stopwords* : Afin de mettre en évidence les mots les plus fréquents dont la charge significative est forte et de pouvoir dégager les différentes thématiques des textes, il a fallu retirer tous les mots-outils, les *stopwords* dont l'importance est faible pour la thématique d'un texte. Nous avons utilisé une liste créée par Mathilde Schwoerer à partir d'une liste proposée par Aurélien Berra

---

1. Dans la bibliothèque en ligne Perseus Scaife Viewer, l'*Heptateuque* est une oeuvre du poète Cyprianus Gallus. Cette attribution est rejetée. Cf. M. R. Petringa, *Il poema dell'Heptateuchos*. Itinera philologica tra tardo antico e alto medioevo, Catane, 2016, p. 19-28 « Introduzione. Un poema anonimo ».

(cf. [stopwords grecs et latins](#)) et modifiée par nos soins en fonction de nos textes. Nous avons créé pour chaque texte une chaîne de caractères destinée à contenir tous les mots, sans ponctuation, ne relevant pas des *stopwords*.

4. *Création d'une liste pour une approche bag of words*<sup>2</sup> : Chaque texte a été découpé en dix morceaux placés dans une liste baptisée « Extraits ».
5. *Transformation en matrice vectorielle* : Chaque document a été transformé en une matrice vectorielle grâce à la fonction *Corpus* qui transforme le document en corpus et à la fonction *VectorSource* qui le transforme en vecteur. Il a fallu installer la *library* tidytext pour effectuer cette transformation.
6. *Création d'un document term matrix (dtm)* : Une DTM<sup>3</sup> a été créée pour chaque document grâce à la fonction *DocumentTermMatrix*.

### 3.2 Analyse des données : fréquence des termes

Une première étape dans l'analyse des données de nos deux textes consiste à observer la fréquence des termes.

1. *Graphe représentant la fréquence des termes* : Nous avons commencé par dessiner un graphe représentant la densité de la fréquence des termes dans le Lévitique de la Vulgate et dans celui de l'*Heptateuque*. Nous avons installé, pour cela, la *library* ggplot2 qui permet de représenter sous la forme d'un graphique des données issues d'un *dataframe*.
2. *Analyse des données* : À l'aide de la fonction *findFreqTerms* à laquelle nous avons attribué différentes valeurs pour le nombre d'occurrences, nous avons ensuite affiché les mots les moins fréquemment utilisés puis les mots les plus fréquemment utilisés (pour l'analyse, cf. section 4.2).

Pour le Lévitique dans l'*Heptateuque*, nous avons également affiché les mots les plus fréquemment associés aux trois mots les plus utilisés (*dominus*, *uir* et *ius1*) grâce à la fonction *findAssocs*.

3. *Nettoyage de la DTM* : Nous avons, pour finir, nettoyé la DTM en éliminant tous les rangs vides *i.e.* tous ceux qui ne contenaient pas de mot.

### 3.3 Topic Modeling

Une fois la fréquence des termes analysée, nous sommes passée au *topic modeling* proprement dit.

1. *Installation de la library pour le topic modeling* : La *library* topicmodels a été installée.

---

2. Voir le script R pour la définition de cette approche.

3. Une DTM est une matrice mathématique décrivant la fréquence des termes qui apparaissent dans une collection de documents.

2. *LDA Latent Dirichlet allocation* : L'allocation de Dirichlet latente est un modèle génératif probabiliste qui permet d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données (cf. [Wikipédia](#)). Le modèle va classer aléatoirement tous les mots du texte en un nombre défini de sujets et s'efforcer d'affiner cette répartition de manière itérative en observant les contextes. Le score  $\beta$  contient la probabilité que chaque mot appartienne à un *topic*.

Pour chacun des textes, nous avons appliqué une répartition en deux puis en trois *topics* en utilisant la fonction *LDA* et en redéfinissant à chaque fois la variable  $k$ . Nous avons ensuite utilisé la fonction *tidy* pour afficher les résultats de notre test dans un *dataframe* récapitulatif.

3. *Paramètres de Gibbs* : Il est possible, grâce aux paramètres de Gibbs, d'affiner le modèle précédent : le score  $\beta$  d'un mot est calculé en s'appuyant sur celui des mots voisins et le nombre optimal de *topics* peut être déterminé.

Après avoir installé la *library* *ldatuning*, nous avons entrepris de déterminer le nombre optimal de thèmes pour chacun de nos deux textes au moyen de la fonction *FindTopicsNumber*. Nous avons utilisé pour cela la méthode Gibbs et les quatre métriques suivantes : Griffiths2004, CaoJuan2009, Arun2010 et Deveaud2014. Le calcul pour le *topic modeling* a ensuite été exécuté : il s'agissait d'une LDA à laquelle nous avons appliqué les paramètres de Gibbs. Enfin, nous avons affiché les premiers résultats pour chacun des modèles et, au moyen de la fonction *tidy* nous avons construit une matrice avec les  $\beta$  des *tokens* afin de montrer la probabilité d'appartenance d'un *token* à un *topic*.

### 3.4 Visualisation

La dernière partie de notre script dans R a été consacrée à la visualisation des calculs effectués pour le *topic modeling*. Diverses visualisations ont été réalisées.

1. *Récupération des mots* : Nous avons tout d'abord récupéré les mots appartenant à chaque *topic*. Il a fallu installer la *library* *dplyr* qui facilite le traitement et la manipulation de données contenues dans une ou plusieurs tables en proposant une syntaxe sous forme de verbes. Puis les mots récupérés ont été affichés dans des diagrammes à barres (*barplots*).
2. *Association des tokens aux topics* : Les *tokens* ont ensuite été associés aux *topics* sous la forme de deux *geom-tile*, des rectangles avec des tuiles dont la couleur indique la probabilité qu'un *token* appartienne à un *topic*. Cette visualisation, réalisée grâce aux fonctions *melt* et *ggplot*, a nécessité l'installation de la *library* *reshape2*.
3. *Observation du score gamma* : Le score gamma, *i.e.* la probabilité qu'un document contienne un sujet, a été affiché dans un *dataframe* pour le Lévitique dans la Vulgate puis dans l'*Heptateuque*.

4. *Nuages de mots* : Grâce aux *libraries* wordcloud, RColorBrewer et wordcloud2, nous avons, pour finir, généré des nuages de mots.

## 4 Description des résultats

### 4.1 Quelques précautions

Les résultats présentés dans ce document sont à prendre avec précaution pour deux raisons :

- Tout d’abord, comme nous l’avons exposé dans la présentation du projet, nous comparons le texte de l’*Heptateuque* à celui de la Vulgate, alors que le poète de l’*Heptateuque* ne travaillait pas avec la Vulgate (cf. section 1).
- Le principal problème vient, toutefois, de la taille des textes. Il s’agit de deux textes (trop) courts pour une analyse en *topic modeling*. La *sparsity* de chacune des DTM le montre bien : celle de la DTM de la Vulgate<sup>4</sup> est de 81% et celle de l’*Heptateuque* est de 87%.

Les résultats auraient été plus probants si nous avions travaillé à l’échelle du poème entier de l’*Heptateuque* et des sept premiers livres de l’Ancien Testament qu’il paraphrase ; mais cela nous intéressait moins pour notre propre travail.

### 4.2 Analyse de la fréquence des termes

Sans surprise, on observe, que ce soit dans la Vulgate ou dans l’*Heptateuque*, la loi de Zipf : beaucoup de termes sont employés très peu de fois et très peu de termes sont employés à maintes reprises.

Ainsi, 1292 mots dans la Vulgate et 895 dans l’*Heptateuque* ont moins de 10 occurrences. On constate que le nombre de mots varie assez peu si l’on augmente le nombre d’occurrences. Ainsi, 1374 mots ont moins de 15 occurrences et 1404 en ont moins de 20 dans la Vulgate. C’est encore plus significatif dans l’*Heptateuque* où 898 mots ont moins de 15 occurrences et 899 en ont moins de 20.

Plus intéressante est l’observation des mots avec une forte fréquence. À l’exception de *dominus*, les mots les plus fréquents dans l’*Heptateuque* ne sont pas les mêmes que dans la Vulgate. Dans la Vulgate, 22 mots apparaissent plus de cinquante fois. On y retrouve des noms de personnes (*Dominus*, *Aaron*, *Moyses*...) ainsi que du vocabulaire relatif au péché ou à la sainteté (*peccatum*, *sanctus*, *immundus*). Seuls quatre mots sont employés plus de cent fois : *dominus*, *filius*, *offero* et *omnis*. Les mots les plus fréquemment employés dans l’*Heptateuque* sont particulièrement intéressants. Six mots sont utilisés plus de dix fois : *annus*, *deus*, *dominus*, *uir*,

---

4. Nous employons à partir de maintenant, par souci de simplification, le terme Vulgate pour désigner le livre du Lévitique dans la Vulgate et celui d’*Heptateuque* pour le Lévitique dans ce poème.

*corpus* et *ius1* (au sens de « droit »). Si le vocabulaire du sacrifice ou de la sainteté est ici absent, ces mots les plus fréquents correspondent quand même à des éléments essentiels du livre du Lévitique : le rapport entre Dieu et l’homme qui se manifeste par la loi/le droit et par une organisation de l’année. Trois mots seulement sont présents plus de douze fois : *dominus*, *uir* et *ius1*. Ces trois mots les plus fréquents (à l’exception, bien sûr, des mots-outils ôtés préalablement) offrent, en quelque sorte, une véritable synthèse du Lévitique : la relation entre Dieu et l’homme définie par le droit. Ainsi, dans l’*Heptateuque*, très peu de mots sont fréquemment utilisés mais ils sont significatifs.

Nous avons cherché quels étaient les mots les plus fréquents associés aux trois termes les plus utilisés dans l’*Heptateuque*. Il en ressort une grande variété de termes. Cela peut s’expliquer par des contraintes métriques (même s’il y a un certain nombre de vers plus ou moins formulaires dans cette oeuvre), mais aussi et surtout par le fait que Dieu (*dominus*) et l’homme (*uir*) sont présents dans tout le Lévitique mais qu’ils ne sont pas, en soi, des thématiques et que la question du droit (*ius1*) est le thème central de ce livre et qu’elle se décline en toute une série de thématiques variées.

### 4.3 Analyse du Topic Modeling

Les paramètres de Gibbs associés à une LDA ont été utilisés pour le *topic modeling*. Le calcul de nombre optimal de *topics* pour chacun des textes a donné les résultats suivants : 7 *topics* pour la Vulgate, 6 *topics* pour l’*Heptateuque*. Si les calculs du *topic modeling* ont été produits à partir de divers nombres de *topics*, les visualisations ont été produites à partir du nombre optimal déterminé de *topics* (cf. dossier Visualisations dans GitHub).

**Pour la Vulgate :** L’association des *tokens* aux *topics*, montre, pour les quatre termes les plus utilisés, que *dominus* est très présent dans trois *topics* mais surtout dans le premier. Le terme *offero* est essentiellement présent dans le quatrième *topic*. Quant aux termes *filius* et *omnis*, ils sont présents de manière plus discrète dans plusieurs *topics*, *filius* se trouvant surtout dans le premier et le cinquième *topic*.

Les *topics* suivant ont pu être déterminés :

- *Topic 1* : Modalités du rapport entre Dieu et les hommes
- *Topic 2* : Péché et purification
- *Topic 3* : Calendrier et législation (manière dont la Loi régit l’année)
- *Topic 4* : Sacrifice du point de vue du fidèle
- *Topic 5* : Tout ce qui est relatif à la prêtrise
- *Topic 6* : Impureté, notamment les maladies de peau
- *Topic 7* : Relations familiales devant Dieu

**Pour l’*Heptateuque* :** L’association des *tokens* aux *topics*, montre, pour les trois termes

les plus fréquents, une nette appartenance à un *topic* précis. Si *dominus* est assez présent dans le cinquième *topic*, il l'est surtout dans le sixième. Le mot *uir* est surtout présent dans le deuxième *topic* tandis que *ius1* l'est essentiellement dans le cinquième.

Il a été beaucoup plus difficile de déterminer la nature de chaque *topic*. En effet, si quelques termes semblent dessiner une thématique, d'autres mots, inclus dans le même *topic*, peuvent évoquer une autre thématique. Nous proposons donc la détermination suivante des *topics*, mais il ne s'agit que de suggestions à prendre avec précaution, comme en témoignent les points d'interrogation :

- *Topic* 1 : Sacrifice
- *Topic* 2 : L'homme par rapport à ses fautes ?
- *Topic* 3 : Les rapports entre époux ?
- *Topic* 4 : Le corps (celui des hommes comme des animaux)
- *Topic* 5 : Calendrier et législation
- *Topic* 6 : Modalités et acteurs dans les rapports entre les hommes et Dieu ?

La difficulté à identifier les *topics* s'explique par la petite taille du texte et nous sommes bien consciente de nous retrouver face à un cas de *Topic Modeling* assez peu probant.

Une comparaison entre les *topics* de la Vulgate et de l'*Heptateuque* montre que ce poème reste très proche, d'un point de vue thématique, de son modèle. Si le poète de l'*Heptateuque* a condensé les 27 chapitres du Lévitique biblique et qu'il a donc, nécessairement, opéré des choix, ces choix semblent relever du détail (par exemple, la mention de tel ou tel sacrifice) et non du maintien ou non d'une thématique générale du Lévitique.