



HAL
open science

Corpus Malherbe : corpus de textes versifiés du XVIIe au début du XXe

Richard Renault

► **To cite this version:**

Richard Renault. Corpus Malherbe : corpus de textes versifiés du XVIIe au début du XXe. Journée d'étude CORLI : Traitements et standardisation des corpus multimodaux et web 2.0, May 2018, Paris, France. . hal-01873911

HAL Id: hal-01873911

<https://hal.science/hal-01873911v1>

Submitted on 8 Oct 2018

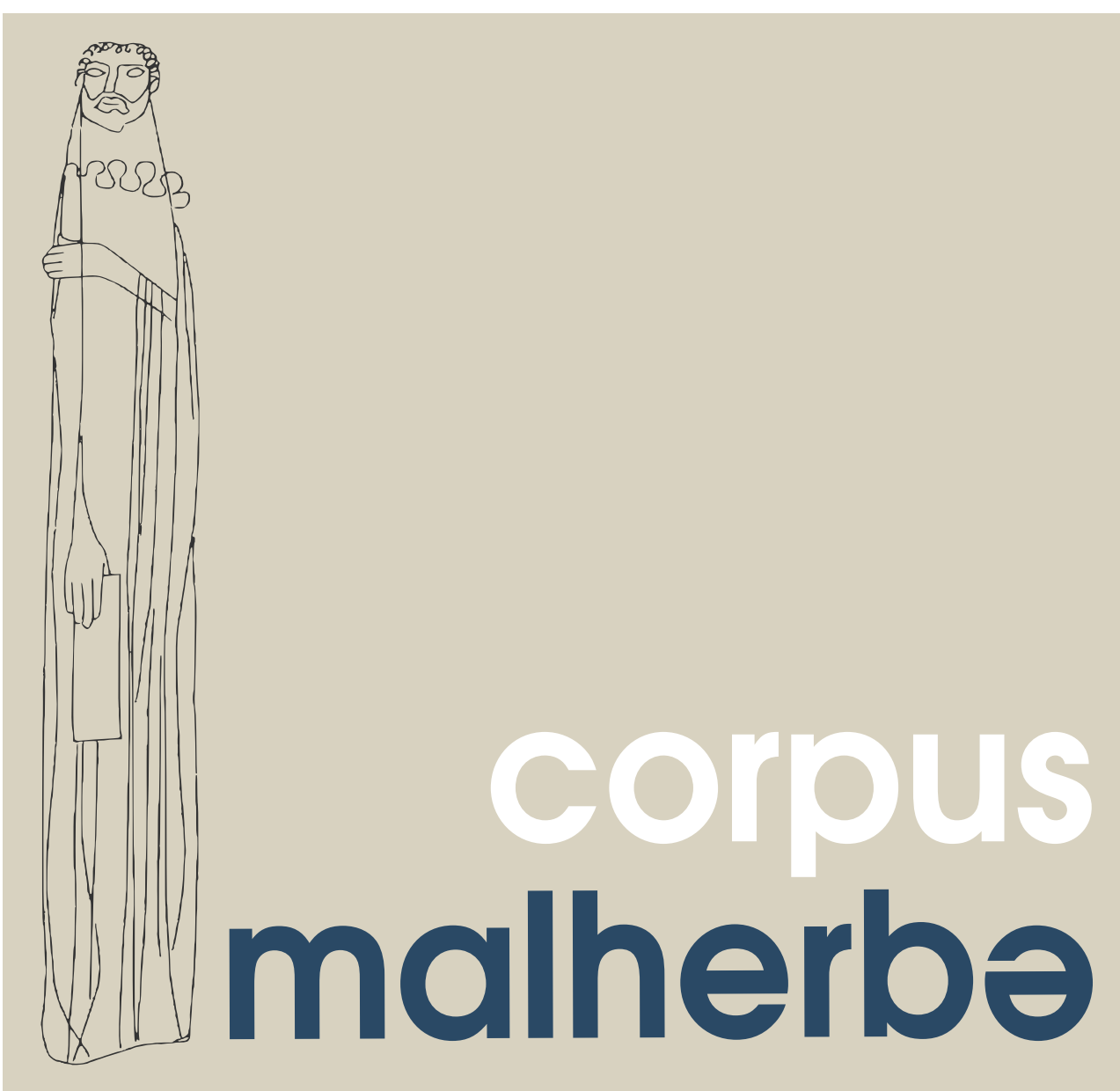
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Corpus de textes versifiés du XVII^e au début du XX^e

projet ANAMÈTRE (traitement automatique de textes versifiés)

responsables du projet : Éliane Delente & Richard Renault



responsable du corpus : Richard Renault

PRÉPARATION DU CORPUS

Exemple d'un texte dont la source est un fichier txt (site Gallica de la BnF)

Version TXT

- Préformatage du texte au moyen de tabulations et d'interlignes
- Vérification orthographique (lettres ligaturées, majuscules accentuées...)
- Normalisation de la ponctuation (espace devant un signe de ponctuation forte, tirets cadratins, guillemets...)
- Introduction des balises XML délimitant les différentes parties du texte au moyen d'expressions régulières

Version XML

- Formatage du texte (italiques, exposants, filets...)
 - Rédaction de l'en-tête éditorial
 - Validation du corpus au moyen d'un schéma XSD*
- * L'élaboration du schéma de validation ainsi que la normalisation de l'ensemble du corpus a bénéficié d'un financement du consortium CORLI-ORTOLANG en 2016.

Post-traitement

- Après une première passe d'analyse métrique automatique des vers, insertion automatique des retraits de vers en fonction de la longueur métrique du vers et de la longueur métrique maximale du poème.
- Insertion d'éléments XML-TEI de formatage métrique :
 - phonétisation de mots incomplets ou de syllabe sans voyelle (*Brrr !*) ...
 - vers sans rime, vers incomplets, vers inanalysables...
- Insertion d'éléments XML-TEI de correction lorsque l'édition imprimée comporte des erreurs.

PAGE HTML DU RECUEIL*

* <http://www.crisco.unicaen.fr/verlaire/>

PAGE HTML D'UN POÈME ANALYSÉ APRÈS TRAITEMENT AUTOMATIQUE

Charles Baudelaire
LES FLEURS DU MAL
1857-1861

SPLEEN ET IDÉAL
LXXVI

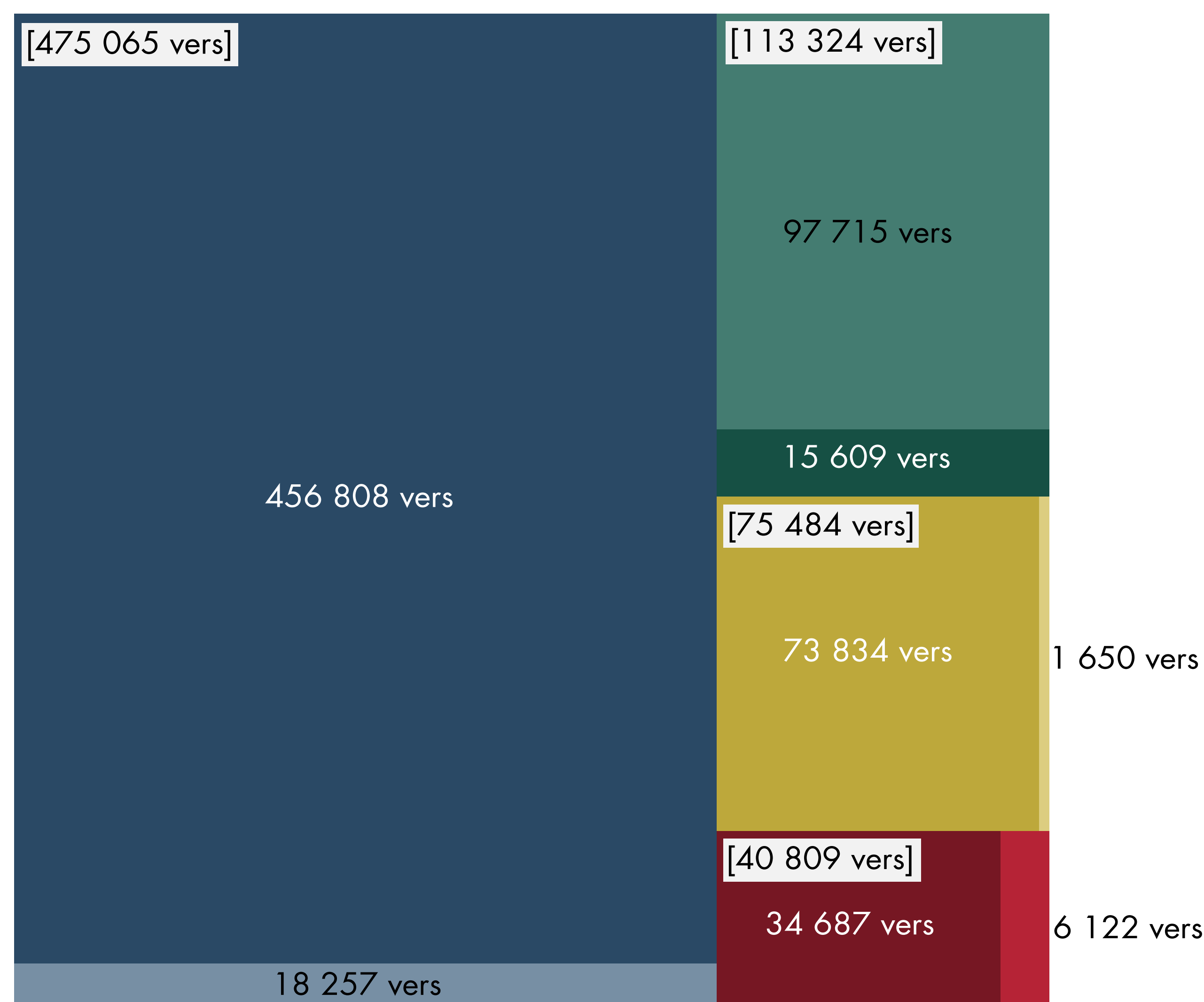
Le Tonneau de la haine

La Haine est le tonneau des pâles Danaïdes ;
La Vengeance éperdue aux bras rouges et f^orts ;
A beau précipiter dans ses ténébres vⁱdes ;
De grands seaux pleins du sang et des larmes des m^orts ;
Le Démon fait des trous secrets à ces abⁱmes ;
Par où furaient mille ans de sueurs et d'eff^orts ;
Quand même elle saurait ranimer ses victⁱmes ;
Et pour les ressaigner ressusciter leurs c^orps ;
La Haine est un ivrogne au fond d'une tav^erne ;
Qui sent toujours la soif naître de la liqu^eur ;
Et se multiplier comme l'hydre de L^erne ;
— Mais les buveurs heureux connaissent leur vain qu^eur ;
Et la Haine est vouée à ce sort lament^able ;
De ne pouvoir jamais s'endormir sous la tab^e.

profil métrique : 6+6
type : sonnet



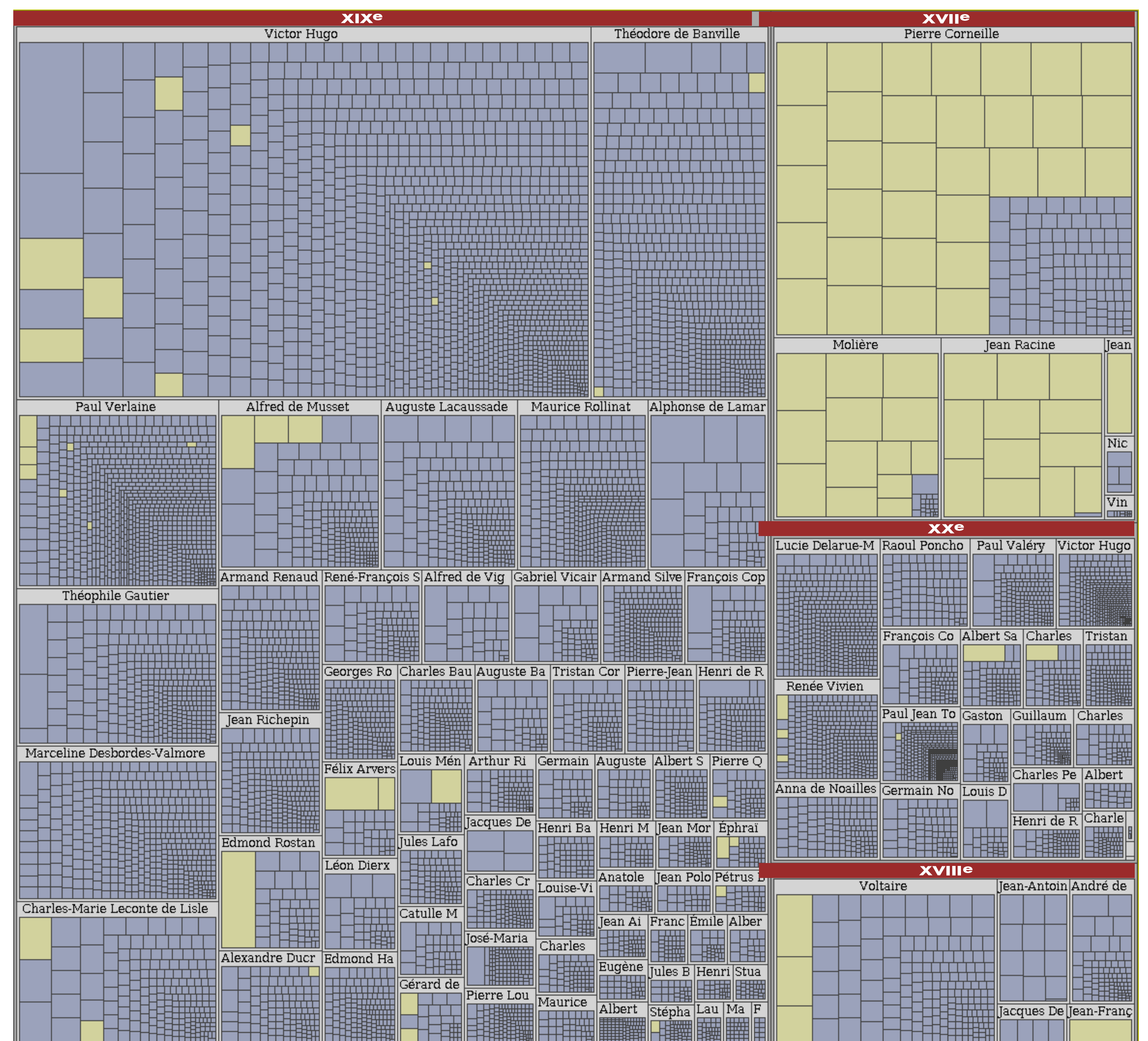
RÉPARTITION DU CORPUS SELON LA PÉRIODE ET LE TYPE DE TEXTE (proportionnellement au nombre de vers)



- XVII^e siècle
 - XVIII^e siècle
 - XIX^e siècle
 - XX^e siècle
 - 87 auteurs
 - 322 textes
 - 26 recueils de poésies
 - 102 pièces de théâtre
 - 12 670 poèmes
 - 704 385 vers
- (version 2.2, janvier 2018)

poèmes
pièces de théâtre
[nombre total de vers]

RÉPARTITION DU CORPUS SELON LA PÉRIODE ET LES AUTEURS (proportionnellement au nombre de vers)



- poème
 - pièce de théâtre
 - 87 auteurs
 - 322 textes
 - 26 recueils de poésies
 - 102 pièces de théâtre
 - 12 670 poèmes
 - 704 385 vers
- (version 2.2, janvier 2018)

poster réalisé par Richard Renault

