Richard Renault & Stéphane Ferrari
CRISCO, EA 4255
Université de Caen Normandie

journée d’étude du 8 juin 2023
CRISCO

TRAITEMENT AUTOMATIQUE DE TEXTES VERSIFIÉS : Bilan et poursuite

1. Bilan

Place du projet dans les opérations du CRISCO

2007
soumission d'un projet à l'ANR :
Coordinatrice du projet : Éliane Delente
Anamètre : Analyse automatique des formes métriques et élaboration d'une base de données de textes poétiques et théâtraux annotés (du XVIIe au début du XXe)

2008-2011
Éliane Delente et Richard Renault
Projet Anamète : Analyse automatique des formes métriques et élaboration d'une base de données de textes poétiques annotés du 17e au 19e siècle

1- Constitution d'un corpus électronique de textes poétiques et théâtraux
du début du XVIIe au début du XXe siècle
2- Outils d'analyse
3 Élaboration d'une base de données de relevés métriques

2012-2015
Éliane Delente et Richard Renault
Traitement automatique des formes métriques

1- Poursuite du projet Anamètre
   a- Constitution d'un corpus électronique de textes poétiques et théâtraux du début du XVIIe au début du XXe siècle
   b- Outils d'analyse
   c- Élaboration d'une base de données de relevés métriques
2- Métrique, linguistique et statistique

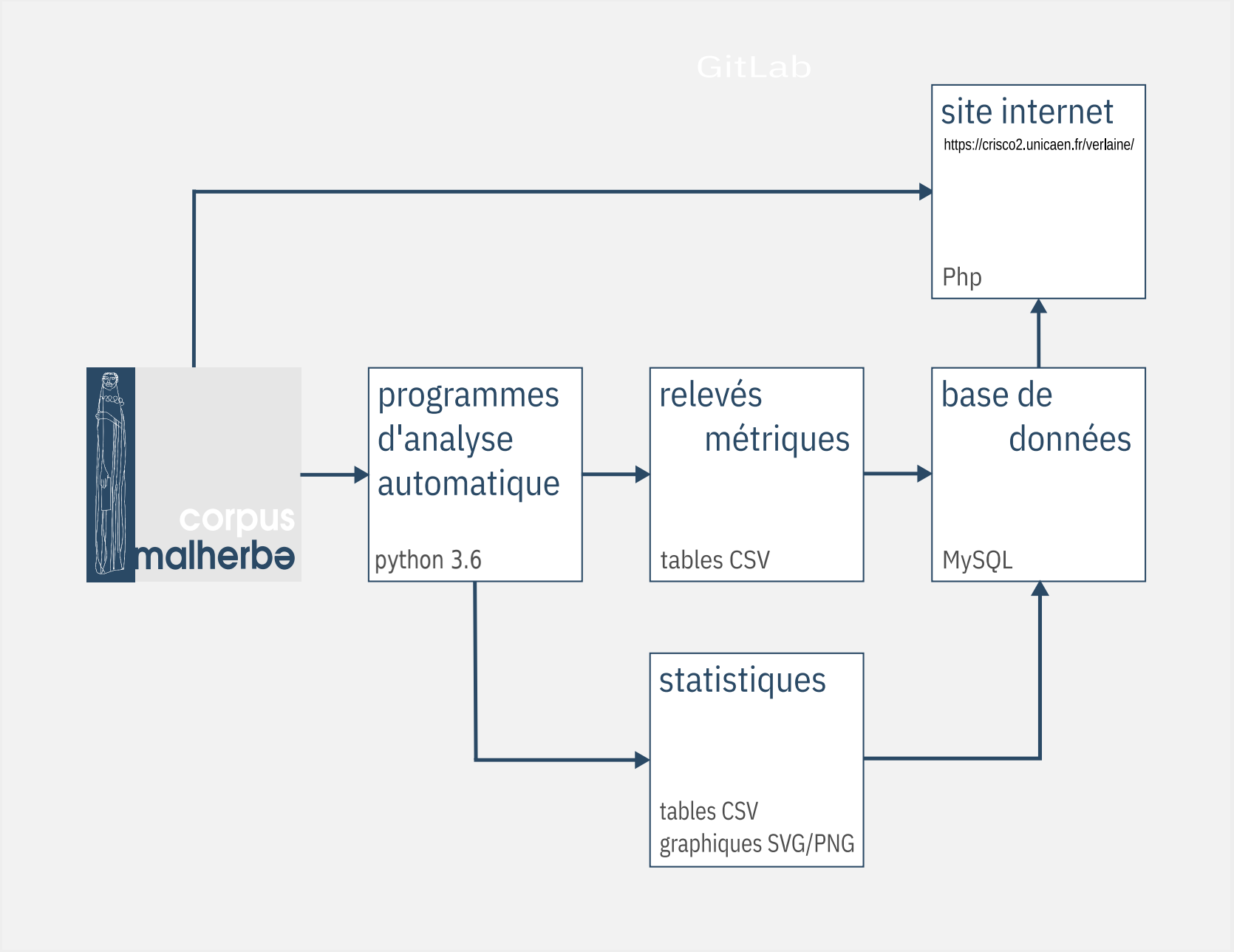
2015-2019
Éliane Delente et Richard Renault
Traitement automatique de textes versifiés

1- Constitution d'un corpus électronique de textes poétiques et théâtraux du début XVIIe au début du XXe siècle.
(Richard Renault)
2- Outils d'analyse
(Richard Renault avec la participation d'Éliane Delente)
3- Relevés métriques
   a- Relevés métriques automatiques (Richard Renault)
   b- Intégration des relevés métriques antérieurs au projet (Éliane Delente et Richard Renault)
4- Métrique, linguistique et statistiques (Éliane Delente avec la participation de Richard Renault et Dominique Legallois)

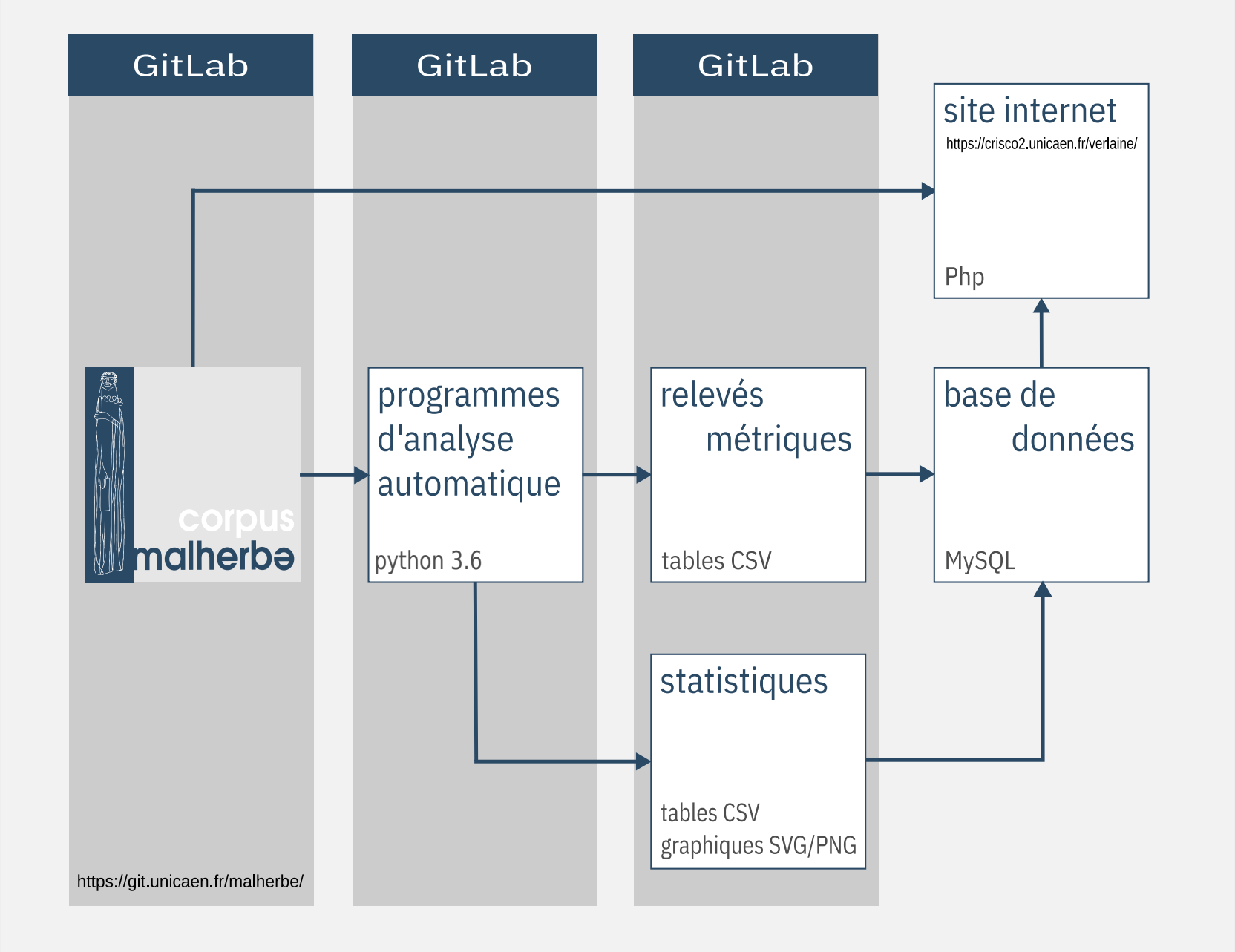
2019-
Richard Renault avec la collaboration de Stéphane Ferrari
Traitement automatique de textes versifiés

1- Constitution d'un corpus électronique de textes poétiques et théâtraux du début XVIIe au début du XXe siècle
(Richard Renault)
2- Outils d'analyse (Richard Renault)
   a- Traitement automatique
   (Richard Renault, avec la participation de Stéphane Ferrari)
   b- Analyse de la convergence
   (Richard Renault et Stéphane Ferrari)

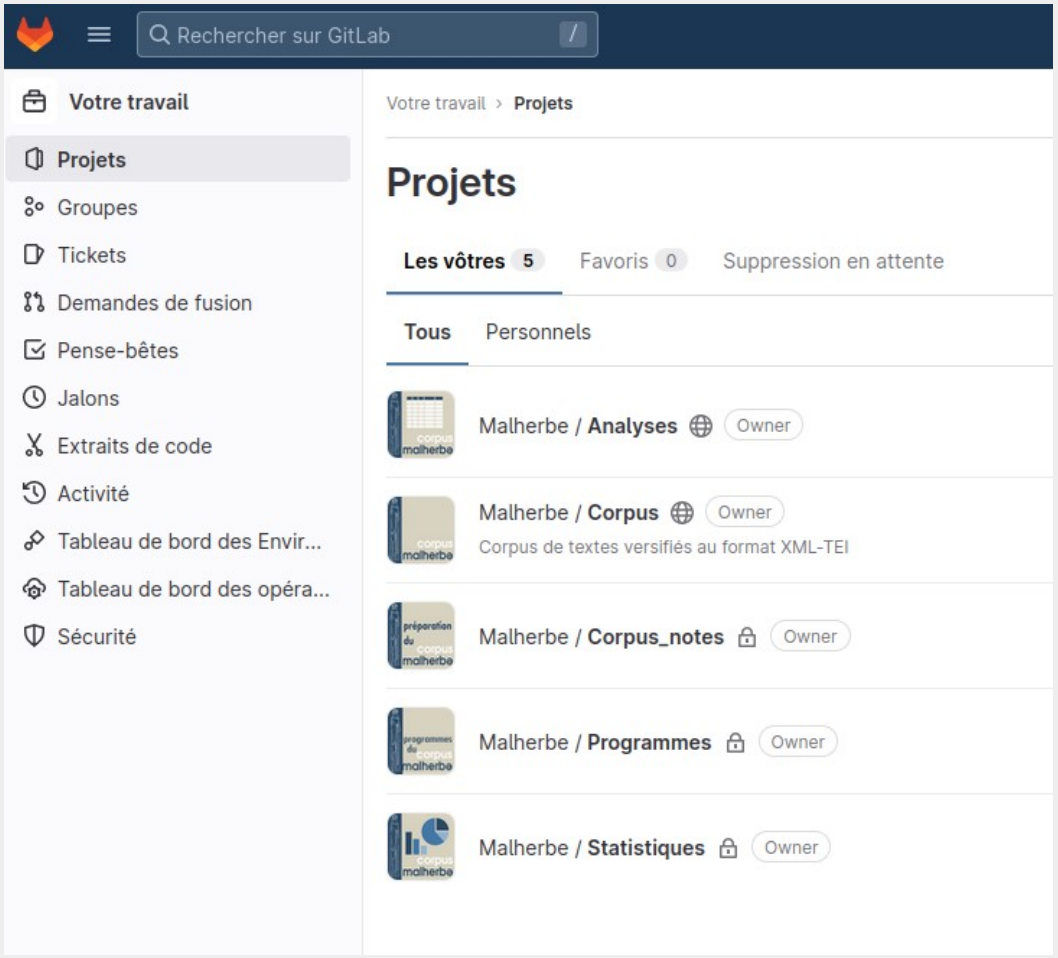
1. Bilan


GitLab
site internet
https://crisco2.unicaen.fr/verlaine/
Php
corpus
malherbe
programmes
d'analyse
automatique
python 3.6
relevés
métriques
tables CSV
base de
données
MySQL
statistiques
tables CSV
graphiques SVG/PNG

1. Bilan

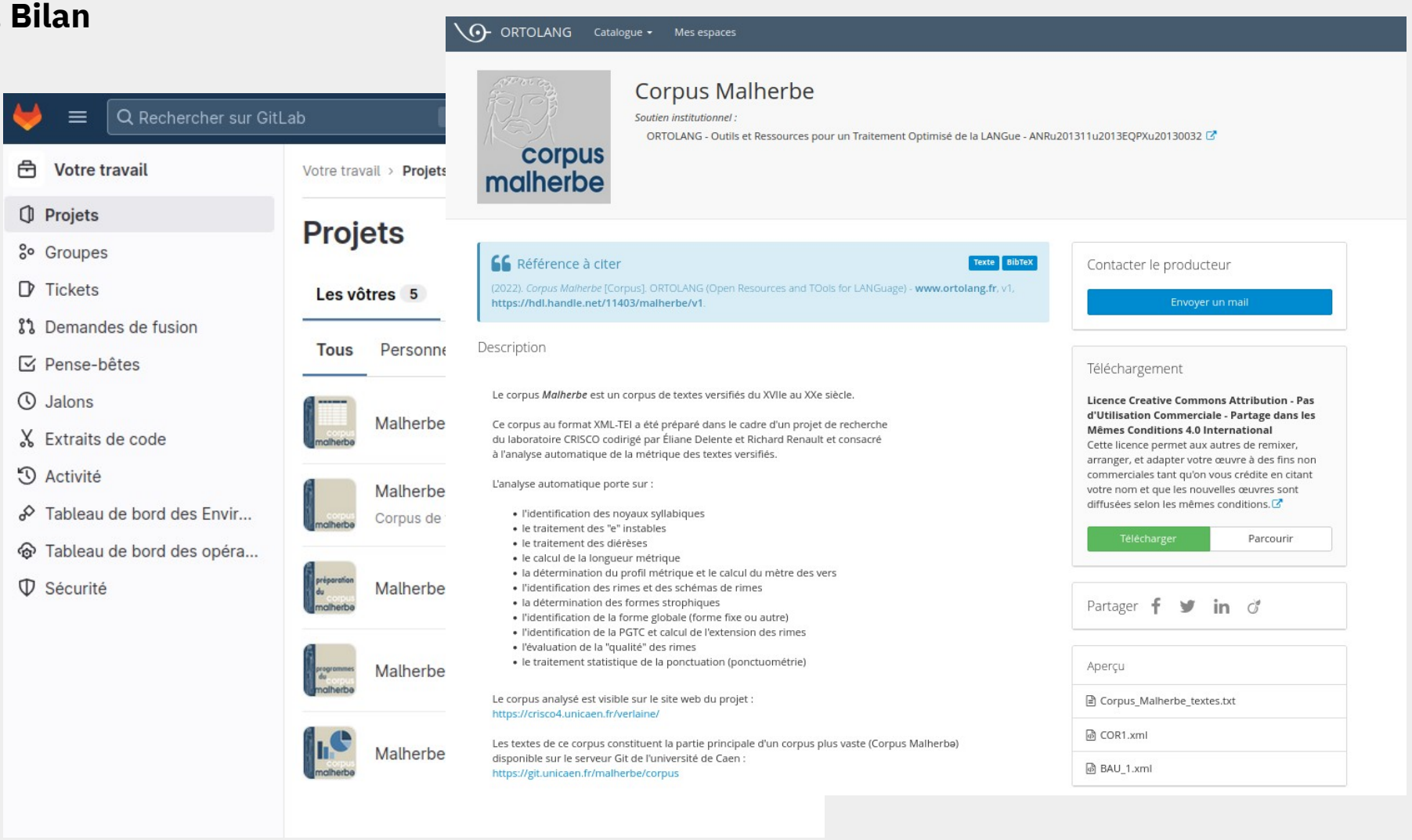

GitLab
GitLab
GitLab
site internet
https://crisco2.unicaen.fr/verlaine/
Php
corpus
malherbe
programmes
d'analyse
automatique
python 3.6
relevés
métriques
tables CSV
base de
données
MySQL
statistiques
tables CSV
graphiques SVG/PNG
https://git.unicaen.fr/malherbe/

1. Bilan



https://crisco4.unicaen.fr/verlaine/
https://crisco4.unicaen.fr/~stage/Verlaine_6/
https://git.unicaen.fr/malherbe/
https://www.ortolang.fr/market/corpora/malherbe

1. Bilan



https://crisco4.unicaen.fr/verlaine/
https://crisco4.unicaen.fr/~stage/Verlaine_6/
https://git.unicaen.fr/malherbe/
https://www.ortolang.fr/market/corpora/malherbe

1. Bilan

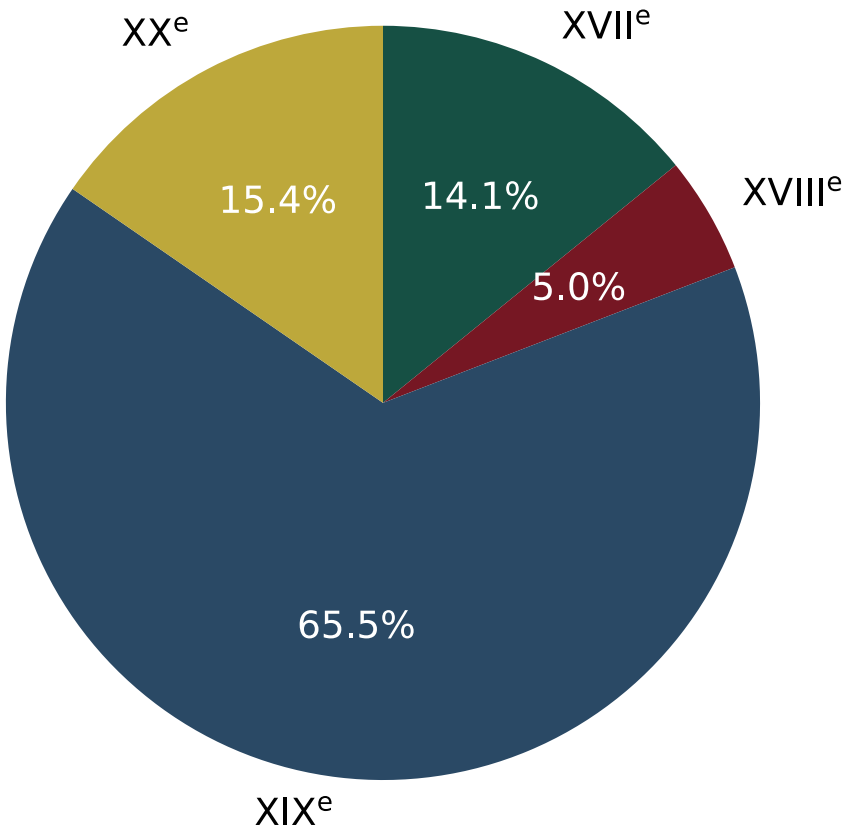
Corpus Malherbə

- 262 auteurs
- 554 recueils de poésies
- 23 047 poèmes
- 140 pièces de théâtre
- 1 169 485 vers

(mai 2023)

RÉPARTITION DU CORPUS SELON LES SIÈCLES
(date d'édition du recueil)

pour un total de 1 179 519 vers



- XVIIe : 166 868 vers
- XVIIIe : 58 961 vers
- XIXe : 772 113 vers
- XXe : 181 577 vers

version : 3.7
mai 2023

1. Bilan

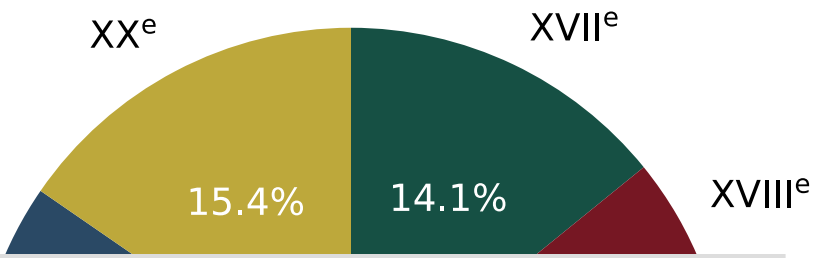
Corpus Malherbə

- 262 auteurs
- 554 recueils de poésies
- 23 047 poèmes
- 140 pièces de théâtre
- 1 169 485 vers

(mai 2023)

RÉPARTITION DU CORPUS SELON LES SIÈCLES
(date d'édition du recueil)

pour un total de 1 179 519 vers



sous-corpus :

- corpus *Malherbe* (CRISCO)
- corpus *Le Rire des vers*

  (Anne-Marie Bories, Université de Bâle)

  https://slw-comicverse.dslw.unibas.ch/prima.php?lang=fr
- corpus *Pamela_Puntel*

  (Pamela Puntel, Universita Degli Studi Di Udine et Université Lumière Lyon 2)

  https://ihrim.ens-lyon.fr/auteur/puntel-pamela

1. Bilan

Les composantes du traitement automatique :

- La nature des noyaux syllabiques
- Le mètre des vers
- Les rimes, les strophes et la forme globale du poème
- La concordance syntaxe/mètre

1. Bilan

Les composantes du traitement automatique :

- La nature des noyaux syllabiques
- Le mètre des vers
- Les rimes, les strophes et la forme globale du poème
- La concordance syntaxe/mètre

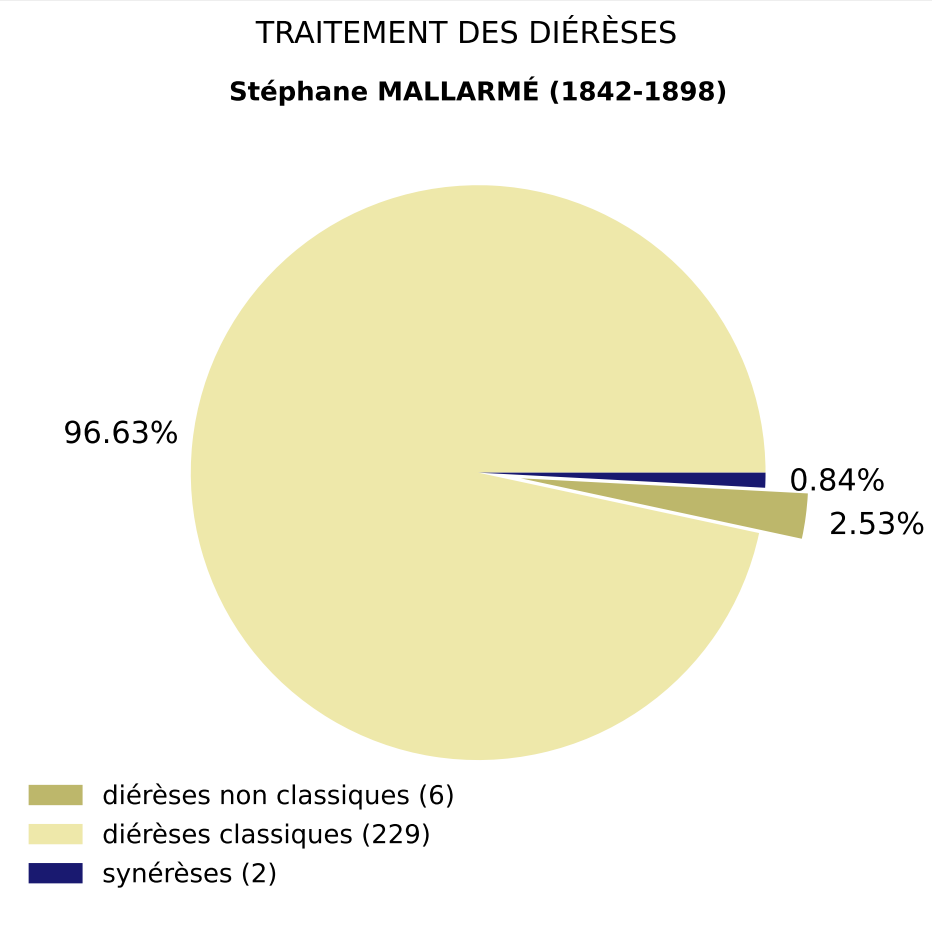
Les étapes du traitement :

1. Découpage en mots (stable)
2. Identification des noyaux syllabiques (stable)
3. Traitement des "e" instables (stable)
4. Traitement des diérèses (stable)
5. Calcul de la longueur métrique des vers (stable)
6. Détermination du profil métrique du poème et du mètre des vers (stable)
7. Appariement des vers en rimes, schémas rimiques des strophes et identification de la forme globale du poème (très avancé)
8. Identification de la PGTC et calcul de l'extension des rimes (avancé)
9. Évaluation de la qualité des rimes (avancé)
10. Évaluation de la concordance syntaxe/mètre (en développement)

1. Bilan

Post-traitement :

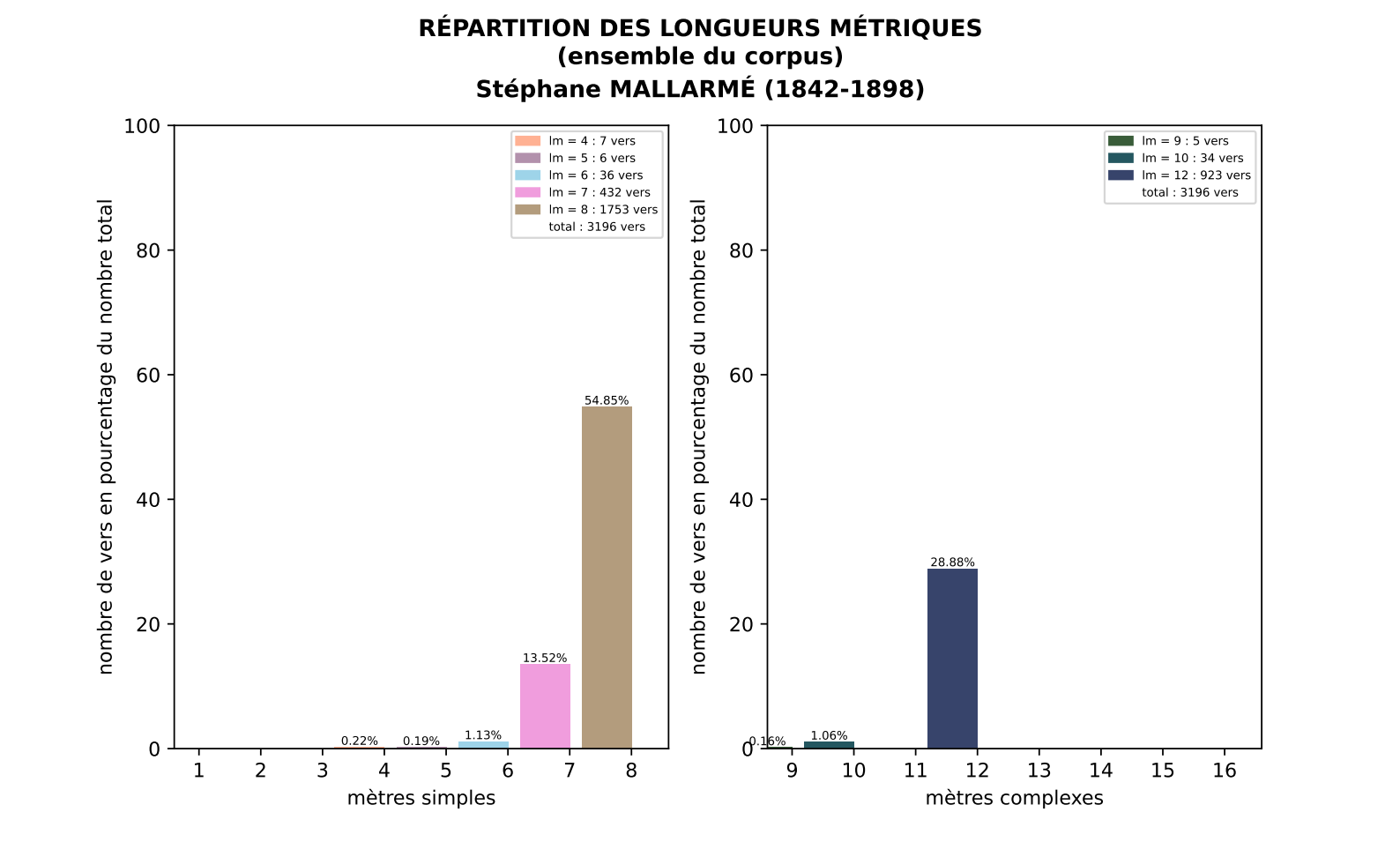
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

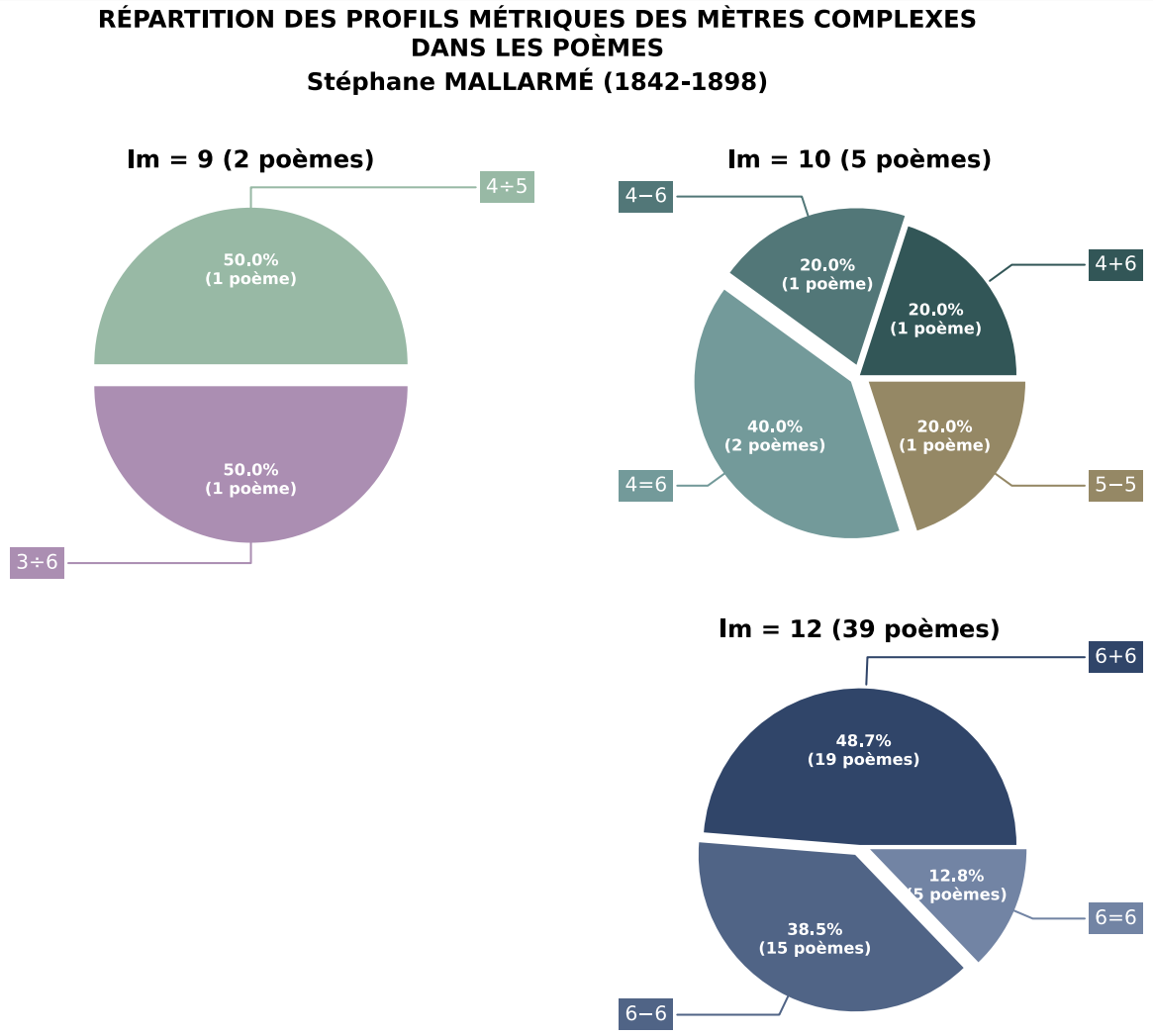
1. Traitement statistique des données analysées (avancé)



1. Le projet : bilan

Post-traitement :

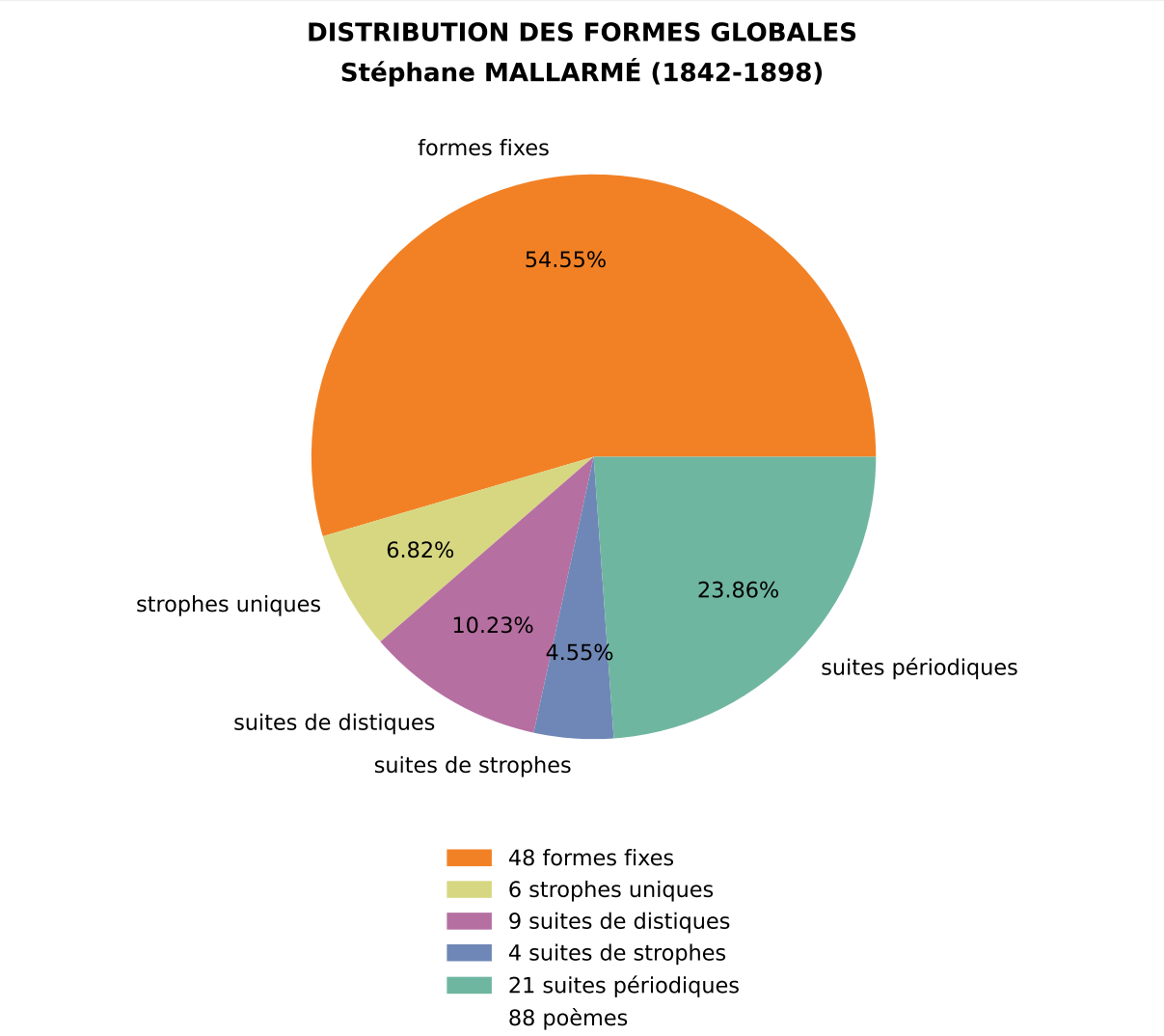
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

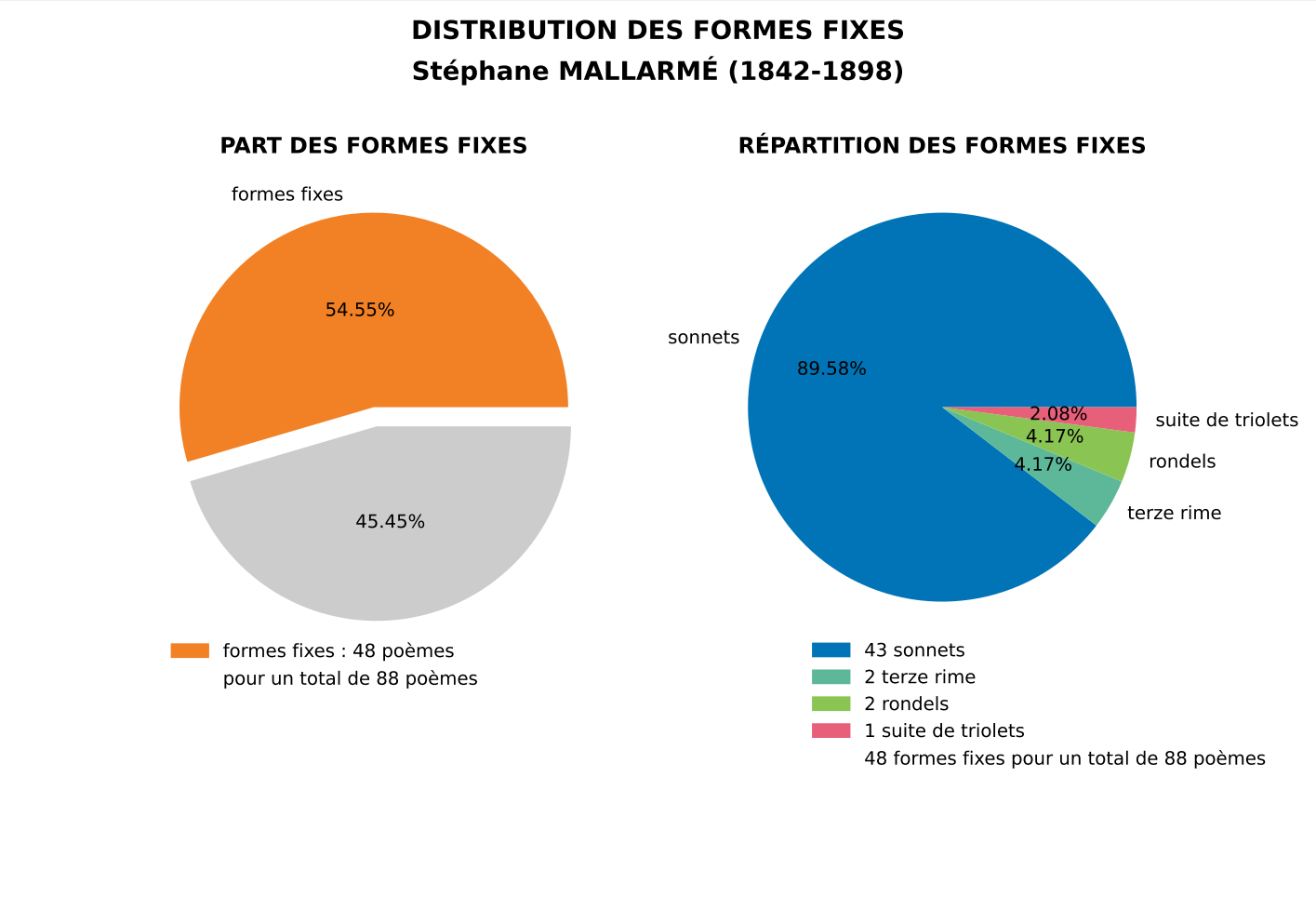
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

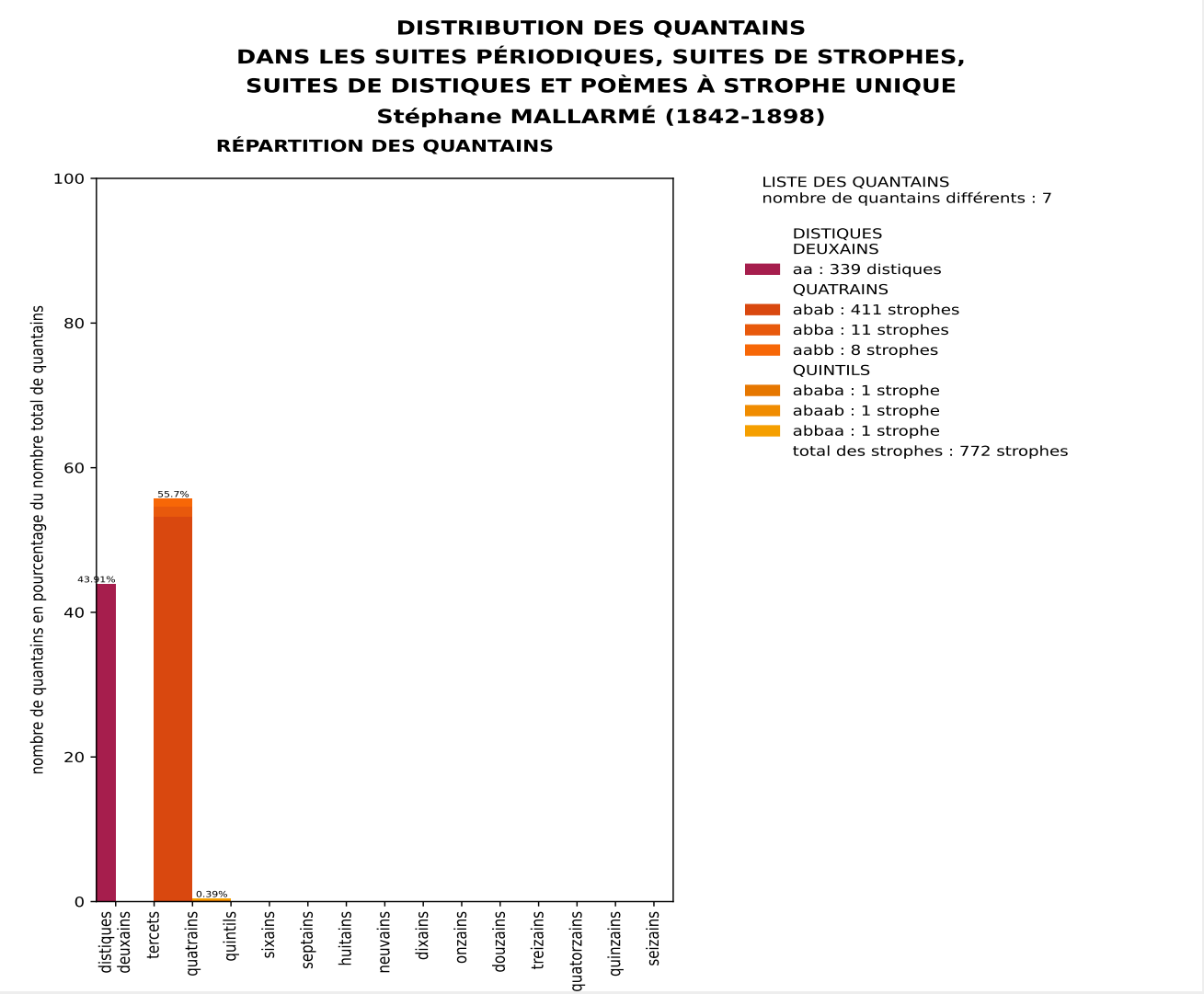
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

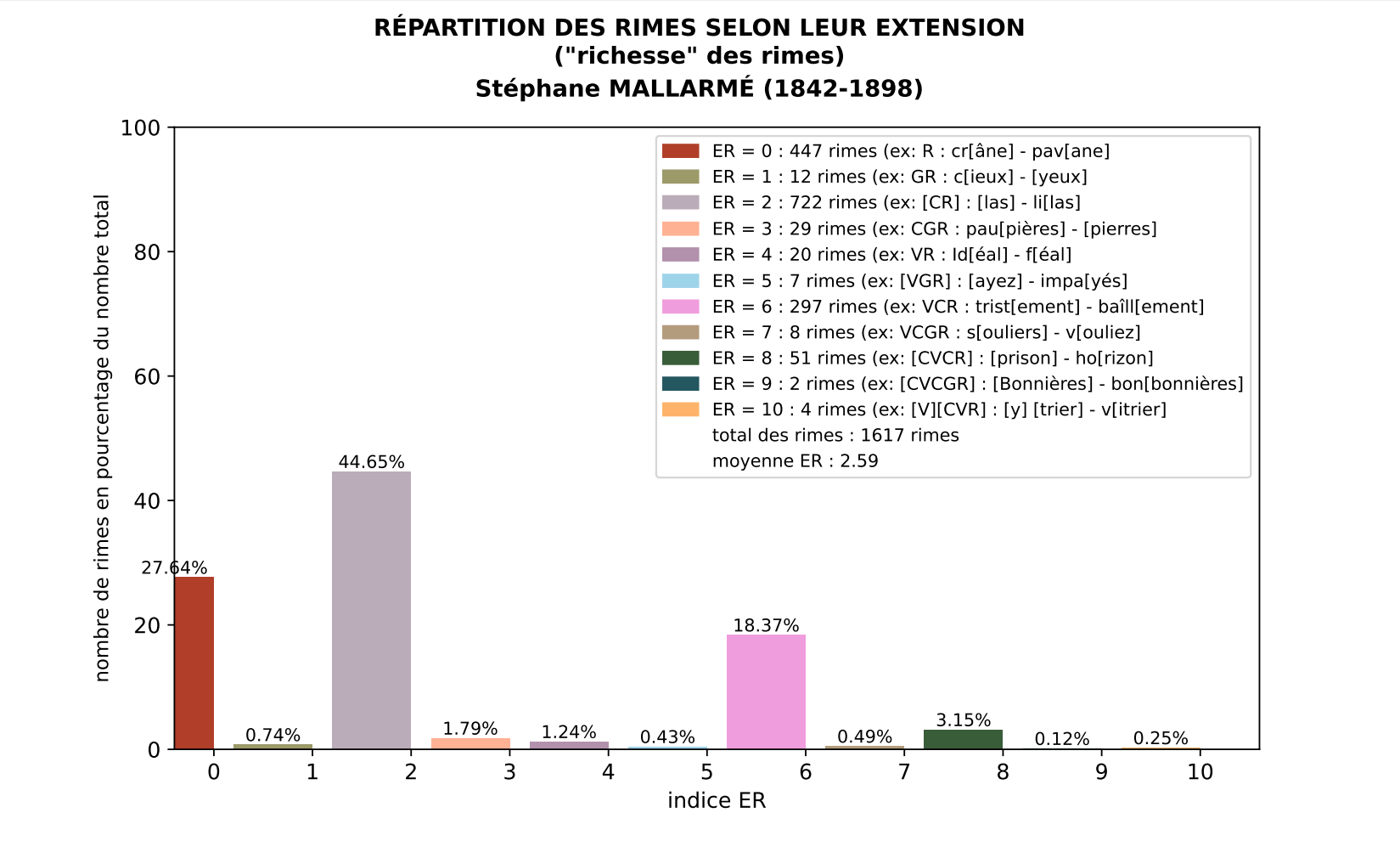
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

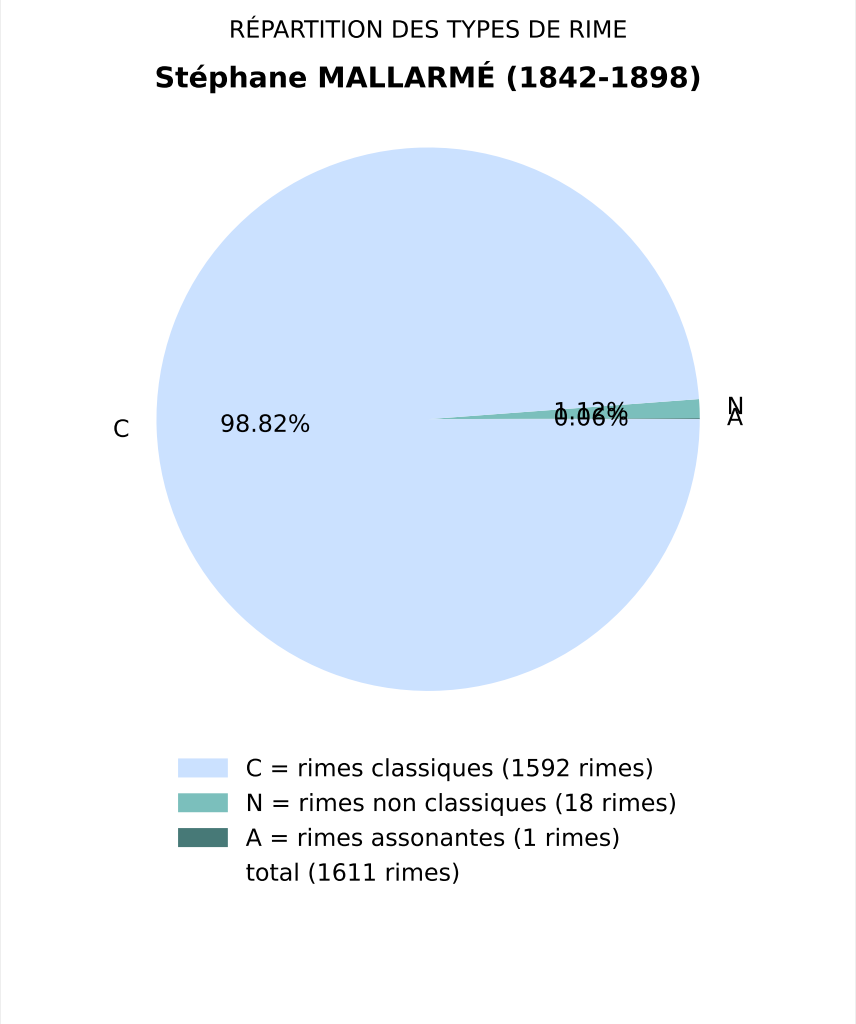
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

1. Traitement statistique des données analysées (avancé)



2. Poursuite : améliorations

Calcul de la richesse de la PGTC (richesse des rimes)

fraternel::éternel

Charles BAUDELAIRE, *Les Sept Vieillards*, LES FLEURS DU MAL, 1857-1861

2. Poursuite : améliorations

Calcul de la richesse de la PGTC (richesse des rimes)

fraternel::éternel

Charles BAUDELAIRE, *Les Sept Vieillards*, LES FLEURS DU MAL, 1857-1861

	PGTC		
	extension de la rime		nombre de rimes
	ER = 0	**moyenne**	
Voltaire	73,33 %	**0,65**	12 960
La Fontaine	65,45 %	**0,89**	21 564
Racine	60,45 %	**1,07**	8 712
Baudelaire	53,47 %	**1,50**	2 001
Valéry	42,13 %	**1,84**	2 466
Mallarmé	31,98 %	**2,45**	1 498

2. Poursuite : améliorations

Calcul de la richesse de la PGTC (richesse des rimes)

fraternel::éternel

Charles BAUDELAIRE, *Les Sept Vieillards*, LES FLEURS DU MAL, 1857-1861

lev**a**::ador**a** (ER=0)

carnav**al**::fin**al** (ER=0)

lev**a**::sau**va** (ER=1)

Vén**us**::an**us**

ven**u**s::rend**u**s

	PGTC		
	extension de la rime		nombre de rimes
	ER = 0	**moyenne**	
Voltaire	73,33 %	**0,65**	12 960
La Fontaine	65,45 %	**0,89**	21 564
Racine	60,45 %	**1,07**	8 712
Baudelaire	53,47 %	**1,50**	2 001
Valéry	42,13 %	**1,84**	2 466
Mallarmé	31,98 %	**2,45**	1 498

2. Poursuite : améliorations

Calcul de la richesse de la PGTC (richesse des rimes)

fraternel::éternel

Charles BAUDELAIRE, *Les Sept Vieillards*, LES FLEURS DU MAL, 1857-1861

leva::adora (ER=0)

carnaval::final (ER=0)

leva::sauva (ER=1)

Vénus::anus

venus::rendus

	PGTC		
	extension de la rime		nombre de rimes
	ER = 0	**moyenne**	
Voltaire	73,33 %	**0,65**	12 960
La Fontaine	65,45 %	**0,89**	21 564
Racine	60,45 %	**1,07**	8 712
Baudelaire	53,47 %	**1,50**	2 001
Valéry	42,13 %	**1,84**	2 466
Mallarmé	31,98 %	**2,45**	1 498

2. Poursuite : améliorations

Reconnaissance de la forme globale : suite périodique

Germain NOUVEAU

Poésies d'Humilis
et Vers Inédits
1872-1881

POÉSIES D'HUMILIS

L'Homme

Homme dont la tristesse | est écrite d'un bout	6+6	a
Du monde à l'autre, et même | aux murs de la campagne,	6+6	b
Forçat de l'hôpital | et malade du bagne ;	6+6	b
Dormeur maussade, à qui | chaque aube dit : « Debout ! »	6+6	a
Voyageur douloureux | qu'attend la Mort, auberge	6+6	a
Où l'on vend le lit dur | et les pleurs blancs du cierge,	6+6	a
Tu gémis, étonné | de te sentir si las ;	6+6	a
Puis un jour tu te dis : | « L'âme est un vain bagage,	6+6	b
Et mon'cœur est bien lourd | pour un pareil voyage !»	6+6	b
Et, sans songer que Dieu | te donne ses lilas,	6+6	a
Tu veux jeter ton cœur, | tu veux jeter ton âme,	6+6	a
Pour alléger ta marche | et mieux porter la Femme ;	6+6	a
Par ta route et ses ponts | fiers de leur parapet,	6+6	a
Compagnon de l'orgueil, | fils des froides études,	6+6	b
Tu vas vers le malheur | et vers les solitudes.	6+6	b
Tout plein des arguments | dont l'esprit se repaît,	6+6	a
Tu fais, pour savourer | ta gloire monotone,	6+6	a
Taire ta conscience | à l'heure où le ciel tonne.	6+6	a

...

mètre	profil métrique : **6+6**
forme globale	type : **suite périodique**
	schéma : **8(abbacc)**

2. Poursuite : améliorations

Reconnaissance de la forme globale : suite périodique

Germain NOUVEAU

Poésies d'Humilis
et Vers Inédits
1872-1881

POÉSIES D'HUMILIS

L'Homme

	Homme dont la tristesse | est écrite d'un bout	6+6	a
	Du monde à l'autre, et même | aux murs de la campagne,	6+6	b
	Forçat de l'hôpital | et malade du bagne ;	6+6	b
	Dormeur maussade, à qui | chaque aube dit : « Debout ! »	6+6	a
	Voyageur douloureux | qu'attend la Mort, auberge	6+6	a
	Où l'on vend le lit dur | et les pleurs blancs du cierge,	6+6	a
	Tu gémis, étonné | de te sentir si las ;	6+6	a
	Puis un jour tu te dis : | « L'âme est un vain bagage,	6+6	b
	Et mon'cœur est bien lourd | pour un pareil voyage !»	6+6	b
10	Et, sans songer que Dieu | te donne ses lilas,	6+6	a
	Tu veux jeter ton cœur, | tu veux jeter ton âme,	6+6	a
	Pour alléger ta marche | et mieux porter la Femme ;	6+6	a
	Par ta route et ses ponts | fiers de leur parapet,	6+6	a
	Compagnon de l'orgueil, | fils des froides études,	6+6	b
15	Tu vas vers le malheur | et vers les solitudes.	6+6	b
	Tout plein des arguments | dont l'esprit se repaît,	6+6	a
	Tu fais, pour savourer | ta gloire monotone,	6+6	a
	Taire ta conscience | à l'heure où le ciel tonne.	6+6	a

...

mètre	profil métrique : **6+6**
forme globale	type : **suite périodique**
	schéma : **8(abbacc)**

Germain NOUVEAU

Poésies d'Humilis
et Vers Inédits
1872-1881

POÉSIES D'HUMILIS

L'Homme

	Homme dont la tristesse | est écrite d'un bout	6+6	a
	Du monde à l'autre, et même | aux murs de la campagne,	6+6	b
	Forçat de l'hôpital | et malade du bagne ;	6+6	b
	Dormeur maussade, à qui | chaque aube dit : « Debout ! »	6+6	a
	Voyageur douloureux | qu'attend la Mort, auberge	6+6	a
	Où l'on vend le lit dur | et les pleurs blancs du cierge,	6+6	a
	Tu gémis, étonné | de te sentir si las ;	6+6	a
	Puis un jour tu te dis : | « L'âme est un vain bagage,	6+6	b
	Et mon'cœur est bien lourd | pour un pareil voyage !»	6+6	b
10	Et, sans songer que Dieu | te donne ses lilas,	6+6	a
	Tu veux jeter ton cœur, | tu veux jeter ton âme,	6+6	a
	Pour alléger ta marche | et mieux porter la Femme ;	6+6	a
	Par ta route et ses ponts | fiers de leur parapet,	6+6	a
	Compagnon de l'orgueil, | fils des froides études,	6+6	b
15	Tu vas vers le malheur | et vers les solitudes.	6+6	b
	Tout plein des arguments | dont l'esprit se repaît,	6+6	a
	Tu fais, pour savourer | ta gloire monotone,	6+6	a
	Taire ta conscience | à l'heure où le ciel tonne.	6+6	a

...

mètre	profil métrique : **6+6**
forme globale	type : **suite de strophes**
	schéma : **8[abba] 8[aa]**

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Principe de meilleure concordance (**Principe de concordance optimale**)

Toutes choses étant égales par ailleurs, on tend spontanément à comprendre un texte en sorte que la structure métrique et le sens, ou du moins la structure grammaticale convergent le mieux possible.

B. de Cornulier, *L'Art Poëtique*, 1995

Implémentation de l'approche de Dell et Benini (2020)

Français Dell et Romain Benini,
La concordance chez Racine,
Rapport entre structure grammaticale et forme métrique dans le théâtre de Racine
Classique Garnier. 2020

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

3 cas de figures à prendre en compte :

- Concordance
- Concordance différée
  - Mazaleyrat, 1990,
  - Roubaud, 1978 (concordance enjambante)
  - Quicherat, 1882
  - Tobler, 1885
  - Mourgues, 1724 (enjambement non vicieux)
  - ...
- Discordance

- **concordance de fin d'hémistiche**

[Elle enivre les cœurs,] [plus forte que le vin.]
Germain Nouveau, Poésies d'Humilis, 1924, *Charité*

- **concordance de fin de vers**

[L'aile de l'hirondelle] [annonce le nuage ;]
[Et le chemin nous aime :] [avec nous il voyage ;]
Germain Nouveau, Poésies d'Humilis, 1924, *Immensité*

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Distinction entre discordance et concordance différée

- **enjambements en fin de vers**

[Bienheureux, j'allongeai] [les jambes sous la table] **discordance**
[Verte : je contemplai] [les sujets très naïfs] **concordance différée**
[De la tapisserie.] [— Et ce fut adorable,]
[Quand la fille aux tétons] [énormes, aux yeux vifs,]

Arthur Rimbaud, *Au Cabaret-Vert, cinq heures du soir*, POÉSIES I, 1869-1970

- **enjambements en fin d'hémistiche**

[Bienheureux, j'allongeai] [les jambes sous la table]
[Verte : je contemplai] [les sujets très naïfs] **concordance différée**
[De la tapisserie.] [— Et ce fut adorable,]
[Quand la fille aux tétons] [énormes, aux yeux vifs,] **discordance**

Arthur Rimbaud, *Au Cabaret-Vert, cinq heures du soir*, POÉSIES I, 1869-1970

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Typologie des enjambements :

- syllabe : typographie
- morphème
  - affixe : typographie
  - base : typographie
- clitique
  - clitique : typographie, listes
  - base : typographie
- composant (mot composé) : typographie, listes
- mot
  - préposition : listes
  - conjonction : listes
  - verbe : catégorie
  - ...
- constituant syntaxique : catégorie, relations syntaxiques

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Typologie des enjambements :

- syllabe : typographie
- morphème
  - affixe : typographie
  - base : typographie
- clitique
  - clitique : typographie, listes
  - base : typographie
- composant (mot composé) : typographie, listes
- mot
  - préposition : listes
  - conjonction : listes
  - verbe : catégorie
  - ...
- constituant syntaxique : catégorie, relations syntaxiques

J'en vis un affamé plus que l'est un moineau,
— Voilà qui tient du diantre ! —
Qui me parut manger tout simplement une o-
-Melette avec son ventre ;
Raoul Ponchon, *Vieux messieurs,* LA MUSE FRONDEUSE, 1971

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Typologie des enjambements :

■ syllabe		: typographie
■ morphème	- affixe	: typographie
	- base	: typographie
■ clitique	- clitique	: typographie, listes
	- base	: typographie
■ composant (mot composé)		: typographie, listes
■ mot	- préposition	: listes
	- conjonction	: listes
	- verbe	: catégorie
	...	
■ constituant syntaxique		: catégorie, relations syntaxiques

Nous, que notre jeu soit super-
Fin, tout autant que le Dimanche
Stéphane Mallarmé, *Triolets*, VERS DE CIRCONSTANCE, 1920

Paraît-il, de ma mine affreuse-
Ment peuple et sans nul galbe exquis
Paul Verlaine, Ecce iterum crispinus, DÉDICACES, 1890

Prête-la moi, je te le rend-
Rai gaillard et digne d'envie.
Paul Verlaine, *Triolets à une vertu pour s'excuser du peu*, FEMMES, 1890

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Typologie des enjambements :

■ syllabe		: typographie
■ morphème	- affixe	: typographie
	- base	: typographie
■ clitique	- clitique	: typographie, listes
	- base	: typographie
■ composant (mot composé)		: typographie, listes
■ mot	- préposition	: listes
	- conjonction	: listes
	- verbe	: catégorie
	...	
■ constituant syntaxique		: catégorie, relations syntaxiques

Pour aimer et chercher le qu'en-
Dira-t-on, et : zut pour ce zeste !
Paul Verlaine, XVI, DANS LES LIMBES, 189

Pour moi le fol Amour, et viens, au clair de la
Lune. Allons vers Cythère ou bien vers Pampelune,
Théodore de Banville, *Variations*, DANS LA FOURNAISE, 1892

Monsieur le... Tiens, au fait, qu'avez-
Vous été sur terre ? — Poète.
Edmond Rostand, *Au ciel,* LES MUSARDISES, 1911

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Typologie des enjambements :

■ syllabe		: typographie
■ morphème	- affixe	: typographie
	- base	: typographie
■ clitique	- clitique	: typographie, listes
	- base	: typographie
■ composant (mot composé)		: typographie, listes
■ mot	- préposition	: listes
	- conjonction	: listes
	- verbe	: catégorie
	...	
■ constituant syntaxique		: catégorie, relations syntaxiques

De *Malacca, Manon Lescaut* avec deux eaux-Fortes (l'une piquée et l'autre disparue)...
Tristan Derème, *CXLIX*, LA VERDURE DORÉE, 1908

Composé de quatre vieillards, d'une demi-Douzaine d'ordinands et du portier, l'usage
Paul Verlaine, *Compliment a un autre magistrat*, INVECTIVES, 1896

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Typologie des enjambements :

■ syllabe		: typographie
■ morphème	- affixe	: typographie
	- base	: typographie
■ clitique	- clitique	: typographie, listes
	- base	: typographie
■ composant (mot composé)		: typographie, listes
■ mot	- préposition	: listes
	- conjonction	: listes
	- verbe	: catégorie
	...	
■ constituant syntaxique		: catégorie, relations syntaxiques

Je me croyais en Décembre,
Ça doit tenir au
Thermomètre de ma chambre
Qui marque zéro.
Raoul Ponchon, *To be or not to be*, LA MUSE VAGABONDE, 1947

Et j'étais parti pour ne plus la revoir ; mais
Comme j'avais gardé la clef, je revenais,
Georges de Porto-Riche, *Elle dormait...*, PRIMA VERBA, 1872

Dans le tourment de sa pensée il regardait
L'épanouissement de ce rêve nocturne ;
Raymond de la TAILHÈDE, *Ombres*, POÉSIES, 1938

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Typologie des enjambements :

- syllabe : typographie
- morphème
  - affixe : typographie
  - base : typographie
- clitique
  - clitique : typographie, listes
  - base : typographie
- composant (mot composé) : typographie, listes
- mot
  - préposition : listes
  - conjonction : listes
  - verbe : catégorie
  - ...
- constituant syntaxique : catégorie, relations syntaxiques

Nos mains se touchaient, le fluide
S'échappait du bout de nos doigts
Raoul Ponchon, *Spiritisme*, LA MUSE FRONDEUSE, 1971

2. Poursuite : développement en cours

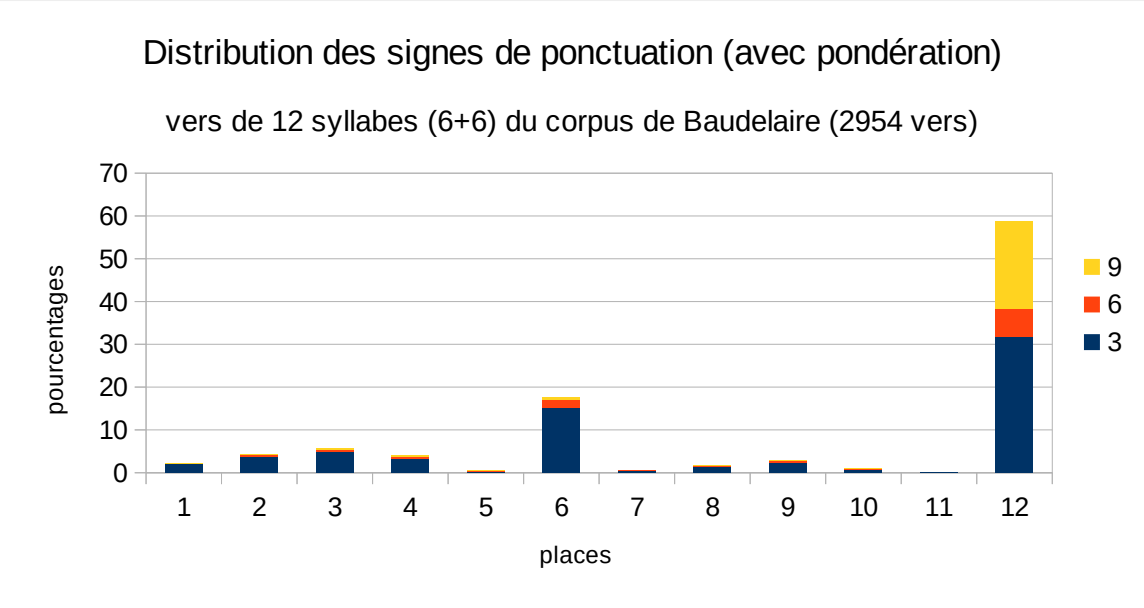
Identification des enjambements et évaluation de la concordance

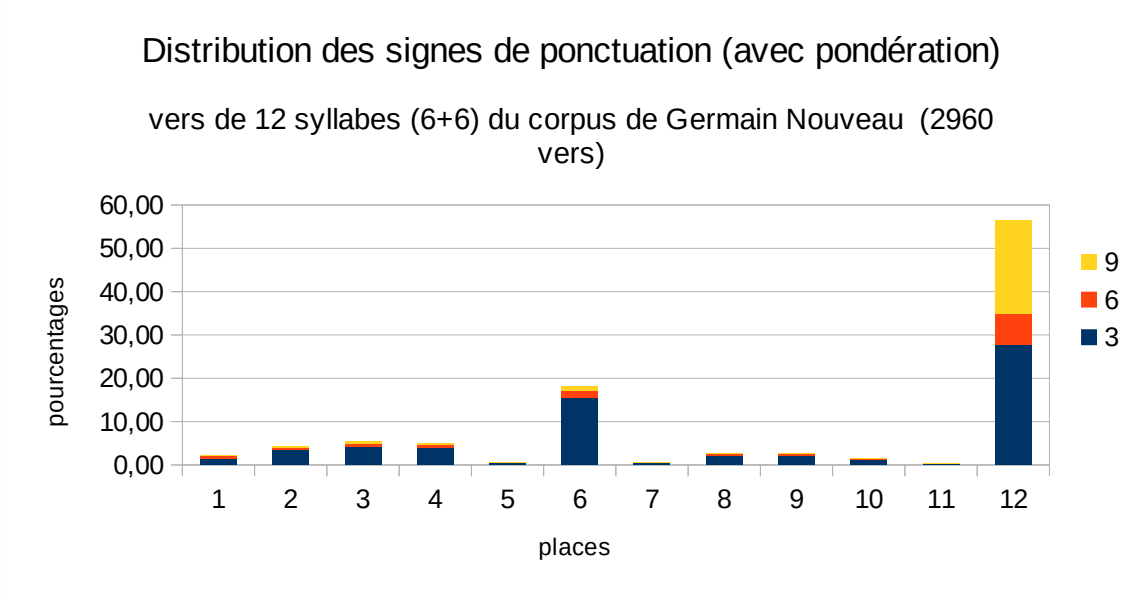
Typologie des enjambements :

■ syllabe		: typographie	
■ morphème	- affixe	: typographie	
	- base	: typographie	
■ clitique	- clitique	: typographie, listes	morphologie
	- base	: typographie	
■ composant (mot composé)		: typographie, listes	
■ mot	- préposition	: listes	
	- conjonction	: listes	
	- verbe	: catégorie	syntaxe
	...		
■ constituant syntaxique		: catégorie, relations syntaxiques	

2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance





2. Poursuite : développement en cours

Identification des enjambements et évaluation de la concordance

Paul VALÉRY

ALBUM DE VERS ANCIENS

1920

César

	César, calme César, | le pied sur toute ch|ose,	6+6	a
	Les poings durs dans la barbe, | et l'œil sombre peu pl|é	6+6	b
	D'aigles et des combats | du couchant contem pl|é,	6+6	b
	Ton cœur s'enfle, et se sent | toute-puissante C|ause.	6+6	a
5	Le lac en vain palpite | et lèche son lit r|ose ;	6+6	a
	En vain d'or précieux | brille le jeune bl|é ;	6+6	b
	Tu durcis dans les nœuds | de ton corps rassem bl|é	6+6	b
	L'ordre, qui doit enfin | fendre ta bouche cl|ose.	6+6	a
	L'ample monde, au delà | de l'immense hor iz on,	6+6	c
10	L'Empire attend l'éclair, | le décret, le t is on	6+6	c
	Qui changeront le soir | en furieuse aur|ore.	6+6	d
	Heureux là-bas sur l'onde, | et bercé du ha s ard,	6+6	e
	Un pêcheur indolent | qui flotte et chante, ign|ore	6+6	d
	Quelle foudre s'amasse | au centre de Cé s ar.	6+6	e

2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

Tagging & chunking

Tagging : analyse des catégories grammaticales

Jeu de catégories proposé pour le français

ABR	abbreviation
ADJ	adjective
ADV	adverb
DET:ART	article
DET:POS	possessive pronoun (ma, ta, ...)
INT	interjection
KON	conjunction
NAM	proper name
NOM	noun
NUM	numeral
PRO	pronoun
PRO:DEM	demonstrative pronoun
PRO:IND	indefinite pronoun
PRO:PER	personal pronoun
PRO:POS	possessive pronoun (mien, tien, ...)
PRO:REL	relative pronoun
PRP	preposition
PRP:det	preposition plus article (au,du,aux,des)
PUN	punctuation
PUN:cit	punctuation citation
SENT	sentence tag
SYM	symbol
VER:cond	verb conditional
VER:futu	verb futur
VER:impe	verb imperative
VER:impf	verb imperfect
VER:infi	verb infinitive
VER:pper	verb past participle
VER:ppre	verb present participle
VER:pres	verb present
VER:simp	verb simple past
VER:subi	verb subjunctive imperfect
VER:subp	verb subjunctive present

2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

Tagging & chunking

Chunking : regroupement en syntagmes « minimaux »

Entraînement sur le French Treebank
http://www.llf.cnrs.fr/en/Gens/Abeille/French-Treebank-fr.php
utilisant les catégories

AP	syntagme adjectival
AdP	syntagme adverbial
COORD	syntagme (ou phrase) coordonné(e)
NP	syntagme nominal
VN	noyau verbal (verbes, clitiques, auxiliaires, faire)
PP	syntagme prépositionnel
PREF	préfixe (le tiret en fait partie)
SENT	phrase indépendante (tout fragment indépendant)
Vppart	proposition participiale (part passé ou part présent)
Vpinf	proposition infinitive (pouvant commencer par une préposition)
Srel	proposition relative (commençant par un pronom relatif ou un PP incluant un Pro rel)
Ssub	proposition subordonnée (complétive, interrogative indirecte, subordonnée circonstancielle)
Sint	proposition conjuguée interne (coordonnée, discours direct, incise)

Le Treetagger semble n'utiliser qu'une partie seulement de ces catégories d'origine

2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

Tagging & chunking

3 étapes prévues pour analyser l'ensemble du corpus

1) Préparation
   Formater les textes du corpus en respectant le découpage en mots déjà effectué
   → on ne fait pas appel au segmenteur intégré à l'outil Treetagger
      risque de perte de performances mais alignement ultérieur facilité

2) Analyse(s)
   tagging puis chunking
   → chunking « instable » actuellement
      possibilité d'intervention sur les résultats avant réalignement avec le corpus d'origine

3) Enrichir les textes du corpus d'origine des résultats de l'analyse
   en cours de développement

2. Poursuite : développement en cours

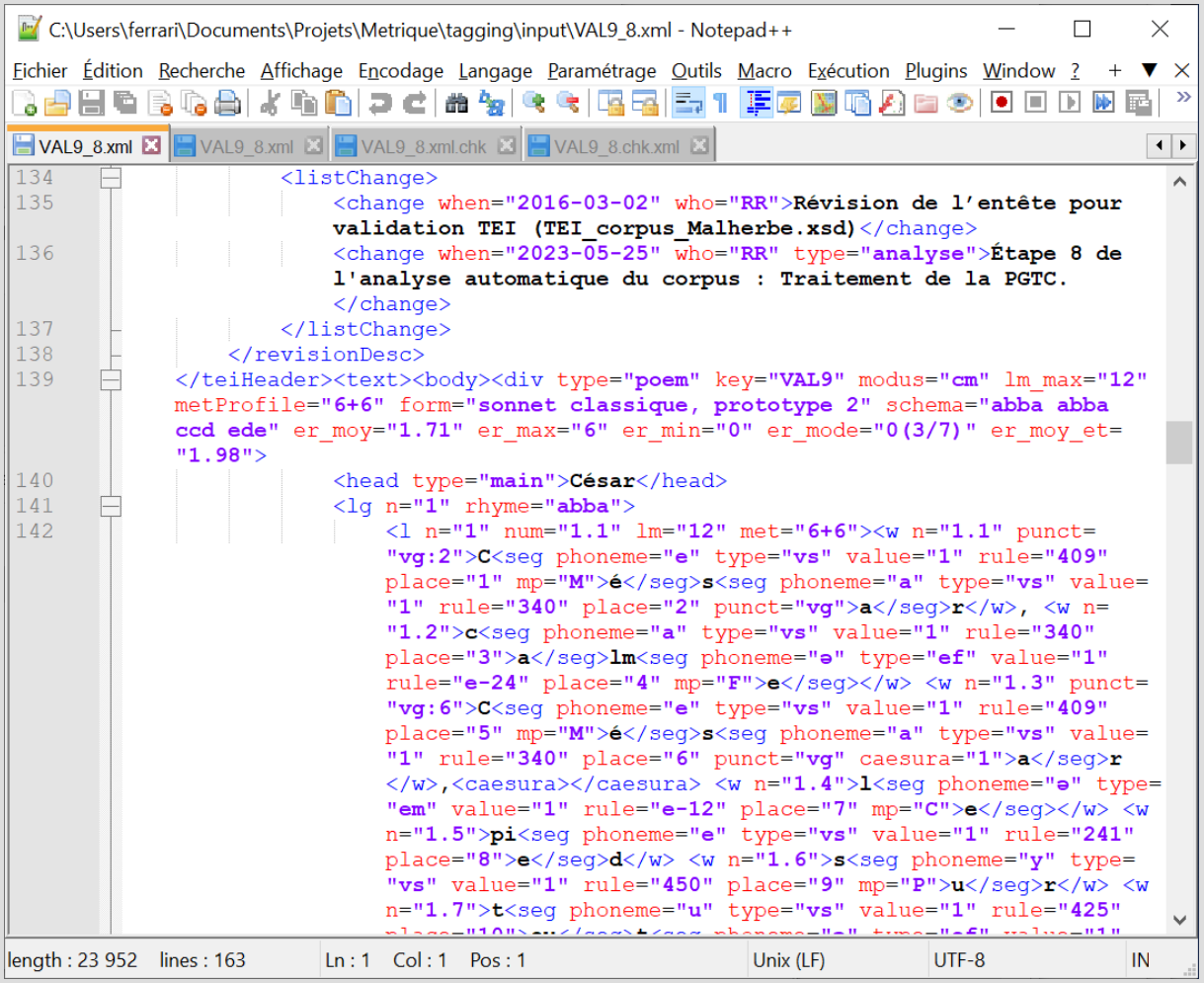
Premiers tests d'analyse avec le treetagger

Tagging & chunking

3 étapes

1) Préparation

Exemple d'entrée



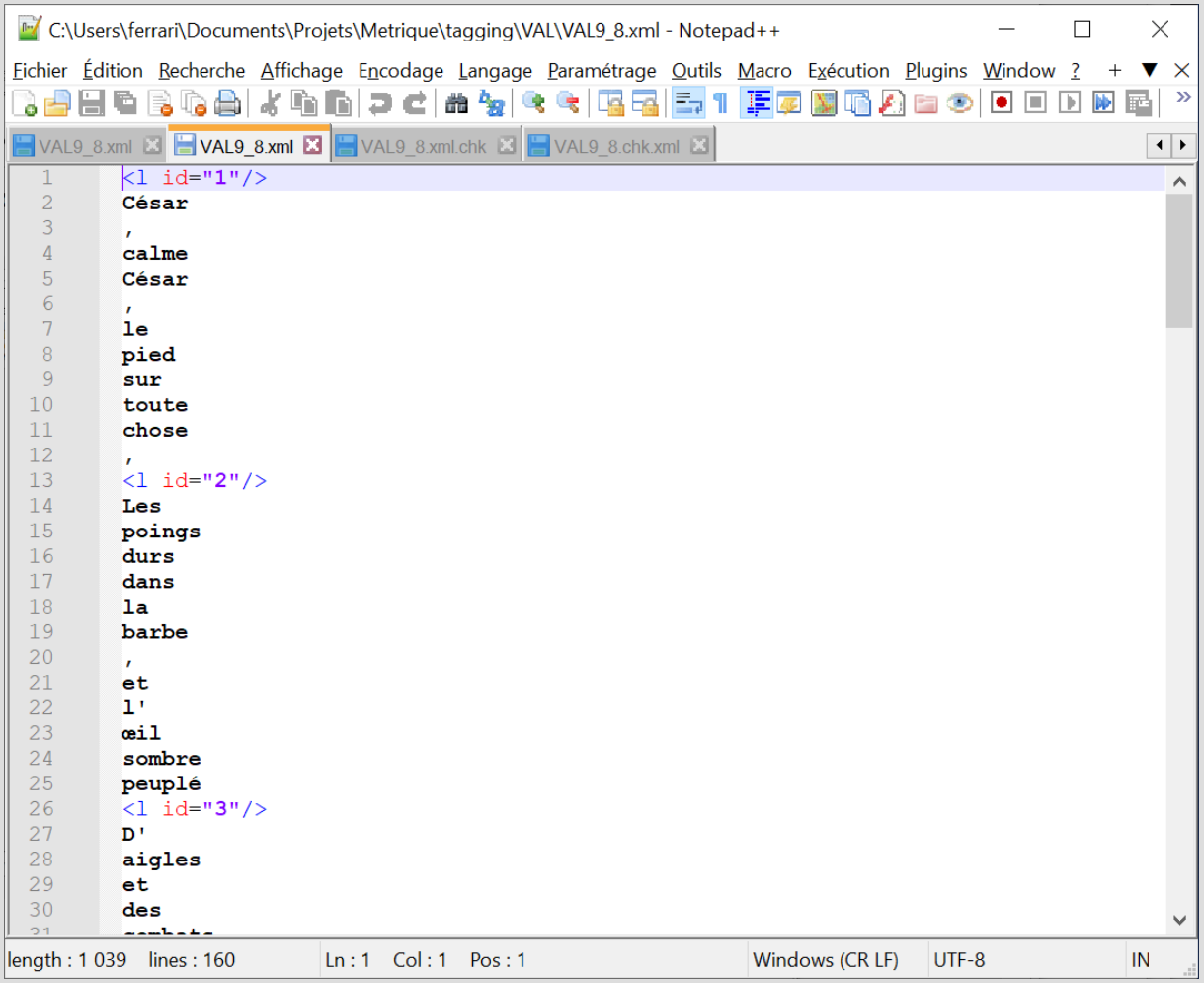
2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

Tagging & chunking

3 étapes

1) Préparation
   Entrée préparée
   (Python + XSLT)



2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

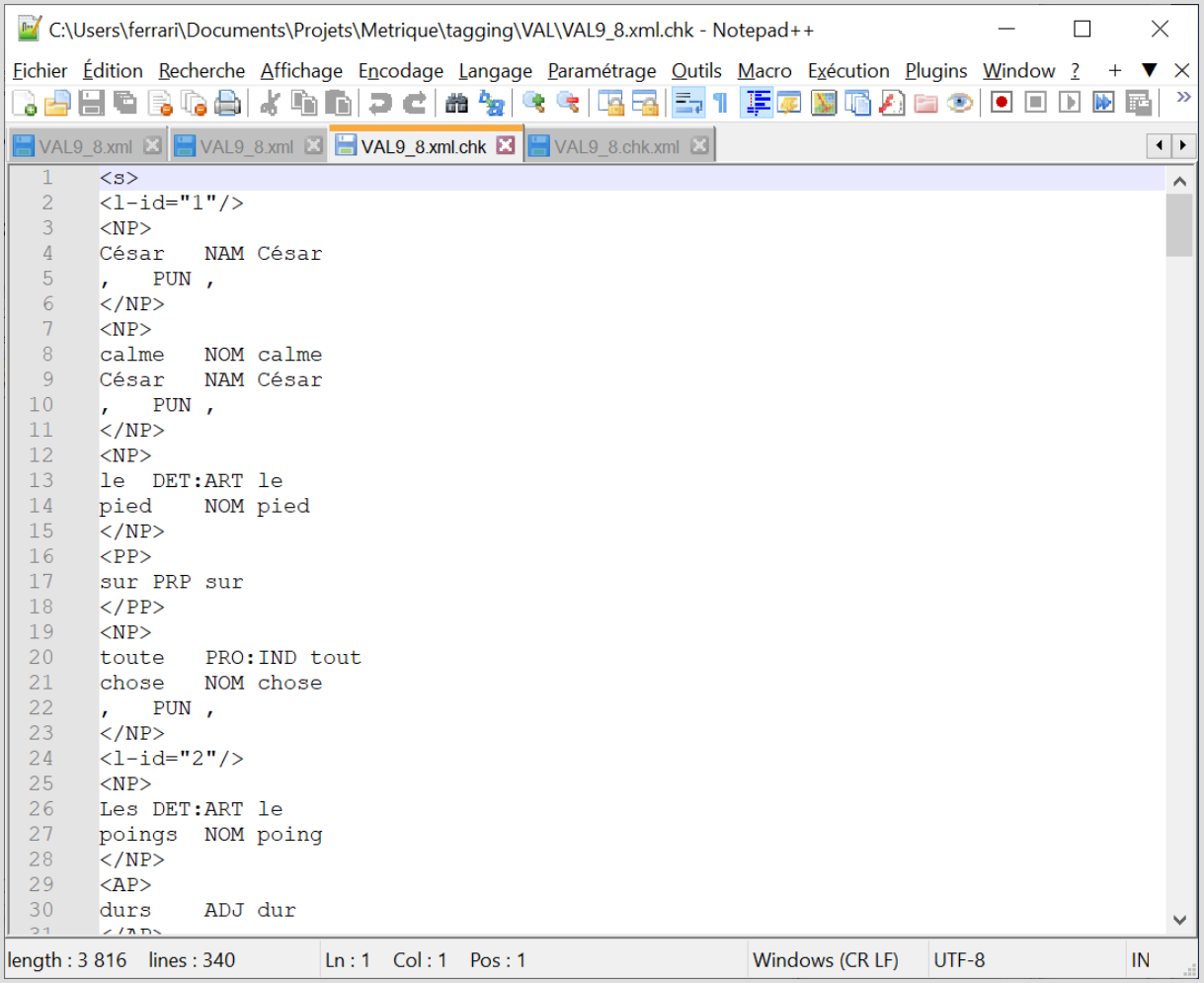
Tagging & chunking

3 étapes

1) Analyse(s)

Exemple de sortie directement issue du Treetagger

mélange de balises et de tabulations (ce n'est pas du XML)



2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

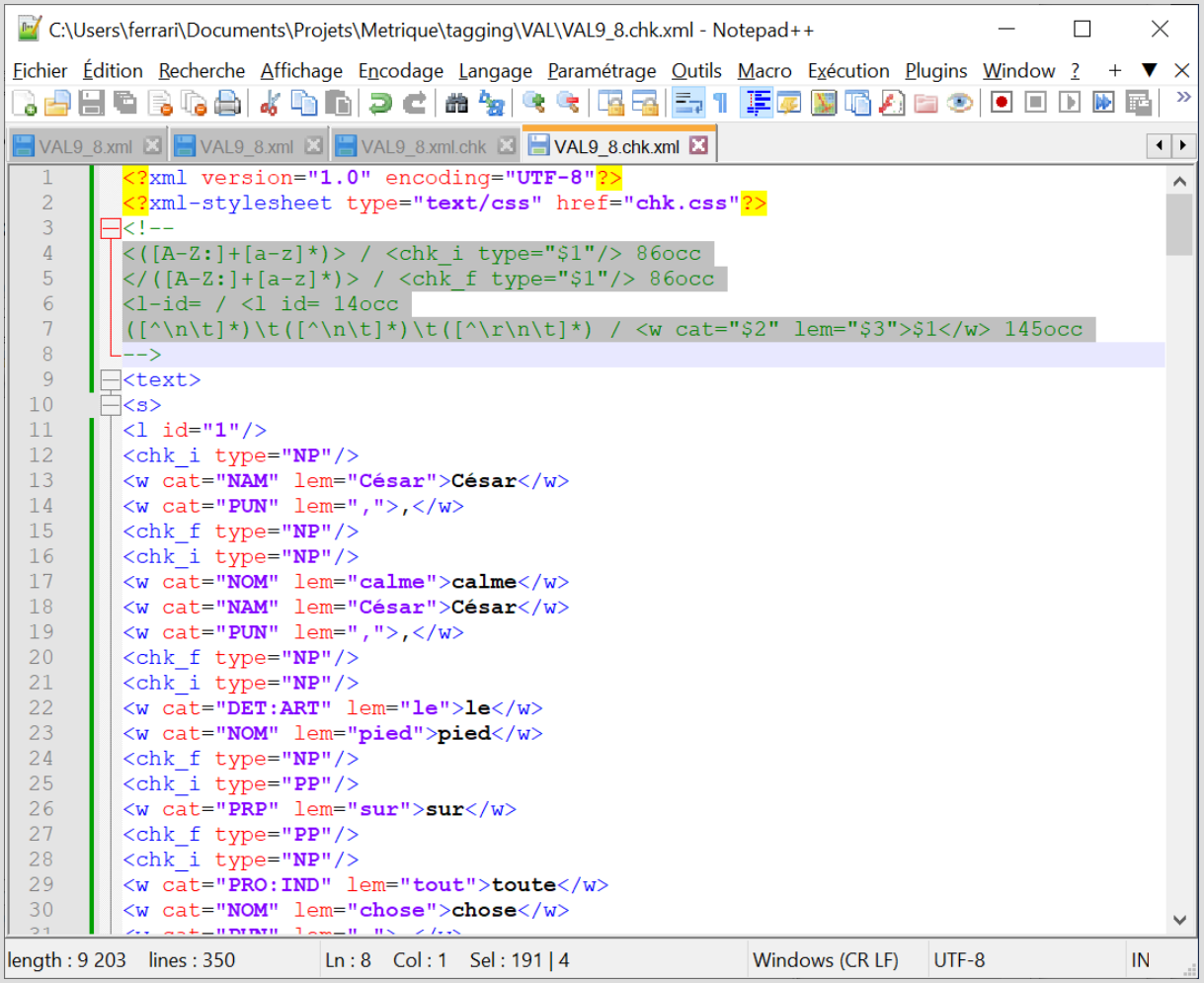
Tagging & chunking

3 étapes

1) Analyses

Exemple de sortie retravaillée pour permettre affichage et « réalignement »

XML



2. Poursuite : développement en cours

Visualisation d'un poème analysé avec le treetagger

l id='1' [César ,] [calme César ,] [le pied] [sur] [toute chose ,]
l id='2' [Les poings] [durs] [dans] [la barbe ,] [et] [l' œil] [sombre] [peuplé]
l id='3' [D'] [aigles] [et] [des] [combats] [du] [couchant] [contemplé ,]
l id='4' [Ton cœur] [s' enfle ,] [et] [se sent] [toute -puissante Cause] .
l id='5' [Le lac] [en] [vain] [palpite] [et] [lèche] [son lit] [rose ;]
l id='6' [En vain] [d'] [or] [précieux] [brille] [le jeune blé ;]
l id='7' [Tu durcis] [dans] [les nœuds] [de] [ton corps] [rassemblé]
l id='8' [L' ordre ,] [qui] [doit enfin] [fendre] [ta bouche] [close] .
l id='9' [L' ample monde ,] [au delà de] [l' immense horizon ,]
l id='10' [L' Empire] [attend] [l' éclair ,] [le décret ,] [le tison]
l id='11' [Qui] [changeront] [le soir] [en] [furieuse aurore] .
l id='12' [Heureux] [là] [-bas] [sur] [l' onde ,] [et] [bercé] [du] [hasard ,]
l id='13' [Un pêcheur] [indolent] [qui] [flotte] [et] [chante , ignore]
l id='14' [Quelle foudre] [s' amasse] [au] [centre] [de] [César] .

2. Poursuite : développement en cours

Visualisation d'un poème analysé avec le treetagger

l id='1' [César ,] [calme César ,] [le pied] [sur] [toute chose ,]
l id='2' [Les poings] [durs] [dans] [la barbe ,] [et] [l' œil] [sombre] [peuplé]
l id='3' [D'] [aigles] [et] [des] [combats] [du] [couchant] [contemplé ,]
l id='4' [Ton cœur] [s' enfle ,] [et] [se sent cat:VER:pres lem:sentir] [toute -puissante Cause] .
l id='5' [Le lac] [en] [vain] [palpite] [et] [lèche] [son lit] [rose ;]
l id='6' [En vain] [d'] [or] [précieux] [brille] [le jeune blé ;]
l id='7' [Tu durcis] [dans] [les nœuds] [de] [ton corps] [rassemblé]
l id='8' [L' ordre ,] [qui] [doit enfin] [fendre] [ta bouche] [close] .
l id='9' [L' ample monde ,] [au delà de] [l' immense horizon ,]
l id='10' [L' Empire] [attend] [l' éclair ,] [le décret ,] [le tison]
l id='11' [Qui] [changeront] [le soir] [en] [furieuse aurore] .
l id='12' [Heureux] [là] [-bas] [sur] [l' onde ,] [et] [bercé] [du] [hasard ,]
l id='13' [Un pêcheur] [indolent] [qui] [flotte] [et] [chante , ignore]
l id='14' [Quelle foudre] [s' amasse] [au] [centre] [de] [César] .

2. Poursuite : développement en cours

Visualisation d'un poème analysé avec le treetagger

l id='1' [César ,] [calme César ,] [le pied] [sur] [toute chose ,]
l id='2' [Les poings] [durs] [dans] [la barbe ,] [et] [l' œil] [sombre] [peuplé]
l id='3' [D'] [aigles] [et] [des] [combats] [du] [couchant] [contemplé ,]
l id='4' [Ton cœur] [s' enfle ,] [et] [se sent] [toute -puissante Cause] .
l id='5' [Le lac] [en] [vain] [palpite] [et] [lèche] [son lit] [rose ;]
l id='6' [En vain] [d'] [or] [précieux] [brille] [le jeune blé ;]
l id='7' [Tu durcis] [dans] [les nœuds] [de] [ton corps] [rassemblé]
l id='8' [L' ordre ,] [qui] [doit enfin] [fendre] [ta bouche] [close] .
l id='9' [L' ample monde ,] [au delà de] PP [l' immense horizon ,]
l id='10' [L' Empire] [attend] [l' éclair ,] [le décret ,] [le tison]
l id='11' [Qui] [changeront] [le soir] [en] [furieuse aurore] .
l id='12' [Heureux] [là] [-bas] [sur] [l' onde ,] [et] [bercé] [du] [hasard ,]
l id='13' [Un pêcheur] [indolent] [qui] [flotte] [et] [chante , ignore]
l id='14' [Quelle foudre] [s' amasse] [au] [centre] [de] [César] .