

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**

Traitement automatique de textes versifiés

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-

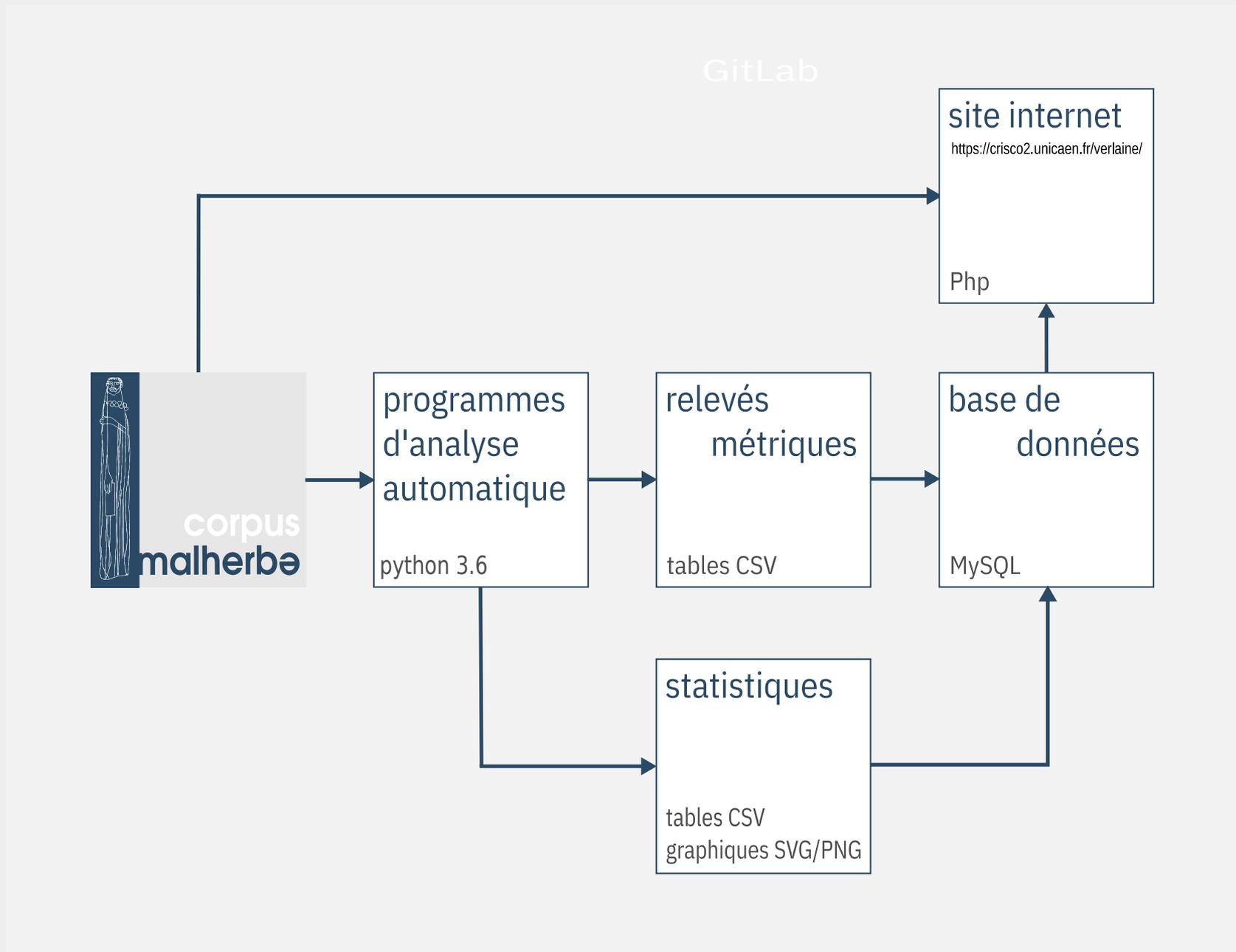
Richard Renault avec la collaboration de Stéphane Ferrari

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)

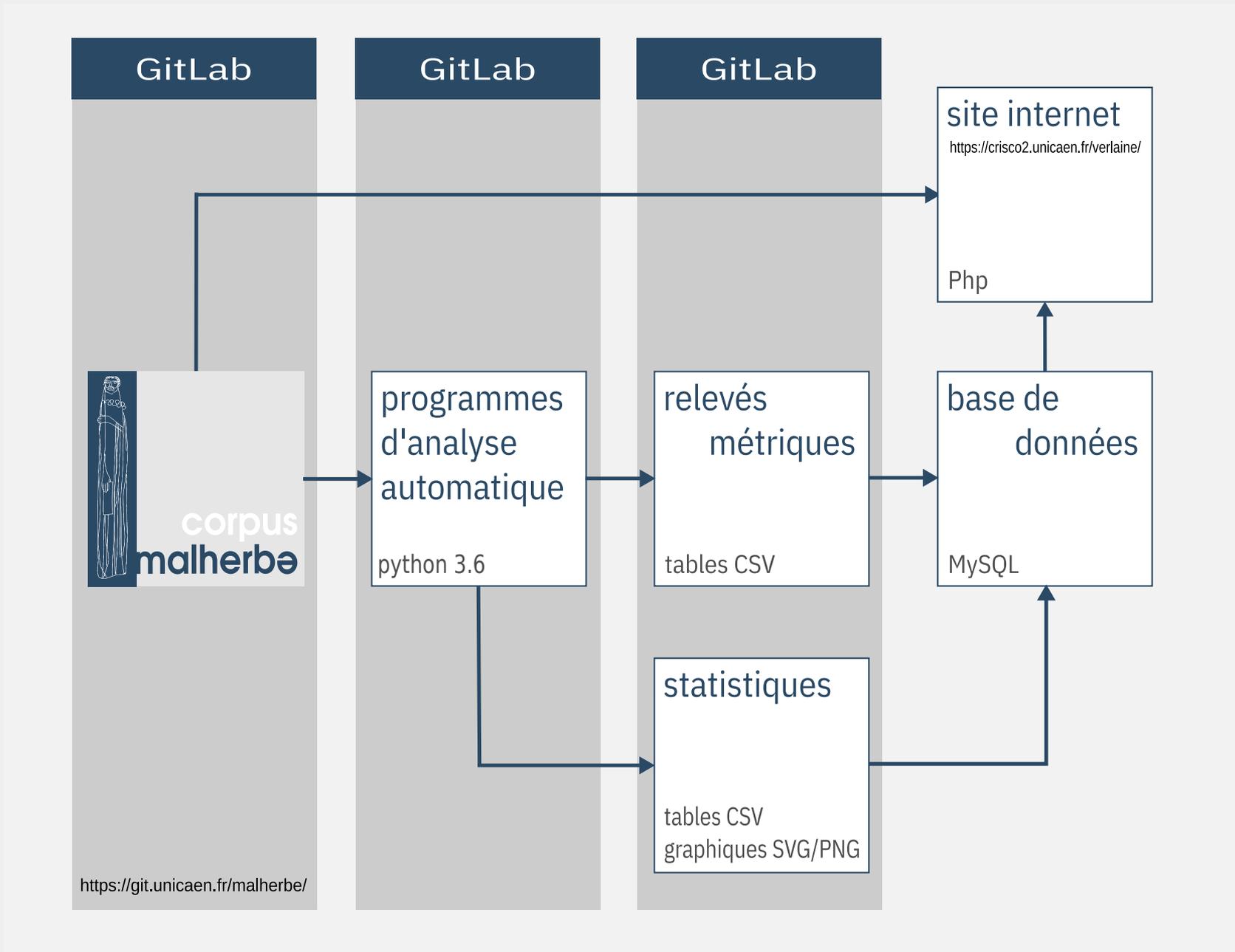
Traitement automatique de textes versifiés

1. Bilan



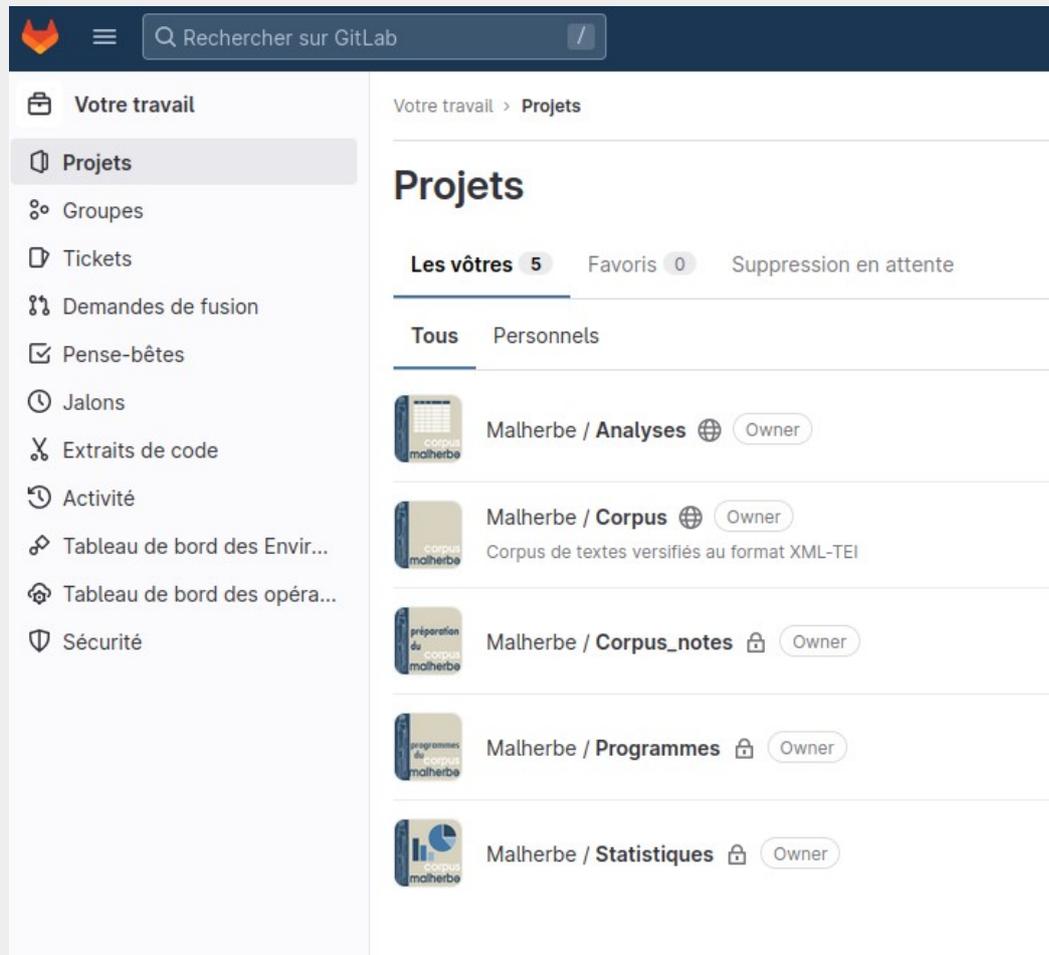
Traitement automatique de textes versifiés

1. Bilan



Traitement automatique de textes versifiés

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>

Traitement automatique de textes versifiés

1. Bilan

The screenshot displays a web application interface. On the left is a sidebar with navigation options: 'Votre travail', 'Projets', 'Groupes', 'Tickets', 'Demandes de fusion', 'Pense-bêtes', 'Jalons', 'Extraits de code', 'Activité', 'Tableau de bord des Envir...', 'Tableau de bord des opéra...', and 'Sécurité'. The main content area is titled 'Corpus Malherbe' and includes a 'Référence à citer' section with a citation for the 2022 Corpus Malherbe project. Below this is a 'Description' section detailing the corpus and its analysis. On the right, there are sections for 'Contacter le producteur', 'Téléchargement' (with a Creative Commons license), and 'Partager'. The bottom of the page lists several links related to the project.

ORTOLANG Catalogue Mes espaces

Corpus Malherbe

Soutien institutionnel :
ORTOLANG - Outils et Ressources pour un Traitement Optimisé de la LANGue - ANRu201311u2013EQPXu20130032

“ Référence à citer Texte BibTeX

(2022). *Corpus Malherbe* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v1, <https://hdl.handle.net/11403/malherbe/v1>.

Description

Le corpus *Malherbe* est un corpus de textes versifiés du XVIIe au XXe siècle.

Ce corpus au format XML-TEI a été préparé dans le cadre d'un projet de recherche du laboratoire CRISCO codirigé par Éliane Delente et Richard Renault et consacré à l'analyse automatique de la métrique des textes versifiés.

L'analyse automatique porte sur :

- l'identification des noyaux syllabiques
- le traitement des "e" instables
- le traitement des diérèses
- le calcul de la longueur métrique
- la détermination du profil métrique et le calcul du mètre des vers
- l'identification des rimes et des schémas de rimes
- la détermination des formes strophiques
- l'identification de la forme globale (forme fixe ou autre)
- l'identification de la PGTC et calcul de l'extension des rimes
- l'évaluation de la "qualité" des rimes
- le traitement statistique de la ponctuation (ponctuométrie)

Le corpus analysé est visible sur le site web du projet : <https://crisco4.unicaen.fr/verlaine/>

Les textes de ce corpus constituent la partie principale d'un corpus plus vaste (Corpus Malherbe) disponible sur le serveur Git de l'université de Caen : <https://git.unicaen.fr/malherbe/corpus>

Contacter le producteur

Envoyer un mail

Téléchargement

Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International
Cette licence permet aux autres de remixer, arranger, et adapter votre œuvre à des fins non commerciales tant qu'on vous crédite en citant votre nom et que les nouvelles œuvres sont diffusées selon les mêmes conditions.

Télécharger Parcourir

Partager

Aperçu

- Corpus_Malherbe_textes.txt
- COR1.xml
- BAU_1.xml

<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

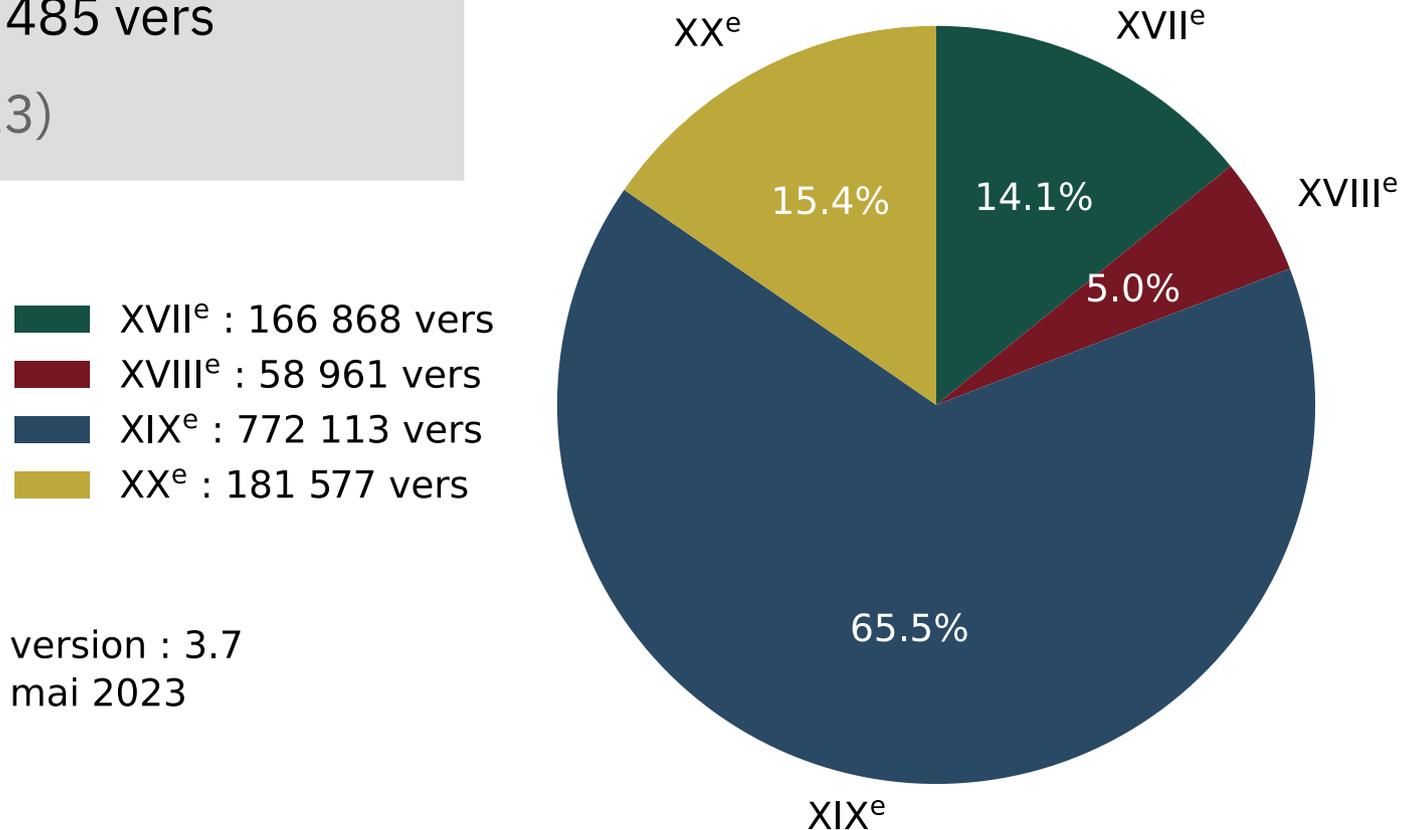
Corpus Malherbè

- 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



version : 3.7
mai 2023

1. Bilan

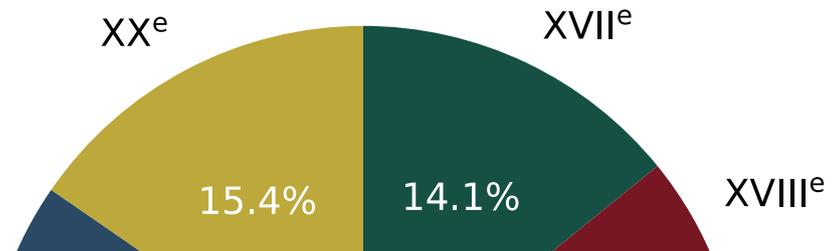
Corpus Malherbæ

- 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, Università 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 :

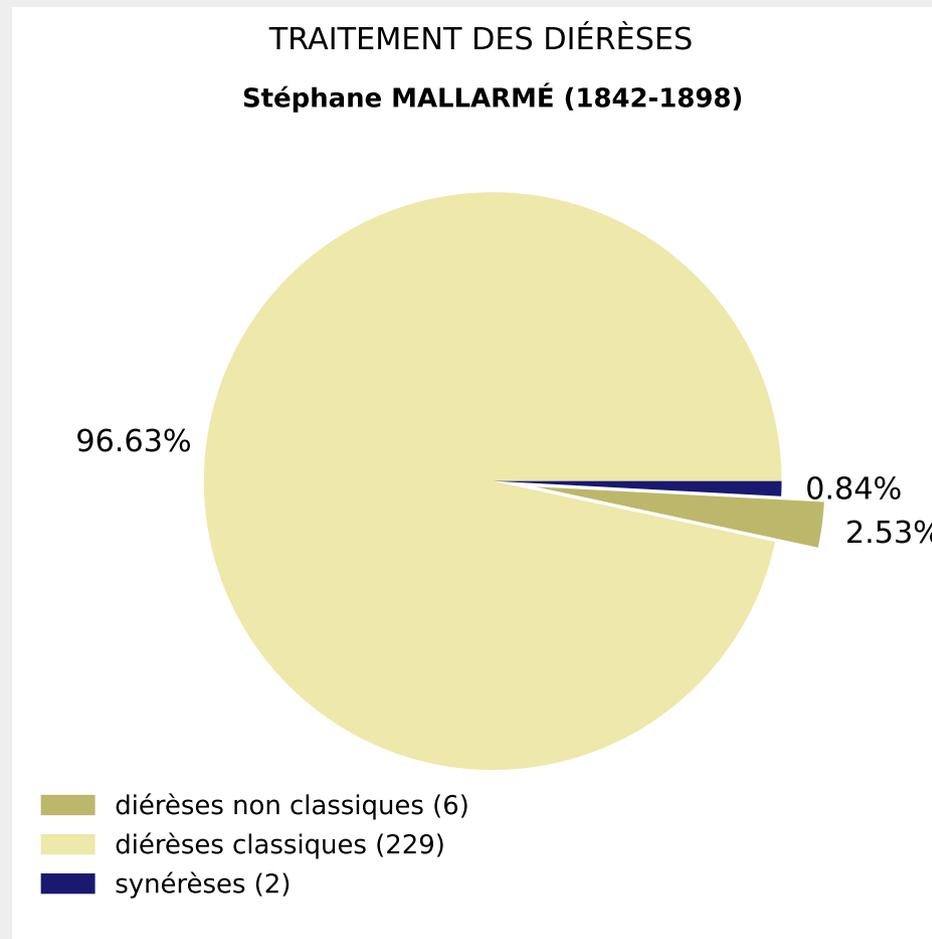
1. Découpage en mots (stable)
2. Identification des noyaux syllabiques (stable)
3. Traitement des "e" instables (stable)
4. Traitement des diérèses (stable)
5. Calcul de la longueur métrique des vers (stable)
6. Détermination du profil métrique du poème et du mètre des vers (stable)
7. Appariement des 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)

Traitement automatique de textes versifiés

1. Bilan

Post-traitement :

1. Traitement statistique des données analysées (avancé)

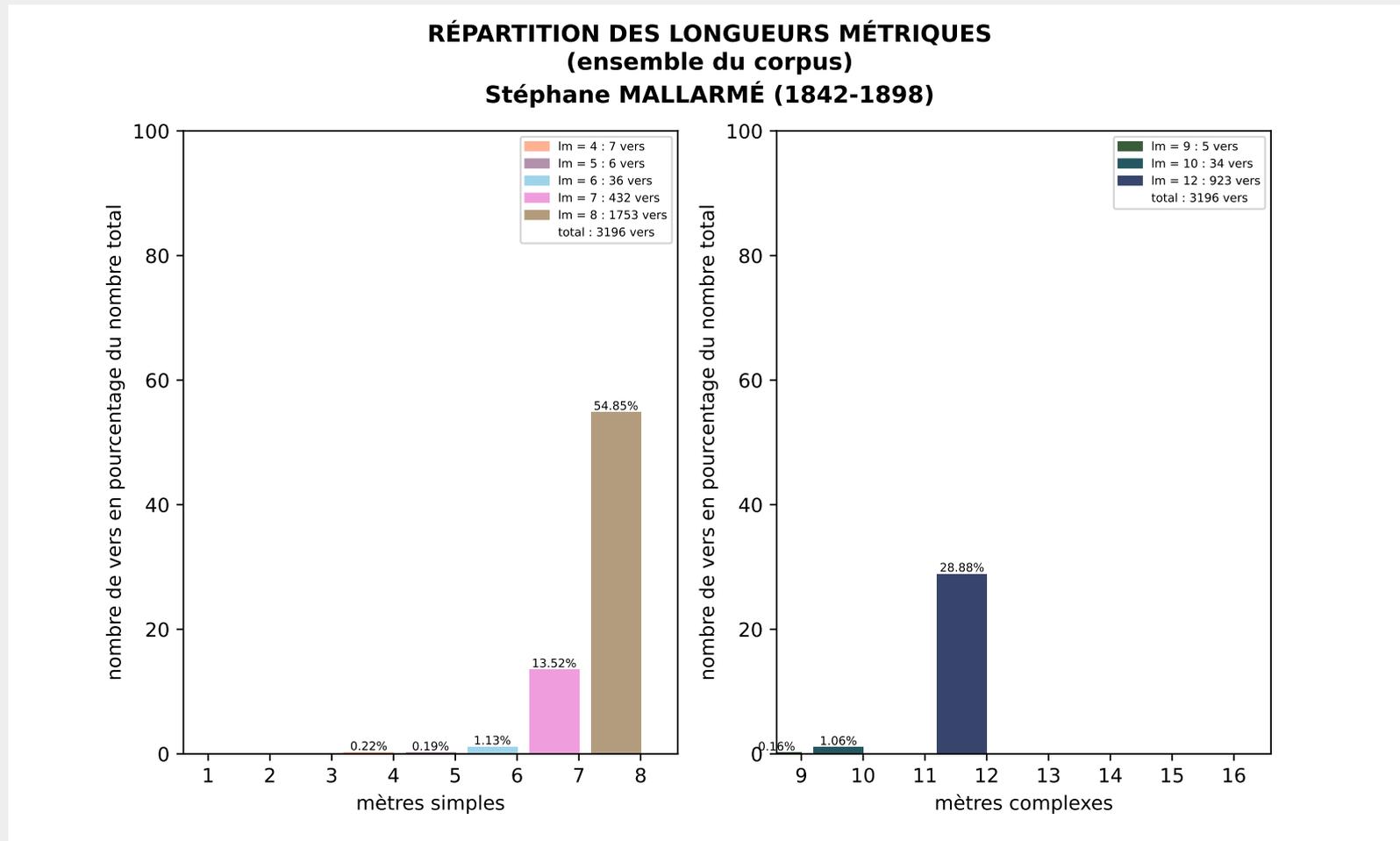


Traitement automatique de textes versifiés

1. Bilan

Post-traitement :

1. Traitement statistique des données analysées (avancé)

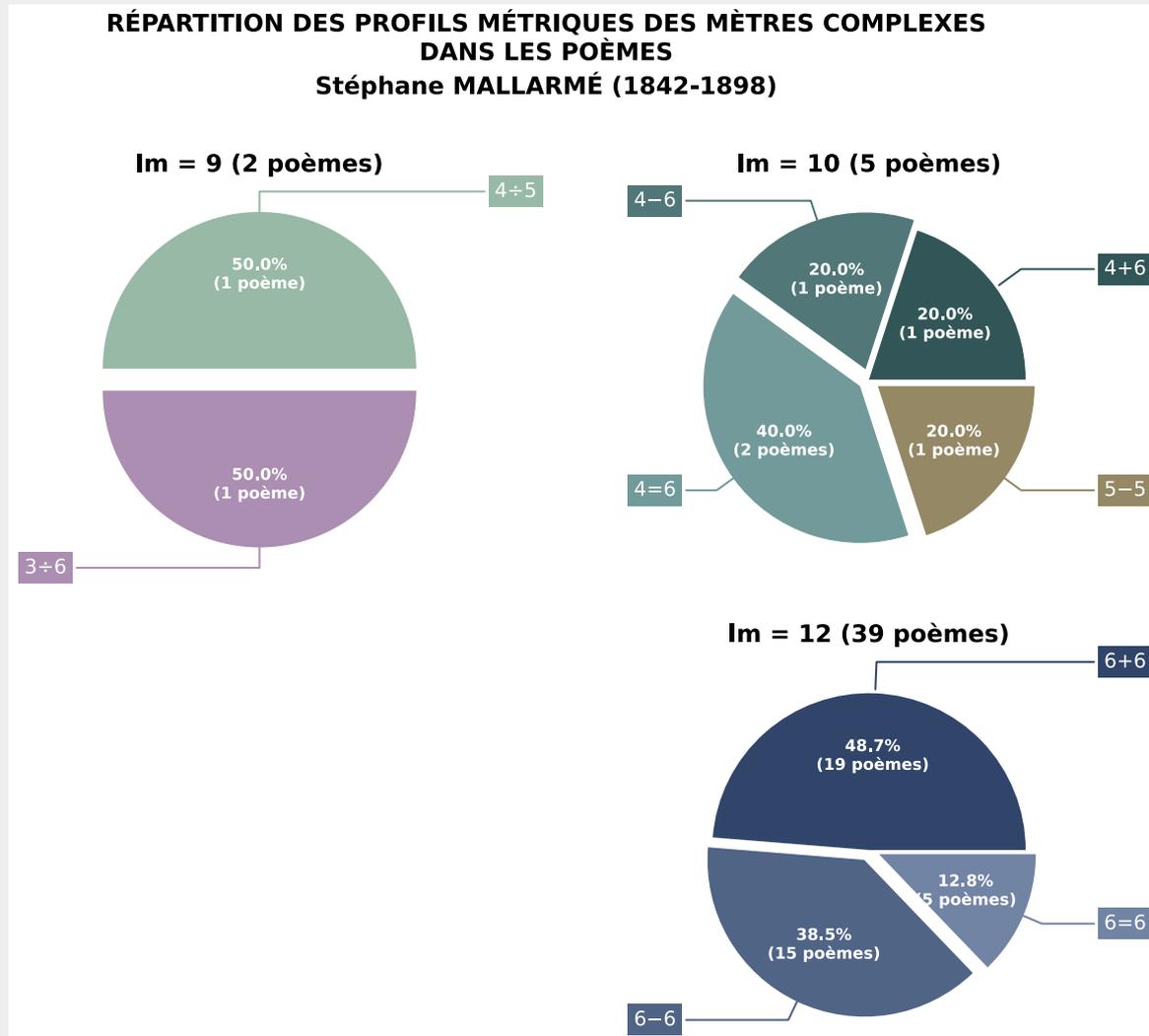


Traitement automatique de textes versifiés

1. Le projet : bilan

Post-traitement :

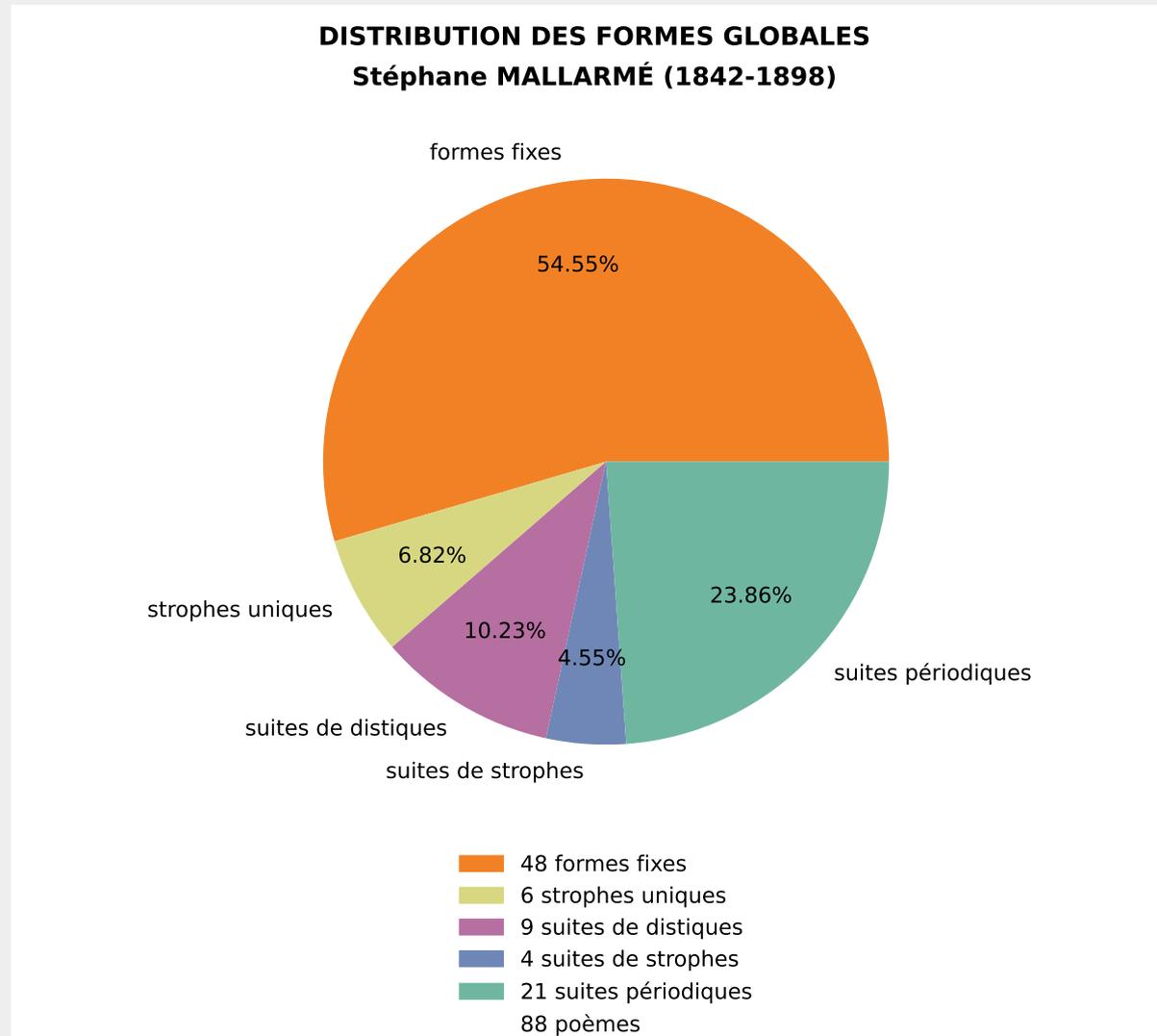
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

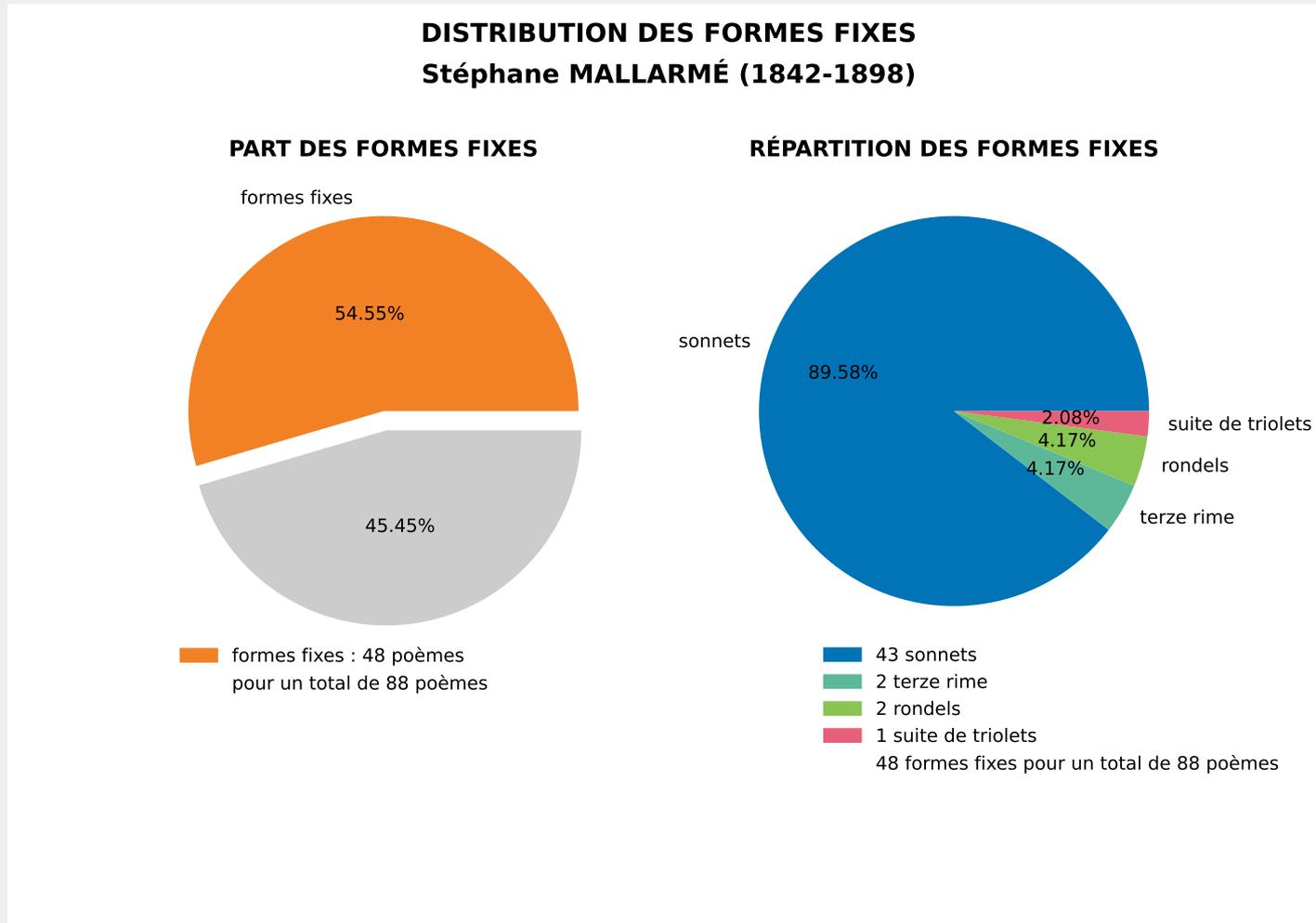
1. Traitement statistique des données analysées (avancé)



1. Bilan

Post-traitement :

1. Traitement statistique des données analysées (avancé)

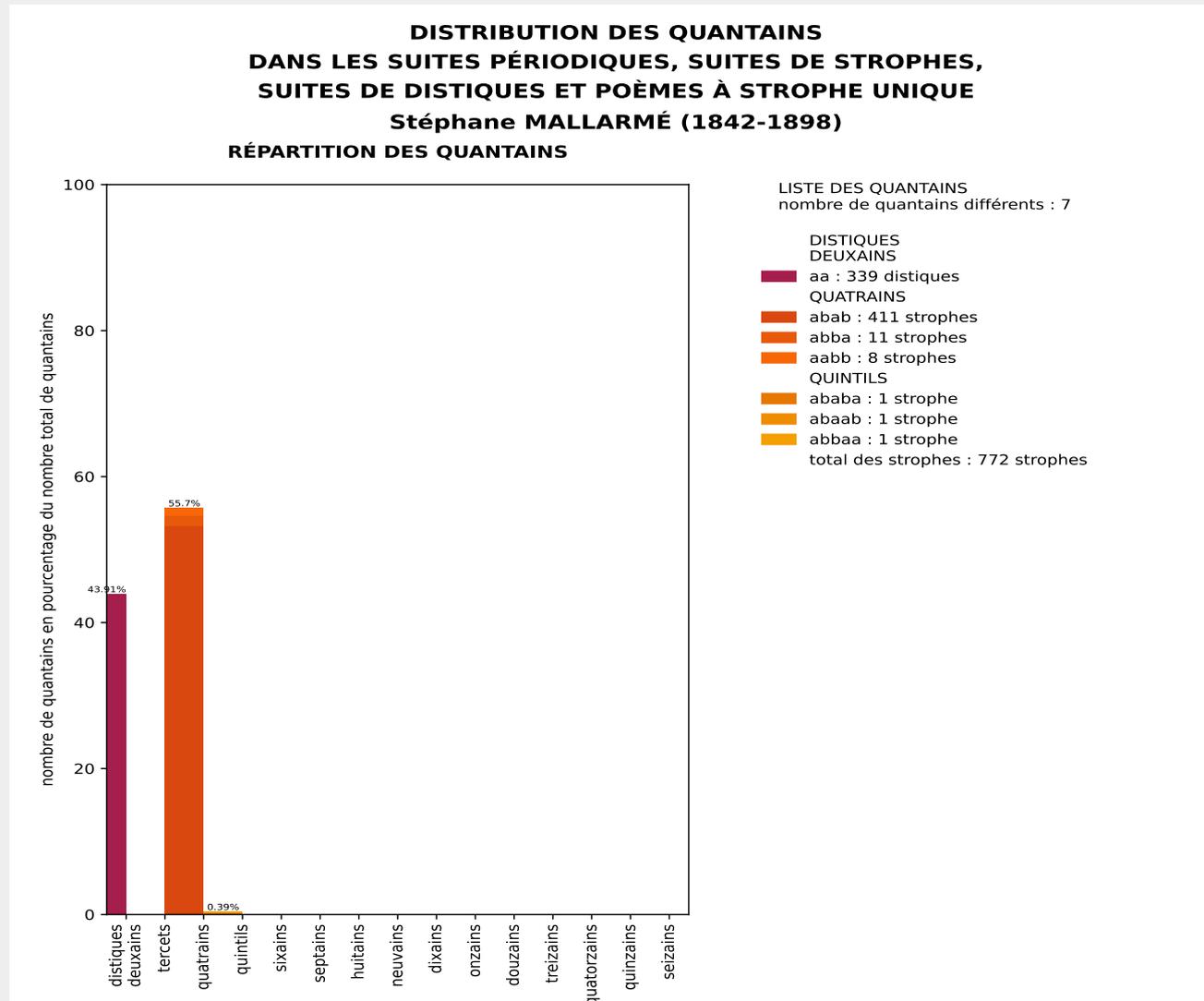


Traitement automatique de textes versifiés

1. Bilan

Post-traitement :

1. Traitement statistique des données analysées (avancé)

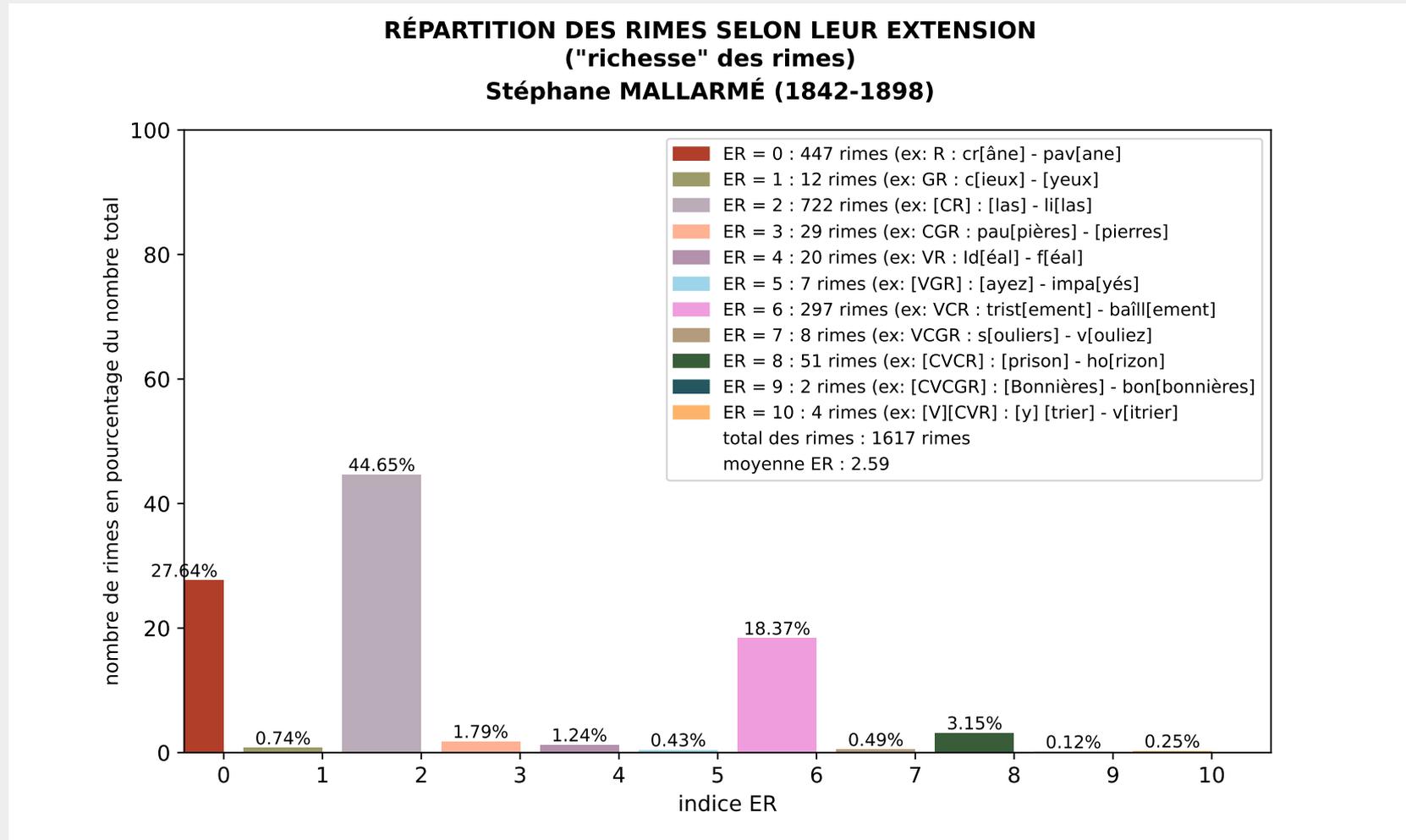


Traitement automatique de textes versifiés

1. Bilan

Post-traitement :

1. Traitement statistique des données analysées (avancé)

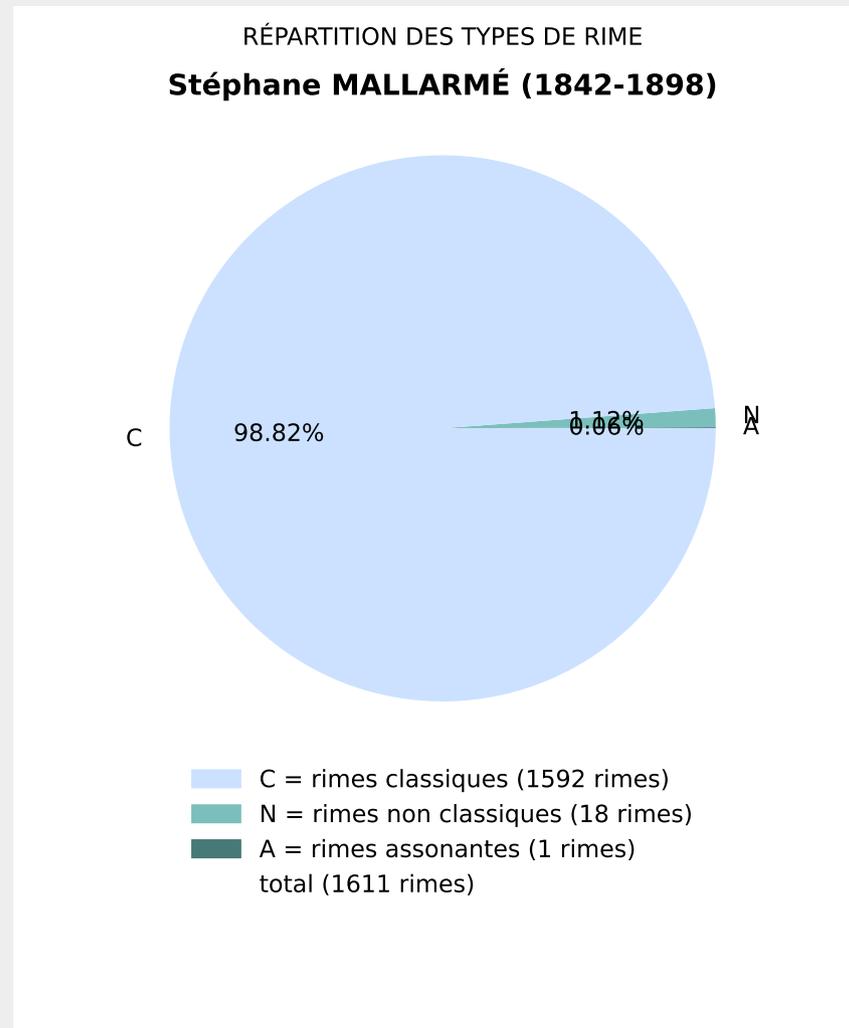


Traitement automatique de textes versifiés

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



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

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 b out	6+6	a
Du monde à l'autre, et même aux murs de la camp agne ,	6+6	b
Forçat de l'hôpital et malade du b agne ;	6+6	b
Dormeur maussade, à qui chaque aube dit : « Deb out ! »	6+6	a
Voyageur douloureux qu'attend la Mort, aub erge	6+6	a
Où l'on vend le lit dur et les pleurs blancs du ci erge ,	6+6	a

Tu gémis, étonné de te sentir si l as ;	6+6	a
Puis un jour tu te dis : « L'âme est un vain bag age ,	6+6	b
Et mon'cœur est bien lourd pour un pareil voy age ! »	6+6	b

10 Et, sans songer que Dieu te donne ses lil as ,	6+6	a
Tu veux jeter ton cœur, tu veux jeter ton âme ,	6+6	a
Pour alléger ta marche et mieux porter la F emme ;	6+6	a

15 Par ta route et ses ponts fiers de leur parap et ,	6+6	a
Compagnon de l'orgueil, fils des froides ét udes ,	6+6	b
Tu vas vers le malheur et vers les solit udes .	6+6	b

Tout plein des arguments dont l'esprit se rep ait ,	6+6	a
Tu fais, pour savourer ta gloire monot one ,	6+6	a
Taire ta conscience à l'heure où le ciel t onne .	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 b out	6+6	a
Du monde à l'autre, et même aux murs de la camp agne ,	6+6	b
Forçat de l'hôpital et malade du b agne ;	6+6	b
Dormeur maussade, à qui chaque aube dit : « Deb out ! »	6+6	a
Voyageur douloureux qu'attend la Mort, aub erge	6+6	a
Où l'on vend le lit dur et les pleurs blancs du ci erge ,	6+6	a
Tu gémis, étonné de te sentir si l as ;	6+6	a
Puis un jour tu te dis : « L'âme est un vain bag age ,	6+6	b
Et mon cœur est bien lourd pour un pareil voy age !»	6+6	b
10 Et, sans songer que Dieu te donne ses lil as ,	6+6	a
Tu veux jeter ton cœur, tu veux jeter ton âme ,	6+6	a
Pour alléger ta marche et mieux porter la F emme ;	6+6	a
Par ta route et ses ponts fiers de leur parap et ,	6+6	a
Compagnon de l'orgueil, fils des froides ét udes ,	6+6	b
15 Tu vas vers le malheur et vers les solit udes .	6+6	b
Tout plein des arguments dont l'esprit se rep ait ,	6+6	a
Tu fais, pour savourer ta gloire monot one ,	6+6	a
Taire ta conscience à l'heure où le ciel t onne .	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 b out	6+6	a
Du monde à l'autre, et même aux murs de la camp agne ,	6+6	b
Forçat de l'hôpital et malade du b agne ;	6+6	b
Dormeur maussade, à qui chaque aube dit : « Deb out ! »	6+6	a
Voyageur douloureux qu'attend la Mort, aub erge	6+6	a
Où l'on vend le lit dur et les pleurs blancs du ci erge ,	6+6	a
Tu gémis, étonné de te sentir si l as ;	6+6	a
Puis un jour tu te dis : « L'âme est un vain bag age ,	6+6	b
Et mon cœur est bien lourd pour un pareil voy age !»	6+6	b
10 Et, sans songer que Dieu te donne ses lil as ,	6+6	a
Tu veux jeter ton cœur, tu veux jeter ton âme ,	6+6	a
Pour alléger ta marche et mieux porter la F emme ;	6+6	a
Par ta route et ses ponts fiers de leur parap et ,	6+6	a
Compagnon de l'orgueil, fils des froides ét udes ,	6+6	b
15 Tu vas vers le malheur et vers les solit udes .	6+6	b
Tout plein des arguments dont l'esprit se rep ait ,	6+6	a
Tu fais, pour savourer ta gloire monot one ,	6+6	a
Taire ta conscience à l'heure où le ciel t onne .	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 Poétique*, 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 MUSARDISSES, 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 à 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
	- base	: typographie
■ composant (mot composé)		: typographie, listes
■ mot	- préposition	: listes
	- conjonction	: listes
	- verbe	: catégorie
	...	
■ constituant syntaxique		: catégorie, relations syntaxiques

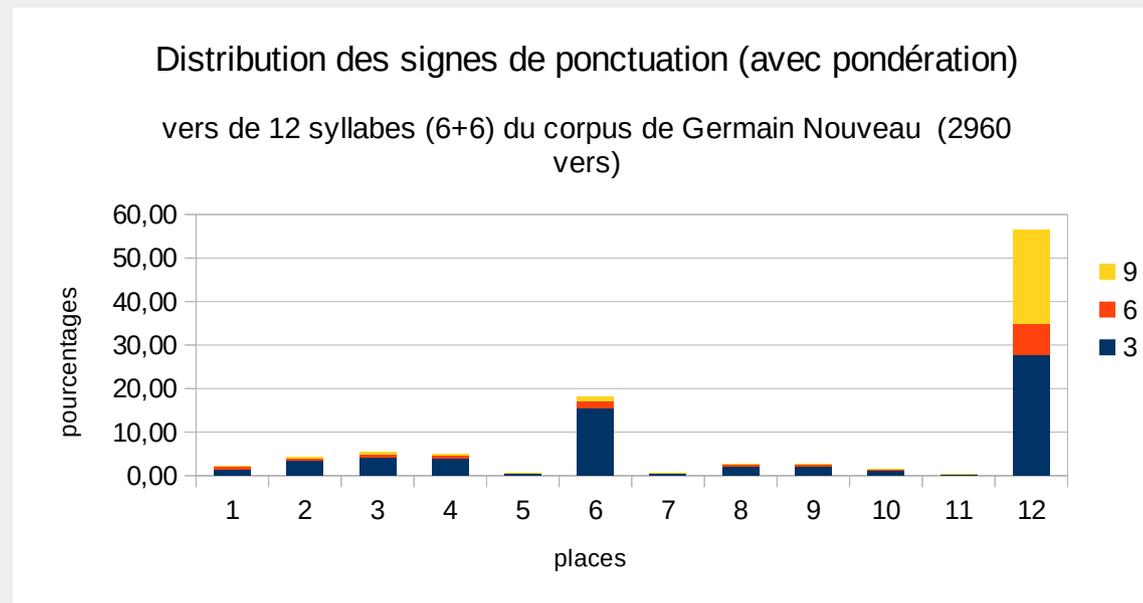
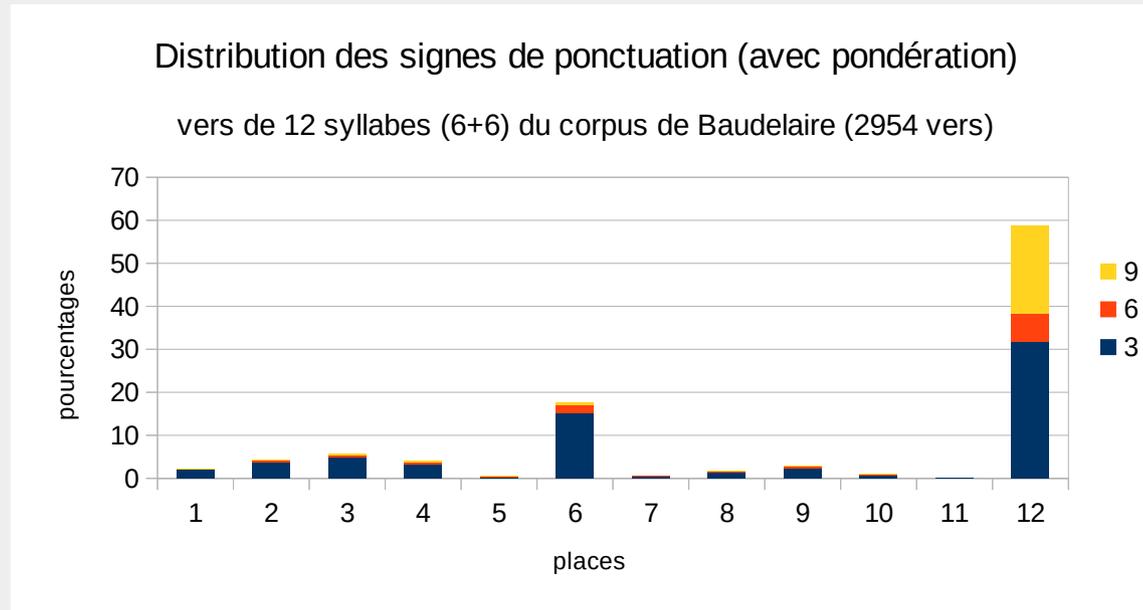
morphologie

syntaxe

Traitement automatique appliqué au corpus

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	plé	6+6	b
	D'aigles et des combats du couchant contem	plé,	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	blé ;	6+6	b
	Tu durcis dans les nœuds de ton corps rassem	blé	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	sard,	6+6	e
	Un pêcheur indolent qui flotte et chante, ign	ore	6+6	d
	Quelle foudre s'amasse au centre de Cé	sar.	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:ppe 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

Traitement automatique appliqué au corpus

2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

Tagging & chunking

3 étapes

1) Préparation

Exemple d'entrée

```
C:\Users\ferrari\Documents\Projets\Metrique\tagging\input\VAL9_8.xml - Notepad++
Fichier  Édition  Recherche  Affichage  Encodage  Langage  Paramétrage  Outils  Macro  Exécution  Plugins  Window  ?  +  ▼  ×
VAL9_8.xml  VAL9_8.xml  VAL9_8.xml.chk  VAL9_8.chk.xml
134  <listChange>
135  <change when="2016-03-02" who="RR">Révision de l'entête pour
validation TEI (TEI_corpus_Malherbe.xsd)</change>
136  <change when="2023-05-25" who="RR" type="analyse">Étape 8 de
l'analyse automatique du corpus : Traitement de la PGTC.
</change>
137  </listChange>
138  </revisionDesc>
139  </teiHeader><text><body><div type="poem" key="VAL9" modus="cm" lm_max="12"
metProfile="6+6" form="sonnet classique, prototype 2" schema="abba abba
ccd ede" er_moy="1.71" er_max="6" er_min="0" er_mode="0 (3/7)" er_moy_et=
"1.98">
140  <head type="main">César</head>
141  <lg n="1" rhyme="abba">
142  <l n="1" num="1.1" lm="12" met="6+6"><w n="1.1" punct=
"vg:2">C<seg phoneme="e" type="vs" value="1" rule="409"
place="1" mp="M">é</seg>s<seg phoneme="a" type="vs" value=
"1" rule="340" place="2" punct="vg">a</seg>r</w>, <w n=
"1.2">c<seg phoneme="a" type="vs" value="1" rule="340"
place="3">a</seg>lm<seg phoneme="e" type="ef" value="1"
rule="e-24" place="4" mp="F">e</seg></w> <w n="1.3" punct=
"vg:6">C<seg phoneme="e" type="vs" value="1" rule="409"
place="5" mp="M">é</seg>s<seg phoneme="a" type="vs" value=
"1" rule="340" place="6" punct="vg" caesura="1">a</seg>r
</w>,<caesura></caesura> <w n="1.4">l<seg phoneme="e" type=
"em" value="1" rule="e-12" place="7" mp="C">e</seg></w> <w
n="1.5">pi<seg phoneme="e" type="vs" value="1" rule="241"
place="8">e</seg>d</w> <w n="1.6">s<seg phoneme="y" type=
"vs" value="1" rule="450" place="9" mp="P">u</seg>r</w> <w
n="1.7">t<seg phoneme="u" type="vs" value="1" rule="425"
place="10">u</seg>t<seg phoneme="e" type="ef" value="1"
rule="409" place="11" mp="M">é</seg>s<seg phoneme="a" type="vs" value=
"1" rule="340" place="12" punct="vg">a</seg>r</w>,</l></lg></body></text></div></div>
length : 23 952  lines : 163  Ln : 1  Col : 1  Pos : 1  Unix (LF)  UTF-8  IN
```

Traitement automatique appliqué au corpus

2. Poursuite : développement en cours

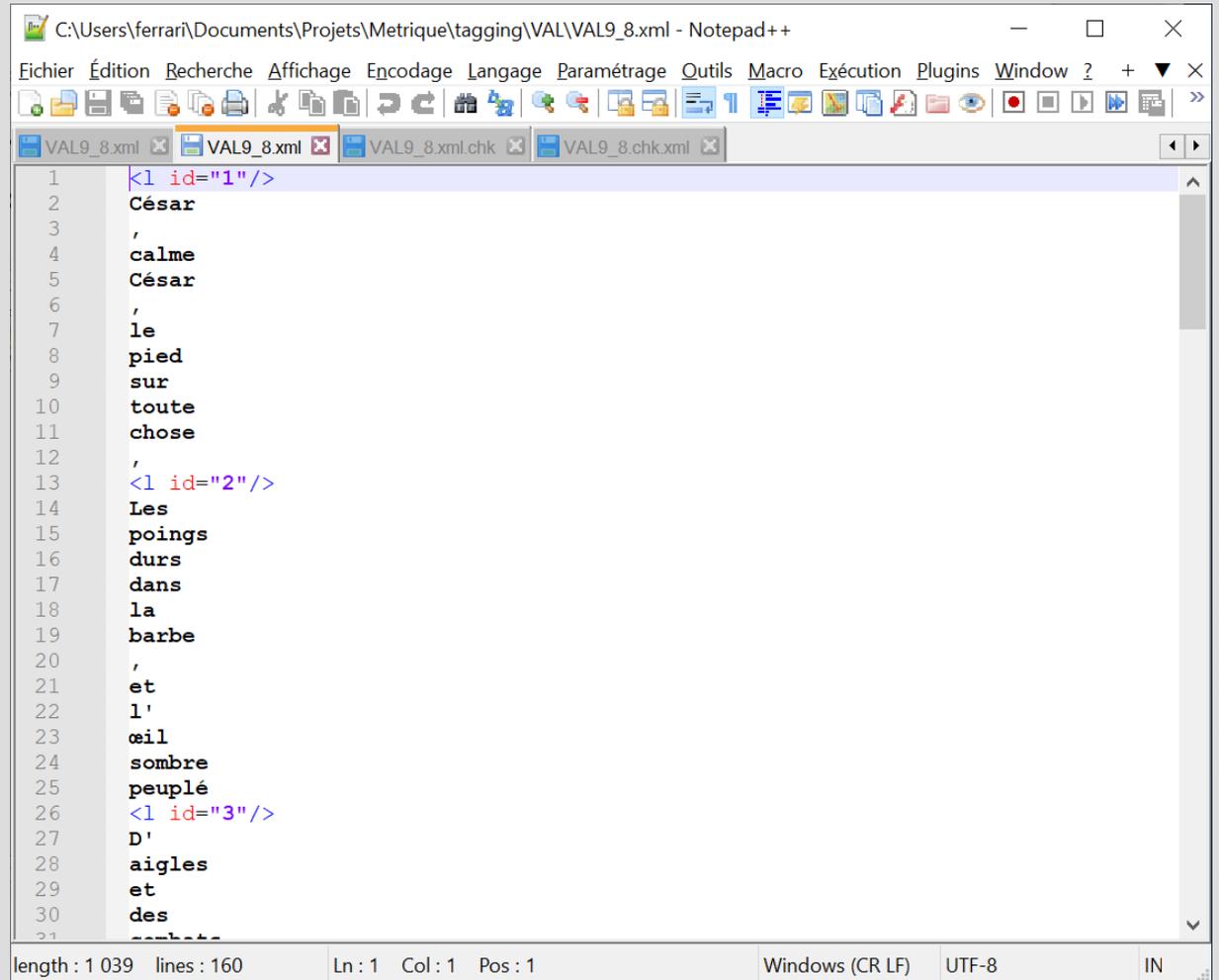
Premiers tests d'analyse avec le treetagger

Tagging & chunking

3 étapes

1) Préparation

Entrée préparée
(Python + XSLT)



The screenshot shows a Notepad++ window with the following content:

```
C:\Users\ferrari\Documents\Projets\Metrique\tagging\VAL\VAL9_8.xml - Notepad++
Fichier Édition Recherche Affichage Encodage Langage Paramétrage Outils Macro Exécution Plugins Window ? + ▼ ×
VAL9_8.xml x VAL9_8.xml x VAL9_8.xml.chk x VAL9_8.chk.xml x
1 <l id="1"/>
2 César
3 ,
4 calme
5 César
6 ,
7 le
8 pied
9 sur
10 toute
11 chose
12 ,
13 <l id="2"/>
14 Les
15 poings
16 durs
17 dans
18 la
19 barbe
20 ,
21 et
22 l'
23 œil
24 sombre
25 peuplé
26 <l id="3"/>
27 D'
28 aigles
29 et
30 des
31 ...
```

length : 1 039 lines : 160 Ln : 1 Col : 1 Pos : 1 Windows (CR LF) UTF-8 IN

Traitement automatique appliqué au corpus

2. Poursuite : développement en cours

Premiers tests d'analyse avec le treetagger

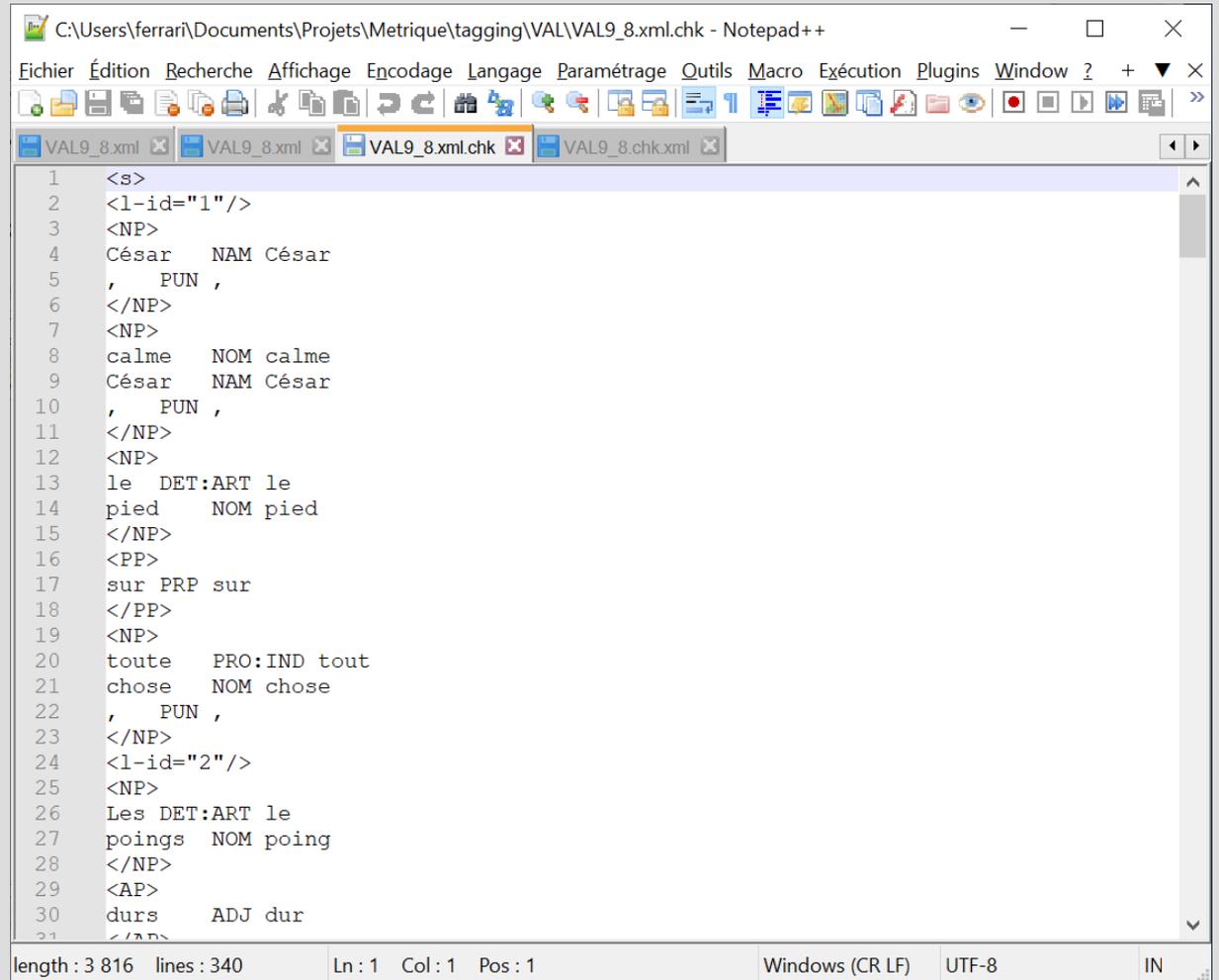
Tagging & chunking

3 étapes

1) Analyse(s)

Exemple de sortie
directement issue
du Treetagger

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



```
C:\Users\ferrari\Documents\Projets\Metrique\tagging\VAL\VAL9_8.xml.chk - Notepad++
Fichier  Édition  Recherche  Affichage  Encodage  Langage  Paramétrage  Outils  Macro  Exécution  Plugins  Window  ?  +  ▼  ×
VAL9_8.xml  VAL9_8.xml  VAL9_8.xml.chk  VAL9_8.chk.xml
1  <s>
2  <l-id="1"/>
3  <NP>
4  César    NAM  César
5  ,        PUN  ,
6  </NP>
7  <NP>
8  calme   NOM  calme
9  César   NAM  César
10 ,        PUN  ,
11 </NP>
12 <NP>
13 le      DET:ART le
14 pied   NOM  pied
15 </NP>
16 <PP>
17 sur    PRP  sur
18 </PP>
19 <NP>
20 toute  PRO:IND tout
21 chose  NOM  chose
22 ,        PUN  ,
23 </NP>
24 <l-id="2"/>
25 <NP>
26 Les    DET:ART le
27 poings NOM  poing
28 </NP>
29 <AP>
30 durs   ADJ  dur
31 </AP>
```

length : 3 816 lines : 340 Ln : 1 Col : 1 Pos : 1 Windows (CR LF) UTF-8 IN

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

```
C:\Users\ferrari\Documents\Projets\Metrique\tagging\VAL\VAL9_8.chk.xml - Notepad++
Fichier  Édition  Recherche  Affichage  Encodage  Langage  Paramétrage  Outils  Macro  Exécution  Plugins  Window  ?  +  ▼  ×
VAL9_8.xml  VAL9_8.xml  VAL9_8.xml.chk  VAL9_8.chk.xml
1  <?xml version="1.0" encoding="UTF-8"?>
2  <?xml-stylesheet type="text/css" href="chk.css"?>
3  <!--
4  <([A-Z:]+[a-z]*)> / <chk_i type="$1"/> 86occ
5  </([A-Z:]+[a-z]*)> / <chk_f type="$1"/> 86occ
6  <l-id= / <l id= 14occ
7  ([^\n\t]*)\t([^\n\t]*)\t([^\r\n\t]*) / <w cat="$2" lem="$3">$1</w> 145occ
8  -->
9  <text>
10 <s>
11 <l id="1"/>
12 <chk_i type="NP"/>
13 <w cat="NAM" lem="César">César</w>
14 <w cat="PUN" lem=",">,</w>
15 <chk_f type="NP"/>
16 <chk_i type="NP"/>
17 <w cat="NOM" lem="calme">calme</w>
18 <w cat="NAM" lem="César">César</w>
19 <w cat="PUN" lem=",">,</w>
20 <chk_f type="NP"/>
21 <chk_i type="NP"/>
22 <w cat="DET:ART" lem="le">le</w>
23 <w cat="NOM" lem="pied">pied</w>
24 <chk_f type="NP"/>
25 <chk_i type="PP"/>
26 <w cat="PRP" lem="sur">sur</w>
27 <chk_f type="PP"/>
28 <chk_i type="NP"/>
29 <w cat="PRO:IND" lem="tout">toute</w>
30 <w cat="NOM" lem="chose">chose</w>
31 <w cat="PUN" lem=",">,</w>
length : 9 203  lines : 350  Ln : 8  Col : 1  Sel : 191 | 4  Windows (CR LF)  UTF-8  IN
```

Traitement automatique appliqué au corpus

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 ] .
```

Traitement automatique appliqué au corpus

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 ] .
```