corpus malherbe

Corpus de textes versifiés

Université Caen Normandie, CRISCO (EA 4255)

1. Projet scientifique

Responsables du projet : Éliane Delente et Richard Renault

Le projet scientifique dans lequel s'inscrit le corpus *Malherbe* a pour objet le traitement automatique des textes versifiés de la fin du XVII^{e} au début du XX^{e} siècle.

Ce projet comporte trois parties :

- La constitution d'un corpus de textes versifiés : poèmes et pièces de théâtre.
- La conception, la réalisation et l'application de programmes d'analyse automatique de textes versifiés. Les différents modules sont :
  - 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 (sonnet, triolet, ballade...), forme périodique ou suite de strophes
  - l'évaluation de l'extension et de la "qualité" des rimes
  - le traitement statistique de la ponctuation (ponctuométrie)
- La constitution d'une base de données de relevés métriques générés automatiquement par application des programmes de traitement automatique. Cette partie du projet inclut également l'intégration de relevés métriques établis antérieurement au projet par des métriciens (relevés métriques produits manuellement).

2. Corpus Malherbe

Responsable du corpus : Richard Renault

Les textes au format XML avec encodage UTF-8 sont structurés au moyen des éléments et attributs définis par la TEI[1]. Le formatage XML des textes est conforme au schéma de validation propre au corpus *Malherbe* (voir plus loin en 2.5).

2.1 Origine des textes

La plus grande partie des textes au format XML-TEI a été préparée à partir de sources électroniques disponibles sur internet. Quelques textes, ou parfois, quelques poèmes en complément d'un recueil, ont été numérisés ou directement saisis par l'équipe du projet. La source des textes disponibles sur internet est diversifiée : *Gallica, CNRTL (Frantext), Enclitt, poesies.net, archive.org, WikiSource, Project Gutenberg, ABU, Bibliothèque électronique de Lisieux...*

2.2 Contenu des fichiers

Les textes du corpus *Malherbe* sont destinés au traitement métrique automatique. L'édition numérique de recueils de poésies ou de pièces de théâtre n'étant pas l'objet du projet, les différentes parties qui relèvent du paratexte des recueils et des pièces de théâtre (préface, avertissement, commentaires, notes de l'éditeur...) ne sont pas reprises dans les textes du corpus. Seules les parties en prose qui relèvent des poèmes ou des pièces sont incluses : dédicaces, citations, arguments, didascalies, notes de l'auteur...

2.2 Préparation des textes

La préparation a consisté à appliquer une mise en forme XML à partir d'une source au format TXT ou PDF. Plus rarement, la préparation s'est limitée à appliquer une transformation XSL lorsque le texte était déjà disponible au format XML. Selon l'origine des textes, cette préparation a demandé plus ou moins de travail car les éditions électroniques utilisées peuvent comporter aussi bien un balisage XML minimal qu'une absence totale de structuration du texte. Bien souvent le texte source au format TXT ne comporte pas de découpage en strophes par exemple. L'orthographe a été vérifiée au moyen du correcteur Aspell[2] afin d'éliminer les éventuelles scories dues à la numérisation. Lors de cette vérification, une attention particulière a été portée aux mots dont l'orthographe diffère de l'usage courant, notamment les formes anciennes (*sentimens*, *souvenois*, *nénufar*...) et les licences poétiques (*encor, certe, avecque, dévoûment, tûra*...).

2.3 Édition de référence

La mise en forme du texte doit d'être conforme à l'édition imprimée de référence ; ce qui ne pose pas de problème lorsque cette édition de référence est celle qui est à l'origine de la version électronique. Mais c'est

1 http://www.tei-c.org/index.xml
2 http://aspell.net

loin d'être toujours le cas, car trop souvent la source électronique ne mentionne pas l'édition imprimée à l'origine de la numérisation. En l'absence d'information sur l'origine du texte source, nous avons ajouté une édition de référence pour le contrôle de la mise en forme du texte, le contrôle des données et pour les corrections métriques qui s'imposaient. Dans la mesure du possible, nous avons choisi pour cette édition de référence celle qui correspond le mieux à la source électronique. Dans certains cas, notamment pour l'œuvre de Victor Hugo et Paul Verlaine, nous avons utilisé comme édition de référence une édition sans rapport avec les sources électroniques. Ce choix est justifié par le fait que nous avons établi la table des matières et la numérotation des poèmes relativement aux relevés métriques exhaustifs fait par des métriciens antérieurement au projet. L'alignement du corpus sur les relevés métriques existants facilite ainsi une analyse comparative entre les relevés métriques des métriciens et ceux produits automatiquement par les programmes d'analyse. Afin d'être conforme au contenu de ces relevés métriques et à l'édition imprimée de référence, nous avons dû parfois ajouter des poèmes.

2.4 Analyse métrique des textes

Tous les textes ont été analysés par les programmes d'analyse automatique ; ce qui a permis de débusquer quelques coquilles et erreurs résiduelles. Il s'agit dans la plupart des cas de coquilles dues à des erreurs de numérisation, mais aussi, plus rarement, d'erreurs présentes également dans la version imprimée de l'édition de référence. Dans ce cas, nous avons eu recours à une autre édition imprimée pour valider les corrections métriques. Ces corrections ont été introduites explicitement au moyen d'éléments et attributs XML appropriés. Il peut cependant rester encore des coquilles ou des erreurs, indécelables par le traitement dans la mesure où elles n'affectent pas les résultats de l'analyse métrique. Les corrections des erreurs qui relèvent de la numérisation ou de l'océrisation du texte ont été enregistrées dans un fichier propre à chaque recueil de poèmes ou pièce de théâtre. Ce fichier texte est disponible sur demande à l'adresse du projet[3]

2.5 Validation XML

Le projet suit les recommandations de la TEI pour le choix des éléments et attributs du formatage XML. Néanmoins, quelques modifications mineures et ajouts ont été faits pour répondre aux spécificités du corpus et pour prendre en compte les exigences relatives à l'analyse métrique. La validation concerne aussi bien le corpus initial (le présent corpus) que le corpus final, après application des programmes d'analyse. Les textes du corpus final sont enrichis par des éléments et attributs qui encodent les propriétés :1) des noyaux syllabiques (type de voyelle, place, phonème...), 2) des vers (longueur métrique, mètre du vers, propriétés métrico-métriques[4], césure...), 3) des rimes (appariement des rimes, PGTC[5], schéma de rimes, valeurs calculées des rimes...), 4) des strophes (type de strophe, schéma de strophe...) 5) du poème ou de la pièce de théâtre (profil métrique, forme globale (forme fixe (sonnet, ballade, triolet, rondeau...), suite périodique avec ou sans alternance, simple suite de strophes...), 6) du texte (ponctuométrie). La version enrichie du corpus XML n'est pas disponible pour le moment de façon systématique mais seulement sur demande et pour une partie du corpus. Le schéma de validation XML-TEI (schéma XSD) est disponible sur le site du projet[6]. Le schéma de validation XSD associé au corpus a été élaboré dans le cadre d'une vacation financée en 2016 par le CNRS sur appel à projet commun au consortium CORLI[7] et à l'Equipex ORTOLANG[8].

2.6 Manuel d'encodage

Le manuel d'encodage du corpus *Malherbe* est en préparation et sera prochainement disponible sur le site du projet.

2.7 Redistribution des textes

Tous les textes du corpus mis à disposition sont hors droits et sont diffusés à des fins de recherche ou d'enseignement. Comme il est d'usage dans ce cas, ces textes peuvent être repris, modifiés et diffusés pour peu que l'entête XML ou un quelconque document d'accompagnement précise et retrace l'origine des textes.

2.8 Extension du corpus

Le corpus *Malherbe* (version 3.5 – septembre 2021) comporte :

- 220 auteurs
- 481 recueils de poésies
- 133pièces de théâtre
- 18 737 poèmes
- 1 001 965 vers

2.9 Disponibilité du corpus

Le corpus est disponible sur le site du projet[9] et sur le serveur Gitlab de l'université de Caen[10].

---

3 crisco.incipit@unicaen.fr.
4 cf. les travaux de Benoît de Cornulier.
5 Plus Grande Terminaison Commune.
6 https://crisco2.unicaen.fr/Verlaine/ressources/TEI_Corpus_Malherbe_1.2.xsd
7 http://www.huma-num.fr/consortiums
8 https://www.ortolang.fr/
9 https://crisco2.unicaen.fr/Verlaine/
10 ****