Le livret d'opéra : établissement et exploration textométrique d'un corpus patrimonial de l'époque classique

Sébastien Jacquot[1], Margareta Kastberg Sjöblom[2]

[1] EA 4661 ELLIADD - Université de Franche-Comté - France

[2] EA 4661 ELLIADD - Université de Franche-Comté - France

Abstract

This article focuses on the establishment and analysis of a corpus made up by a patrimonial heritage which includes the entire repertoire of the first generation of French Opera librettos performed at the *Royal Music Academy* at *Palais Royal* in Paris between 1673 and 1732.
The constitution of such a corpus is far from easy. Texts from the seventeenth century and the first half of the eighteenth century still show significant spelling variations and the old typography makes automation unreliable. The aim of our contribution is to show how adequate normalization of the corpus and structured tagging, according to the TEI procedures, enhance the quality of the final study of the corpus and the interpretation of the statistical results.

Keywords: corpus establishment, normalization, tagging, lexicometry, textometry, historical text

Résumé

Cet article s'intéresse à l'établissement et l'exploration d'un corpus représentant un genre particulier et une époque particulière de l'histoire de France, la naissance de l'opéra français, constitué à partir de livrets couvrant l'ensemble du répertoire de la première génération de livrets d'opéra françaises réalisées à l'Académie de musique royale au Palais-Royal à Paris entre 1673 et 1732.
La constitution d'un tel corpus est loin d'être facile. Les textes de l'époque classique montrent encore des variations orthographiques importantes et l'ancienne typographie rendent l'automatisation peu fiable. Le but de notre contribution est de montrer comment une normalisation adéquate du corpus et un codage structuré et nivelé, conformément aux procédures TEI, permettent de cibler l'analyse du corpus et d'affiner l'interprétation des résultats statistiques.

Mots-clés : établissement de corpus, normalisation, encodage, philologie numérique, textométrie

1. Introduction

Ce travail s'intéresse à l'établissement et l'exploration d'un corpus représentant un genre particulier et une époque particulière de l'histoire de France. Il s'agit de la naissance de l'opéra français, sous le régime de Louis XIV. *La tragédie en musique* est le nom qu'on donne au genre naissant, l'opéra français, à la deuxième moitié du XVII^{e} siècle lorsqu'elle est introduite et présentée par Lully et Quinault en 1673. Ce n'est qu'au milieu du XVIII^{e} siècle que la dénomination *tragédie lyrique* s'impose.

La tragédie en musique, et plus généralement l'opéra, est un genre particulièrement complexe qui obéît à des contraintes génériques précises et aux XVII^{e} et XVIII^{e} siècles ces règles étaient particulièrement strictes.

Les contraintes imposées par le genre sont nombreuses ; la tragédie en musique est composée de trois éléments, paroles, musique et danse ; la partie textuelle, « les paroles », doivent donc respecter les contraintes musicales, le rythme et la versification. *La tragédie en musique* doit aussi être conforme aux idéaux et à l'esthétique baroque, et l'élément tragique, qui prend, le plus souvent, sa source dans la mythologie grecque, doit permettre l'élévation vers un genre noble, à côté de la tragédie classique. *La tragédie en musique* est aussi porteuse d'un message politique très fort, caractéristique d'une époque où le pouvoir central de la monarchie absolue était extrêmement important. Néanmoins, la tragédie en musique doit divertir, le public était avide de spectacles et de sensations fortes, dans une époque où Louis XIV et ses courtisans étaient les premiers spectateurs.

Le genre de la *tragédie en musique* diffère des autres genres de l'époque tels que le ballet, la comédie-ballet, le ballet-pastoral, le pastoral héroïque, le ballet héroïque, opéra-ballet et le divertissement. Il est calqué sur le modèle de la tragédie, avec un prologue suivi de cinq actes qui est la caractéristique générique formelle, au moins jusqu'en 1732 et l'arrivée de Rameau.

2. Constitution du corpus

La délimitation du corpus est un préalable à toute étude statistique lexicale ou textométrique. Nous avons tenu à ce que notre corpus ait une cohérence institutionnelle aussi bien que poétique. La cohérence institutionnelle se réfère à la scène de l'Académie Royale de musique de Paris (notre corpus ne contient aucune œuvre italienne ou performance de province)[1].

Notre corpus s'étend sur près de cinquante ans, du 27 avril 1673 au 6 novembre 1732, et représente la totalité de la production de tragédies en musique de la période de Lully à Rameau données sur la scène de l'Académie Royale de musique durant cette période. Le corpus englobe 75 livrets et respecte scrupuleusement l'ordre chronologique ; il contient 654.809 occurrences distribuées sur 16.489 lemmes et les divers sous-corpus sont de tailles relativement homogènes, chaque livret comptant, en moyenne, 8731 occurrences. On y trouve le répertoire complet de livrets de tragédies en musique répertoriés dans le *RECUEIL GENERAL DES OPERA, representez par l'ACADEMIE ROYALE DE MUSIQUE, DEPUIS SON ETABLISSEMENT,* édité en 1703 par Christophe Ballard, imprimeur du Roy pour la musique, rue St. Jean de Beauvais, au Mont-Parnasse, avec Privilège de sa Majesté.

La constitution de ce genre de corpus est loin d'être aisée. En effet, dans le cas de ces livrets, l'automatisation de la saisie et un traitement d'océrisation se sont avérés presque impossibles. Une large partie des textes a par conséquent été saisie manuellement[2]. Dans les textes du XVII^{e} et de la première moitié du XVIII^{e} siècle, on est en effet confronté à une graphie non encore stable et à des variations orthographiques importantes, même à l'intérieur d'un seul livret. La typographie ancienne rend elle aussi l'automatisation peu fiable.

Dans les textes du XVII^{e} et de la première moitié du XVIII^{e} siècle, avec l'imprimerie bien mise en place, on se trouve face à des textes bien plus lisibles que les textes médiévaux. Toutefois, on est toujours confronté à une graphie non encore stable et à des variations orthographiques importantes même à l'intérieur d'un livret donné. La typographie ancienne rend aussi l'automatisation peu fiable. Le *s* long « *ſ* », par exemple, est confondu avec le « *f* »

[1] C'est pourquoi ce corpus ne commence pas en 1669 avec Cambert, mais bien en 1673 avec l'œuvre de Lully sur livret de Quinault, et il termine en 1732, ce qui est cohérent du point de vue musical avec l'arrivée sur la scène de Rameau et la « grande querelle » entre « Lullistes » et « Ramistes ». Nous n'ignorons pas que la tragédie en musique perdure après 1732, mais elle prend de plus en plus souvent le nom d'opéra.

[2] Saisie en partie financée par l'obtention d'un BQR en 2012-2013.

(f) lors de la reconnaissance optique, c'est pourquoi l'ampleur du travail de correction manuelle après traitement OCR a incité à saisir les textes manuellement. S'ajoute également la difficulté des voix multiples et simultanées, caractéristique du chant, qui demande un soin particulier, étant donné que le texte n'est pas toujours linéaire.

3. Normalisation des données

Les logiciels textométriques sont construits selon des principes différents. *Hyperbase*[3] par exemple ne fait pas appel à un codage XML, il traite le corpus comme un continuum textuel, sans indications hiérarchiques, prenant comme jalons les limitations naturelles des différents livrets. La première version du corpus numérisé, celle qui respecte la graphie originale avec ses variantes et fluctuations, a été intégrée au logiciel *Hyperbase*. Il s'agit là d'un travail coûteux car la fluctuation orthographique est permanente et capricieuse, ce qui demande un grand effort manuel. Ce choix d'établir un corpus respectant les variantes de langue permet l'accès au texte original mais également de voir l'évolution orthographique durant cette période, bien que l'objectif de ce travail ne soit pas l'évolution de la langue du XVII^{e} et XVIII^{e} siècles.

La base a été lemmatisée par *Cordial*[4], en employant une technique qui tient compte de la versification. Le traitement du texte en vers est en effet relativement contraignant car il ne faut pas, comme dans l'analyse du texte en prose, tenir compte des limitations selon les paragraphes (pour lesquelles sont généralement conçus les logiciels textométriques), mais des vers et des strophes[5].

Etant donné la disparité graphique qui rend le regroupement selon les lemmes difficile, il convient, lorsque l'on a affaire à un corpus de ce genre, d'effectuer une normalisation et une homogénéisation des formes afin d'acquérir davantage de robustesse et de fiabilité dans l'analyse statistique.

Le caractère particulier de ce corpus, avec ses nombreuses variantes graphiques, demande en effet à être normalisé afin de mener des analyses rigoureuses sur le vocabulaire. Pour la normalisation et la création d'une version normalisée du corpus selon les codes contemporains, nous avons eu recours au logiciel *DiaTag*[6] qui est un ensemble modulaire destiné à répondre aux besoins d'élaboration et de préparation de bases textuelles en français, révisé pour être compatible avec les recommandations de la TEI[7].

DiaTag procède par comparaison et reconnaissance à partir de son propre dictionnaire, à l'origine strictement dédié au français moderne et contemporain. Afin de détecter les items lexicaux anciens, considérés comme « étrangers » au dictionnaire de référence, nous avons confronté le corpus au dictionnaire de référence.

[3] E. Brunet, *Manuel d'Hyperbase, version 9.0*, Université de Nice-Sophia Antipolis, 2014.
[4] Il s'agit d'une lemmatisation « simple » qui n'homogénéise pas les variantes orthographiques.
[5] E. Brunet, « Enquête sur la langue poétique au XVI^{e} siècle », *in Comptes d'auteurs, Tome 1, Etudes statistiques de Rabelais à Gracq,* Paris, Honoré Champion, 2009, p. 19-39. Article publié dans *Du Bellay, antiquités et nouveaux mondes dans les recueils romains,* Université de Nice Sophia-Antipolis.
[6] Outil élaboré par Jean-Marie Viprey au sein du laboratoire ELLIADD, Université de Franche-Comté. Cf. V. Lethier, J.-M. Viprey, « Annotation linguistique de corpus : vers l'exhaustivité par la convivialité », *in* S. Heiden et B. Pincemin (éds.), *JADT'09, 9èmes Journées internationales d'Analyse statistique des Données Textuelles,* Lyon, Presses Universitaires de Lyon, 2008.
[7] TEI, Text Encoding Initative.

L'application et l'analyse de *DiaTag* permet ainsi l'identification initiale des items inconnus dans le corpus. Une fois les formes reconnues, une programmation permet d'introduire des automatisations dans le traitement. Ainsi, les *moy, toy, soy* et *roy* se convertissent en *moi, toi, soi* et *roi*, les paradigmes du verbe *sçavoir* s'accordent avec celles de *savoir*, les terminaisons en *-ois, oit* et *-oient* deviennent *-iais, ait* et *-aient*, les *meme, mesme* et *même* se trouveront sous l'unique forme *même*, etc.

Certains problèmes résistent toutefois aux automatismes de traitement. Par exemple la terminaison *-ez* qui est difficile à gérer. Le *-ez* jusqu'à la fin du XVII^{e} siècle correspondent aussi bien à la conjugaison verbale de la deuxième personne à tous les temps qu'au participe passé ou à l'adjectif verbal. L'outil *DiaTag* permet de travailler soit de façon automatique, soit en combinant le traitement informatique avec une annotation manuelle. En effet, étant donné qu'un corpus étiqueté n'est pas forcément totalement désambiguïsé et qu'il reste certains problèmes non résolus pendant l'opération d'étiquetage automatique, *DiaTag* fait partie de ces logiciels qui aboutissent à une phase interactive qui permet le dialogue entre homme et machine pour lever ces ambiguïtés, ce qui est extrêmement utile dans le travail d'un corpus comme le nôtre.

Une autre difficulté, cette fois-ci liée au genre, est posée par les nombreuses références à la mythologie grecque. La tragédie en musique fourmille en effet de noms propres grecs, on fait très souvent référence aux peuples anciens habitant les rives de la Méditerranée dans la mythologie ainsi qu'à des lieux ensevelis dans l'oubli. Ainsi, on est en présence de chœurs de *Lidiens*, de *l'isle de Cytere*, etc. S'ajoute à cette difficulté la disparité orthographique de ces formes, avec des variantes qui portent facilement à confusion. En effet, la francisation de termes et de noms grecs n'est pas encore accomplie aux XVII^{e} et XVIII^{e} siècles. On remarque même une volonté chez les librettistes de « revenir en arrière », vers la source grecque, au milieu du XVIII^{e} siècle. S'ajoute enfin une autre particularité générique : dans certaines pièces les chœurs chantent des textes italiens, et évidemment il s'agit d'un italien vieilli et désuet.

La normalisation du corpus aboutit en fin de traitement à une nouvelle version du corpus, qui permet une analyse thématique plus fiable. En effet, il est aujourd'hui indispensable de constituer et explorer les trois différentes versions d'un corpus, basées sur des données respectant la forme graphique, des données lemmatisées et des données normalisées. Dans l'exploration textométrique il s'avère nettement que ces différentes versions de corpus peuvent répondre à des questions différentes et qu'elles sont complémentaires.

4. Encodage du corpus

L'étiquetage d'un corpus n'est pas une activité neutre, le choix du format d'encodage est déterminant pour la pérennité du corpus et pour la finalité de la recherche.

La constitution d'un corpus demande donc une attention toute particulière quant à l'encodage et aux annotations. Dans l'objectif de création d'un corpus de référence, au niveau français aussi bien qu'au niveau international, le choix du format est crucial. Seul l'encodage TEI répond aujourd'hui aux critères internationaux[8]. Le modèle théorique s'adapte aujourd'hui au

[8] Le consortium international de la TEI réunit 70 membres de 16 pays différents et il élabore des recommandations pour le balisage de textes et de ressources linguistiques. Pour des informations plus détaillées voir : http://www.tei-c.org/

langage XML[9]. Les recommandations du consortium définissent ainsi une syntaxe recommandée pour le format et un métalangage pour la description des structures d'encodage de texte.

Les principes de la TEI – aujourd'hui considérés comme la référence de « l'état de l'art » en encodage international – doivent être la ligne directrice de l'encodage de tout corpus patrimonial dans le domaine de la philologie, ou plus largement, des humanités numériques.

Pour la création de la base de données de *la tragédie en musique*, les livrets ont été soumis à un codage respectant les normes TEI afin de permettre une accessibilité aisée ainsi que la possibilité d'une comparaison adéquate avec d'autres corpus patrimoniaux. Il s'agit d'un fonds numérisé des livrets qui autorise la création, d'une base de fac-similés, d'une base plein-texte respectant la graphie originale et d'une version normalisée toutes deux conformes aux standards XML TEI, permettant ainsi une analyse statistique, hypertextuelle et multidimensionnelle.

Ce projet s'associe en partie aux travaux de l'équipe ICAR de l'ENS Lyon, dans le cadre du projet TXM (Plateforme Textométrie), dans le sens où le corpus encodé respecte les contraintes de TXM et qu'il sera consultable dans le portail TXM (TXM Web).

L'étiquetage XML intégral du corpus selon les normes de la TEI permet en outre une répartition hiérarchique du corpus, selon sa particularité générique, et la possibilité de travailler sur des partitions pour ainsi cibler et affiner l'analyse.

En effet, le codage TEI permet aussi une meilleure maniabilité d'un corpus comme la création d'une certaine hiérarchie dans les textes pour ainsi extraire des parties distinctes, par exemple les didascalies ou les « approbations du roi ». Pour l'analyse linguistique, notamment textométrique, ce codage rend aussi possible le travail sur les seules répliques prononcées (sachant que la redondance par exemple de l'annonce des personnages peut fausser les statistiques) et permet ainsi de créer des sous-corpus pertinents et de procéder à des analyses précises et nivelées.

4.1. Encodage structurel selon le critère générique

La structure hiérarchique de la tragédie en musique ressemble à celle du théâtre, spécifique aux genres dramatiques. Nous avons donc élaboré un encodage qui tient compte de cette spécificité et de ce caractère hiérarchique en nous inspirant des principes de la TEI et des recommandations pour encoder le texte théâtral TEI P5-*Drama*.

Le module TEI-*Drama* traite de l'encodage des textes destinés à être représentés sur scène ou des textes des transcriptions de spectacles parmi lesquelles les recommandations de la TEI ne distinguent pas les diverses modalités de représentation scéniques[10].

Le schéma de l'encodage respecte dans les grandes lignes celui du théâtre classique, à savoir la hiérarchie et les partitions suivantes :

[9] Pour plus de détails sur le langage XML, *cf.* T. Boulanger, *XML par la pratique - Bases indispensables, Concepts et cas pratiques,* St Herblain, Editions ENI (3e édition), 2015.

[10] Il est important de noter que le TEI-*Drama* est destiné spécifiquement à la transcription des aspects structurels des écrits (ou imprimés) des textes dramatiques. Pour la transcription des spectacles réels tels que textes parlés, les recommandations de la TEI se réfèrent à d'autres éléments définis dans son module oral, *Transcriptions de discours*.

Dédicace au roi (le cas échéant) • Personnages du prologue • Prologue • Personnages de la tragédie • Premier acte ▪ première scène ▪ deuxième scène ▪ troisième scène ▪ quatrième scène ▪ etc. • Deuxième acte ▪ première scène ▪ deuxième scène ▪ troisième scène ▪ etc.	• Troisième acte ▪ première scène ▪ deuxième scène ▪ etc. • Quatrième acte ▪ première scène ▪ deuxième scène ▪ etc. • Cinquième acte ▪ première scène ▪ deuxième scène ▪ troisième scène ▪ quatrième scène ▪ etc. • Approbation du roi

L'encodage permet ainsi une identification efficace des partitions des textes. L'exemple suivant est extrait de *Cadmus & Hérmione*[11] :

```
  </sp>
</div2>
<div2 n="2" name="SCÈNE TROISIÈME" type="scene">
  <head>
    <w orig="SCENE">SCÈNE</w> <w orig="TROISIEME">TROISIÈME</w>.</head>
  <stage type="setting">HERMIONE, <w orig="CHARITE">CHARITÉ</w>., AGLANTE, LA NOURRICE D'HERMIONE, UN PAGE.</stage>
  <sp who="HERMIONE">
    <speaker>HERMIONE.</speaker>
    <lg>
      <l>Cet aimable <w orig="sejour">séjour</w>
      </l>
      <l>Si paisible &amp;amp; si sombre,</l>
      <l>Offre du silence et de l'ombre,</l>
      <l>A qui veut éviter le bruit, &amp;amp; le grand jour ;</l>
      <l>Ah ! que n'est-il aussi facile</l>
      <l>De trouver un <w orig="azile">asile</w>
      </l>
      <l>Pour éviter l'Amour !</l>
      <pb facs="01_CADMUS/164.png" n="164">$164</pb>
      <l>L'Impitoyable tyrannie,</l>
      <l>Dont je <w orig="sui">suis</w> les barbares <w orig="loix">lois</w>,</l>
      <l>Ne <w orig="deffend">défend</w> pas d'aimer le chant &amp;amp; l'harmonie.</l>
      <l>Vous qui me faites compagnie</l>
      <l>Répondez à ma voix.</l>
    </lg>
  </sp>
  <sp who="AGLANTE">
    <speaker>AGLANTE.</speaker>
    <lg>
      <l>On a beau fuir l'Amour, on ne peut l'éviter,</l>
      <l>On n'oppose à ses traits qu'une <w orig="deffense">défense</w> vaine :</l>
```

Ce balisage hiérarchique permet d'isoler par la suite dans les paroles chantées par les acteurs chantants des indications pour l'organisation scénique, différentes présentations d'acteur, des prologues, dédicaces, approbations, etc.

4.2. Développement d'un parseur dédié

Pour l'encodage du corpus, un parseur dédié à l'automatisation de la conversion des textes source vers le standard XML TEI P5-*Drama* a été développé en Java. Le parseur prend en entrée des fichiers au format HTML, tente de maximiser la détection des diverses entités, puis produit des fichiers XML TEI P5 en se basant sur les recommandations TEI-*Drama*. Ces fichiers peuvent ensuite aisément être importés dans des logiciels prenant en charge le format XML TEI, comme TXM par exemple.

[11] Le lecteur attentif notera que les formes graphiques anciennes sont complétées avec leurs correspondants normalisés en français contemporain, nous y reviendrons dans le passage suivant.

Lors de l'analyse des lignes et des caractères des fichiers source, le parseur utilise une liste de règles pour tenter de définir si la partie du texte en train d'être lue peut, ou ne peut pas être, de tel ou tel type. Bien qu'ayant fait l'objet d'une analyse préalable assez approfondie, ces règles ont été affinées de manière empirique au cours du développement, procédant ainsi beaucoup par essai et erreur. Actuellement, le parseur permet notamment d'auto-détecter et d'encoder, depuis un fichier HTML « bien construit », les éléments suivants :

– Acte, Prologue, Approbation, autres (*head* et *div1@type*)

– Scène (*div2@type*)

– Orateur (*speaker* et *sp@who*)

– vers et groupe de vers (*l* et *lg*)

– Différents types de didascalies (*stage@type*) : didascalie de début d'acte ou de scène ; didascalie entre les répliques et didascalie liée à un *speaker*

Le parseur permet ainsi, par exemple, d'auto-détecter les lignes contenant le nom de l'orateur et d'en extraire l'orateur et les éventuelles didascalies qui y sont accolées. Il prend également en charge la normalisation des graphies d'un texte en fonction d'une table de remplacement spécifiée lors de l'exécution. Une interface Web a également été développée en PHP et JS permettant la vérification manuelle de l'auto-détection des sections, des orateurs, de la structuration hiérarchique et de la normalisation. Elle permet également la mise en évidence de formes non présentes dans un dictionnaire préalablement fourni au parseur.

5. Exploration textométrique

La normalisation du corpus thématique est une condition importante, déjà évoquée dans ce travail, pour une analyse fiable du texte en langue classique. Ce n'est que par l'homogénéisation des variantes orthographiques qu'on peut obtenir des résultats qui reflètent réellement le vocabulaire et les thèmes dans un corpus.

En effet, la lemmatisation et la normalisation des formes graphiques permettent une recherche des thématiques d'un corpus solide et fiable[12]. Toutefois il convient d'être prudent et de tenir compte des particularités génériques avant de lemmatiser, surtout du texte ancien. Par exemple, si on lemmatise la forme *plaisirs* en l'accordant avec *plaisir* dans un corpus d'opéra de la Renaissance français on oublie que les fontaines et les jardins, de Versailles notamment, sont peuplés de *nymphes*, de *naïades* et de *plaisirs* qui chantent la gloire du Roi Soleil. On ne peut pas réellement ici parler d'homonymie, non plus que pour les *amours* (au sens de Cupidon), il s'agit dans les deux cas d'une personnification mythologique, d'un sens dérivé, caractéristique du genre.

L'analyse de la concordance de la forme graphique *plaisirs* permet ainsi de constater que tantôt on a affaire aux *plaisirs qui volent* et *qui prennent leur place*, tantôt on jouit des *doux plaisirs.*

Toutefois, les réserves concernant la lemmatisation étant émises, il convient de souligner que c'est bien la normalisation qui assure des résultats fiables à l'étude thématique.

[12] La recherche thématique en textométrie est un calcul de spécificité particulier, où on recherche une relation privilégiée entre les mots eux-mêmes, ce que mesure aussi le calcul de corrélation.

5.1. Exploration thématique d'un corpus normalisé

La violence est omniprésente dans la tragédie classique comme dans la tragédie en musique. Nous avons par ailleurs pu montrer que le vocabulaire « tendre » diminue de manière générale pour laisser place à un vocabulaire qui fait référence à la violence[13].

Le sang coule en effet dans l'opéra baroque[14]. Le substantif sang compte à lui seul 504 occurrences. Sa distribution est révélatrice (coefficient de corrélation +0.634) et la tendance est claire, le public est attiré par l'élément de la violence physique et le sang coule de plus en plus dans la tragédie en musique à partir de 1712, qui marque l'évolution dans le corpus de l'élément violent et tragique, prédominant dans la troisième période de la première génération de la tragédie en musique.

D'autres items lexicaux qui font référence à la violence ont également une fréquence significative dans le corpus, par exemple : *périr, victime, frapper, dévorer, poignard, tomber, trembler, effroi, trépas, sacrifice* et *cruel*. L'histogramme ci-dessous illustre l'addition[15] de ces items avec *sang* et l'évolution chronologique de ces items relatifs à la violence :

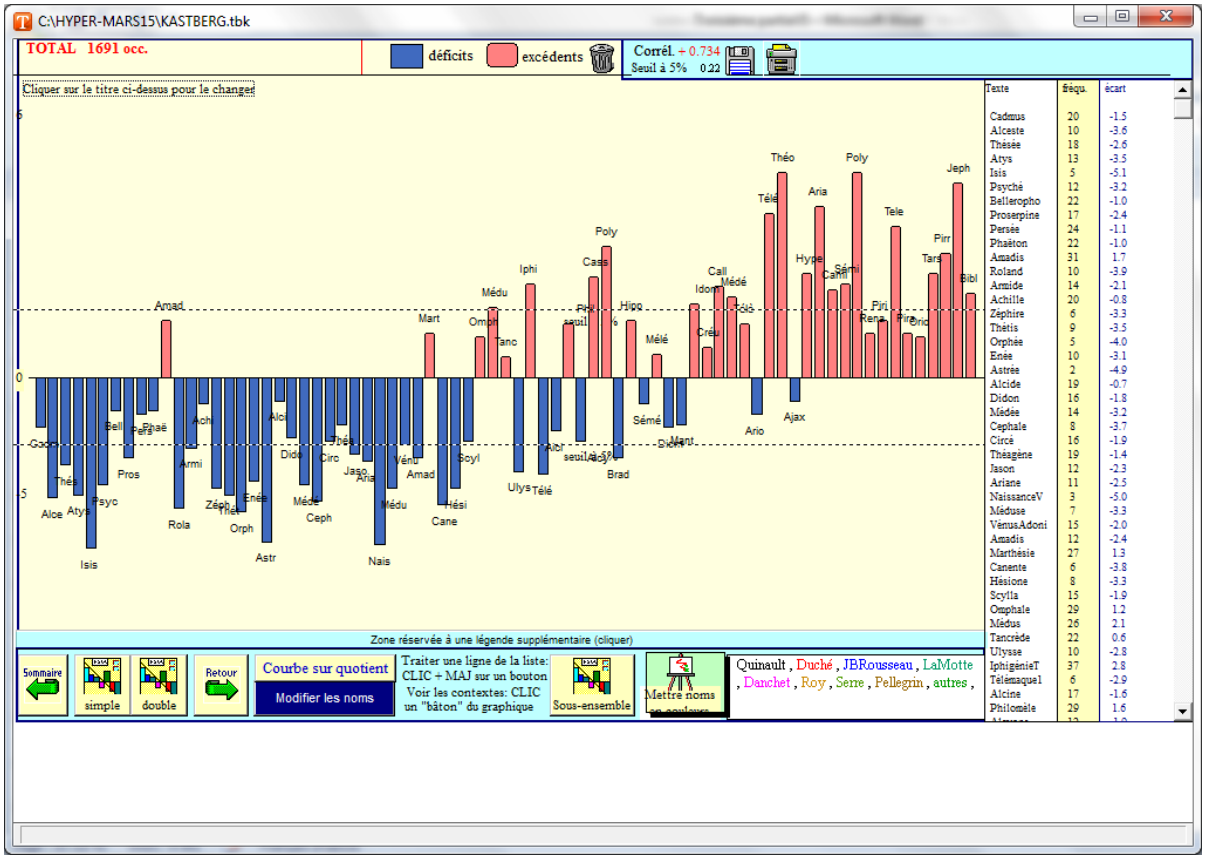

Figure n°1 : La distribution relative de l'ensemble des items : sang, périr, victime, frapper, dévorer, poignard, tomber, trembler, effroi, trépas, sacrifice *et* cruel

[13] *Cf.* M. Kastberg Sjöblom, *La langue de la tragédie en musique, France 1673-1732* (250 pages), Paris, Champion en cours de publication,.

[14] *Cf.* Bouissou Sylvie, *Crimes, cataclysmes et maléfices dans l'opéra baroque français,* Paris, Minerve, 2011.

[15] Il s'agit ici d'un traitement (calcul hypergéométrique, puis conversion en écart réduit) à partir d'un tableau constitué de ces items dans le logiciel Hyperbase. Un tableau de probabilités est calculé parallèlement et transformé en écarts. C'est sur ces écarts que se fondent ensuite les histogrammes de ce type.

La constatation est ici encore plus nette, le coefficient de corrélation de +0.734 confirme bien que la fréquence des items qui relèvent de la violence augmente de façon significative à partir du début du XVIII^{e} siècle dans un climat belliqueux à la Cour, plus précisément vers 1706 lorsque la guerre de succession d'Espagne bat son plein.

5.2. L'exploration du corpus après encodage structuré : l'exemple des didascalies

L'encodage au format TEI-*Drama,* adapté à la tragédie en musique pour l'étude des livrets, permet comme décrit auparavant d'extraire un sous-corpus selon une hiérarchie textuelle spécifique, ici liée au genre particulier.

Une particularité de la structure générique de l'opéra (et du théâtre) est constituée par les didascalies qui opèrent à un niveau bien distinct de celui des strophes et des vers chantés par les personnages de la pièce.

Ces didascalies posent une réelle difficulté dans le codage structuré d'un corpus composé de livrets d'opéra. En effet, les didascalies dans la tragédie en musique sont de différents types et les librettistes ne font pas tous le même usage de ces parties destinées à l'organisation scénique. On observe également que les indications scéniques deviennent de plus en plus détaillées au fur et à mesure que le XVIII^{e} siècle avance.

Nous avons identifié trois types de didascalies que nous avons distinguées par un codage différencié. Les trois types de didascalies sont présents tout au long du corpus, avec un fonctionnement bien distinct les uns des autres, s'éloignant du théâtre classique, mais en gardant des formes souvent archaïques, loin des didascalies d'opéra ou de théâtre de nos jours.

L'investissement important dans la normalisation et l'encodage du corpus donne une meilleure connaissance, une meilleure gestion, et une maniabilité accrue d'un corpus de taille relativement importante. La finalité ultime de ce processus relativement coûteux, pour cette étude en tout cas, est de pouvoir partitionner le texte mais aussi d'en extraire des sous-corpus pour une étude textométrique ciblée d'un niveau hiérarchique précis.

Dans l'exploration textométrique il s'avère très utile de pouvoir cibler l'analyse et créer des partitions selon les besoins scientifiques. Une fois les catégories *didascalies* (*stage*), les *textes périphériques*, *les prologues* et les *lignes introductives de personnages* écartés, on cible ici uniquement la partie textuelle occupée par la catégorie *paroles prononcées (chantées) par les personnages* (*act)*.

Au-delà de la normalisation, le codage TEI permet en effet une analyse nivelée qui tient compte de la structure particulière du genre pour affiner l'analyse textométrique. Nous espérons ainsi ouvrir une porte pour la même méthode dans d'autres domaines tels que le théâtre où la structure hiérarchique du texte est très semblable à celle de la tragédie lyrique. La même technique s'appliquerait également dans le domaine des sciences sociales où l'extraction de la parole individuelle d'un récit ou d'un dialogue, à l'intérieur d'une enquête « longue », excluant les questions de l'enquêteur, les titres etc., peut être très utile pour aisément effectuer une analyse sur un corpus constitué à partir d'un niveau hiérarchique choisi.

7. Conclusion

Avec ce travail nous avons voulu montrer jusqu'à quel point la préparation du corpus au préalable est déterminante pour son établissement et jusqu'à quel point, par la suite, l'établissement du corpus est essentiel pour l'exploration et l'analyse.

L'exploration et l'analyse de ce corpus par différents logiciels permet par la suite d'optimiser l'analyse textométrique et d'approfondir l'étude. En effet, on ne peut pas assez souligner l'importance dans ces cas de croiser les analyses, d'établir un multiple jeu de versions de corpus et de construire ce que nous appelons un « parcours textométrique » qui ainsi permet une interprétation nuancée et fine.

L'interprétation repose en large partie ici sur la connaissance du corpus mais aussi sur des résultats statistiques issus d'un corpus normalisé selon des critères génériques et historiques.

Etudier un corpus, normalisé ou non, exige en effet une connaissance profonde du corpus que l'on analyse, de son intertexte et du contexte temporel et social qui l'entoure pour bien mener l'analyse textométrique ; cette contextualisation l'inscrit très naturellement dans la démarche philologique et interprétative, vers le retour au texte original, et incite à lire le texte tel qu'il était lu à l'époque. C'est ainsi que le pas de la lexicométrie vers la textométrie prend tout son sens.

8. Références

Bouissou S. (2011), *Crimes, cataclysmes et maléfices dans l'opéra baroque en France*, Paris, Minerve, Musique ouverte.

Boulanger T. (2015), *XML par la pratique - Bases indispensables, Concepts et cas pratiques,* St Herblain, Editions ENI (3e édition).

Brunet E. (2009), « Enquête sur la langue poétique au XVIe siècle », *in Comptes d'auteurs, Tome 1, Etudes statistiques de Rabelais à Gracq,* Paris, Honoré Champion, p. 19-39. Article publié dans *Du Bellay, antiquités et nouveaux mondes dans les recueils romains,* Université de Nice Sophia-Antipolis.

Brunet E. (2015), *Hyperbase, Manual of reference, version 10.0.*, Université de Nice-Sophia Antipolis.

Kastberg Sjöblom M. (2015) *Le vocabulaire de la tragédie en musique, France 1673 – 1732*, Paris, Champion, en cours de publication.

Kintzler C., *Théâtre et opéra à l'âge classique,. Une familière étrangeté,* Paris, Fayard, collection Les chemins de la musique, 2004.

Lethier V.,Viprey J.-M., « Annotation linguistique de corpus : vers l'exhaustivité par la convivialité », *in* S. Heiden et B. Pincemin (éds.), *JADT'09, 9èmes Journées internationales d'Analyse statistique des Données Textuelles,* Lyon, Presses Universitaires de Lyon, 2008.

Néraudau J.-P. (1991), *La tragédie lyrique*, Versailles, Edition Cicéro, Théâtre des Champs Elysées.