Vers une analyse génétique de textes assistée par l'informatique et le TAL : contextes et pistes exploratoires

Claire Lemaire

Vers une analyse génétique de textes assistée par l'informatique et le TAL : contextes et pistes exploratoires

LEMAIRE
Claire

UFR SCIENCES DU LANGAGE

Mémoire de master 2 recherche – 30 crédits – Modélisation et traitements en industries de la langue : parole, écrit, apprentissage

Spécialité ou Parcours : TALEP

Sous la direction de : Thomas LEBARBÉ

Année universitaire 2009–2010

Remerciements

Tout d'abord, je souhaiterais remercier tous les enseignants de ce master qui pendant deux années universitaires ont fait preuve de patience et d'ouverture et m'ont permis de croire en cette reconversion professionnelle.

Ensuite je voudrais exprimer ma gratitude envers ceux qui parmi eux m'ont secondée dans les aspects pratiques (prêts d'ordinateur, de bureau, vacations, etc.) : Cécile Meynard, Virginie Zampa, Olivier Kraif et Thomas Lebarbé.

Je tiens aussi à remercier le groupe de recherche des Manuscrits de Stendhal pour avoir mis à ma disposition leur érudition et leur corpus stendhalien, et pour m'avoir accueillie dans leurs séminaires de recherche.

Merci également aux chercheuses en littérature qui ont essayé de m'initier à la génétique textuelle, et m'ont communiqué leur enthousiasme pour l'étude des processus de création d'œuvres littéraires : Cécile Meynard et Muriel Bassou.

À Lucy Garnier, Cécile Meynard, Muriel Bassou et Thomas Lebarbé pour les discussions scientifiques et passionnées sur la littérature ou la linguistique, exactement ce que je recherchais en quittant le monde de l'industrie et en frappant à la porte de la recherche universitaire.

Au DIP en général pour la bonne ambiance de travail ; à Roseline Magalhaes en particulier pour sa disponibilité, son efficacité et ses conseils maternels.

Un grand merci ensuite aux doctorants (du bureau I110 mais pas seulement) qui, en m'acceptant parmi eux et en m'invitant en dehors du cadre universitaire ont représenté pour moi un immense soutien pratique et psychologique : à Agnès, Aïcha, Aurélie, Auriane, Isabelle, Laurence, Lucy, Mathieu, Myriam et les autres.

Merci à mon fan-club extra-universitaire, qui a tout fait pour me libérer des soirées et des week-ends, m'a persuadée du bien-fondé de cette reconversion et m'a en revanche dissuadée les deux fois où j'ai failli abandonner le Master abattue par le manque de sommeil et les virus d'hiver.

Enfin, je remercie chaleureusement tous les relecteurs qui ont pris du temps pour lire ce mémoire et me faire part de leurs remarques pertinentes et constructives.

Avertissement

Le sujet de ce mémoire est interdisciplinaire. La réflexion à son origine est menée par la volonté de réussir un travail collaboratif entre des équipes qui ne connaissent pas forcément la discipline de leur voisin. Contrairement à ce que nous avons fourni au cours des trois autres semestres de ce master, nous ne trouverons pas ici de recherches techniques, de démonstrations ni d'études quantitatives à renfort de statistiques, mais plutôt l'exploration d'un nouveau champ de recherches commun dans lequel construire un dialogue entre la littérature, la linguistique et l'informatique.

Par ailleurs, ou par conséquent, au fur et à mesure des lectures successives de ce mémoire par des spécialistes d'un domaine ou d'un autre, nous avons développé des explications ou ajouté des notes qui pourront sembler naïves à certains lecteurs mais qui seront fort utiles pour d'autres.

Dédicace

Aux doctorantes, docteur, visiteuses et visiteurs du I110.

Sommaire

Introduction

Au commencement était le projet *Manuscrits de Stendhal*. Il consistait et consiste toujours à transcrire dans un format exploitable textuellement 20 000 feuillets conservés à la Bibliothèque Municipale de Grenoble. Le but de cette transcription est de permettre aux chercheurs en littérature et en linguistique d'utiliser les outils informatiques pour parcourir les manuscrits et travailler sur un contenu textuel plutôt que sur des photos d'écriture. Un outil de recherche par mots-clés développé récemment par l'équipe du projet permet déjà d'effectuer des requêtes simples à travers 1 200 feuillets. Ces feuillets ne sont pas tous rédigés proprement ni tous ordonnés, comme ils le seraient dans le cadre de manuscrits aboutis, mais revêtent plutôt l'aspect de brouillons : entre autres problèmes, nous pouvons nous trouver face à plusieurs versions d'un même texte, éparpillées dans différents registres du fonds. Il arrive aussi que sur une même page, cohabitent des croquis, des réflexions littéraires, des listes de livres à lire, des traductions, des répliques à insérer dans ses projets de pièces de théâtre, des notes diaristes, etc.

L'analyse génétique de textes interprète ces manuscrits de travail dans le but de découvrir les influences, les méthodes de travail ou les pratiques scripturales d'un auteur. L'étude relatée dans ce mémoire détermine dans quelle mesure l'informatique et le traitement automatique des langues (TAL) pourraient faciliter cette analyse.

Dans une première partie, nous tenterons de caractériser le contexte littéraire, ce qu'est la génétique de textes, la façon dont travaillent les généticiens et les institutions sous lesquelles ils se sont regroupés. Dans une seconde partie, consacrée au domaine linguistique et informatique, nous chercherons comment les modèles linguistiques peuvent assister la génétique et nous présenterons des outils informatiques conçus pour l'étude de manuscrits. Enfin, dans une troisième et dernière partie nous proposerons une perspective de recherche comportant trois niveaux de granularité différents. À un niveau macrogénétique, une recherche sur les bibliothèques d'auteurs, à un niveau génétique plus fin, une étude sur les pièces de théâtre vues ou lues par Stendhal et à un niveau microgénétique, une analyse sur le changement de langue au sein d'un même texte.

Partie 1

Essai de caractérisation de la génétique littéraire

« La longue marche des chercheurs à travers les manuscrits les a menés,
sans qu'ils s'en rendent toujours compte, de la description d'un objet, le manuscrit,
à la compréhension d'un processus, celui de l'écriture. »

Louis Hay

Chapitre 1 – L'analyse génétique de textes

1.1 Les définitions

1.1.1 L'analyse et la critique

Zola peignait son siècle, le nez sur les tableaux de Delacroix, Renoir ou Monet, Proust transformait les biscottes en madeleines et Stendhal ajoutait à ses propres écrits des annotations tellement volumineuses qu'elles pouvaient par la suite donner naissance à de nouveaux livres.[1,2] Qu'est-ce que l'analyse génétique ? Interrogée sur le sujet par un matin de février, voici ce que Cécile Meynard, maître de conférences en génétique et édition savante de manuscrits du XIXe siècle, répondit :

> « C'est le fait d'étudier comment se construit l'œuvre, l'embryon d'œuvre, quelles sont les différentes étapes de remplissage de la page, et éventuellement tout ce travail de correction, de rature, de remords, de rajout, de suppression, etc. C'est compliqué car chaque auteur a son processus de genèse propre. Dans les romans de Stendhal par exemple c'est surtout un processus d'écriture par amplification. »
>
> Cécile Meynard

Nous apprenons donc que la génétique est la découverte de la façon dont a été écrite une œuvre littéraire à travers l'observation de ses brouillons, et par la même occasion qu'il existe une génétique différente par auteur. Nous risquons donc de devoir faire face très rapidement à plusieurs types de systèmes.

Continuons avec Pierre-Marc de Biasi, directeur de recherche au CNRS : la génétique est « l'interprétation de l'œuvre à la lumière de ses brouillons ou de ses documents préparatoires »[3].

Voilà un deuxième aspect de l'analyse génétique : d'une part nous sommes face à un travail objectif consistant à répertorier de façon factuelle les éléments physiques qui ont été nécessaires à l'écrivain pour rédiger son œuvre, et d'autre part à un travail

[1] [Doubrovsky, 2007] citant [Lejeune, 1971]
[2] [Meynard, à paraître]
[3] [de Biasi, 2000]

subjectif d'interprétation littéraire de l'œuvre en fonction de ce que l'étude des manuscrits aura permis de découvrir.

Passons à Philippe Lejeune, maître de conférences honoraire à l'Université de Paris-Nord, arrivé en génétique via l'autobiographie :

> « Elle [La génétique] va faire l'histoire de la production. Comme toute science historique, elle aura ses exigences et ses méthodes : l'établissement de toutes les traces laissées par le processus de production, leur description méticuleuse, leur chronologisation. C'est un travail si énorme qu'on peut s'y perdre, et perdre de vue le but, qui est de comprendre. »
>
> [Lejeune, 2007]

Nous retrouvons dans sa définition la dimension diachronique[4] de l'étude citée plus haut. Pour lui le but ultime réside dans la compréhension de l'œuvre. Or « comprendre » ne signifie-t-il pas « interpréter » cette œuvre avec une grille d'analyse de plus en plus fine au fur et à mesure de l'accumulation de connaissances sur le processus de création ?

Ce qu'il y a de commun à ces trois définitions reste la découverte des « secrets de fabrication » de l'œuvre par l'étude des traces de sa création. Le sujet, même réduit à sa dimension objective et factuelle d'étude de traces physiques de la création d'une œuvre, reste très vaste.

1.1.2 La génétique pour notre problématique

Nous laisserons donc aux chercheurs en littérature la critique en elle-même et nous nous intéresserons à l'analyse des supports d'écriture et à l'écriture elle-même en tant que signifiant et non signifié. Comment classer ces milliers de signes (mots, symboles ou dessins), quels systèmes trouver pour les organiser efficacement ?

Sous cet angle, l'analyse génétique ressemble à une recherche fastidieuse et minutieuse de la vérité. Pour Louis Hay, « à la différence de tant d'autres, la génétique n'est pas

[4] En linguistique, l'approche diachronique s'intéresse à l'évolution de la langue à travers l'histoire par opposition à l'approche synchronique qui s'intéresse à la langue à un temps T.

fille d'une théorie. Elle est née d'une expérience empirique et garde toujours dans sa méthode les traits d'une pratique ».[5] Notre but est de comprendre la génétique et les généticiens, et de proposer de rapprocher leurs méthodes de travail des possibilités qu'offrent l'informatique et le TAL. Alors à quoi ressemblent ces brouillons, quand sont-ils apparus, sous quelle forme ? Regardons rapidement l'histoire de notre objet d'étude à travers les âges.

1.2 L'historique

Cet aperçu de l'évolution du support de l'écrivain nous aidera également à restreindre la période concernée par notre problématique.

1.2.1 Du parchemin au manuscrit moderne

La première époque du manuscrit littéraire en Occident est celle du scribe qui écrivait sur papyrus ou parchemin. Les écrits sont alors recopiés, chaque copie étant unique et légèrement différente des autres. Jusqu'à l'époque médiévale, il sera impossible de retrouver l'origine précise d'une œuvre[6]. En ce temps, le parchemin coûteux était utilisé plusieurs fois ; les caractères de l'ancien texte étaient grattés et un nouveau texte copié. Avec l'arrivée du papier fabriqué en Italie dès 1250 un nouvel usage apparaît : celui de la rature, premier indice de la genèse d'un texte. Or à cette époque deux copies d'un manuscrit, réalisées par des scribes (religieux ou laïques), pouvaient différer sans qu'il soit possible de retrouver l'original exact. C'est avec Gutenberg en 1455 que les premiers livres imprimés font leur apparition (des bibles au nombre de cent quatre-vingt)[7]. Progressivement les écrivains ont abandonné les parchemins pour faire écrire les scribes ou les copistes sur des feuilles de papier. Puis les copistes ont peu à peu disparu et les écrivains se sont mis à écrire eux-mêmes. Cependant, les manuscrits utilisés pour la création sont détruits, le talent étant plus valorisé que le travail, il ne fallait pas montrer ses coups d'essai[8].

[5] [Hay, 2007]
[6] [Cerquiglini, 1989]
[7] [Martin, 2010]
[8] [de Biasi, 2000]

Le regard des écrivains sur leurs manuscrits change. Les auteurs prennent conscience de la valeur de leur travail artisanal, original, manuel par opposition aux objets manufacturés qui apparaissent de plus en plus suite à la révolution industrielle[9]. Les manuscrits commencent à être protégés et classés selon une stratégie semblant indiquer l'intention d'offrir à la postérité des objets de recherche. Ainsi, Victor Hugo conserve avec soin ses manuscrits et ne donne aux imprimeurs qu'une copie allographe[10] du texte définitif. Il couche sur son testament le don de tous ses manuscrits à la Bibliothèque de Paris[11]. C'est suite à ce don que fut créé le *Département des manuscrits modernes* de la Bibliothèque de Paris et par la même occasion que naît ce qui est appelé le « manuscrit moderne », objet d'étude auquel se consacre aujourd'hui la génétique de textes[12].

1.2.3 Ces traces qui nous intéressent

Reprenons notre histoire de l'écriture en nous focalisant sur les preuves physiques qu'elle a laissées derrière elle.

Les premières traces de brouillon que nous ayons retrouvées sont des tablettes de cire. L'auteur n'écrivait pas lui-même mais dictait un texte à un scribe, se le faisait relire, puis dictait des corrections. Une version définitive était recopiée sur du parchemin, toujours par le scribe. Enfin, les tablettes de cire étaient lissées pour servir à nouveau.

Nous n'avons par conséquent aucune trace de la genèse de ces textes.

Avec l'ère du papier, comme nous l'avons vu, les pratiques changent. En effet, le travail étant considéré par la Bible comme une malédiction infligée à l'homme par Dieu, il fallait surtout se débarrasser de ces traces.[13]

Les manuscrits sont donc détruits avec soin. Ceux de Stendhal, né en 1783 et mort en 1842, étaient détruits après leur publication. En revanche, il reprenait ses œuvres après leur parution et ajoutait des notes en marge de ses propres textes (appelées *marginalia*) en fonction des critiques qu'il avait lues ou des idées qui lui étaient venues.[14] Ainsi nous savons qu'il sera impossible de retrouver parmi les 20 000 feuillets les manuscrits

[9] [de Biasi, 2000]
[10] Écrit par une autre main que celle de l'auteur
[11] Aujourd'hui Bibliothèque Nationale de France
[12] [de Biasi, 2000]
[13] [de Biasi, 2000]
[14] [Meynard, à paraître]

originaux de *Le Rouge et le Noir* ou de *La Chartreuse de Parme*, car ils ont été détruits après la parution des romans.

D'autres études génétiques s'intéressent également aux manuscrits beaucoup plus récents, tapés à la machine (parfois appelés *tapuscrits*) ou rédigés avec des logiciels de traitement de textes. Par exemple il existe aujourd'hui des logiciels recueillant les hésitations des auteurs en enregistrant toutes les corrections effectuées sur un logiciel de traitement de textes lors de la rédaction d'un document.

Nous nous contenterons pour notre étude, des manuscrits papiers, ce qui correspond donc à une période de deux ou trois siècles.

Chapitre 2 – Le métier de généticien

2.1 Les documents utilisés en génétique et leur terminologie

2.1.1 Le manuscrit définitif

Le manuscrit qui nous vient naturellement à l'esprit lorsque nous ne sommes pas spécialistes en génétique est le manuscrit définitif, recopié par l'auteur en fin de rédaction pour fournir un document de référence à l'imprimeur[15].

2.1.2 Le dossier génétique

Aujourd'hui, le plus intéressant pour le chercheur en génétique textuelle est l'ensemble des éléments de création d'une œuvre, le plan, le scénario, les personnages décrits sur des cahiers, des carnets, des livres, etc. C'est pourquoi, en tout premier lieu, le généticien va recueillir toutes ces informations et constituer avec un *dossier génétique*. Le dossier génétique est « l'ensemble matériel des documents et manuscrits se rapportant à la genèse que l'on entend étudier »[16]. C'est donc le fruit d'un travail préliminaire long et minutieux. Si le dossier comprend les documents de création de l'œuvre, il peut également contenir des informations extérieures à la création de l'œuvre mais précieuses pour l'analyse.

2.1.3 Le livre annoté et les bibliothèques d'auteurs

Ainsi la bibliothèque personnelle de l'écrivain intéresse au plus haut point le chercheur en génétique, non seulement car elle aide à comprendre les influences générales de l'auteur mais aussi parce que les livres sont bien souvent annotés avec des allusions explicites à la future œuvre en cours de création. Par exemple sur le site des manuscrits de Flaubert[17] il est possible de feuilleter virtuellement les pages d'un livre écrit par le journaliste socialiste Auguste-Jean-Marie Vermorel sur les événements de

[15] Ce sont justement ces manuscrits définitifs qui ont été détruits dans le cas de *Le Rouge et le Noir* et de *La chartreuse de Parme*.
[16] [de Biasi, 2000]
[17] http://flaubert.univ-rouen.fr/manuscrits/

1848 et y trouver les annotations de Flaubert.[18] Ce livre entre en l'occurrence dans le dossier génétique de *Bouvard et Pécuchet*. De plus en plus de centres d'archives recensent un à un les volumes de ces bibliothèques d'auteurs et relèvent leurs traces d'écriture pour donner aux généticiens des indications sur leur mode de travail. Un autre exemple est celui du Centre Dürrenmatt à Neuchâtel qui, en collaboration avec les Archives littéraires suisses, catalogue actuellement la bibliothèque de philosophie et de sciences naturelles de Friedrich Dürrenmatt[19]. Un autre exemple encore est celui de la bibliothèque Sormani à Milan sur les livres annotés de Stendhal[20].

2.1.4 L'avant-texte

Une fois que les documents de genèse sont réunis dans un dossier, un travail de recensement peut alors commencer au cours duquel chaque pièce du dossier génétique est déchiffrée, datée et classée. Un *avant-texte* est le résultat de ce travail : il correspond à un dossier de genèse dont chacun des éléments a été non seulement déchiffré et daté mais « organisé » à l'intérieur d'un regroupement logique, souvent par ordre d'apparition chronologique mais pas systématiquement[21]. Pour Lebrave et Grésillon, l'avant-texte est le dossier de genèse rendu « utilisable » pour une interprétation, il désigne « l'ensemble des documents écrits qui portent témoignage de l'élaboration progressive du texte »[22].

2.1.5 Le brouillon

Dans cet avant-texte, certains documents représentent la réflexion de l'auteur, d'autres sont le témoignage des premières productions textuelles, ce sont les *brouillons.* Les brouillons correspondent à l'ensemble de documents relatifs au travail rédactionnel, au travail de textualisation, c'est-à-dire de mise en phrases. Un brouillon comporte certes des ratures ou des ajouts spectaculaires mais les phrases sont généralement déjà syntaxiquement correctes. Ce sont ces documents qui sont au cœur de l'analyse

[18] Vermorel (Auguste-Jean-Marie), *Les Hommes de 1848*, 2e éd., Paris, Décembre-Alonnier, 1869.
[19] http://www.bundesmuseen.ch/cdn/00126/00167/index.html?lang=fr
[20] http://www.digitami.it/stendhal/ricerca/postille.php
[21] [de Biasi, 2000]
[22] [Lebrave & Grésillon, 2009]

génétique, et ce sont justement ces traces de formulation, ratures, reformulation, ajouts, etc. qui permettront au généticien de comprendre comment l'œuvre est née.

2.2 La méthode d'analyse codicologique

Dans le métier de généticien, une grande place est faite à la *codicologie*. Il s'agit de l'étude matérielle des manuscrits en tant qu'objets physiques par l'étude des matériaux servant à leur confection et leur utilisation.

En codicologie, nous commençons par inventorier les supports qui composent les différents volumes ou liasses. Les ensembles, reliés ou non, peuvent être constitués de différents matériaux : cahiers, carnets ou feuilles volantes, formats in-folio[23], in-quarto[24], etc., qualité des papiers, modes d'assemblage. Nous distinguerons également les documents autographes des interventions de copistes et nous relèverons d'éventuelles marques d'origine privée ou institutionnelle concernant le classement du corpus : foliotation, pagination, annotations marginales, feuillets intercalaires, tampons de bibliothèques, numérotation de liasses par des bibliothécaires. Puis nous repérons les différents types de papier utilisés, chaque feuillet étant mesuré et décrit dans le détail, y compris, le cas échéant, son filigrane[25], afin de regrouper progressivement sur une même fiche toutes les occurrences d'un même type de papier ; et les traces de coutures également peuvent apporter de précieuses informations. Ensuite les écritures en tant qu'éléments physiques font l'objet d'une étude à leur tour, couleur, nature des instruments d'écriture, caractéristiques de la disposition spatiale. Enfin, le chercheur recourt aux diverses sources qui lui permettent de mieux situer les phases de rédaction dans l'espace et le temps : des sources biographiques, des ouvrages sur l'histoire du papier, des listes de filigranes etc.

[23] Forme de livre où la feuille a été pliée une fois, donnant ainsi deux feuillets soit quatre pages.
[24] La feuille a été pliée deux fois, soit huit pages.
[25] Empreinte d'identification d'un papier obtenue en plaçant une forme en cuivre ou en laiton formant un dessin ou une inscription, fixée sur la forme destinée à recevoir la pâte à papier.

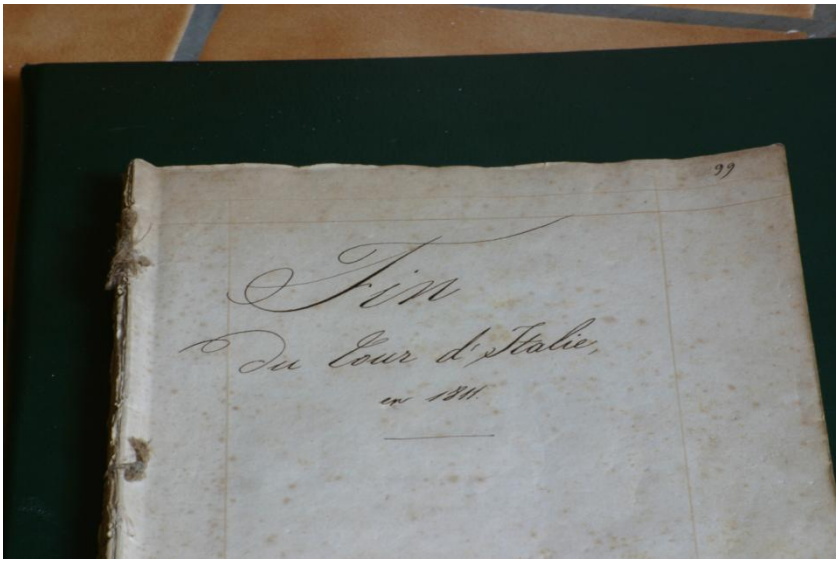


Illustration 1 – Exemple de manuscrit de Stendhal : copie Arbelet « Fin du Tour d'Italie », cahier n°50, propriété de M. Arbelet

La codicologie, par son analyse physique des manuscrits, permet bien souvent de compléter efficacement le dossier génétique ou de résoudre des énigmes. Par exemple[26], les éditeurs ont toujours considéré que les deux documents de Stendhal « *For the two men* » (réflexion pour son projet de pièce *Les Deux hommes*) et *Style* étaient totalement indépendants. Ils les ont publiés l'un avec le théâtre de Stendhal, et l'autre avec ses réflexions littéraires, bien qu'il y ait une continuité par le contenu, *Style* étant une réflexion sur le style en lien avec le projet de pièce de théâtre des *Deux hommes*. L'analyse codicologique confirme l'unité de l'ensemble : ces deux documents ont été écrits par Stendhal dans un même cahier, et en observant les feuillets, nous remarquons qu'il y a sur un document une remarque générale de la même encre que le deuxième document, ces deux documents sont donc liés matériellement. Ce détail est déterminant pour la politique éditoriale. En effet la prochaine édition *Journaux et Papiers*[27] de Stendhal pourrait donc également comprendre le document *For the two men*.

Un autre exemple d'apport de la codicologie est celui du projet de roman *Lamiel*. Serge Linkès, à partir de l'analyse codicologique et génétique a établi une nouvelle chronologie de ce roman en démontrant que ce qui passait pour la dernière version était

[26] Exemple donnée par Cécile Meynard, en mai 2010
[27] Titre provisoire

en fait l'avant-dernière, et ce grâce à l'identification du copiste et à l'analyse du papier. Stendhal ne pouvait être en possession d'un certain type de papier en Italie, donc il aurait écrit la *vraie* dernière version à Paris. Le tome 3 des *Œuvres romanesques* de Stendhal dans la collection Pléiade, à paraître en 2010, fournira ainsi une édition tout à fait différente des éditions actuelles.

2.3 La méthode d'analyse du processus de genèse

Quelles que soient les méthodes utilisées[28], elles ont en commun de commencer par découper le processus de genèse en quatre phases : les phases prérédactionnelle, rédactionnelle, pré-éditoriale et éditoriale.

2.3.1 La phase prérédactionnelle

La phase prérédactionnelle est le premier pas vers la création de l'œuvre. C'est au cours de ce travail de réflexion préliminaire que l'auteur utilise sa bibliothèque et constitue des dossiers de notes. Le généticien doit prendre en considération tous les documents accumulés pour cette occasion. À ce stade un contrat d'édition par exemple peut jouer un rôle structurant ; c'est donc une trace à ne pas négliger. Outre les livres et les contrats, nous prenons en compte les plans, les scénarios, les bribes de rédaction, nous séparons bien les éléments autographes des éléments allographes, et nous ajoutons les notes de projets, d'idées et de voyages.

2.3.2 La phase rédactionnelle

La phase rédactionnelle est la phase représentant le cœur du travail de genèse. La collection de pièces correspondant à cette deuxième phase, la phase dite *rédactionnelle*, est constituée essentiellement des brouillons de l'œuvre. Dans cette phase d'exécution du projet a lieu la textualisation : les documents rassemblés précédemment sont consultés et se met en place le processus intellectuel de mise en phrases syntaxiquement correctes. À un niveau structurel, les plans et scénarios de la phase prérédactionnelle sont alors développés, des chronologies, des généalogies et des argumentaires sont couchés sur le papier, seul support utilisé.

[28] [Hay & Naguy, 1982], [de Biasi, 2000] et [Derrida *et al.*, 1995]

2.3.3 La phase pré-éditoriale

Dans cette troisième étape, l'auteur part d'un texte déjà rédigé. Il le relit, l'améliore encore et encore, et parfois des changements spectaculaires ont lieu au cours de cette période. Cependant se construit « petit à petit un système de contraintes qui fait entrer l'avant-texte dans un processus indiscutablement finalisant »[29]. Le texte est recopié, c'est une *mise au net*. La perspective d'une édition est présente à l'esprit de l'auteur à ce moment du processus d'écriture. Les traces produites par ce travail sont un *manuscrit prédéfinitif*, puis un *manuscrit définitif*, dernier état autographe de l'avant-texte. Le manuscrit définitif également appelé *Manuscrit* avec un M majuscule revêt une valeur symbolique particulière. C'est celui qui représente l'œuvre et qui est de plus en plus conservé à partir du milieu du XIXe siècle[30]. Le Manuscrit est alors recopié par un copiste, ce *manuscrit du copiste*, servira de base aux travaux d'impression. Stendhal est un cas à part dans ce processus génétique puisqu'il lui arrive de dicter directement son texte à un copiste : le manuscrit allographe est alors le premier dans le dossier génétique, et Stendhal n'intervient que dans un second temps[31]. Le copiste peut introduire des fautes qui seront vues et corrigées ou non par l'auteur. Avec ou sans fautes, ce sera le manuscrit qui servira de référence et dans le cas de la mort de l'auteur plus aucune modification ne sera réalisée, les éventuelles fautes étant rééditées à chaque fois. Ce manuscrit est alors envoyé à l'imprimeur qui renvoie en retour les *épreuves corrigées* qui seront relues par l'auteur. L'auteur a encore la possibilité de reprendre le texte mais dans le cadre d'un accord avec l'imprimeur : Flaubert ne changeait plus rien alors que Balzac effectuait un travail énorme d'amplification prévu par l'imprimeur qui lui laissait des marges de part et d'autre des pages de ces épreuves corrigées. Enfin, à partir du moment où l'auteur écrit de sa main « Bon à tirer » sur l'épreuve corrigée celle-ci devient le *bon à tirer* et ne peut plus subir de modification. La genèse est terminée.

[29] [de Biasi, 2000]
[30] Stendhal (1783-1842) nous l'avons vu, n'adhéra pas à cette pratique nouvelle mais fit au contraire détruire des Manuscrits ou les détruisit lui-même
[31] Voir par exemple la deuxième partie d'*Histoire d'Espagne*, ou les pages de *Féder*

2.3.4 La phase éditoriale

Si cette phase nous intéresse moins puisque le processus génétique est terminé, l'œuvre peut subir encore de nombreuses modifications. Dans cette phase éditoriale, nous trouvons par exemple des suppressions pour cause de censure ou au contraire des développements de textes demandés par l'éditeur pour adapter l'œuvre à un autre format : roman, recueil poétique, etc. D'ailleurs, à ce sujet, Stendhal est intéressant puisqu'il retravaille ses œuvres publiées en vue de rééditions : *Rome, Naples et Florence en 1817, Promenades dans Rome* (1829).

2.3.5 Une synthèse et une simplification

Dans le but de déboucher sur un outil, l'illustration suivante présente une synthèse des éléments à prendre en compte dans notre étude. Il s'agit d'une simplification du tableau de Pierre-Marc de Biasi[32, 33]. Pour notre problématique, ce tableau nous servira de fil conducteur.

Phase	**Documents de genèse**
Prérédactionnelle	Bibliothèques d'auteurs Carnets Cahiers Dessins Feuillets
Rédactionnelle	Feuillets
Pré-éditoriale	Manuscrit prédéfinitif Manuscrit définitif
Éditoriale	Bon à tirer (ou épreuve corrigée)

Illustration 2 – Tableau des documents de genèse pour notre problématique

[32] [de Biasi, 2000]

[33] Les spécialistes en génétique proposent en effet d'affiner le regroupement de ces documents en termes plus précis que ce que nous proposons. Par exemple une distinction peut être faite entre l'esquisse, le croquis et le dessin.

2.4 La méthode d'analyse des pièces

2.4.1 La datation des différents jets

Pour aborder les aspects méthodiques, nous avons interviewé deux chercheuses en littérature[34]. Il existe différentes méthodes et il semble difficile de n'en avoir qu'une seule, normalisée, pour tout le monde. La raison principale est qu'il existe des processus de genèse différents selon les auteurs, comme nous l'avons vu. Cependant des aspects communs peuvent être dégagés.

Le premier aspect commun à toutes les méthodes est le fait de partir des manuscrits et de repérer le premier jet. Le second aspect est l'identification de la deuxième campagne d'écriture, par exemple les annotations de Stendhal qui sont ajoutées. C'est là que se complique l'affaire, car il peut y avoir un troisième jet et un quatrième jet, ou d'autres encore. Quelques indices peuvent cependant nous aider, par exemple il peut y avoir des couleurs d'encres différentes. Deux façons de procéder semblent alors exister suivant les chercheurs, suivant les auteurs, voire suivant les dossiers génétiques : soit par ordre chronologique, en remontant de la dernière écriture à la première, soit par ordre anti-chronologique en essayant au contraire de partir du plus loin, du tout premier jet et en avançant vers l'état final de la page. Par ailleurs, gardons bien à l'esprit qu'il s'agit de provisoire. Par exemple, les manuscrits de Stendhal que nous possédons, nous restent parce qu'ils n'ont pas été publiés. Quand cela a été le cas, nous l'avons vu, Stendhal les a fait disparaître. Il semble donc qu'un outil d'aide devrait absolument contenir le critère de datation ou du moins une identification des étapes, même imprécise lorsqu'il n'y pas de date, ainsi qu'une fonctionnalité permettant un enchaînement des pièces du dossier à la fois chronologique qu'« anti-chronologique ».

Cette interview des deux chercheurs en littérature nous a permis d'aborder le quotidien du généticien lorsqu'il étudie les éléments de son dossier. Revenons maintenant en amont au rassemblement de ces pièces, lesquelles choisir, comment les choisir et comment les organiser en systèmes à la granularité peut-être moins fine que la datation de chaque document.

[34] Cécile MEYNARD et Muriel BASSOU. Interview du 25 février 2010, retranscrite partiellement en Annexe 1.

2.4.2 La constitution du corpus

La première démarche consiste à créer un corpus[35], c'est-à-dire à collecter le maximum de pièces que l'écrivain a utilisées pour concevoir et rédiger son texte et qui ont échappé à la destruction. Elles sont peut-être dispersées dans plusieurs institutions, plusieurs collections privées, plusieurs pays. Après avoir réuni le plus de documents possible, en général des reproductions numériques, le généticien doit dater chacune des pièces, déchiffrer les écritures et authentifier le ou les scripteurs : s'agit-il de l'auteur, d'un ami écrivain avec qui il écrivait à quatre mains, d'un copiste à qui il dictait ? Chaque élément fait alors l'objet d'un premier classement sommaire, habituellement en suivant les quatre phases citées précédemment[36].

2.4.3 Le déchiffrement

Inclus dans le classement des brouillons, se trouve le déchiffrement de chaque page d'écriture. En effet, dès qu'une pièce est déchiffrée, elle peut être alors rapprochée d'autres pièces. Parallèlement, ce rapprochement permet de reconnaître des mots illisibles qui ne l'avaient pas été lors d'un premier passage mais que nous pouvons deviner, en observant la version postérieure sur le brouillon suivant.

2.4.4 La transcription

Le troisième volet de ce traitement des brouillons, également simultané aux deux autres démarches, est la transcription. Elle met à disposition de la communauté littéraire le support de l'analyse génétique et chaque critique pourra se reporter à cette version transcrite pour effectuer ses travaux de recherche sans refaire le travail de déchiffrement ni de classement. Il existe plusieurs types de transcription plus ou moins fidèles à l'original, aucune ne pouvant l'être intégralement. Notons simplement les plus utilisées. La transcription *linéarisée codée* consiste à ajouter à la suite les différentes versions des morceaux de phrase qui ont été biffés. La transcription *semi-diplomatique codée* consiste à respecter la mise en page et à ajouter des codes conventionnés précisant s'il

[35] Au sens littéraire du terme. En linguistique, nous parlerions plutôt de « collection de textes »
[36] [de Biasi, 2000]

s'agit d'un ajout en interligne supérieure ou inférieure, etc. La transcription *diplomatique* respecte la mise en page et est donc très coûteuse en place. La transcription *diachronique linéarisée* n'est pas codée et décrit explicitement ce qui appartient à la première, deuxième ou troisième campagne d'écriture.

Le métier de généticien apparaît donc comme une activité méticuleuse consistant surtout en un long travail de préparation de l'objet à étudier. L'analyse critique de l'œuvre sous la lumière de sa genèse n'intervient en effet qu'après une multitude de tâches telles que la collection de pièces, l'analyse des supports papiers, le déchiffrage de l'écriture, la transcription en un format exploitable et enfin l'identification des campagnes d'écriture.

Chapitre 3 – Les institutions de la génétique

Nous ne pourrions pas terminer cette partie sur la génétique littéraire sans citer ses deux instances françaises, l'une étant la fille de l'autre, capitales dans le paysage des généticiens : le Centre d'analyse des manuscrits (CAM) et l'Institut des textes et manuscrits modernes (ITEM).

3.1 Le Centre d'analyse des manuscrits

Le CAM a été fondé par Louis Hay en 1974. Le centre regroupe alors des chercheurs du Centre national de la recherche scientifique (CNRS) d'origine très diverse : linguistique, sociocritique, psychanalyse, etc. Les spécialistes travaillaient alors à des projets de corpus d'écrivains : Aragon, Flaubert, Heine, Joyce, Nerval, Proust, Sartre, Zola, etc. Cette interdisciplinarité a vite porté ses fruits, les écrivains eux-mêmes ont apprécié ces recherches. Aragon par exemple a même légué ses manuscrits au CAM en 1976.

3.2 L'Institut des textes et manuscrits modernes

En 1982, le CAM a été transformé en laboratoire du CNRS à part entière et associé à la Bibliothèque nationale de France (BNF) et à l'École Normale Supérieure (ENS) et rebaptisé à cette occasion *Institut des textes et manuscrits modernes*. De nouveaux pôles de recherche ont alors été créés : codicologie, supports et tracés, édition génétique, hypertexte et multimédia, histoire des écritures, théories de la genèse textuelle, genèse et sciences cognitives et archives de la création. Aujourd'hui l'ITEM est une Unité mixte de recherche entre le CNRS et l'ENS de la rue d'Ulm (UMR 8132) répartie en trois sites parisiens, deux dans le 5^{e} arrondissement et un dans le 17^{e}, et compte 22 équipes de littéraires et de linguistes[37]. Deux équipes notamment travaillent sur des sujets proches de notre problématique : l'équipe « Manuscrit - Linguistique - Cognition » et l'équipe « Génétique et théories linguistiques ». Elles visent en effet à élargir le champ de la génétique en replaçant l'écriture littéraire dans le contexte global

[37] Voir la liste exhaustive en Annexe 1

de la production verbale écrite dans son ensemble[38]. De nombreux chercheurs sont des membres associés, par exemple des professeurs, des maîtres de conférences, des ingénieurs de recherches, etc.[39] L'ITEM édite trois périodiques : *Textes & manuscrits* : collection de critique génétique créée par Louis Hay ; *Genesis*, revue internationale de critique génétique et le *Bulletin d'informations proustiennes (BIP).*

[38] http://www.item.ens.fr/index.php?identifier=equipes
[39] 108 membres sur 361. Pour le détail, voir le site http://www.item.ens.fr/index.php?identifier=annuaire dernière vérification le 13 mai 2010.

Conclusion

Pour clore cet essai de caractérisation de l'analyse génétique, notons que les supports qui nous intéressent sont les manuscrits provisoires, ceux qui témoignent de la fabrication de l'œuvre, et qui appartiennent aux XVIIe, XVIIIe et XIXe siècles. Nous avons compris que chaque auteur avait sa propre genèse, et en observant les méthodes des généticiens, nous avons découvert la codicologie, les quatre phases de création d'une œuvre et le rassemblement des pièces de corpus, leur déchiffrement, leur transcription, leur classement et la datation des différents jets. Enfin nous avons appris l'existence de l'ITEM qui regroupe une majorité de généticiens en France.

Nous allons maintenant regarder les modèles linguistiques pour éventuellement trouver des aspects méthodiques intéressants pour la génétique, puis nous présenterons des outils informatiques qui pourraient nous inspirer.

Partie 2

Les propositions linguistiques et informatiques

« La linguistique génétique *a encore du pain sur la planche et la recherche sur la production verbale écrite ne fait que commencer… »*

Jean-Louis Lebrave
et Almuth Grésillon

Chapitre 4 – Des modèles linguistiques peu adaptés

Bien avant l'informatique, la linguistique s'est intéressée à l'analyse génétique de textes. Au départ les deux disciplines se sont longuement ignorées et tout n'est pas la faute de la littérature comme nous le rappelle Roland Barthes :

> « La linguistique elle-même adhérait parfaitement à l'image séparatiste que la littérature voulait donner elle-même ; soumise à un sur-moi scientifique très fort, elle ne se reconnaissait pas le droit de traiter de la littérature, parce que pour elle la littérature se situait en grande partie en dehors du langage (dans le social, l'historique, l'esthétique). »
>
> [Barthes, 1968, p. 4]

Puis, les colloques de Cluny « Linguistique et Littérature », organisés à la fin des années soixante ont donné le jour à une collaboration entre les deux disciplines pour, entre autres, répondre par la linguistique à la problématique de la genèse[40]. Comment s'y prendre pour aborder scientifiquement le processus de création d'œuvres à partir des brouillons d'auteurs ? Quels modèles proposent les sciences du langage pour traiter ces textes particuliers ? Nous allons tenter de répondre à ces questions en parcourant rapidement certaines théories linguistiques.

4.1 Les théories linguistiques

4.1.1 La grammaire générative

Dans le modèle chomskyen, la grammaire générative propose une résolution logique mathématique de la description d'ensembles textuels[41]. Une langue naturelle (aux mots polysémiques) ou formelle (aux mots univoques) pourrait être décrite par une grammaire qui permet d'énoncer l'exhaustivité de l'ensemble des phrases de cette

[40] [Ablali & Kastberg Sjöblom, 2010]
[41] [Lebrave & Grésillon, 2009]

langue. Or l'unité textuelle de ce modèle n'est que la phrase et non pas l'ensemble de phrases qu'est le texte littéraire.

Les travaux de Duhem et de Lang[42] ont tenté d'élargir ce modèle phrastique au texte en imaginant d'ajouter aux arbres de phrases un nœud supérieur T (pour texte) qui réunirait les phrases d'un même texte[43].

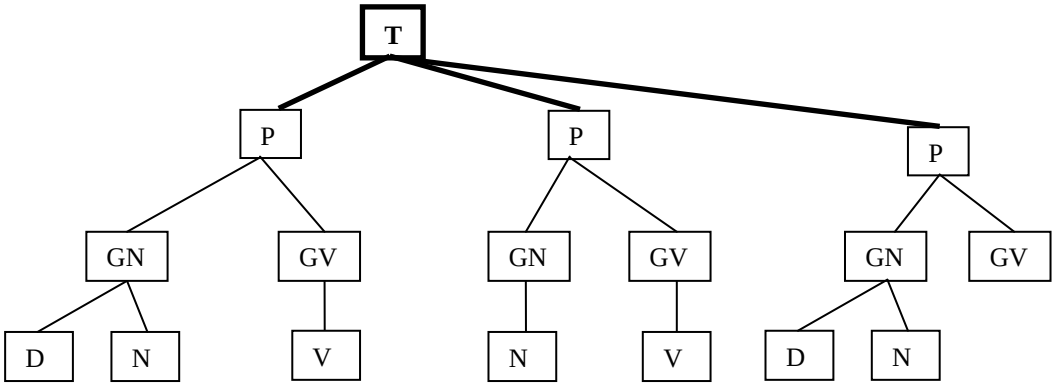


Illustration 3 – Essai de représentation de l'arbre au nœud supérieur T de Duhem

Ces travaux ont rapidement été mis à mal car les données purement inhérentes au texte ne pouvaient être prises en compte par un tel modèle. Ce dernier étant démontré comme inefficace pour la description de données textuelles dépassant la phrase[44], il l'était par conséquent *a fortiori* pour des textes avec ratures, ajouts ou suppressions que contiennent les brouillons.

4.1.2 Les opérations énonciatives

Avec la théorie des opérations énonciatives d'Antoine Culioli, nous obtenons quelques éléments de réponse à notre problématique. Le résultat écrit, l'énoncé, peut être séparé du moment de sa création, l'énonciation[45]. L'énoncé, de nature matérielle, peut être détaché de l'acte qui sera traité à part. Les énoncés ont donc une histoire

[42] [Duhem, 1972] et [Lang, 1972]
[43] [Duhem, 1972]
[44] [Lang, 1972]
[45] [Lebrave & Grésillon, 2009]

propre s'inscrivant dans une temporalité, dont certaines propriétés peuvent être modélisées.

Nous avons ici finalement des outils linguistiques se rapportant à l'écrit. Ce dernier n'est plus considéré comme la transcription d'un oral (Saussure) mais comme une forme propre avec ses spécificités.

4.1.3 Le modèle d'Hayes et Flower

La particularité principale de l'écrit est donc l'indépendance totale entre les deux champs que sont la production du message et la réception du message. Ces deux champs étant parfaitement autonomes, une possibilité nouvelle apparaît : celle de modifier le message plusieurs fois, avant de le transmettre. Les travaux des psycholinguistes Hayes et Flower proposent un modèle, décrit dans l'illustration ci-dessous, tenant compte de cette non-simultanéité.

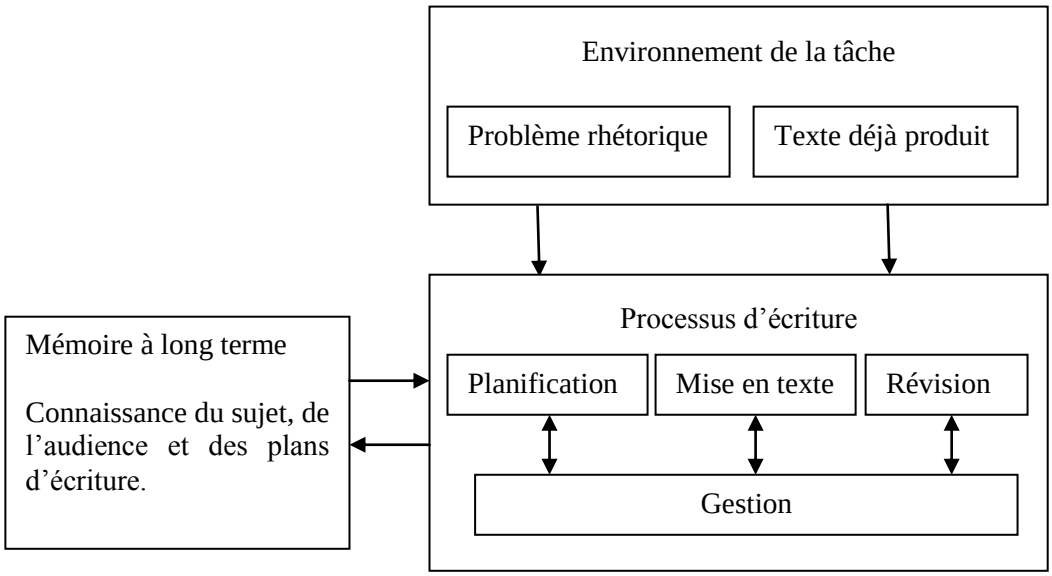


***Illustration 4 – Modèle de processus d'écriture selon Hayes et Flower*[46]**

Le modèle propose trois éléments dissociés : l'*environnement de la tâche*, la mémoire à long terme et le *processus de production écrite*. L'environnement de la tâche comprend le destinataire, les thématiques abordées et l'écrit en cours de création jusqu'à sa

[46] [Hayes & Flower, 1981]

production finale, moment où il se dissocie de l'environnement pour devenir un objet à part entière. La mémoire à long terme correspond à toutes les connaissances déclaratives et procédurales du scripteur, lexique, plans d'écriture des scénarios ainsi que tous les souvenirs que le scripteur peut utiliser. Par comparaison, chez Stendhal par exemple cette étape peut coexister dans le temps avec celle de la rédaction proprement dite, ou même lui succéder : il rédige *Lucien Leuwen*, et indique en marge ou sur des feuillets à part, le plan qu'il a suivi ou qu'il suit ; il fait aussi des portraits de personnages ou des commentaires.

Ce modèle est plus prévu pour modéliser la pensée mais cette mémoire pourrait correspondre pour notre problématique aux notes, plans, généalogies, etc. de l'écrivain. Le processus d'écriture, lui-même subdivisé en trois objets (planification, mise en texte et révision) entre en interaction.

Selon Hayes et Flower, il n'y a pas de hiérarchisation des trois objets. Ce modèle psycholinguistique nous permet bien d'appréhender par certains aspects l'œuvre en devenir, mais reste insuffisant pour notre problématique car celle-ci inclut également un grand nombre d'autres données qui ne sont pas modélisables simplement par l'étude du processus d'écriture. En effet, le modèle qu'ils proposent sert à identifier les différents mécanismes simultanés entrant en jeu dans la production de l'écrit[47]. Conçu au départ pour déceler les origines de troubles de la rédaction, ce modèle présente des aspects pouvant être repris pour notre problématique.

À la recherche d'un modèle d'analyse génétique de textes, nous n'avons trouvé que des modèles trop éloignés de notre problématique. Observons maintenant les outils de la discipline, peut-être pourrons nous en utiliser quelques-uns.

[47] [Tognotti, 1997]

4.2 Les outils linguistiques

4.2.1 La substitution

Au niveau du vocabulaire, le structuralisme propose la *substitution*. Le terme est repris par des généticiens mais avec une autre acception ou plus précisément un élargissement de sens[48]. Pour le structuraliste, la substitution est symétrique et indépendante du temps alors que pour le généticien le changement dans l'écriture est nécessairement ordonné chronologiquement et la substitution n'est en rien symétrique (même s'il arrive que l'écrivain raye la substitution pour proposer une troisième solution qui sera en fait le premier terme). Le scripteur écrit d'abord « x », puis le remplace par « y ». Pour pouvoir traiter les quatre cas de réécriture, à savoir, le remplacement, l'ajout, la suppression et le déplacement, une nouvelle notion linguistique a été inventée : la variable ø (ensemble vide)[49]. Ainsi, les autres cas de réécriture sont représentés ainsi :

1. le remplacement comme « x → y », (x devient y) ;
2. l'ajout comme « ø → x », (rien devient y) ;
3. la suppression comme « x → ø », (x devient rien) ;
4. le déplacement comme « abcd→bcda » (abcd devient bcda).

4.2.2 La variante

Toujours au niveau du champ lexical, le structuralisme a apporté un autre terme, la *variante*. Pour le structuraliste, elle signifie deux réalisations dans le même contexte d'une même unité linguistique (par exemple deux sons ou deux morphèmes) mais elle ne modifie pas la valeur de l'unité linguistique elle-même : par exemple le *r* roulé et le *r* grasseyé. Pour le généticien la variante devient un peu le contraire dans la mesure où elle modifie la valeur du texte. Deux unités textuelles (par exemple deux mots, groupes syntaxiques ou phrases) sont considérés comme variantes si dans un même contexte ils engendrent une différence de sens. Si l'écrivain biffe un mot pour le remplacer par un synonyme, il ne s'agit pas d'une variante ; en revanche s'il le remplace par un mot

[48] [Grésillon *et al.*, 1990]
[49] C'est-à-dire rien, absence de quelque chose

entraînant une *différence de sens pertinente*[50] il s'agit bel et bien d'une variante : par exemple : *une ~~femme~~ dame jouait du violon* n'est pas une variante de *une femme jouait du violon* mais est une variante de *~~une femme~~ un jeune garçon jouait du violon*.

4.2.3 La variante liée versus la variante libre

Le linguiste danois Louis Hjelmslev avait introduit la notion pour le morphème[51] de lié et de libre : le morphème est dit lié s'il ne se manifeste pas comme lemme[52] et n'existe jamais à l'état libre, mais est toujours rattaché à un autre morphème appelé base : comme *-ons* dans *ouvr-ons* ou *re-* dans *re-faire* ; le morphème est dit libre s'il peut constituer un mot : *le* ou *beau* sont libres.

Les linguistes généticiens proposent de reprendre cette notion et de l'adapter à la variante[53]. La variante sera dite liée quand elle est due à des contraintes de langue, morphologie, lexique, syntaxe ou si elle n'est que l'effet grammaticalement nécessaire d'une variante première ; la variante est dite libre si elle n'est pas liée.

4.2.4 La variante d'écriture versus la variante de lecture

La variante peut également être dite « d'écriture », auquel cas il s'agira d'une correction au fil de la plume, ou de lecture si elle a été insérée après coup et non immédiatement. Cette différence est souvent repérable sur les manuscrits grâce à sa position par exemple, soit dans l'interlinéaire, soit dans la marge, soit sur d'autres feuillets ou grâce à une autre encre utilisée lors de la relecture ; parfois il n'est pas possible d'identifier formellement, avec certitude la propriété de la variante.[54]

4.2.5 Le texte non variant versus le texte variant

Enfin, la notion de variante a également été utilisée pour décrire le texte à un niveau de granularité supérieur. Le texte non variant représente le texte inchangé tout au long de la genèse et le texte variant, celui qui subi des modifications.

---

[50] [Lebrave & Grésillon, 2009]
[51] Un morphème est la plus petite unité porteuse de sens qu'il soit possible d'isoler dans un énoncé.
[52] Un lemme est une suite de caractères formant une unité sémantique et pouvant constituer une entrée de dictionnaire. Souvent synonyme de « mot », ou « terme ». Les lemmes sont constitués de morphèmes.
[53] [Grésillon *et al.*, 1990]
[54] [Grésillon & Lebrave, 1982]

4.2.6 Le texte, le méta-texte et le non-texte

Outre le texte proprement dit, les manuscrits contiennent deux types d'autres données : du « méta-texte » et du « non-texte ». Par méta-texte nous entendons des informations liées au texte comme par exemple les ratures, les soulignements, les marques d'insertions, les différentes couleurs d'encre, etc. ; par non-texte nous comprenons les croquis, dessins, rébus, etc. Ces deux autres types de données sont bien souvent essentiels à la compréhension du processus de genèse.

Ainsi cette analyse du matériel linguistique nous a permis de retenir pour notre problématique le modèle d'Hayes et Flower, la notion de substitution et de variantes liées ou non, d'écriture ou non, de texte variant ou non ainsi que la nécessité d'inclure en plus du texte, des méta-données et des illustrations.

Chapitre 5 – Des outils informatiques trop spécifiques

Les enjeux de l'informatique dans notre problématique se situent à plusieurs niveaux. Comme nous l'avons vu pour la constitution de l'*avant-texte*, nous jonglons avec des livres annotés, des cahiers ou carnets, et des feuillets. Nous avons besoin d'étiqueter et de dater chaque document, et chaque page du document. Puis si nous descendons au niveau de la page, quelle que soit sa provenance, nous trouvons du texte et du graphisme. Nous pouvons représenter et stocker des informations scripturales mais aussi visuelles, et surtout effectuer après-coup des recherches pertinentes sur ces informations. Le modèle doit également présenter suffisamment de souplesse pour que le généticien puisse effectuer des allers-retours entre les différents maillons : nous l'avons vu, il s'agit de commencer par récupérer et préclasser les éléments constitutifs de l'avant-texte avec une première lecture approximative, puis de déchiffrer le feuillet avec précision, et le cas échéant, le référencer plus finement, voire modifier sa référence.

Avant de se lancer dans le montage d'un modèle répondant à toutes ces contraintes, effectuons un rapide tour d'horizon des modélisations existantes. Cette fois-ci, il ne s'agit pas de chercher dans l'histoire de l'informatique des modèles adaptables à l'analyse génétique, la différence entre les deux objets observés étant trop importante, mais plutôt d'inspecter les outils existants. Nous avons trouvé trois outils qui pourraient nous aider dans notre problématique.

5.1 Les logiciels EDITE et MEDITE

5.1.1 Un programme de comparaison de styles et de versions

Le programme EDITE[55] a initialement été créé pour faciliter le travail du chercheur dans la comparaison de plusieurs versions d'un même texte, comprendre la nature du travail de réécriture afin de comparer deux styles d'écriture et éventuellement de déceler automatiquement l'auteur de lignes. EDITE a par exemple été utilisé pour une analyse sur un corpus constitué des manuscrits de Romain Gary et des manuscrits

[55] Pour *Étude Diachronique et Interprétative du Travail de l'Écrivain*

écrits sous son pseudonyme Émile Ajar[56]. Romain Gary (1914-1980), romancier français originaire de Pologne, est le seul double lauréat du Prix Goncourt. Il a obtenu les prix sous deux noms différents, une fois Romain Gary, une fois Emile Ajar. L'analyse devait permettre de déterminer si réellement Romain Gary était parvenu à créer un vocabulaire et un style propres à son pseudonyme Émile Ajar, ou si des méthodes scientifiques d'attribution d'auteur auraient permis le démasquer[57]. À l'époque tous les critiques littéraires avaient été trompés.

5.1.2 Un algorithme en trois étapes

EDITE repose sur un algorithme en trois étapes. Au départ, pour découvrir automatiquement les corrections à partir de la comparaison de deux versions, le programme conçu par Jean-Gabriel Ganascia[58] cherchait à réutiliser des algorithmes utilisés en informatique et en bioinformatique[59]. La tâche s'est rapidement avérée irréalisable pour des raisons de complexité algorithmique : ce qui s'adaptait à deux séquences de génomes courts et très semblables par ailleurs ne s'adaptait pas à des chaînes de caractères longues et très différentes. Ganascia a donc créé un nouvel algorithme spécifique qu'il a nommé MEDITE. Le programme repère les unités linguistiques soumises aux quatre modifications suivantes : suppression, insertion, remplacement et déplacement. Il repère également les marques de style : déplacement d'un adverbe, remplacement d'un mot par un hyperonyme ou un hyponyme[60], suppression ou ajout d'un adjectif etc. L'algorithme fonctionne en trois phases successives, la première consiste à détecter des blocs communs, la deuxième à identifier les déplacements et la dernière, à calculer les insertions, les suppressions et les remplacements (les déplacements ne sont pas traités).

[56] [Chepiga, 2008]
[57] [Labbé, 2008]
[58] Du Laboratoire d'informatique de Paris 6 (Pierre et Marie Curie), le LIP6
[59] Pour le détail des algorithmes, cf. [Crochemore & Rytter, 1994] et [Sankoff & Kruskal, 1983] cités dans [Ganascia *et al.*, 2004]
[60] Hyperonyme : terme dont le sens inclut celui d'un ou de plusieurs autres (mot générique) : véhicule est l'hyperonyme de voiture. Voiture est l'hyponyme de véhicule.

5.1.2.1 Détection des blocs communs

La détection des blocs communs fait appel à des algorithmes de recherche de sous-séquences communes dans des séquences plus longues[61]. Les chaînes de caractères communes peuvent se situer à l'intérieur de mots, ou si elles font plusieurs mots de long voire plusieurs lignes, couper des mots à leurs extrémités (en début et en fin de chaîne). C'est pourquoi les signes de ponctuation et les blancs ont été utilisés pour couper les séquences et donner des blocs dits *disjoints.* Il est à noter que la taille minimale du bloc peut être choisie par l'utilisateur et que par défaut elle est de quatre caractères (donc deux lettres comprises entre deux signes de ponctuation ou des blancs).

Nous avons maintenant deux listes de blocs. Par exemple :

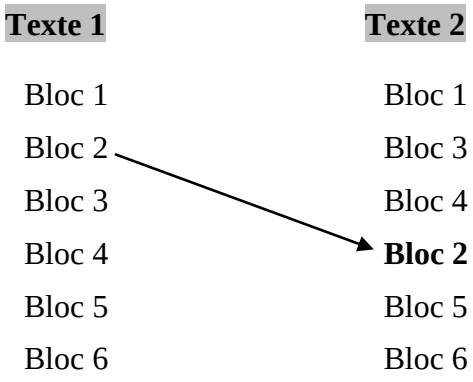


Illustration 5 – MEDITE : identification de blocs communs

5.1.2.2 Identification des déplacements et des pivots

Pour repérer les blocs déplacés, le logiciel va maintenant repérer le Bloc 2 et l'extraire dans la liste des « blocs déplacés ». Il est à noter qu'il est possible que l'auteur n'ait pas du tout déplacé le Bloc 2 mais plutôt les Blocs 3 et 4 comme ceci :

[61] Karp R M, Miller R E and Rosenberg A L, 1972 ; Landraud A, Avril J-F and Chrétienne P, 1989

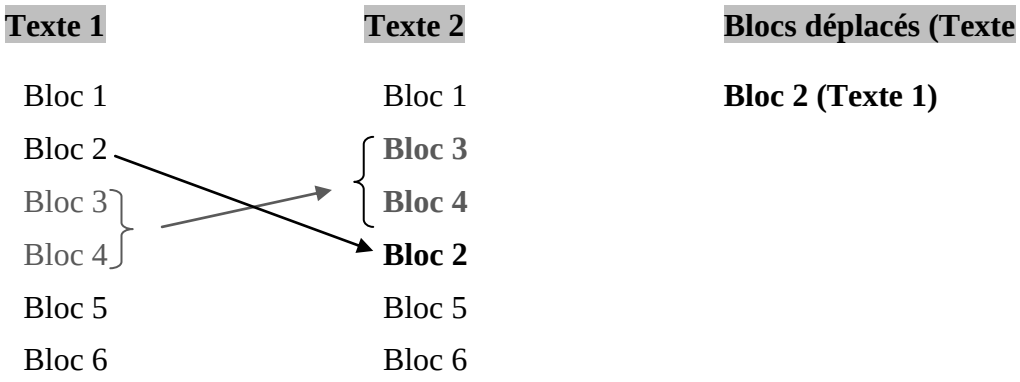


Illustration 6 – MEDITE : identification de blocs communs

Les blocs ici ne correspondent en rien à des paragraphes mais sont justes des chaînes de quatre caractères ou plus. Des deux solutions, l'algorithme choisit le plus petit nombre de caractères à déplacer. Les blocs qui apparaissent dans le même ordre dans les deux textes, les inchangés, sont appelés les *blocs pivots* de la comparaison. La fin d'un bloc pivot est forcément là où il y a différence avec l'autre texte.

5.1.2.3 Calcul des insertions, des suppressions et des remplacements

Maintenant identifions les insertions, les suppressions et les remplacements. Lorsque deux blocs pivots sont consécutifs dans la première version, par exemple les Bloc 4 et Bloc 5, la chaîne qui sépare le Bloc 4 du Bloc 5 dans la deuxième version correspond à une insertion. Ensuite, lorsque deux pivots Bloc 1 et Bloc 3 sont consécutifs dans la deuxième version, la chaîne qui sépare Bloc 1 et Bloc 3 dans la première version correspond à une suppression.

Texte 1 : Pivots	Texte 2 : Pivots
Bloc 1	Bloc 1
	Bloc 3
Bloc 3	Bloc 4
Bloc 4	
Bloc 5	Bloc 5
Bloc 6	Bloc 6

Illustration 7 – MEDITE : calcul des suppressions

Enfin, lorsque deux pivots Bloc 1 et Bloc 4 ne sont consécutifs ni dans la première version, ni dans la deuxième version, il est dit qu'il y a remplacement de la chaîne comprise entre Bloc 1 et Bloc 4 dans la première version, par la chaîne comprise entre Bloc 1 et Bloc 4 dans la deuxième version.

5.1.3 La description du logiciel

Le logiciel est programmé dans le langage Python et fonctionne sous les systèmes d'exploitation Windows et LINUX et Mac OS X. Le traitement automatique du texte se fait après un travail préalable de transcription linéarisée (c'est-à-dire de transcription de la résultante du processus d'écriture) de chaque état du manuscrit. Sur l'exemple qui suit, les deux états les plus distants ont été analysés : la toute première version et le manuscrit final. Toutes les campagnes de réécritures intermédiaires ont donc été ignorées dans cet exemple. Notons que la toute première version a été transcrite à partir d'un manuscrit tapé à la machine et non écrit à la main. Le logiciel propose une interface de visualisation avec trois fenêtres. En plus des deux fenêtres présentant chacune une des deux versions (texte source, texte corrigé), une troisième fenêtre décrit le type de modification.

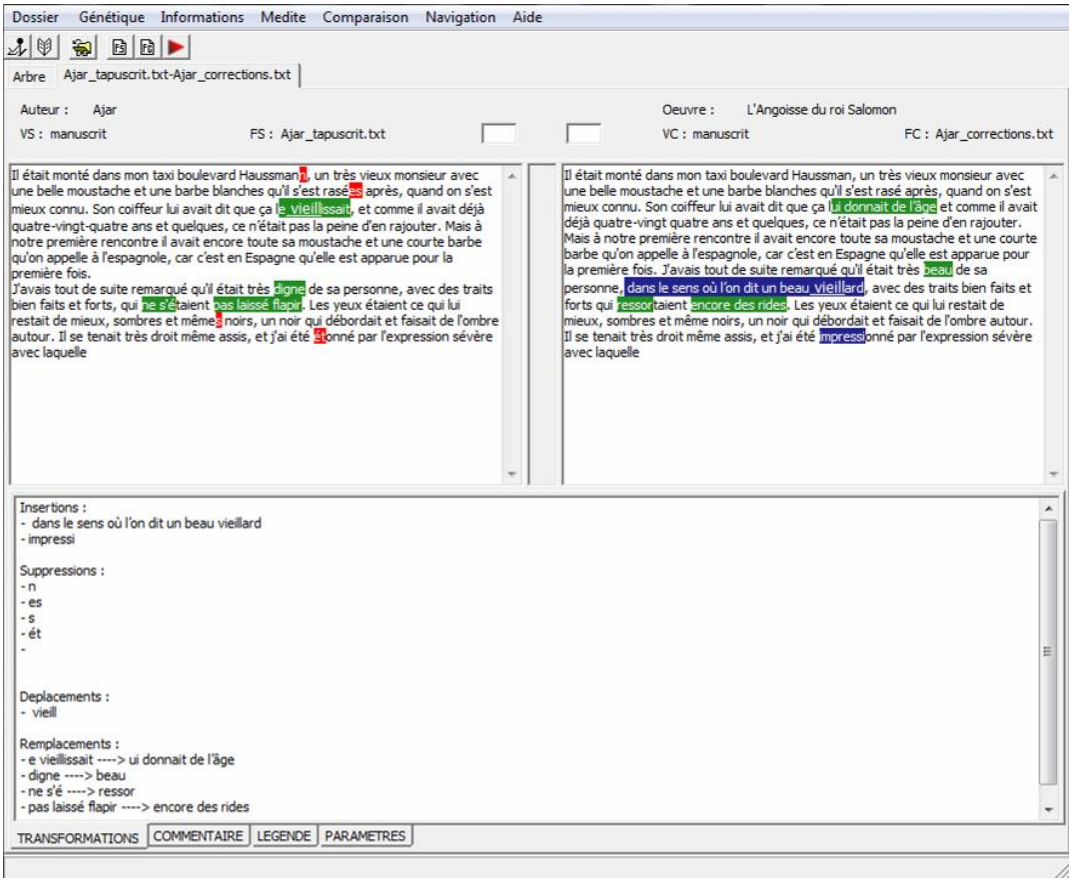


Illustration 8 – Interface de comparaison du logiciel MEDITE

Le logiciel MEDITE a donc une interface accessible, les différents onglets permettant au généticien de jongler avec les commentaires ou les statistiques.

5.1.4 MEDITE et notre problématique

Pour notre problématique, ce logiciel est trop spécifique. Premièrement, nous ne pouvons comparer que deux versions. Deuxièmement les deux versions comparées doivent être assez proches, donc il s'agira de la fin de la genèse. Par rapport à notre problématique générale, les fonctionnalités présentées sont situées un peu en aval, en fin de processus. En effet, si nous reprenons notre tableau, nous sommes ici dans la phase rédactionnelle.

Phase	**Documents de genèse**	**Outil informatique**
Prérédactionnelle	Bibliothèques d'auteurs	
	Carnets	
	Cahiers	
	Dessins	
	Feuillets	
Rédactionnelle	Feuillets	**MEDITE**
Pré-éditoriale	Manuscrit prédéfinitif	**MEDITE**
	Manuscrit définitif	**MEDITE**
Éditoriale	Bon à tirer (ou épreuve corrigée)	

Illustration 9 – Correspondance de l'outil MEDITE aux documents de genèse

5.2 Le logiciel MUSE

La base de données relationnelle MUSE (Manuscrits, Usage des Supports et de l'Ecriture) a été conçue dans le cadre du programme du CNRS « Archives de la création » par Claire Bustarret et Serge Linkès[62] de l'ITEM pour la recherche en codicologie et en génétique.

5.2.1 Un programme de codicologie

Ce logiciel permet la saisie d'une description matérielle systématique et détaillée de tout corpus manuscrit du XVIIIe au XXe siècle : identification des papiers filigranés, des papiers sur lesquels a été apposé un timbre sec, étude de la composition des supports, de la répartition des scripteurs et des instruments d'écriture ou encore repérage par la datation du papier de certaines campagnes de rédaction. C'est donc un programme utilisé seulement à des fins codicologiques. Pourtant le principe utilisé pour ce programme est très commun en informatique puisqu'il s'agit de la base de données relationnelle.

5.2.2 Le principe de la base de données relationnelle

Le principe de la base de données relationnelle est très simple. Nous disposons de *sets d'informations* que nous pourrions comparer à des fiches de bibliothèque. Sur l'une d'elle nous pourrions avoir par exemple : le titre de l'ouvrage, son type, son auteur, sa date de parution, sa cote, etc. Imaginons que nous dessinions un tableau dans lequel chaque ligne corresponde à une fiche. Nous obtiendrions le tableau suivant :

Fiche de bibliothèque	N° de fiche	Titre	Type	Auteur
	n° 1	Le Rouge et le noir	Roman	Stendhal
	n° 2	La Chartreuse de Parme	Roman	Stendhal

Illustration 10 – MUSE : Exemple 1 d'une table de base de données

La fiche de bibliothèque 1 correspond au roman de Stendhal *Le Rouge et le noir*.

[62] [Bustarret & Linkès, 2008]

Dans une base de données chaque ligne correspond à un *enregistrement* et chaque tableau à une *table*. Dans MUSE, il ne s'agit pas de répertorier des ouvrages mais des *feuillets*. Chaque *feuillet* comporte deux pages (un recto et un verso). Chaque description de page est appelée *fiche*. Par exemple une table pourra s'appeler *Papier* et comporter la liste de toutes les pages en papier (de feuillets, d'in-quarto, etc.).

Une autre table pourra par exemple s'appeler *Timbre sec* et comporter la liste de toutes les fiches ayant un timbre sec.

Timbre sec	N° de fiche	Identifiant papier	Identifiant timbre sec
	n° 1	n° 3	n° 63
	n° 78	n° 2	n° 28
	n° 263	n° 3	
	n° 374	n° 3	n° 5

Illustration 11 – MUSE : Exemple 2 d'une table de base de données

Pour occuper moins de place dans la table *Timbre sec*, seul un code d'identifiant papier est créé, un numéro, et non une série de colonnes dédiées à la description du papier. Une deuxième table est donc créée s'appelant *Identifiant papier* et correspondant aux colonnes qui n'ont pas été insérées dans la table *Timbre sec* pour ne pas la surcharger.

Identifiant papier	N° de papier	Type de papier	[…]
	n° 1	feuillet	feuillet
	n° 2	in-folio	in-folio
	n° 3	in-quarto	in-quarto
	n° 374	n° 7	n° 7

Illustration 12 – MUSE : Exemple 3 d'une table de base de données

Les deux tables sont reliées entre elles pour croiser les informations.

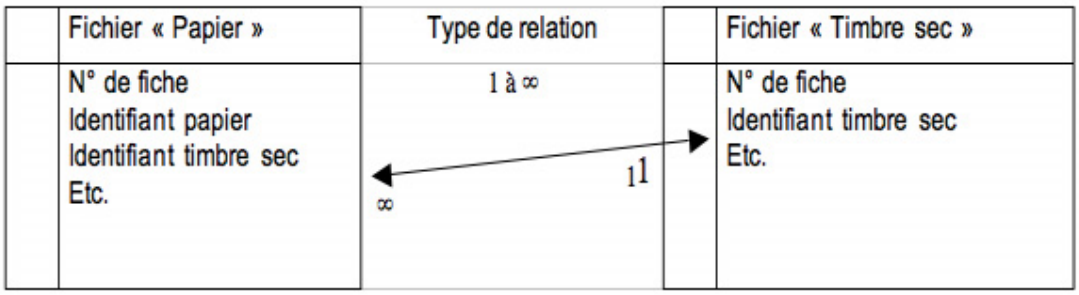


Illustration 13 – MUSE : Exemple d'une relation entre deux tables de base de données[63]

Sur l'illustration ci-dessus, nous apercevons le nom de la table et le titre de chaque colonne de la table (par exemple les colonnes *N° de fiche*, *Identifiant papier*, *Identifiant timbre sec*). La liste des tables et les critères de tri *Timbre sec*, *Filigrane*, *Papier*, etc. peut s'allonger en fonction des besoins.

Ainsi, la base de données relationnelle permet d'éviter la saisie redondante d'un même renseignement et de recouper des informations lors de requêtes multiples. Par exemple il est possible d'afficher seulement les entrées qui sont à la fois du type *in-folio* (type de papier n°2 sur notre exemple) et à la fois n'ayant *pas de timbre sec* (case vide sur notre exemple). Il s'agit là d'une simple intersection entre l'ensemble 1 *in-folio* et l'ensemble 2 *pas de timbre sec*. Cette logique de base de données relationnelle est ainsi souple et adaptable à de très nombreuses problématiques.

5.2.3 La description du logiciel

La base de données MUSE a été conçue avec un logiciel commercialisé de gestion de base de données fonctionnant sous les systèmes d'exploitation Windows et Mac OS X : *File Maker Pro*®. Le but était d'utiliser un logiciel commercialisé ne risquant pas de disparaître du jour au lendemain comme c'est souvent le cas dans le monde informatique[64]. MUSE procède en deux étapes. La première consiste à cataloguer dans la base de données tous les feuillets et les informations les concernant. La seconde consiste à trier les feuillets en parcourant les tables et en donnant plusieurs critères comme nous l'avons vu avec l'exemple des tables *Papier* et *Timbre sec*. En fonction des résultats obtenus, le généticien peut par exemple dater certains fragments

[63] [Bustarret & Linkès 2008]
[64] [Bustarret & Linkès, 2008]

qui ne l'étaient pas ou attribuer un feuillet isolé à un ensemble documentaire. L'utilisation de la base de données relationnelle permet non seulement de stocker un nombre de données dépassant nos capacités intellectuelles mais également l'exploitation de données chiffrées (mesures d'épaisseur et de rugosité donnant lieu à des calculs automatiques) qui modifient les données de base. Ainsi par exemple, si un feuillet isolé est rattaché à un ensemble documentaire, les données de base sont retouchées et les tables correspondantes sont mises à jour (retour à l'étape 1) en fonction des résultats trouvés. Voici une représentation du processus cette fois-ci du point de vue du feuillet. Un type de feuillet est créé qui reste provisoire mais qui est utilisé pour faire avancer les travaux ; puis lorsque nous parvenons à des résultats par mesures physiques et mathématiques, nous créons éventuellement un nouveau type de feuillet[65].

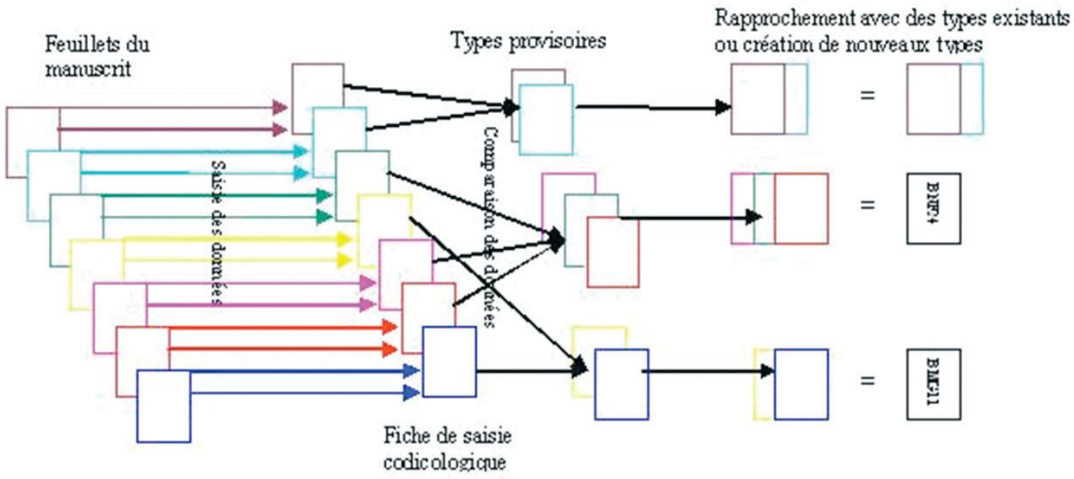


Illustration 14 – MUSE : Processus d'analyse des papiers

Les manuscrits de Stendhal ont déjà fait l'objet de plusieurs enquêtes codicologiques avec le logiciel MUSE entre 1997 et 2000. L'analyse codicologique de 719 feuillets de brouillon dont la rédaction s'étend d'avril 1839 à mars 1842 a permis de rectifier la datation de diverses campagnes de rédaction grâce à des recoupements entre la provenance italienne ou française des supports et les interventions successives des copistes romains, de l'auteur lui-même lors de ses séjours à Civittavecchia et d'un copiste parisien (Bonavie).

[65] [Bustarret & Linkès, 2006]

Le logiciel MUSE est donc spécialisé pour l'étude codicologique.

5.2.4 MUSE et notre problématique

La logique de base de données relationnelle, souple et adaptable à de très nombreux sujets convient parfaitement pour notre problématique. L'aspect codicologique en revanche l'est moins, bien qu'il ait déjà été utilisé sur les manuscrits de Stendhal. En effet, nous ne disposons actuellement pas d'un accès libre aux manuscrits papiers mais seulement à des numérisations ; par ailleurs nous n'avons actuellement dans l'équipe ni compétence ni outils d'analyse en papeterie qui nous permettrait une étude sérieuse en codicologie. Nous ne retiendrons par conséquent pas la méthode spécifique utilisée par MUSE mais conserverons le principe de la base de données.

Si nous classons maintenant MUSE dans notre tableau de correspondance, nous le placerons en face de plusieurs phases. En effet, le logiciel de codicologie peut correspondre à tous les feuillets de genèse textuelle, qu'ils appartiennent aux phases prérédactionnelle, ou rédactionnelle.

Phase	**Documents de genèse**	**Outil informatique**
Prérédactionnelle	Bibliothèques d'auteurs	
	Carnets	
	Cahiers	
	Dessins	
	Feuillets	**MUSE**
Rédactionnelle	Feuillets	MEDITE **MUSE**
Pré-éditoriale	Manuscrit prédéfinitif	MEDITE
	Manuscrit définitif	MEDITE
Éditoriale	Bon à tirer (ou épreuve corrigée)	

Illustration 15 – Correspondance de l'outil MUSE aux documents de genèse

5.3 Les logiciels HyperNietzsche et Nietzsche Source

Un autre travail autour de manuscrits sont les projets HyperNietzsche[66] et Nietzsche Source[67] portés par Paolo d'Iorio[68] et ayant pour but de valoriser les manuscrits de Nietzsche et toutes les publications sur ces manuscrits.

5.3.1 Une gigantesque bibliothèque

Au départ le projet a consisté à numériser les carnets de notes et manuscrits de l'auteur et à proposer les images au public ; puis, peu à peu des éditions génétiques ont été mises en ligne[69]. HyperNietzsche est organisé avec la logique d'une bibliothèque, chaque objet étant répertorié dans un catalogue général. Au fur et à mesure de leur arrivée sur le site, les nouveaux objets ont été référencés. La navigation et les recherches dans les différents documents s'effectuent classiquement par la recherche textuelle mais elle est doublée d'une *mise en contexte hypertextuelle.* Cette mise en contexte consiste à ajouter des liens (via des balises) dans chaque document vers des articles scientifiques correspondants au document d'origine. C'est une équipe dédiée, des membres du projet réunis en association loi 1901 pour l'occasion qui évaluent les références contenues dans les articles des chercheurs et qui ajoutent les balises. Ces balises permettent ensuite de naviguer d'un document à un autre.

5.3.2 Le système de calques

Dans le modèle mathématique sous-jacent, chaque élément est lié à d'autres selon des critères de sens attribué par des humains et non automatiquement. Ainsi la navigation a lieu suivant des relations sémantiques et il ne s'agit plus de parcourir des occurrences de chaînes de caractères. Le codage des balises est réalisé en *HyperNietzscheMarkupLanguage*(HNML), un langage de balisage créé sur mesure, proche de la TEI et fondé sur l'XML.

[66] http://www.hypernietzsche.org/base.html
[67] http://www.nietzschesource.org/
[68] [D'Iorio, 2008]
[69] *Le Voyageur et son Ombre* et *Aurore*

Par ailleurs, un système développé par D'Alfonso et Saller[70] pour l'analyse génétique permet une étude à plusieurs niveaux. Sur une base contenant les informations communes à tous les niveaux vont se superposer une à une des couches comme autant de calques qui seraient posés sur un dessin technique, chaque calque contenant une étape génétique. S'il y a eu quatre jets d'écriture par exemple, nous aurons une base et trois calques superposés. Ce système permet d'isoler chaque jet et de le référencer à loisir.

5.3.3 D'HyperNietzsche à Nietzsche Source

La navigation dans HyperNietzsche est élaborée mais l'équipe a vite remarqué que l'utilisateur ne se repérait pas facilement sur le site. C'est pourquoi l'équipe décide en 2007 de repenser totalement l'interface. C'est ainsi que naît l'interface de navigation Nietzsche Source[71], permettant une navigation en étoile à partir d'un point source. Le point source est toujours un manuscrit original de Nietzsche. L'utilisateur peut se perdre dans sa navigation, il aura toujours un point de repère, l'objet manuscrit d'où il a commencé sa navigation. Ainsi 30 000 pages ont été numérisées en 2003, dont environ 6 000 ont été publiées sur le site *HyperNietzsche*, correspondant au dossier génétique complet de deux œuvres de Nietzsche (*Le Voyageur et son ombre* et *Aurore*). Puis en 2009 ces premières 6 000 pages ont été transférées sur le nouveau site *Nietzsche Source* et un deuxième lot de 3 300 pages a été transféré à son tour correspondant à 86 cahiers manuscrits et aux premières éditions des œuvres[72].

5.3.4 La description du logiciel

Dans cette nouvelle interface, deux modes séparées et accessibles sous deux liens différents sont proposés : le premier, le *mode simple* sans contextualisation et le second, le *mode savant* avec une contextualisation.

[70] [D'Alfonso & Saller, 2007]
[71] http://www.nietzschesource.org/
[72] http://www.item.ens.fr/index.php?id=377179

5.3.4.1 Le mode simple

Dans le mode simple, cinq onglets sont proposés. Ils suivent les phases génétiques vues dans la première partie. L'utilisateur part de l'épreuve, donc de la phase éditoriale, et remonte à travers la genèse jusqu'aux cahiers et carnets.

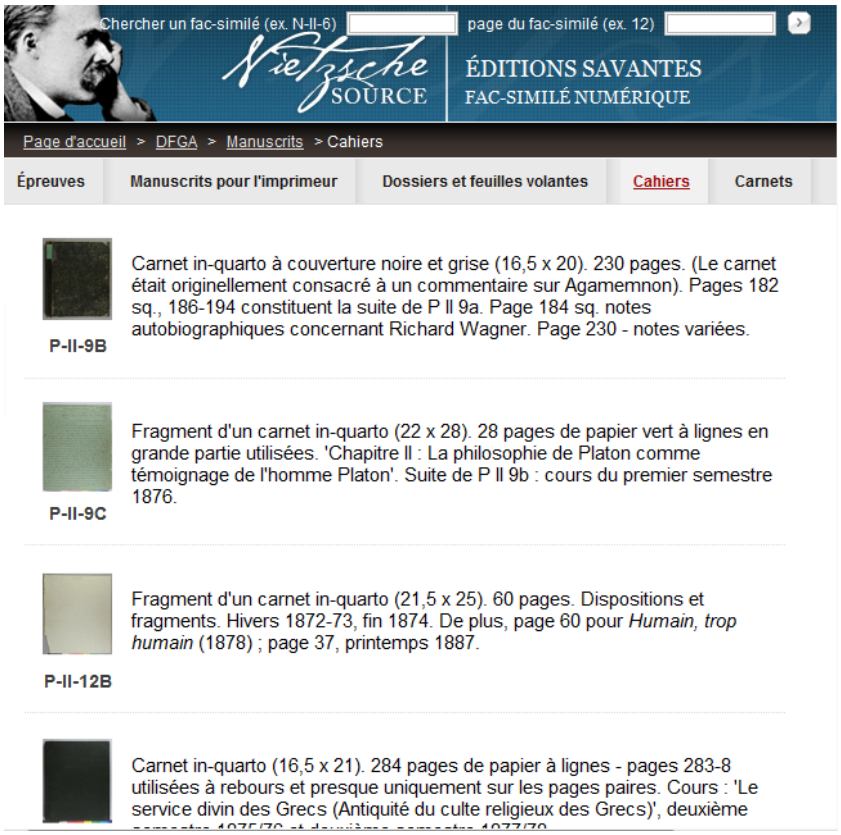


Illustration 16 – Nietzsche Source : épreuves et cahiers

Sur l'illustration ci-dessus, nous voyons une liste de pièces manuscrites de Nietzsche. Il s'agit de carnets ou de fragments de carnets in-quarto. Il est possible (en cliquant sur l'image) de visualiser chacune des pages à l'intérieur de chaque document.

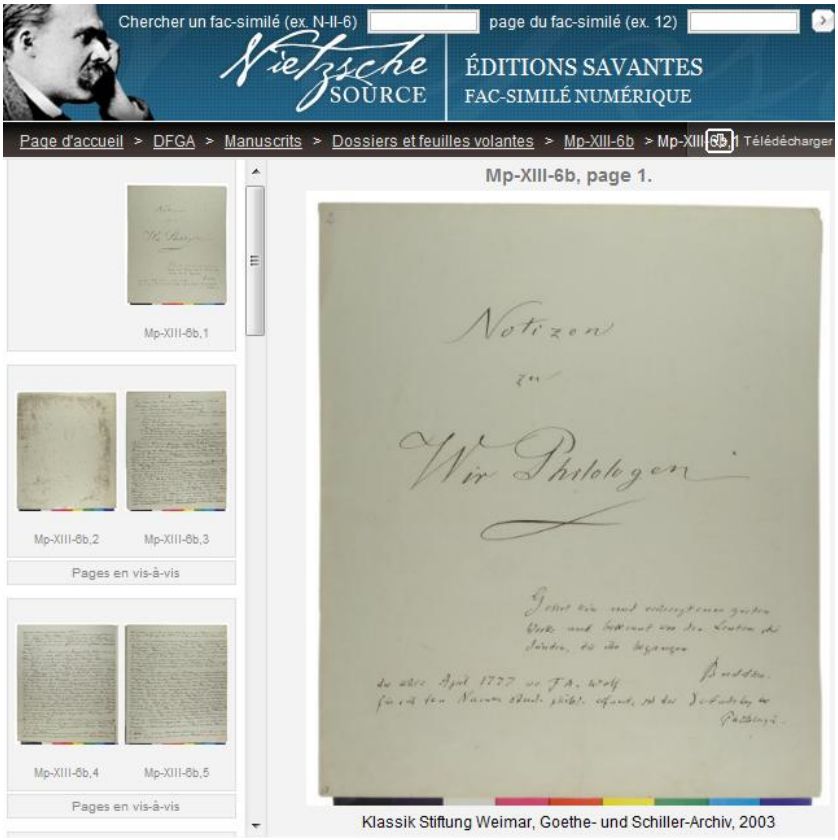


Illustration 17 – Nietzsche Source : exemple de cahier de genèse

Par exemple sur l'illustration ci-dessus, nous avons un cahier avec des notes de genèse. Il s'agit de notes prises en prévision de la création d'une œuvre qui s'intitulera : *Wir Philologen*[73]. Sur la première page, nous pouvons lire *Notizen zu Wir Philologen*, ce qui signifie *Notes pour « Nous les philologues »*.

Cet environnement informatique est moins spécifique que ne le sont MEDITE et MUSE ; il est accessible à tout public et facile à utiliser. Nous n'avons toutefois ici qu'une numérisation et pas de transcription, ce qui ne permet pas d'effectuer de recherches textuelles.

5.3.4.2 Le mode savant

C'est ce mode savant qui propose un environnement riche et complet consacré à la publication de contenus scientifiques concernant l'œuvre et la vie de Friedrich Nietzsche et dont les publications sont validées par des pairs. L'environnement est placé sous la responsabilité de la *Nietzsche Source Organization*, une association à but non

[73] Nous les philologues

lucratif, dont le siège est situé à l'École normale supérieure de Paris et qui vise à réunir des spécialistes de Nietzsche à travers le monde. Le travail d'édition, de commentaire et d'interprétation de l'œuvre de Nietzsche est intégré au fur et à mesure sur le site.

Illustration 18 – Nietzsche Source : mode savant

Une édition critique numérique des œuvres complètes et de la correspondance (appelée eKGWB)[74] de Nietzsche est proposée sur le site. Elle regroupe le travail de collation du texte numérisé et transcrit mot à mot et de comparaison avec chaque mot de l'édition imprimée.

[74] Pour *e-Kritische Gesamtausgabe Werke und Briefe* (e- de *elektronische*)

Illustration 19 – Nietzsche Source : corrections génétiques

L'utilisateur peut consulter, parcourir et copier le texte, chercher des mots ou des phrases dans l'édition ou dans certaines de ses parties ou encore imprimer des passages (voir illustration). Les corrections génétiques retrouvées dans les différents volumes de critique de l'édition imprimée ont été intégrées dans le texte qui apparaît et mises en évidence (surlignés) ; la source de la correction apparaît en commentaire.

5.3.5 Nietzsche Source et notre problématique

En premier lieu, l'organisation elle-même du site est intéressante pour notre problématique. En effet elle permet de proposer à la fois la numérisation des originaux et un format exploitable textuellement. En second lieu le système sous-jacent développé par D'Alfonso et Saller[75] pour l'analyse génétique semble très pratique. Ce système de superposition de calques, chaque calque correspondant à une campagne d'écriture, pourrait nous permettre de jongler avec les différents jets d'écritures sans se perdre dans le foisonnement d'une édition diplomatique.

[75] [D'Alfonso & Saller, 2007]

Phase	**Documents de genèse**	**Outil informatique**
Prérédactionnelle	Bibliothèques d'auteurs	
	Carnets	**Nietzsche Source**
	Cahiers	**Nietzsche Source**
	Dessins	**Nietzsche Source**
	Feuillets	MUSE **Nietzsche Source**
Rédactionnelle	Feuillets	MEDITE **MUSE** **Nietzsche Source**
Pré-éditoriale	Manuscrit prédéfinitif	MEDITE
	Manuscrit définitif	MEDITE
Éditoriale	Bon à tirer (ou épreuve corrigée)	**Nietzsche Source**

Illustration 20 – Correspondance de l'outil Nietzsche Source aux documents de genèse

Nietzsche Source couvre toutes les étapes génétiques et pratiquement tous les documents de genèse.

Conclusion

Pour terminer sur les propositions linguistiques et informatiques, nous ne reprendrons ni la grammaire générative, ni les opérations énonciatives. En revanche, le modèle d'Hayes et Flower se rapproche de notre problématique et nous pouvons par conséquent nous en inspirer. Par ailleurs, deux aspects de théorie linguistique seront également à retenir lors du développement de notre modèle : la notion de substitution et celle de variante. En ce qui concerne les logiciels utilisés peu ou prou pour la génétique, nous avons, à travers trois d'entre eux, couvert plusieurs facettes de notre problématique : une comparaison de deux versions génétiques d'un même texte, une organisation en base de données relationnelle, et une organisation partant à la base de la numérisation de chaque document pour naviguer vers d'autres informations attenantes. Il s'agit maintenant de rajouter l'élément « bibliothèques d'auteurs », puis de réussir à définir un modèle réunissant toutes les autres fonctionnalités et permettant leur utilisation en harmonie. Nous proposons maintenant de regarder l'outil déjà créé pour l'exploitation des manuscrits de Stendhal puis nous lancerons la réflexion vers une recherche à trois niveaux de granularité : une étude sur les bibliothèques d'auteurs, une sur les influences théâtrales et enfin une sur les traductions dans la génétique.

Partie 3

Vers une recherche à trois niveaux

« Les opérations qui interviennent au cours d'une genèse sont si nombreuses et souvent si complexes que l'approche directe ne peut porter que sur des corpus assez restreints.

En revanche, l'outil informatique rend possible le traitement de corpus de n'importe quelle dimension. »

Pierre-Marc de Biasi

Chapitre 6 – Le domaine expérimental : la base documentaire CLELIA

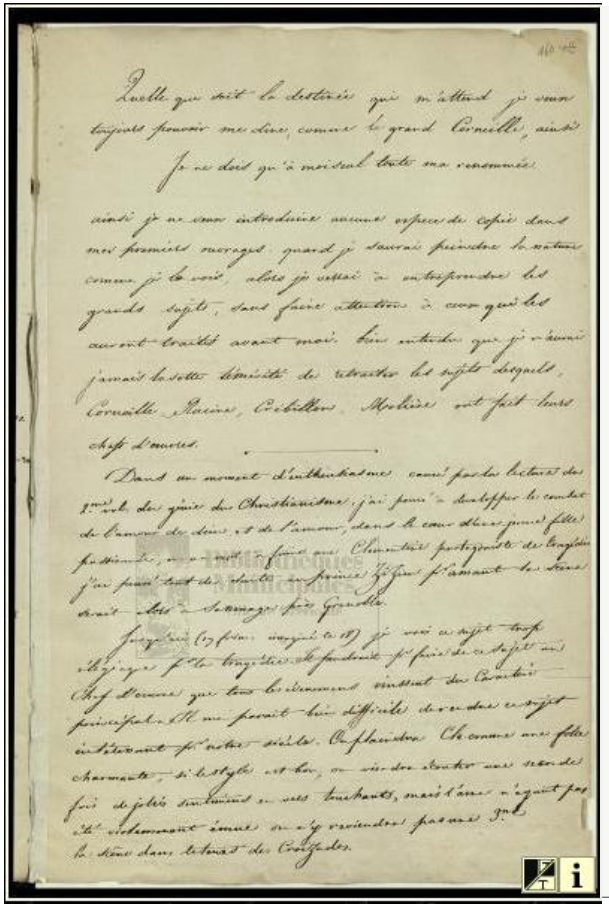
L'outil sur lequel nous nous proposons de construire un modèle pour l'analyse génétique est une base documentaire nommée CLELIA comprenant les images numériques des pages des manuscrits laissés par Stendhal à sa mort ainsi que la transcription au format XML de certaines de ces pages.

6.1 Des textes photographiés puis transcrits

Entre 2007 et 2009 la Bibliothèque municipale de Grenoble a numérisé la quasi totalité des manuscrits, soit environ 20 000 feuillets, soit 40 000 pages. Simultanément, des chercheurs de l'Université de Grenoble, regroupés sous le projet *Manuscrits de Stendhal*, ont entrepris la transcription des pages dans un format textuel exploitable avec des outils de la vie quotidienne.[76] L'outil propose une opération de transcription avec un fichier au format XML correspondant à une page de manuscrit et trois points d'entrée : un regroupement physique en volumes, un regroupement logique en textes et un moteur de recherche. Avec le moteur de recherche par exemple, il est possible de rechercher le mot « Grenoble » et d'obtenir les images (et leurs transcriptions) de toutes les pages contenant le mot « Grenoble » (voir les illustrations pages suivantes). Trois représentations sont proposées : l'image de la page originale, une transcription proche d'une transcription diplomatique, appelée par le projet *transcription pseudo-diplomatique*[77] et une troisième transcription, une transcription linéarisée[78] enrichie d'informations visant à rendre le texte plus accessible au lecteur non averti. En outre un outil pointu de réorganisation des écrits est en cours de construction[79]. L'illustration suivante présente les images d'une page de manuscrit numérisée ainsi que de ses transcriptions et des différentes informations sur son contenu textuel. Nous avons

[76] Le copier-coller dans Word ou dans Excel par exemple devient possible
[77] Le pseudo diplomatique reproduit autant que possible la mise en page et la graphie de l'auteur
[78] Comme nous l'avons vu, la transcription linéarisée est la transcription de la résultante du processus d'écriture
[79] Travaux dans le cadre d'une thèse de doctorat à paraître en 2011 : Aïcha Touati, laboratoire LIDILEM de l'Université Stendhal Grenoble 3.

sélectionné cette page grâce au moteur de recherche intégré en effectuant une recherche sur le mot « Grenoble ».



160
12

Quelle que soit la destinée qui m'attend je veux
toujours pouvoir me dire, comme le grand Corneille, ainsi

Je ne dois qu'à moi seul toute ma renommée.

ainsi je ne veux introduire aucune espece de copie dans
mes premiers ouvrages. quand je saurai peindre la nature
comme je la vois, alors je verrai à entreprendre les
grands sujets, sans faire attention à ceux qui les
auront traités avant moi. bien entendu que je n'aurai
jamais la sotte témérité de retraiter les sujets desquels,
Corneille, Racine, Crébillon, Molière, ont fait leurs
chefs d'œuvres.

Dans un moment d'enthousiasme causé par la lecture du
2. me vol. du génie du Christianisme, j'ai pensé à developper le combat
de l'amour de dieu et de l'amour, dans le cœur d'une jeune fille
passionnée, en un mot à faire une Clementine protagoniste de Tragédie
j'ai pensé tout de suite au prince Zizim p. r amant la scène
serait lors à Sassenage près Grenoble.

Jusqu'ici (19 frim. imaginé le 18) je vois ce sujet trop
élégiaque p. r la tragédie. Il faudrait p. r faire de ce sujet un
Chef d'œuvreque tous les évenemens vinssent du Caractère
principal. Il me parait bien difficile de rendre ce sujet
intéressant p. r notre siécle . On plaindra Cle. comme une folle
charmante, si le style est bon, on viendra écouter une seconde
fois de jolies sentimens en vers touchants, mais l' ame n'ayant pas
été violemment émue on n'y reviendra pas une 3. me
la scêne dans le tems des Croizades .

160

Quelle que soit la destinée qui m'attend je veux toujours pouvoir me dire, comme le grand Corneille, ainsi

Je ne dois qu'à moi seul toute ma renommée.

ainsi je ne veux introduire aucune espèce de copie dans mes premiers ouvrages. quand je saurai peindre la nature comme je la vois, alors je verrai à entreprendre les grands sujets, sans faire attention à ceux qui les auront traités avant moi. bien entendu que je n'aurai jamais la sotte témérité de retraiter les sujets desquels, Corneille, Racine, Crébillon, Molière, ont fait leurs chefs-d'œuvre .

Dans un moment d'enthousiasme causé par la lecture du 2. me vol[ume] du Génie du Christianisme, j'ai pensé à développer le combat de l'amour de dieu et de l'amour, dans le cœur d'une jeune fille passionnée, en un mot à faire une Clémentine protagoniste de Tragédie j'ai pensé tout de suite au prince Zizim p[ou]r amant la scène serait lors à Sassenage près Grenoble.

Jusqu'ici (19 frim[aire] imaginé le 18) je vois ce sujet trop élégiaque p[ou]r la tragédie. Il faudrait p[ou]r faire de ce sujet un Chef-d'œuvre que tous les événements vinssent du Caractère principal. Il me paraît bien difficile de rendre ce sujet intéressant p[ou]r notre siècle . On plaindra Clé[mentine] comme une folle charmante, si le style est bon, on viendra écouter une seconde fois de jolis sentiments en vers touchants, mais l' âme n'ayant pas été violemment émue on n'y reviendra pas une 3. me la scène dans le temps des Croisades .

Illustration 21 – Page des manuscrits de Stendhal : 1) document numérisé ; 2) transcription pseudo-diplomatique ; 3) transcription linéarisée.

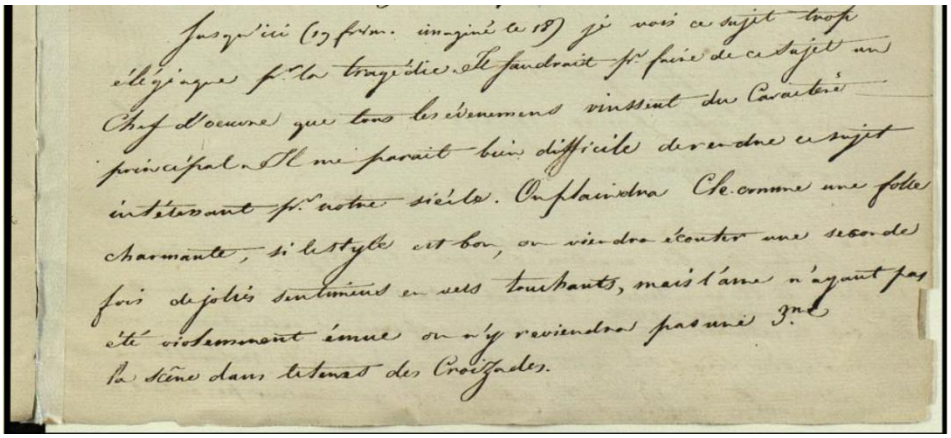


Illustration 22 – Extrait de page

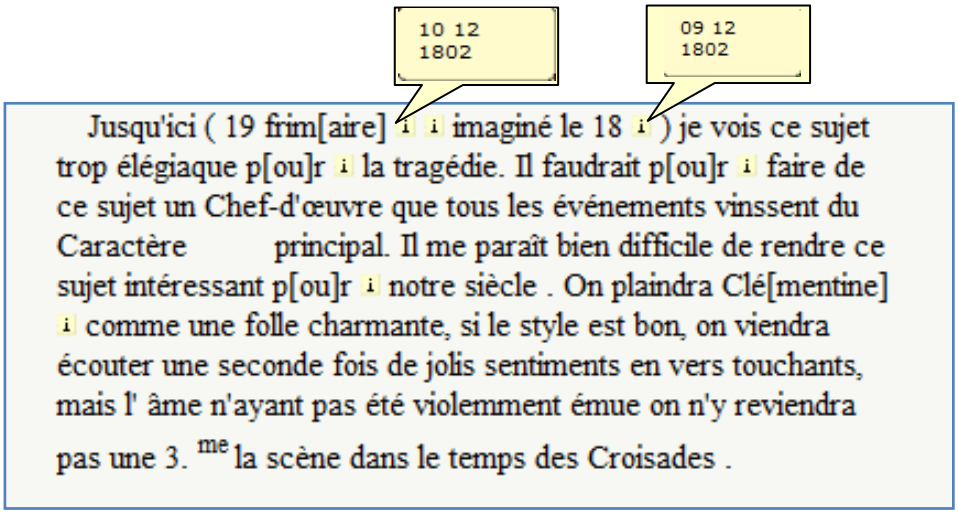


Illustration 23 – Transcription linéarisée de l'extrait de page

6.2 Un modèle indépendant de CLELIA

Nous utiliserons dans notre modèle les fonctionnalités existantes de la base documentaire CLELIA. Cependant le modèle que nous envisageons sera adaptable à d'autres problématiques d'analyse génétique que celle des manuscrits de Stendhal. Nous partons d'un corpus existant pour économiser plusieurs années de travail mais si, comme nous l'avons vu, il existe une genèse par auteur, il n'existe pas nécessairement un modèle d'outil d'analyse génétique par auteur. Nous prévoyons donc de développer un outil d'après notre modèle sur la base CLELIA puis de le transporter sur d'autres bases le cas échéant.

6.3 Une base SQL et des fichiers XML

CLELIA repose sur une base de données relationnelle comme le programme MUSE que nous avons vu. Au départ, nous ne partons pas d'un ensemble documentaire déjà formé (comme le serait un paquet de 60 cahiers par exemple). En effet, puisque les feuillets ont été mélangés, il a été choisi de prendre justement le feuillet (ou plus exactement la page recto et la page verso du feuillet) comme centre de la logique de la base de données relationnelle. Nous avons des tables reliées entre elles, chaque table contenant des informations spécifiques, ce qui permet de regrouper ou de recouper les informations qui nous intéressent ou de le faire faire à un programme informatique. Voici l'organisation de la base de données relationnelle CLELIA organisée autour de la table « Page » correspondant au feuillet, point de départ du modèle[80].

[80] Image donnée par T. Lebarbé le 25 mai 2010

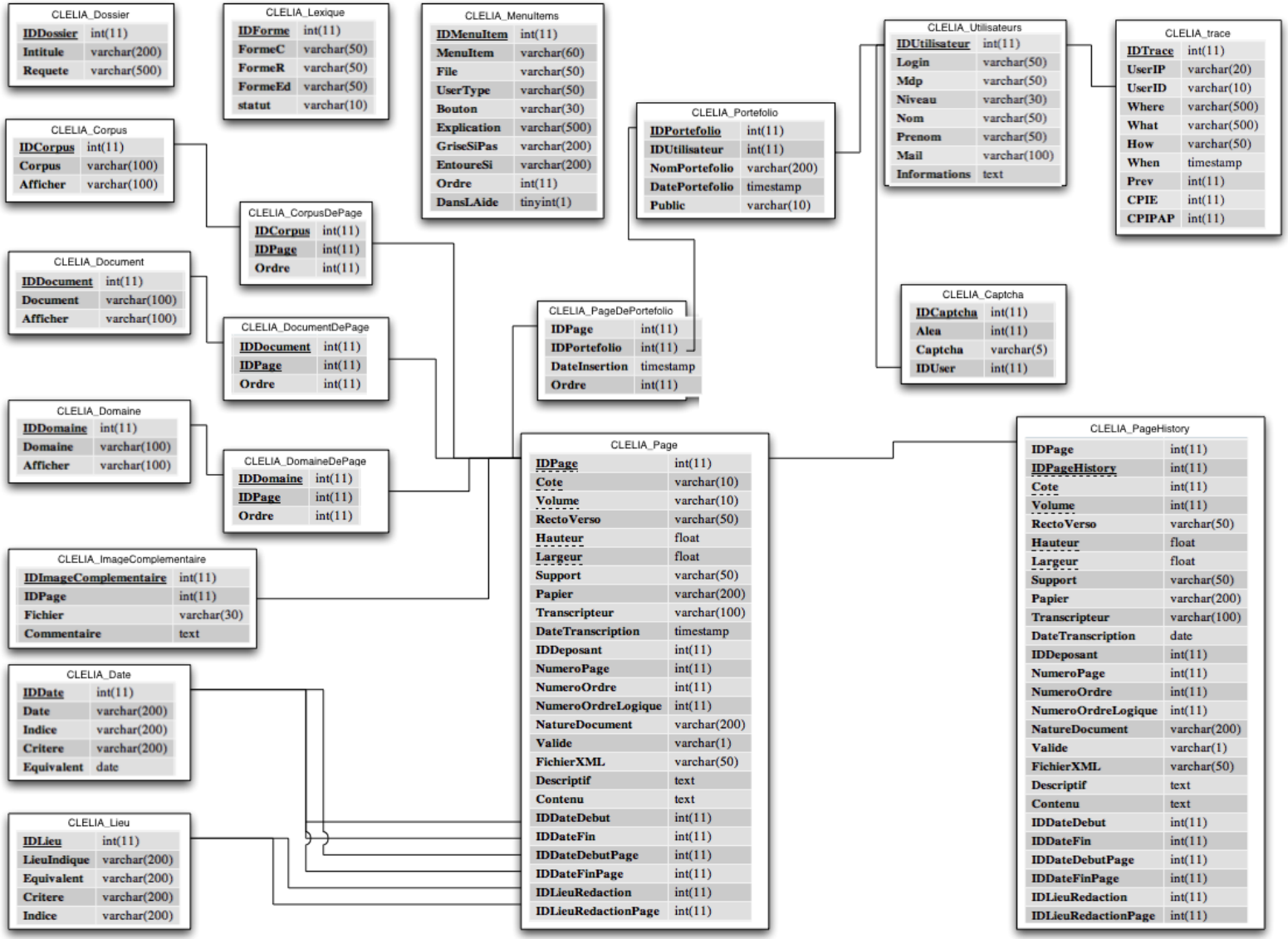


Illustration 24 – Organisation de la base CLELIA

Le bloc de données textuelles accompagnant chaque feuillet se décompose en deux parties principales : un en-tête de méta-données permettant le référencement et donnant des renseignements sur la page, et un corps contenant la transcription elle-même. Chaque page est décomposée en neuf parties : la zone centrale principale, les quatre coins de la page, les quatre côtés (marges latérales, supérieure et inférieure). Chacune de ces zones peut contenir des éléments textuels (blocs de texte, blocs de citation, paginations, foliotations, marginales, notes, etc.) qui se décomposent en entités identifiables visuellement (paragraphes, lignes, interlignes, figures, tableaux, etc.) et en entités de mises en forme (biffe, calligraphie, etc.). Tous ces éléments peuvent être enrichis d'annotations d'ordre critique (commentaires pour le grand public, pour les spécialistes, pour les membres de l'équipe, identification du scripteur, datation, localisation géographiques, etc.). Un travail codicologique minimal recense des informations matérielles : dimensions des papiers, trous de couture permettant d'identifier des cahiers, etc. Les informations sont stockées dans des tables et permettent des regroupements de documents présentant des caractéristiques communes, selon le principe de base de données relationnelle vu dans le chapitre 5.

Chaque transcription d'une page de feuillet manuscrit correspond à un fichier au format XML. Ainsi les méta-données sont organisées suivant une logique rigoureuse et unique pour toutes les pages de la base. Cette logique est matérialisée par une grammaire réunissant l'ensemble des règles de gestion et de description des pages. Il est possible de rajouter le cas échéant des méta-données qui n'ont pas encore été utiles jusqu'à aujourd'hui mais qui pourraient l'être dans le cadre de nouveaux travaux, ce qui est très intéressant pour notre problématique. Le modèle est donc suffisamment souple pour évoluer potentiellement en fonction de nouveaux besoins. Or justement, nous souhaitons implémenter de nouvelles fonctionnalités requérant des informations qui ne sont pas encore présentes dans la base.

6.4 Des fonctionnalités à ajouter

Pour notre problématique nous partirons donc de la base CLELIA et ajouterons quelques fonctionnalités. Nous préciserons en premier lieu la langue utilisée, puisque Stendhal change de langue en cours de rédaction. Voici exactement comment il est possible d'ajouter cette fonctionnalité en fonction de l'organisation et de la nature de la

base de données CLELIA. Nous proposons pour cela de modifier la grammaire (DTD) dont un extrait[81] est présenté par l'illustration ci-dessous. Un bloc traduction peut être défini dans le texte comme contenant du texte et de sa mise en forme ; nous y associssions une traduction du passage en langue étrangère et trois niveaux de commentaires.

```
<!ELEMENT script %contenu_de_ligne;>
   <!ATTLIST script commentaire CDATA #IMPLIED>
   <!ATTLIST script commentaire_public CDATA #IMPLIED>
   <!ATTLIST script commentaire_scientifique CDATA #IMPLIED>

<!ELEMENT calligraphie %contenu_de_ligne;>
   <!ATTLIST calligraphie commentaire CDATA #IMPLIED>
   <!ATTLIST calligraphie commentaire_public CDATA #IMPLIED>
   <!ATTLIST calligraphie commentaire_scientifique CDATA #IMPLIED>

<!ELEMENT traduction  %contenu_de_ligne;>
   <!ATTLIST traduction traduction CDATA #IMPLIED>
   <!ATTLIST traduction commentaire CDATA #IMPLIED>
   <!ATTLIST traduction commentaire_public CDATA #IMPLIED>
   <!ATTLIST traduction commentaire_scientifique CDATA #IMPLIED>

<!ELEMENT italique  %contenu_de_ligne;>
   <!ATTLIST italique commentaire CDATA #IMPLIED>
   <!ATTLIST italique commentaire_public CDATA #IMPLIED>
   <!ATTLIST italique commentaire_scientifique CDATA #IMPLIED>
```

Illustration 25 – Extrait de la DTD des Manuscrits de Stendhal

Cette grammaire comporte 700 lignes. Or il suffira d'insérer une seule ligne pour récupérer systématiquement cette information génétique capitale. Nous ajouterons « traduction langue » et ajouterons au fur et à mesure qu'une langue est rencontrée, un nouvel élément à la liste des langues. Cet exemple sera suivi de nombreux autres en fonction des recherches génétiques souhaitées.

Pour notre problématique nous avons donc avec la base documentaire CLELIA un outil évolutif et pouvant servir de support à différents travaux. Nous allons maintenant exposer les perspectives de travaux ; ils se situent à trois niveaux d'analyse génétique différents, nous procéderons par ordre de granularité, de la plus grosse à la plus fine.

[81] Extrait de la DTD des Manuscrits de Stendhal, la DTD complète se trouve en annexe

Chapitre 7 – Le niveau 1 ou la macrogénétique : les bibliothèques d'auteurs

Comme nous l'avons vu, la bibliothèque personnelle de l'écrivain intéresse grandement le généticien car elle l'aide à comprendre les influences générales de l'auteur. Nous avons retrouvé cet intérêt dans l'équipe des Manuscrits de Stendhal :

> « Au sujet de la bibliothèque de Stendhal, nous ne disposons pas d'un inventaire enrichi d'informations. Il serait intéressant d'établir la liste des œuvres qu'il a lues, ou vues avec si possible une date, même approximative, de sa lecture. »
>
> Cécile Meynard

Ce sont d'une part les lectures elles-mêmes qui représentent une information importante, et d'autre part les annotations de la part du lecteur-écrivain dont les livres regorgent, qui constituent de précieux éléments à recueillir. L'objectif est maintenant de proposer un modèle efficace de bibliothèque virtuelle pour permettre ces recherches que nous qualifierons de *macrogénétiques*.

7.1 Les sources

Tout d'abord nous proposons de constituer virtuellement un groupe de livres comprenant des ouvrages issus de plusieurs sources. Il peut agit de livres empruntés à une bibliothèque ou bien de la collection privée du lecteur-écrivain. Puis pour chaque source, nous observerons le format : s'agit-il d'une liste sous format HTML, d'un fichier PDF, etc. À titre d'exemple voici deux extraits de listes bibliographiques des lectures de Flaubert.

LA BIBLIOTHÈQUE D'*HÉRODIAS*[1]

AGOSTINI (Leonardo) : *Les Pierres antiques*, Rome, 1657-1669, 2 volumes
ARAGO (François) : *Astronomie populaire par François Arago… publiée d'après son ordre sous la direction de M. J.-A. Barral,* Paris, Gide ; Leipzig, T. O. Weigel, 1854-1857, 4 volumes.

BAILLET (Adrien) : *Les Vies des saints… disposées selon l'ordre des calendriers et des martyrologes avec… l'histoire des autres festes de l'année*, Paris, Librairie Roulland, 1701-1703.

BANIER (Abbé Antoine) : *Cérémonies et coutumes religieuses de tous les peuples du monde*, Frères Chatelain, Amsterdam, 1743.

BASNAGE DE BEAUVAL (Jacques) : *Antiquitez judaïques, ou Remarques critiques sur la Républiques des Hébreux*, Paris, 1713[2].

CHAMPAGNY (François-Joseph-Marie-Thérèse Nompère, Comte Franz de) : *Rome et la Judée au temps de la chute de Néron*, Paris, Lecoffre, 1858.

CHAMPAGNY (François-Joseph-Marie-Thérèse Nompère, Comte Franz de) : *Les Antonins, ans de J.-C. 69-180,* Paris, A. Bray, 1863, 3 volumes.

CHASSAN (P.) : *Essai sur la symbolique du Droit*, Paris, Vidocq, 1847.

[1] Par la *Correspondance*, on sait que la rédaction de *Trois Contes* est contemporaine de la relecture par Flaubert de « ses » grands classiques, Rabelais, Voltaire et Shakespeare, auxquels on pourrait ajouter le Marquis de Sade, Hugo, Racine et d'autres encore que l'on retrouve, par effets d'intertextualité (parodie ou imitation) dans les trois nouvelles. Raymonde Debray Genette retrouve dans *Hérodias* divers éléments empruntés à l'*Odyssée* d'Homère (épisode de la grotte de Calypso) et réemployés pour la description des écuries de Machærous (Raymonde Debray Genette, « Les écuries d'*Hérodias* ; genèse d'une descriptions », *in Genesis* n°1, ITEM / CNRS, Jean-Michel Place, Paris, 1982, p. 103 et n. 28).

Illustration 26 – Bibliothèque de Flaubert – 1

L'illustration ci-dessus correspond à un ouvrage imprimé, tandis que l'illustration suivante correspond à une liste mise en ligne (d'ouvrages empruntés par Flaubert à la Bibliothèque Nationale entre 1870 à 1880[82]).

[82] Pour la liste complète, aller à http://flaubert.univ-rouen.fr/bibliotheque/05bnf.php

Date emprunts/ retour	Mentions portées aux registres de prêts et titres des ouvrages entièrement restitués.	Emargements
6 janv. 1870 R 10 mars	*S^ti Epiphani opera gr. lat. rec. Petavius... Col, 1682.* [Τοῦ... Ἐπιφανίου... ἅπαντα τὰ σῳζόμενα Sancti.. Epiphanii opera omnia... Dionysius Petavius... recensuit, latine vertit et animadversionibus illustravit... Editio nova... cui accessit vita Dyonisii Petavii ab H. Valesio ... – Coloniae (Lipsiae), sumptibus J. Schrey et H.J. Meieri, 1682. 2 vol.in-fol.]	G. Flaubert
4 janv. 1878 R 10 janv.	*Daunou. Cours d'études histor.* 1842, *T.* 1-2. [DAUNOU (Pierre-Claude-François). – Cours d'études historiques... [publié par A Taillandier, etc.] – Paris, 1842-1845 20 vol.in-8°.] – *B. P.*, 154.	G. Flaubert
4 janv. 1878 R 10 janv.	*Rollin. Œuvres.* 1824. *T.* 17, 18. [ROLLIN (Charles). – Œuvres complètes de Rollin. Nouvelle édition... par M. F. Guizot... – Paris, E.-A. Lequien, 1824. In-°. – *Les tomes XVII et XVIII contiennent* : l'Histoire romaine.] – *B. P.*, 150.	G. Flaubert
4 janv. 1878 R 10 janv.	*Anquetil. Hist. de France, 2^e éd.* 1829. *T.* 1-2. [ANQUETIL (Louis-Pierre). – Histoire de France depuis les Gaulois jusqu'à la mort de Louis XVI... Continuée jusqu'au sacre de S.M. Charles X par M. Léonard Gallois,... [et jusqu'à l'avènement de Louis-Philippe par M. N.-A. Dubois]... 2e éd. – Paris, Jubin, 1829-1831. 15 vol. in-8°.] – *B. P.*, 146.	G. Flaubert
11 janv. 1878 R 22 janv.	*Legendre. Traité de l'opinion. 1733. T.* 1-6. [LE GENDRE DE SAINT AUBIN (Mis Gilbert-Charles). – Traité de l'opinion, ou Mémoires pour servir à l'histoire de l'esprit humain [par Legendre de Saint-Aubin] – Paris, C. Osmont, 1733. 6 vol. in-8°.]	G. Flaubert

Illustration 27 – Bibliothèque de Flaubert – 2

En fonction des résultats nous préférerons soit n'avoir qu'un seul format auquel cas nous convertirons certaines sources, soit créer un modèle acceptant plusieurs formats comme par exemple celui que nous avons vu avec Nietzsche Source.

7.2 Une typologie de relations entre auteurs et ouvrages

Nous créerons une typologie de relations entre auteurs et ouvrages en nous appuyant sur les travaux d'un projet collaboratif qui se met actuellement en place. Ce projet, qui n'a pas encore de nom, se met en place dans une collaboration entre le

GERCI de Grenoble (Ch. Del Vento), l'ENSSIB Lyon (R. Mouren), le Conseil général Nord-Pas-de-Calais (I. Westeel) et le LIDILEM de Grenoble (T. Lebarbé). Il vise à centraliser et mutualiser des bibliothèques d'auteurs. L'illustration suivante, qui nous montre comment relier auteurs et ouvrages, ainsi que son explication détaillée sont de Thomas Lebarbé.

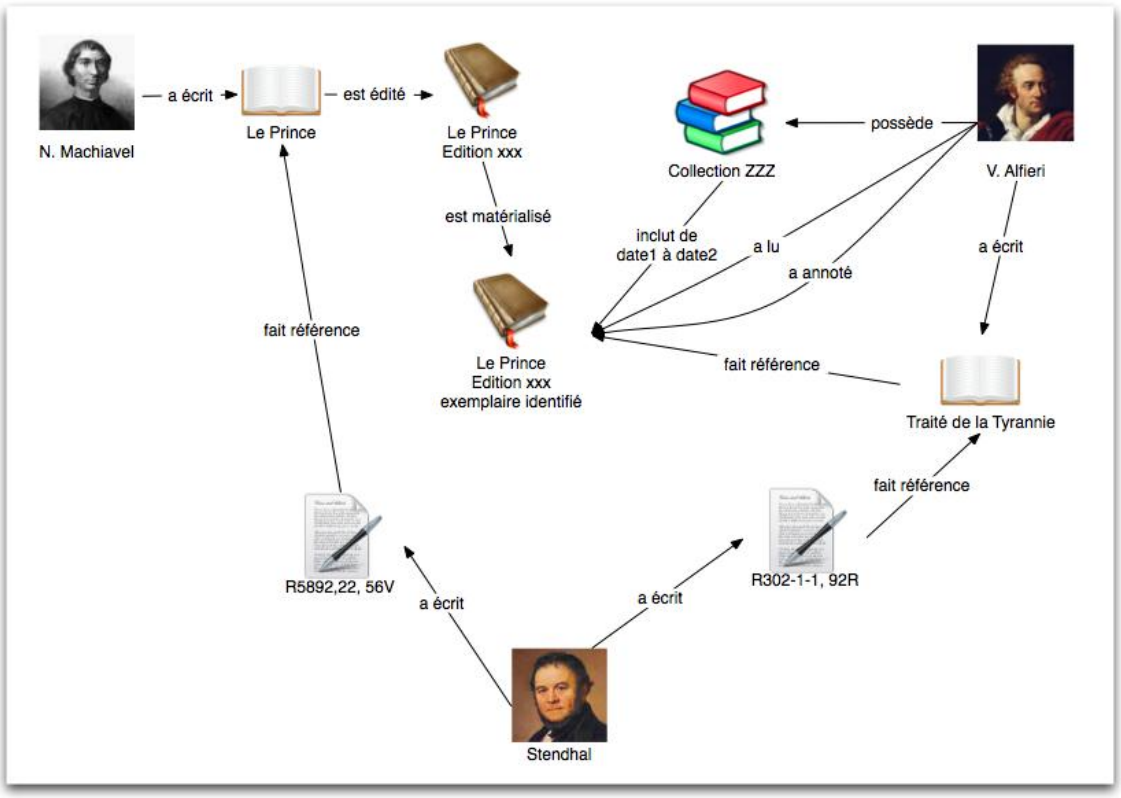


Illustration 28 – Typologie de relations entre auteurs et ouvrages

Explication détaillée de l'illustration :

> « La relation entre auteurs et ouvrages s'effectue en plusieurs étapes. Tout d'abord, un spécialiste enregistre que :
>
> - Machiavel a écrit une œuvre intitulée *Le prince* (notion conceptuelle de l'œuvre)
>
> - Éditée par XXX (notion matérielle de l'œuvre – trace de l'objet)
>
> - Matérialisée par un exemplaire en particulier (géolocalisé).
>
> Puis un spécialiste enregistre que :

- Alfieri possède une collection qui, de telle à telle date contient un exemplaire de la précieuse édition xxx du *Prince*

- Qu'il l'a lue et annotée (les annotations sont elles-mêmes enregistrées)

- Qu'en écrivant son traité de la tyrannie il fait explicitement référence au Prince

D'autre part, est ajouté le fait que Stendhal, dans ses manuscrits, fait d'une part référence au Prince de Machiavel (mais nous ne savons pas quelle édition) et d'autre part à *La Tyrannie* d'Alfieri (pourrons mettre à jour cette information ultérieurement en fonction de nouvelles découvertes).

Enfin, le réseau met en évidence des coïncidences et permet de poser des hypothèses (que les spécialistes vérifieront), par exemple de causalité : Stendhal a-t-il lu le prince suite à sa lecture du *Traité de la Tyrannie* ? Ou le *Traité de la Tyrannie* lui a-t-il donné une autre lecture du prince ce qui aurait eu une influence sur sa *Constitution voulue par le peuple*[83] ? »

Thomas Lebarbé

Pour notre propre problématique nous identifierons les ouvrages lus mais non annotés, les ouvrages annotés, les ouvrages empruntés à une bibliothèque ou à un ami ou bien les ouvrages possédés par le lecteur-écrivain. De même, nous constituerons une liste de livres « cités » par le lecteur-écrivain. Puis dans cette première typologie, nous pourrons inclure d'autres typologies en fonction de la répartition déjà réalisée par l'écrivain s'il y en existait une d'identifiable : ordre alphabétique, thématique, chronologique, ordre de préférence, etc. Enfin, sans effectuer de mise en abyme, nous voulons également établir une liste des ouvrages cités par les personnages de notre lecteur-écrivain. Nous pourrons nous inspirer à ce propos des travaux de l'Université de Rouen dans la cadre du projet *Hyper Flaubert*[84] qui consiste à réaliser une bibliothèque virtuelle de l'écrivain.

[83] Feuillet 5896-5, 6R
[84] http://flaubert.univ-rouen.fr/bibliotheque/

7.3 Le catalogue

Lorsque nous aurons déterminé les critères de répartition nous constituerons un catalogue. Nous le réaliserons en organisant les informations en fonction de : l'auteur, le titre, l'éditeur, la date, le nombre de pages, le format en centimètres, la collection, les conditions physiques de l'ouvrage, tout ce qui atteste de la propriété de l'œuvre par le lecteur-écrivain (reçus, dédicaces, etc.), le lieu où se trouve actuellement le livre, la liste des pages où se trouvent des annotations et la description de telles annotations.

Nous pourrons à ce sujet nous appuyer sur les travaux effectués sur le lecteur-écrivain Nietzsche. En effet, il existe un catalogue publié à Berlin dans la collection *Supplementa Nietzscheana* contenant la liste de toutes les publications dont Nietzsche a été possesseur. Elle est constituée de : l'inventaire raisonné de la bibliothèque personnelle de Nietzsche conservée à la *Herzogin Anna Amalia Bibliothek* de Weimar, du déchiffrement des reçus des libraires relatifs aux achats de Nietzsche conservés aux archives *Goethe-Schiller* de Weimar, et de la consultation des catalogues précédents et des registres internes à la *Herzogin Anna Amalia Bibliothek*.

7.4 La navigation ou l'exploitation

Lorsque toutes ces options auront été établies et seulement lorsqu'elles l'auront été, nous créerons le modèle d'exploitation du catalogue. Nous choisirons comment nous voulons naviguer et quelles sont exactement les informations que devra prendre en compte l'outil de recherche. Nous observerons de nouveau comment travaille le généticien sur cet aspect précis de la bibliothèque d'auteur. Or pour que le généticien le sache, nous proposerons le catalogue sous un format provisoire, puis travaillerons en collaboration et à force d'allers retours entre la programmation des fonctionnalités et de leurs essais par les chercheurs en littérature, nous affinerons la navigation dans la bibliothèque jusqu'à obtenir la pertinence suffisante.

Pour ce premier niveau de recherches macrogénétique que représentent les bibliothèques d'auteurs, le caractère le plus structurant semble être la variété des sources, et ce quelles que soient les informations précises à glaner ou le format sous lequel nous souhaitons les exploiter. C'est pourquoi, pour réaliser un tel travail, nous envisageons non seulement un travail collaboratif entre différents laboratoires de recherches mais encore des enquêtes auprès de nombreuses bibliothèques et autres instances dans toute la France voire dans plusieurs pays d'Europe.

Chapitre 8 – Le niveau 2 ou la génétique : les influences théâtrales

A un niveau plus fin, que nous qualifierons de *génétique* tout simplement, nous avons noté le besoin de chercheurs en littérature de retrouver la liste des pièces de théâtre que Stendhal aurait lues ou vues. Le champ de recherche est plus restreint que pour les bibliothèques d'auteurs car le corpus est ici constitué seulement des 20 000 feuillets de la Bibliothèque municipale. Il s'agit maintenant d'extraire des titres avec l'aide d'un programme informatique.

8.1 L'extraction manuelle de titres d'œuvres théâtrales

Comment s'y prendre pour faire rechercher des titres de pièces de théâtre à un programme informatique ? Cherchons tout d'abord à effectuer la manœuvre à la main et notons les étapes.

8.1.1 L'étape 1 : recherche avec une amorce de départ

Tout d'abord, parcourons le corpus avec le moteur de recherches afin qu'il pointe une pièce de théâtre. Par exemple si nous recherchons les mots « j'ai lu », il y a des chances que ces derniers soient suivis du nom d'une œuvre, qu'elle soit théâtrale ou non. Nous établissons ainsi une liste de plusieurs types d'œuvres dont nous pouvons stocker les titres. Contenant certainement aussi des titres d'ouvrages non relatifs au théâtre, cette liste nous permettra au passage de compléter une éventuelle bibliothèque d'auteur.

Avec la recherche de la séquence « j'ai lu », nous obtenons :

- occurrence n°1 : ***J'ai lu*** *l'Art poétique d'Horace* (trouvée sur le recto du feuillet 185 du tome 1 du volume 1 du Registre 302) ;

- occurrence n°2 : ***J'ai lu*** *le 22 Floréal an 11 pour la première fois l'Œdipe de Sophocle* (trouvée sur le recto du feuillet 52 du volume 27 du Registre 5896).

Nous avons maintenant deux titres à stocker : *Art poétique* et *Œdipe*. Nous appellerons ces mots-clés des *amorces* puisqu'ils précèdent l'objet de la recherche, pour ne pas les confondre avec des mots-clés *directs* qui eux ne précèdent pas la recherche, mais la constituent en eux-mêmes, comme par exemple le mot-clé « Grenoble » que nous avons

vu dans le Chapitre 6. Nous créons avec notre résultat une liste de titres, voire d'auteurs s'ils sont mentionnés.

- ***Titre*** *(Auteur)*
- **Art poétique** (Horace)
- **Œdipe** (Sophocle)

Illustration 29 – Liste de titres

Recherchons maintenant, sans amorce mais directement, les termes *Art poétique* et *Œdipe* dans l'espoir d'allonger notre liste.

8.1.2 L'étape 2 : recherche directe avec les premiers titres

Par exemple, choisissons *Œdipe* comme mot-clé pour une recherche directe. Nous lançons la recherche et obtenons deux occurrences :

- occurrence n°1 : *Je sors d'**Œdipe** suivi de l'Amant bourru.* (trouvée sur le verso du feuillet 46 du volume 22 du Registre 5896) ;

- occurrence n°2 : *J'ai lu le 22 Floréal an 11 pour la première fois l'**Œdipe** de Sophocle* (trouvée sur le recto du feuillet 52 du volume 27 du Registre 5896).

Nous pouvons compléter notre liste. Nous ajoutons l'occurrence n°1 (pour laquelle nous n'avons pas d'auteur) et les références du feuillet où nous avons trouvé l'occurrence pour ne pas la confondre avec son homonyme. Ajoutons également la date lorsque nous la trouvons, puisque comme nous l'avons vu, il s'agit d'une information cruciale en génétique. Nous avions déjà trouvé l'occurrence n°2.

- ***Titres***	*(Auteurs)*	*(Feuillet)*	*(Date)*
- Art poétique	(Horace)	(-)	(-)
- Œdipe	(Sophocle)	(-)	(-)
- Œdipe	(-)	(R. 302 Rés., vol1, tome 1, feuillet 185, recto)	(1804 04 26)

Illustration 30 – Liste de titres 2

Maintenant, effectuons la fouille des contextes de ces occurrences dans le but de trouver d'autres amorces pour chercher de nouveaux titres qui eux ne sont pas précédées de *j'ai lu*[85].

8.1.3 L'étape 3 : fouille des contextes

Nous trouvons dans le contexte gauche de l'occurrence n°1 la nouvelle amorce « je sors » et dans son contexte droit le nouveau titre « Amant bourru ». Notons qu'à ce moment de la réflexion il est trop tôt pour choisir la taille et l'unité de mesure des contextes. Nous pouvons compléter notre tableau de titres.

Titres	(Auteurs)	(Feuillet)	(Date)
- Art poétique	(Horace)	(-)	(-)
- Œdipe	(Sophocle)	(-)	(-)
- Œdipe	(-)	(R. 302 Rés., volume1, tome 1, feuillet 185, recto)	(1804 04 26)
- Amant Bourru	(-)	(R. 302 Rés., volume1, tome 1, feuillet 185, recto)	(1804 04 26)

Illustration 31 – Liste de titres 3

Par ailleurs nous pouvons également commencer une liste d'amorces que nous pourrons réutiliser par la suite.

Amorces

- J'ai lu

- Je sors

Illustration 32 – Liste d'amorces

Par ces trois étapes manuelles réitérées cycliquement, nous pouvons extraire une longue liste d'œuvres. Le principe est simple, au fur et à mesure que nous trouvons de nouvelles amorces, nous pouvons relancer des recherches de titres.

[85] Notons que l'outil de recherche est insensible à la casse donc « J'ai lu » et considéré comme « j'ai lu ».

Nous souhaitons maintenant faire faire la recherche à un programme informatique.

8.2 L'extraction automatique : schématisation et contraintes

8.2.1 La schématisation de l'extraction automatique

Reprenons les trois étapes de notre cycle de recherche. L'illustration de l'étape 1 représente notre recherche automatique avec l'amorce de départ j'ai lu. Nous obtenons deux résultats *Art poétique* d'Horace et *Œdipe* de Sophocle que nous stockons. Nous extrayons l'amorce « j'ai lu » et nous la stockons dans la table d'amorces puis nous extrayons les deux titres que nous stockons dans la table des titres.

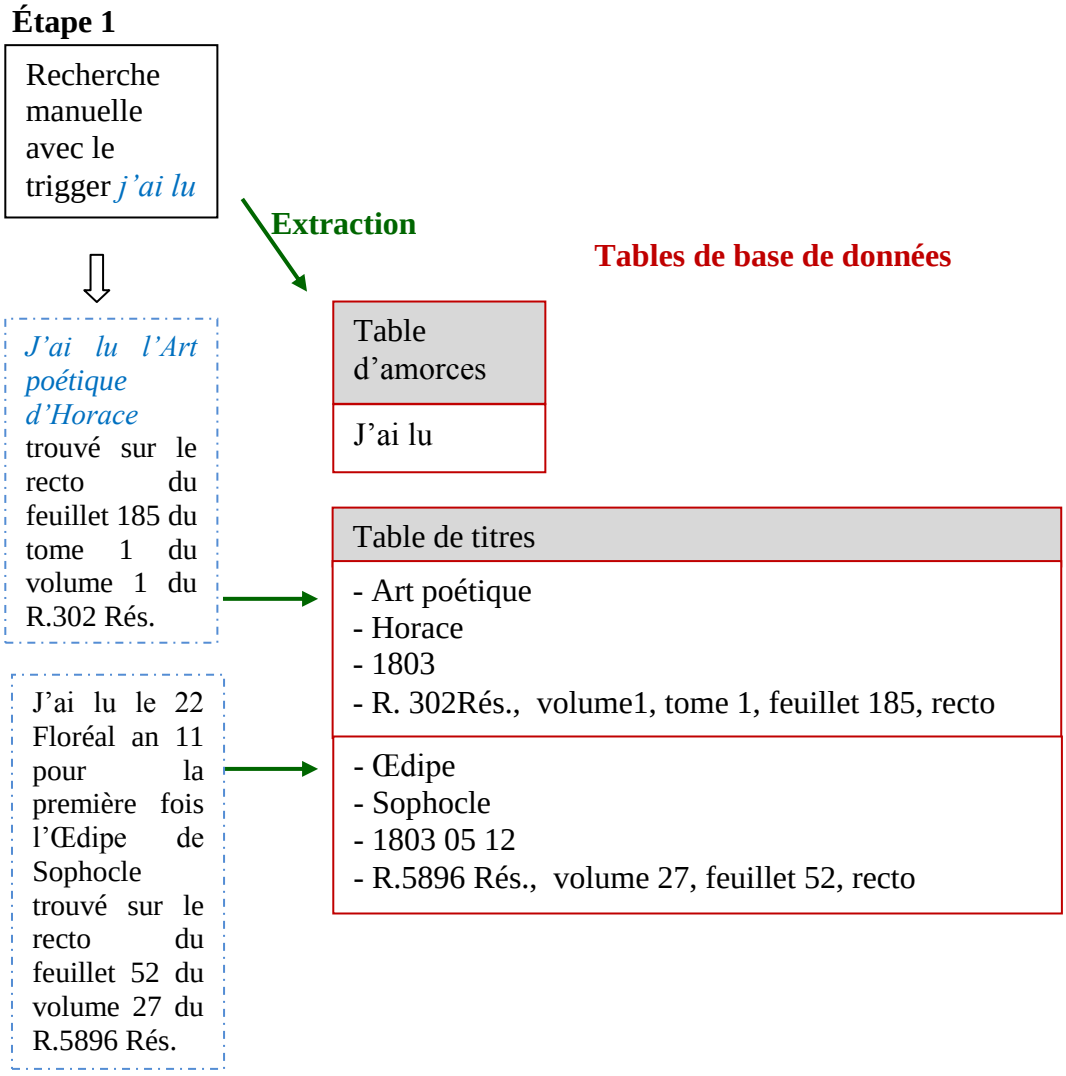


Illustration 33 – Extraction manuelle de titres d'œuvres théâtrales – étape 1 : recherche avec une amorce de départ

L'illustration suivante décrit la deuxième étape. Nous lançons une deuxième recherche mais cette fois directe (et non par amorce) avec le mot-clé *Œdipe*. Nous obtenons *Je sors d'**Œdipe** suivi de l'Amant bourru* et *J'ai lu le 22 Floréal an 11 pour la première fois l'Œdipe de Sophocle*. Nous extrayons les deux *Œdipe* et stockons seulement le deuxième (puisque nous avons déjà le premier) dans la table de titres.

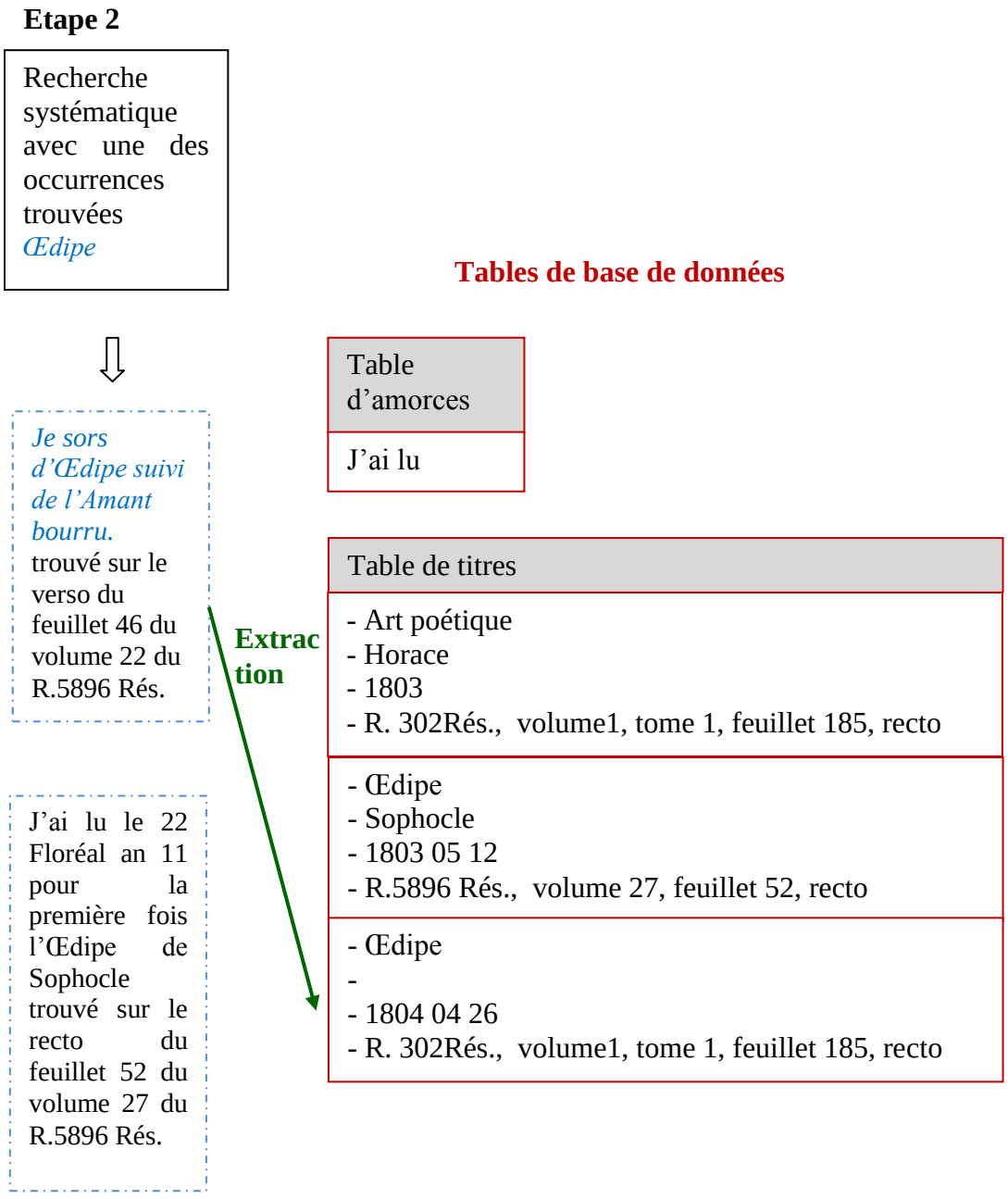


Illustration 34 – Extraction manuelle de titres d'œuvres théâtrales – étape 2 : recherche directe avec les premiers titres

L'illustration de l'étape 3 décrit la fouille des contextes. Nous fouillons le contexte gauche et obtenons *Je sors* puis le contexte droit et obtenons *Amant bourru*. Nous extrayons les occurrences et les stockons dans leurs tables respectives.

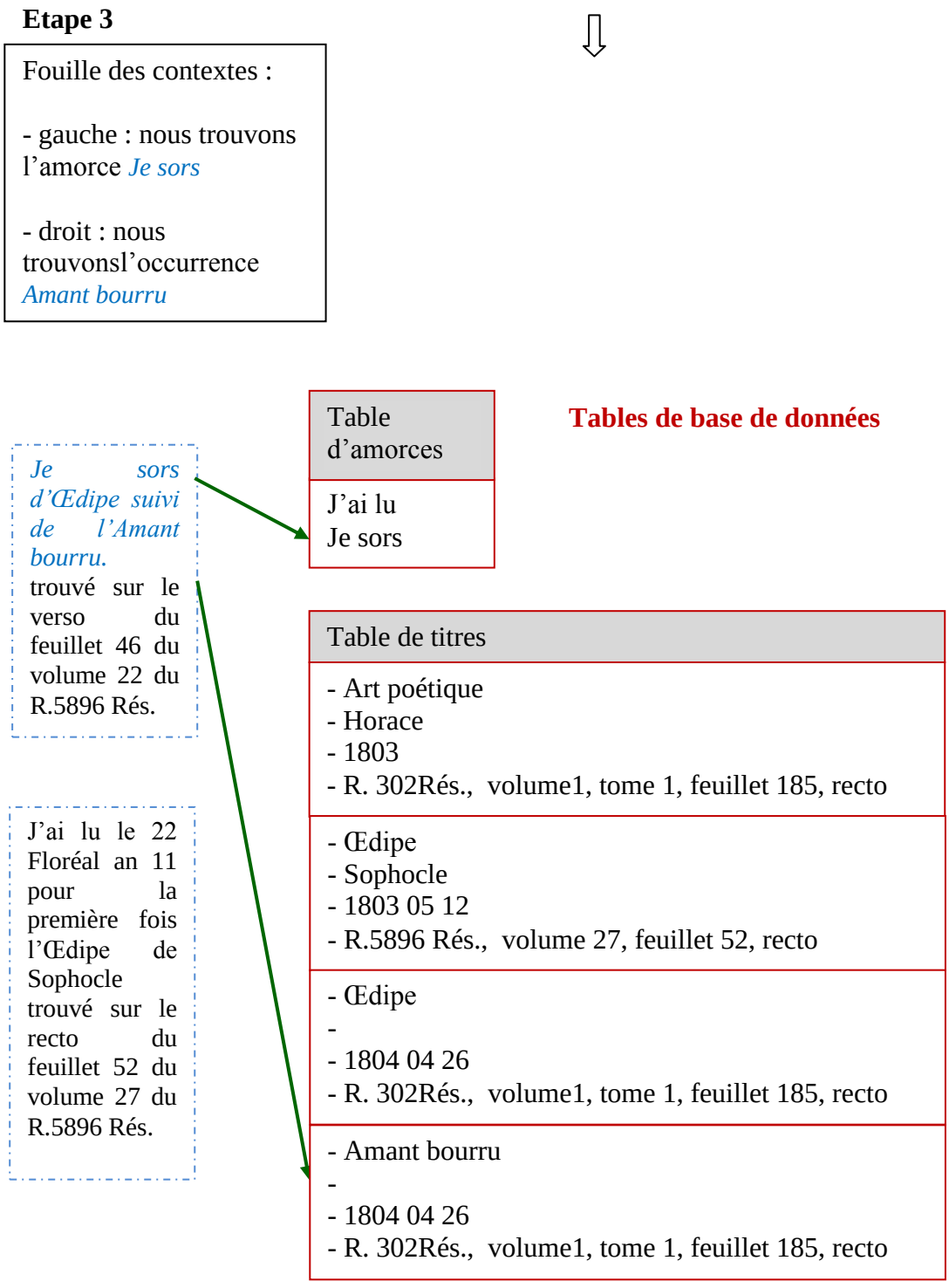


Illustration 35 – Extraction manuelle de titres d'œuvres théâtrales – étape 3 : fouille des contextes

Nous venons de déterminer de façon naturelle ce qui constitue une amorce et ce qui constitue un titre d'œuvre. Or ce qui semble évident pour un œil humain (et encore pas toujours) est difficile à décrire avec des règles univoques nécessaires à un programme informatique. Ce que nous venons de réaliser spontanément n'est pas possible automatiquement en l'état actuel de l'outil. Nous proposons pour résoudre ce problème de poursuivre les travaux selon deux axes.

8.2.2 Les contraintes de l'extraction automatique

Pour informatiser l'extraction que nous venons d'effectuer manuellement, plusieurs contraintes seront à prendre en compte. Reprenons les trois étapes de l'extraction mais cette fois, dans une perspective d'automatisation.

Dans la première étape, nous pouvons aisément écrire un programme (un script en l'occurrence) qui stockera la première amorce dans une table de base de données et qui effectuera une recherche avec les termes « j'ai lu ». Nous déterminerons comment définir le titre. Pour réaliser cette tâche nous devrons déterminer comment repérer le titre. Nous commencerons par répondre à plusieurs questions : faut-il rechercher après l'amorce le premier mot commençant par une majuscule puis extraire les mots suivants ? Dans ce cas, combien de mots exactement ? Et que faire dans le cas de deux titres ? Et s'il y a plusieurs majuscules ? De même comment extraire avec certitude le nom de l'auteur ? Un nom d'auteur commence par une majuscule, mais un titre aussi et qu'en est-il si Stendhal ne met pas de majuscule ?

Pour poursuivre les travaux en ce sens, nous proposons tout d'abord de nous orienter vers un outil d'*aide* à l'extraction automatique de titres d'œuvres (puis de pièces de théâtre) plutôt que d'opter vers le tout automatique. C'est-à-dire que nous proposons de prévoir une vérification humaine. En effet celle-ci permettrait, d'une part de superviser l'identification correcte des titres, auteurs et dates, (vérification faite, la *référence* quant à elle se trouvant dans une balise tout à fait identifiable ne devrait pas poser de problème) et d'autre part d'éviter que le programme ne s'emballe. Ce risque existe en effet puisque le programme se nourrit automatiquement de nouvelles amorces et relance les recherches avec les amorces trouvées.

Ensuite, nous proposons de chercher les références puis de lire les travaux de M. Ehrmann, G. Jacquet ou encore C. Roux (société Xerox) et ceux de T. Lebarbé (laboratoire LIDILEM) sur le sujet.

Chapitre 9 – Le niveau 3 ou la microgénétique : les traductions

Le troisième niveau de recherche se situe à la hauteur du texte cette fois-ci, à un niveau *microgénétique* donc, puisqu'il s'agit de la rédaction de l'écrivain. En effet, après des années à étudier les manuscrits de Stendhal, une chercheuse a eu les intuitions suivantes : le sabir, c'est-à-dire le fait d'exprimer sa pensée dans une autre langue participerait de la genèse, et serait très souvent à la première étape de cette genèse.[86] Nous voudrions vérifier cette hypothèse ou du moins initier les recherches en s'aidant de l'informatique et du traitement automatique des langues.

9.1 Une méthode d'extraction dans des méta-données

Il s'agit de trouver une méthode d'extraction automatique d'informations présentes dans les méta-données de notre corpus en vue de la conception d'un outil utilisable par les spécialistes en littérature pour l'analyse de la genèse d'œuvres littéraires. En effet, comme nous l'avons vu dans le Chapitre 6, les informations concernant les traductions se trouvent dans notre corpus au niveau des méta-données, c'est-à-dire des informations ajoutées pour enrichir les transcriptions. Nous allons donc développer un outil de repérage des balises <traduction/>.

9.2 Le sabir stendhalien

Pour rechercher des informations sur le sabir, encore faut-il avoir une idée précise de ce dernier. Nous avons donc commencé par consulter le Trésor de la Langue Française Informatisé[87]. Nous y avons trouvé que le sabir en linguistique est une « langue mixte, généralement à usage commercial, née du contact de communautés linguistiques différentes »[88]. Puis nous avons interrogé T. Lebarbé : « il faudrait chercher si le sabir pour les linguistes correspond au phénomène linguistique

[86] Cécile MEYNARD & Muriel BASSOU. Interview du 25 février 2010, retranscrite en Annexe 1.
[87] http://atilf.atilf.fr/dendien/scripts/tlfiv4/showps.exe?p=combi.htm;java=no
«Sabir», *Trésor de la Langue Française Informatisé* [En ligne].
http://www.universalis-edu.com/corpus2.php?recherche_mode=alpha&nref=L110231#07000000
[88] Cf. définition complète en annexe

d'alternance codique ou de code-switching ». Enfin, après consultation de l'*Encyclopaedia universalis*[89], nous avons trouvé dans un article rédigé sous la houlette d'Andrée Tabouret-Keller, professeur à l'Université Louis-Pasteur de Strasbourg, que le code-switching correspondait à l'alternance de langues[90]. Nous considérerons comme « sabir » toute alternance de langues.

9.3 Des informations connexes

Pour établir la liste des informations connexes nécessaires au chercheur littéraire, nous avons interrogé Cécile Meynard et Lucy Garnier (post doctorante du laboratoire Traverses 19-21 de l'Université Stendhal, Grenoble 3). La liste des informations à extraire en plus du sabir serait constituée de la langue du sabir, le ou les corpus, les domaines, le nom du document, la ou les dates de rédaction du document ainsi que le ou les lieux de rédaction. En tant que linguiste il nous semble pertinent de donner également les contextes gauche et droit. Pour extraire toutes ces informations, nous avons donc développé un outil de traitement informatique des fiches transcrites.

9.4 L'outil de traitement

Le programme informatique que nous avons développé[91] effectue plusieurs tâches successivement. Tout d'abord, le programme lit ligne par ligne le fichier XML représentant la fiche manuscrite de Stendhal retranscrite et enrichie d'informations supplémentaires. Puis le programme parcourt la ligne et « capte » les informations suivantes :

- la langue du sabir utilisée
- le nom de la fiche de manuscrit
- le nom du transcripteur
- le nom du corpus (sous-domaine, entre "domaine" et "document")
- le nom du deuxième corpus, le cas échéant
- le nom du troisième corpus, le cas échéant
- le nom du document
- le nom de la date de la période probable de rédaction ou de la seule date

[89] http://www.universalis-edu.com/article2.php?napp=&nref=O142331
[90] Cf. extrait d'article en annexe
[91] Cf. l'intégralité du programme en annexe

- le nom de la date de fin de période probable de rédaction, le cas échéant
- le nom du lieu de rédaction estimé ou avéré
- la traduction en français du texte étranger rédigé par Stendhal
- le texte de Stendhal en étranger
- le contexte intéressant.

Ce programme ou plus exactement ce script est écrit dans le langage Perl. En sortie nous obtenons un fichier au format texte avec les informations pour chaque sabir repéré. Chacune des informations est isolée dans une colonne. Les informations sont aisément exploitables avec un tableur. Sur les 1 811 fiches analysées, 79 contenaient un ou plusieurs sabirs. Il est à noter qu'un mot exprimé en langue étrangère, même s'il ne s'agit pas d'une alternance de langue, sera considéré par le programme comme un sabir, puisque le mot fera l'objet d'une balise traduction. Par exemple lorsque Stendhal parle dans la fiche *5896-16-024.xml* d'une discussion qu'il a eue sur la comparaison entre le Corn beef et le roti (sans accent circonflexe dans le texte), le terme *Corn beef* est extrait par le programme comme un sabir, alors qu'il ne s'agit manifestement pas là d'une alternance de langue mais de la simple évocation d'un plat anglais.

9.5 La méthode utilisée

Nous remarquons que le programme est capable de récolter toutes les informations sauf une : la langue du sabir ; or cette information est capitale. Nous pensons que la simple fouille des fiches transcrites dans leur état actuel, c'est-à-dire sans la langue ne constitue pas une méthode suffisante pour recueillir les informations nécessaires au chercheur en littérature pour le sabir. *A fortiori* pour une étude sur la genèse d'œuvres littéraires, pour laquelle une multitude d'informations complexes sont nécessaires, cette méthode d'extraction d'information ne nous semble pas adéquate. Nous proposons en conséquence de faire appel aux méthodes de traitement automatique des langues et d'utiliser un outil de détection automatique des langues.

Dans ce chapitre, nous avons constaté qu'il s'avère insuffisant de se contenter d'outils simples, comme la détection automatique de balises de traduction, pour pouvoir un jour élargir la portée de notre arsenal linguistique à l'analyse génétique d'œuvres littéraires. Nous ajouterons par conséquent pour notre problématique des outils de

traitement automatique des langues. Cela signifie qu'il conviendra de changer de démarche linguistique et de définir un nouveau paradigme à la fois plus puissant et plus spécifique à la recherche de traduction. Pour ce faire, nous nous tournerons dans un premier temps, de nouveau vers les spécialistes littéraires pour reprendre la caractérisation de leurs besoins, mais cette fois-ci en ayant une idée très précise de l'organisation informatique de chacune des informations dont nous disposons. Notons également que cette étude microgénétique s'inscrira dans le cadre d'un travail collaboratif et interdisciplinaire entre lettres, linguistique et informatique.

Conclusion

Pour conclure cette partie sur les perspectives de recherches notons que nous avons, avec la base documentaire CLELIA, un corpus d'expérimentations suffisant pour nos travaux déjà doté d'un moteur de recherche. Cet outil d'extraction assistée par le TAL permettra aux niveaux macrogénétique et génétique de constituer facilement une bibliothèque d'auteurs et d'extraire une foule d'informations intéressants les généticiens : des titres d'œuvres, de pièces de théâtre, de noms d'auteurs, des dates, des lieux, etc. Au niveau microgénétique de fouille dans les méta-données, nous proposons une amélioration simple mais efficace de l'outil et un enrichissement en informations nouvelles réutilisables pour d'autres travaux en littérature, en linguisique et en traitement automatique des langues.

Conclusion

Parmi les supports manuscrits étudiés, nous nous intéresserons aux manuscrits provisoires, ceux qui témoignent de la fabrication de l'œuvre, et qui appartiennent aux XVIIe, XVIIIe et XIXe siècles. Nous avons vu que chaque auteur avait sa propre genèse, et en observant les méthodes des généticiens, nous avons découvert la codicologie, les quatre phases de création d'une œuvre et le rassemblement des pièces de corpus ainsi que leur déchiffrement, leur transcription, leur classement et la datation des différents jets, et nous avons également pris connaissance de l'existence de l'Institut des textes et manuscrits modernes qui regroupe une majorité de chercheurs en génétique en France. De l'étude des modèles linguistiques nous retiendrons comme source d'inspiration lors du développement de notre outil, le modèle d'Hayes et Flower et deux aspects de théorie linguistique : la notion de substitution et celle de variante. En ce qui concerne les logiciels dédiés à la génétique étudiés, nous avons, à travers trois d'entre eux, répondu à plusieurs besoins de notre problématique : une organisation en base de données relationnelle et une organisation partant à la base de la numérisation de chaque document pour naviguer vers d'autres informations attenantes. Nous avons testé trois pistes de recherches en nous appuyant sur la base documentaire CLELIA et en utilisant comme domaine expérimental les Manuscrits de Stendhal. Ce corpus est suffisamment conséquent pour les travaux que nous avons prévus et son moteur de fouille de texte actuel permettra d'effectuer nos recherches aux niveaux macrogénétique et génétique : constitution d'une bibliothèque d'auteurs et études d'influences d'autres auteurs en général et d'influences théâtrales en particulier. En revanche, la base documentaire en l'état actuel ne répond pas encore à nos besoins au niveau microgénétique de fouille dans les méta-données. Pour pallier le manque, nous proposons une amélioration simple mais efficace de l'outil ainsi qu'un enrichissement en informations nouvelles indispensables à notre problématique et par ailleurs réutilisables pour d'autres travaux dans les domaines littéraires, linguistiques et informatiques. Notre modèle se situe en effet au centre de recherches pluridisciplinaires. Si notre réflexion est partie de besoins de chercheurs en littérature, en la menant, nous avons découvert que les fonctionnalités de recherche et d'extraction à trois niveaux de granularité créées pour nos propres besoins pourraient être réutilisés pour d'autres recherches : en TAL avec la fouille de titres, en linguistique avec la pensée en langue étrangère versus la pensée en langue

maternelle, et en littérature avec la création de bibliothèques virtuelles de différents auteurs reliées entre elles.

Bibliographie

[Ablali & Kastberg Sjöblom, 2010]
Driss ABLALI & Margareta KASTBERG SJÖBLOM (2010). *Linguistique & Littérature. Cluny, 40 ans après*. Besançon : Presses universitaires de Franche-Comté, coll. "Annales littéraires de l'université de Franche-Comté", 2010, 344 p.

[Barthes, 1968]
Roland BARTHES (1968). Revue Langages, n°12, Décembre 1968, « *Linguistique et Littérature* », td Didier, Larousse, pp. 3–8.

[Bourdin & Duhem, 1972]
Jean-François BOURDIN & Pierre DUHEM (1972). *La grammaire de texte en pays de langue allemande*. Langages, 7e année, n° 26. *La grammaire générative en pays de langue allemande*. pp. 59–74.

[Bustarret & Linkès, 2008]
Claire BUSTARRET & Serge LINKÈS (2008).*Un nouvel instrument de travail pour l'analyse des manuscrits : la base de données* MUSE, [En ligne], mis en ligne le : 16 mars 2008. Dernière consultation : 17 mai 2010.

[Bustarret & Linkès, 2006]
Claire BUSTARRET & Serge LINKÈS (2006) *De Muse en Argolide, ou la codicologie à l'ère du numérique.* Dans *Editer et valoriser des fonds de manuscrits : l'apport (et les limites ?) du numérique*, Université Grenoble III, Presses Universitaire de Grenoble, Grenoble, 2008.

[Cerquiglini, 1989]
Bernard CERQUIGLINI (1989). *Eloge de la variante*, Seuil, coll. Des travaux, 1989.

[Chepiga, 2008]
Valentina CHEPIGA (2008). *Des unités de traitement au style d'un auteur. EDITE : une méthodologie informatisée de comparaisons génétiques Corpus contrastif Gary / Ajar*. Colloque thématique du Cercle belge de linguistique. Bruxelles 22–24 mai 2008 « Nouvelles approches en linguistique textuelle ».

[Crasson, 2006]
Aurèle CRASSON (2006). *Représenter l'illisible*. Genesis 27, 2006, p. 163–164.

[Crochemore & Rytter, 1994]
Maxime CROCHEMORE & Wojciech RYTTER (1994). *Text Algorithms. Approximate pattern matching : 237–251*. Non lus, cités dans [Ganascia *et al.*, 2004].

[D'Alfonso & Saller, 2007]
Matteo D'ALFONSO & Harald SALLER (2007). *Kodierung und Darstellung von Schreibschichten in der elektronischen Edition des Druckmanuskripts zu „Der Wanderer und sein Schatten", Literatur und Literaturwissenschaften auf dem Weg zu den digitalen Medien – Eine Standortbestimmung.* M. Stolz, L. Marco G. & J. Loop (éd.), Zürich, Germanistik.ch, 2007, p. 117–126.

[D'Iorio, 2008]
Paolo D'IORIO (2008). *L'île des savoirs choisis*. *De HyperNietzsche à Scholar source : pour une infrastructure de recherche sur le Web*. In *Recherches & Travaux*, n° 72/2008, Grenoble, Ellug, 2008, pp. 279–301.

[D'Iorio & SIMON-RITZ, 2001]
Paolo D'IORIO & Frank SIMON-RITZ (2001). *Le catalogue multimédia de la bibliothèque de Nietzsche*. In Paolo D'Iorio et Daniel Ferrer (éds.), *Bibliothèques d'écrivains*, Paris, éditions du CNRS, 2001, pp. 145-169.

[de Biasi, 2000]
Pierre-Marc DE BIASI (2000). *La génétique des textes*, Paris, Armand Colin, 128, 2000.

[Derrida *et al.* 1995]
Jacques DERRIDA, Daniel FERRER, Michel CONTAT, Jean-Michel RABATÉ et Louis HAY (1995). *Une discussion avec Jacques Derrida*. Archive et brouillon. Table ronde du 17 juin 1995 » In *Pourquoi la critique génétique ? Méthodes et théories. Sous la direction de Michel Contat et Daniel Ferrer*. CNRS Éditions, 1998, pp. 189–209.

[Doubrovsky, 2000]
Serge Doubrovsky (2000) *La place de la madeleine : écriture et fantasme chez Proust*, Grenoble : Ellug, 2000. 164 p.

[Fuchs *et al.*, 1987]
Catherine FUCHS, Jean-Louis LEBRAVE, Josette REY-DEBOVE, Jean PEYTARD, Almuth GRÉSILLON (1987). *La Genèse du texte : les modèles linguistiques*, C.N.R.S. éd., Paris, 1987

[Ganascia *et al.*, 2004]
GANASCIA J.-G., FENOGLIO I. (2004), *EDITE MEDITE, un logiciel de comparaison de versions*. In Le poids des mots, Actes des 7èmes journées internationales d'analyse statistique des données textuelles (JADT 04), UCL Presses Universitaires de Louvain.

[Grésillon & Lebrave, 1982]
Almuth GRÉSILLON & Jean-Louis LEBRAVE, Les manuscrits comme lieu de conflits discursifs. *La genèse du texte : les modèles linguistiques*. Paris, Éditions du CNRS, 1982, p. 129.

[Grésillon *et al.*, 1990]
Almuth GRÉSILLON, Jean-Louis LEBRAVE et Catherine VIOLLET, *Proust à la lettre. Les intermittences de l'écriture*, Tusson, Du Lérot, 1990.

[Hay & Naguy, 1982]
Louis HAY & Péter NAGUY (1982). *Avant-texte, texte, après-texte*, Budapest : Akadémiai Kiadô, Maison d'édition de l'Académie des sciences de Hongrie, Paris : Éd. du CNRS, 1982, 217 p. ISBN 963-05-2910-6 (Akadémiai Kiadô). ISBN 2-222-02895-7 (CNRS).

[Hay, 2007]
Louis HAY (2007). *Qu'est-ce que la génétique ?*, [En ligne], Mis en ligne le: 2 avril 2007, Disponible sur: http://www.item.ens.fr/index.php?id=44566. Dernière consultation : 9 mai 2010.

[Hayes & Flower, 1981]
John R. HAYES & Linda FLOWER (1981). *Identifying the organization of writing processes.* In *Cognitive processes in writing* L. W. Gregg & E. R. Steinberg (Eds.).*College Composition and Communication*, Vol. 32, No. 4 (Dec., 1981), pp. 365–387

[Labbé, 2008]
Dominique LABBÉ (2008). *Note adressée à Benoît Peeters et Michel Lafon pour leur livre « Nous est un autre ».* 15 mai 2008, HAL : hal-00279663, version 1

[Lang *et al.*, 1972]
Ewald LANG, Danièle CLÉMENT et Yves SCHWARTZ. *Quand une « grammaire de texte » est-elle plus adéquate qu'une « grammaire de phrase » ?*. Langages, 7e année, n° 26. *La grammaire générative en pays de langue allemande*. pp. 75–80.

[Lebrave & Grésillon, 2009]
Jean-Louis LEBRAVE & Almuth GRÉSILLON(2009). *Linguistique et génétique des textes : un décalogue*. [En ligne]. Mis en ligne le: 16 février 2009
Disponible sur: http://www.item.ens.fr/index.php?id=384099. Dernière consultation le 11 mai 2010.

[Lejeune, 1971]
Philippe LEJEUNE (1971). *Écriture et sexualité. Europe*, février-mars 1971, p. 113–143 (Sur Proust).

[Lejeune, 1998]
Philippe LEJEUNE (1998). *Anne Frank. Pages retrouvées. La Faute à Rousseau*, n° 19, octobre 1998, p. 61–64.

[Lejeune, 2007]
Philippe LEJEUNE (2007). *Génétique et autobiographie.* Extrait de la communication lors de la session CLELIA 2007, *Lalies, n° 28*, pp.169–187

[Linkès, 2003]
Serge LINKES (2003). *Un nouvel instrument de travail : la base de données MUSE* , article en collaboration avec Claire Bustarret, in *GENESIS* n° 21, Editions Jean-Michel Place, Paris, 2003.

[Martin, 2010]
Henri-Jean MARTIN (2010). *Encyclopaediauniversalis*http://www.universalis-edu.com/encyclopedie/gutenberg/Dernière consultation le 10 mai 2010.

Cécile MEYNARD & Muriel BASSOU. Interview du 25 février 2010, retranscrite partiellement en Annexe 1.

[Meynard, à paraître]
Cécile MEYNARD. *L'exemplaire « Serge André » des « Promenades dans Rome » : Stendhal critique de Stendhal.* Communication le 2 juin 2007 lors de la journée d'étude du séminaire Stendhal organisé à l'ENS Ulm-Sèvres par Xavier Bourdenet et François Vanoosthuyse, à paraître aux ELLUG (Grenoble), dans la collection « Bibliothèque stendhalienne et romantique » dirigée par Marie-Rose Corredor et Chantal Massol.

[Sankoff & Kruskal, 1983]
David SANKOFF & Joseph Bernard KRUSKAL (1983). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading. Non lu, cités dans [Ganascia *et al.*, 2004].

[Tabouret-Keller & Gardner-chloros, 2010]
Andrée TABOURET-KELLER et Pénélope GARDNER-CHLOROS (2010).*Plurilinguisme*. Dans *Encyclopaedia Universalis* [En ligne]. http://www.universalis-edu.com/corpus2.php?recherche_mode=alpha&nref=L110231#07000000 (Dernière consultation le 11 mai 2010)

[Tognotti, 1997]
Sandrine TOGNOTTI (1997). *Etude d'un dispositif de coopération rédacteur - lecteur pour l'apprentissage de la rédaction technique.* Mémoire de DES STAF de Sciences et Technologie de l'Apprentissage et de la Formation, Faculté de Psychologie et de Sciences de l'Education, Université de Genève, octobre 1997.

Citations en en-tête de partie

Partie 1
Louis HAY (2007). *Qu'est-ce que la génétique ?*, [En ligne], Mis en ligne le: 2 avril 2007. Disponible sur: http://www.item.ens.fr/index.php?id=44566. Dernière consultation : 9 mai 2010.

Partie 2
Jean-Louis LEBRAVE & Almuth GRÉSILLON(2009). *Linguistique et génétique des textes : un décalogue*. [En ligne]. Mis en ligne le: 16 février 2009. Disponible sur: http://www.item.ens.fr/index.php?id=384099. Dernière consultation : 11 mai 2010.

Partie 3
Pierre-Marc DE BIASI (2000). *La génétique des textes*, Paris, Armand Colin, 128, 2000.p. 68.

Table des annexes

Annexe 1
Interview de deux chercheuses

EXTRAIT D'INTERVIEW DU JEUDI 25 FEVRIER 2010
CECILE MEYNARD, MURIEL BASSOU

Claire : Existe-t-il une méthodologie propre à l'analyse génétique ?

Muriel :
Oui il existe une méthodologie de l'analyse génétique, c'est assez mis au point même avec un centre, l'ITEM qui travaille là-dessus. Il y a des revues.

C'est l'étude des manuscrits, c'est l'étude du support. C'est assez codifié. Voilà.

Claire : Est-ce que je peux trouver des ouvrages sur la méthodologie en analyse génétique ?

Cécile :
Oui , Pierre-Marc de Biasi (de l'ITEM) a écrit dans la collection 128, une synthèse sur la génétique des textes mais le livre n'est plus édité.

L'analyse de la genèse des textes littéraires, c'est comment se construit l'œuvre, l'embryon d'œuvre, donc les différentes étapes, qui s'empilent sur une page (donc on parle de papier). Et éventuellement tout ce travail de correction, de rature, de remords, de rajout, de suppression, etc. Et c'est là que c'est compliqué car chaque auteur à son processus de genèse propre.
Chez Stendhal c'est surtout l'amplification. Processus d'écriture par amplification.

Il existe différentes méthodologies développées par les chercheurs de l'ITEM, en sachant que comme il y a justement des processus de genèse différents selon les auteurs, c'est difficile d'avoir une seule méthodologie normalisée pour tout le monde.

Donc je dirais qu'il y a quand même des normes : il faut partir des manuscrits, il faut essayer d'identifier le premier jet. Donc premier jet chez Stendhal, par exemple quand c'est une page qu'il a dictée à un copiste et bien on reconnaît l'écriture du copiste, en général tu as un interligne large.

Ensuite identifier la deuxième intervention, deuxième campagne d'écriture, donc par exemple les annotations de Stendhal qui sont ajoutées et alors après, où ça se complique, c'est comment identifier ce qui est deuxième jet, ce qui est troisième jet, ce qui est quatrième jet, parfois c'est compliqué. On va dire et bien s'il y a une différence de couleur d'encre donc on peut supposer que c'est quelque chose qui a été encore rajouté après.
Mais la notion de deuxième intervention, troisième, quatrième, cinquième, dixième est parfois difficile à identifier.

Donc, voilà, je pense qu'il faut avoir une approche la plus rigoureuse possible, c'est ça la méthodologie je dirais, prendre la page de manuscrit avec tout l'empilement des couches, parfois c'est illisible et essayer de repartir de zéro : qu'est ce que je peux identifier comme le premier état de la page. Ensuite comment ça s'est ajouté et ça permet finalement d'identifier ce qui est devenu un entremêlement de formes.
Je te montrerai les pages, où tu te dis, « mais je ne comprends pas ; il y a un mot qui suit un autre mot qu'est-ce que ca veut dire ? », et en fait c'est juste qu'il y a différentes couches qui se sont ajoutées et tu mets tout à plat avec la mise en page.

Donc je pense que voilà, je dirai que c'est la méthodologie commune à tous les chercheurs en génétique.
C'est vraiment d'essayer de remonter, alors soit en remontant soit en essayant au contraire de partir du plus loin et en se ramenant de l'état final de la page.

En sachant qu'on est souvent dans du provisoire. Les manuscrits qui nous restent de Stendhal, pourquoi il nous en reste, c'est parce que ils n'ont pas été publiés. Quand ils ont été publiés il les a fait disparaître.

Pour d'autres auteurs, pour Flaubert, pour des auteurs plus tardifs, on a souvent les manuscrits, du manuscrit de départ jusqu'à l'état publié, une première édition.

Donc là c'est intéressant, tu as toute la genèse, du projet, jusqu'à l'œuvre, du premier jet jusqu'à l'œuvre, alors que dans le cas de Stendhal, on ne peut pas parler d'œuvre puisque ce n'est pas terminé, ce n'est pas final, ce n'est pas abouti.

Donc Lucien Leuwen, en fait on ne devrait pas dire, c'est une œuvre, on devrait dire, c'est un projet d'œuvre. En réalité, quand on dit, le roman Lucien Leuwen, non, c'est le projet de roman Lucien Leuwen qui n'a pas été fini.

L'ITEM est le centre de référence, c'est l'Institut des Textes et Manuscrits Modernes. Donc c'est vraiment le centre de référence pour ce qui est de l'analyse génétique des textes, quand on parle de génétique les gens croient tout de suite que l'on va utiliser les gènes, le génome et tout, non « génétique textuelle », donc c'est l'institut de référence en France à l'échelle internationale pour ce qui est de l'analyse des textes littéraires donc qui travaille aussi bien sur Joyce que sur des auteurs français, sur des inconnus que sur des auteurs qui ont publiés, c'est vraiment très vaste, comme champ d'activité, sur le cinéma, sur ... tout, tout, toute lettre, tout graphisme est objet d'étude pour eux. Ce n'est pas forcément un texte littéraire, c'est, c'est la genèse du théâtre, ils travaillent sur tout.

Nous avons une certaine approche. Mais d'autres généticiens pourront avoir d'autres approches que nous. Nous on a par exemple cette perspective de mise en ligne avec pour certaines pages, l'affichage des différentes couches, d'autres généticiens ne vont pas du tout procéder pareil.
C'est pareil, on parlait méthodologie tout à l'heure, on n'a pas forcément les mêmes, ce qu'on appelle les signes diacritiques, qu'on va ajouter sur la page, certains vont faire une transcription linéarisée en ajoutant des signes, par exemple des chevrons et tout, pour indiquer comment la phrase a été corrigée, se poursuit etc.
Donc on peut avoir des normes qui peuvent être différentes selon les instituts, selon les chercheurs, selon les universités.

[…]

Au sujet du sabir
Cécile :
A force de regarder des textes où il y a du sabir, j'ai l'impression que le sabir est là tout de suite, dès le premier jet. Donc du coup, il n'y a pas forcément un phénomène de réécriture, mais bon dans ce cas là, ça participe à la genèse effectivement, mais je dirais qu'il pense tout de suite en sabir, il écrit en sabir, ce n'est pas le mot qui est écrit en français et puis ensuite je le transforme, parfois on a des phénomènes de cryptage, ça existe, mais c'est souvent tout de suite ...

Muriel :
Sauf dans la Chartreuse de Parme par exemple quand il écrit « amie de mon cœur » après il le retranscrit en « amica di cuore ».

Cécile :
Donc ça c'est dans le projet de réécriture, c'est ça ? Là on est presque dans de la post-genèse.

Claire :
C'est une œuvre déjà publiée c'est ça ?

Cécile :
Voilà. De toute façon, on n'a pas le manuscrit initial de la Chartreuse, tout ça a été détruit, donc c'est dans ses projets de réédition suite aux critiques de Balzac.
Donc il envisage de transformer éventuellement son roman, de réorienter des choses, et là on a effectivement du sabir qui est construit par exemple. Mais cela n'aboutit pas.
Ce serait dommage de le transformer.

Claire :
Mais vous n'avez pas en tête de cas où il est parti d'un premier jet en sabir et ensuite il l'a corrigé pour pouvoir le publier dans une version ultérieure où il l'a traduit en fait vers le français.

Cécile :
Je n'ai pas l'impression. Je ne connais pas tous les manuscrits de Stendhal. Je ne peux parler que de ce que je connais. Dans le journal, c'est clair tu as tout un phénomène de cryptage, le sabir est nécessaire pour l'oralisation de ce qu'il a à dire. Dans les romans il y a peu de sabir mais c'est souvent maintenu.
En revanche, je ne sais pas s'il y a du sabir dans *L'histoire de la peinture en Italie* et si dans la perspective de publication, il a supprimé le sabir.
Donc je dirais oui, que le sabir participe de la genèse, mais qu'il est très souvent à la première étape dans la genèse.

Claire :
Et au niveau du journal on n'a qu'un premier jet ou on a déjà deux ou trois versions ?
Cécile :
Dans la plupart des journaux on n'a que le premier jet, puisque c'est la dimension « journal personnel, je ne le tiens que pour moi ».

On a essentiellement un cas qui est le journal de 1811, le « Tour through Italy » où il envisage en 1813, sans doute, on n'a pas de preuve, de le publier, donc il le retravaille. Là il va crypter encore plus des noms, il va les supprimer, il laisse uniquement des petites croix, il raye les noms, donc on a un phénomène de disparition parce que tu ne peux pas parler de choses concernant des personnes existantes.

Claire :
Mais c'est plus de l'anonymisation, c'est moins du sabir que je traduis en français dans une perspective d'édition.

Cécile :
Je n'ai pas l'impression. J'ai transcrit un certain nombre de pages, je n'ai pas l'impression qu'il y ait de phénomène de transformation du sabir en français pour la clarté en vue de publication. Je dirai que nous sommes dans l'analyse génétique certainement puisque justement on identifie le sabir et là, dès l'origine mais en revanche il n'y a pas forcément de transformation du sabir en français ou dans une autre langue. […]

AUTORISATION DE DIFFUSION DE CECILE MEYNARD

Formulaire type à compléter puis faire remplir et signer

En vue de l'enregistrement audio/vidéo et l'exploitation des données enregistrées

Nom et coordonnées du responsable de l'équipe ou du projet

Claire Lemaire
Université Stendhal, bureau I 111, BP 25, 38040 GRENOBLE Cedex 9

Présentation de l'enquête, ses objectifs, ses modalités,

Enquête sur la méthodologie en analyse de textes génétique

Autorisation à faire remplir et signer par le locuteur ou le représentant légal du locuteur

Je soussigné(e) Cécile MEYNARD

1. autorise par la présente ~~Mr~~/Mme Claire Lemaire
à procéder à l'enregistrement de Cécile Meynard
lieu, date(s), horaire(s) : Université Stendhal, le 25 février 2010, de 10h30 à 11h30

2. prends acte que les données enregistrées et exploitées seront anonymées avant toute diffusion, ce qui signifie :
   a) que dans la transcription des paroles, toute information devant rester confidentielle (nom, adresse ou autre information permettant d'identifier une personne ou un groupe) sera supprimée ou modifiée ;
   b) que les enregistrements audio seront bruités si nécessaire (lors de la mention d'une information devant rester confidentielle) ;
   c) que les enregistrements vidéo seront floutés si nécessaire (lors de la vision d'une information devant rester confidentielle).

3. autorise la diffusion de ces données sous leur forme audio ou vidéo aussi bien que sous leur forme transcrite et anonymée :
   a) à des fins de recherche scientifique (mémoire, thèse, article ou exposé scientifique) ☒
   b) à des fins d'enseignement universitaire (cours, séminaire) ☒
   c) pour diffusion auprès de la communauté des chercheurs (moyennant convention entre équipes) ☒
   d) pour diffusion et archivage sur un site dédié à la recherche ou hébergeant des corpus ☒

4. souhaite que la contrainte supplémentaire suivante soit respectée :

Lieu et date:

Signature :

AUTORISATION DE DIFFUSION DE CECILE MEYNARD, MURIEL BASSOU

Formulaire type à compléter puis faire remplir et signer

En vue de l'enregistrement audio/vidéo et l'exploitation des données enregistrées

Nom et coordonnées du responsable de l'équipe ou du projet

Claire Lemaire
Université Stendhal, bureau TH1, BP 25, 38040 Grenoble Cedex 9

Présentation de l'enquête, ses objectifs, ses modalités,

Enquête sur la méthodologie en analyse génétique textuelle.

Autorisation à faire remplir et signer par le locuteur ou le représentant légal du locuteur

Je soussigné(e) Muriel Bassou

1. autorise par la présente ~~Mr~~/Mme Claire Lemaire
Muriel Bassou
à procéder à l'enregistrement de
lieu, date(s), horaire(s) : Université Stendhal, Grenoble, le 25 février 2010, de 10h30 à 11h30

2. prends acte que les données enregistrées et exploitées seront anonymées avant toute diffusion, ce qui signifie :
a) que dans la transcription des paroles, toute information devant rester confidentielle (nom, adresse ou autre information permettant d'identifier une personne ou un groupe) sera supprimée ou modifiée ;
b) que les enregistrements audio seront bruités si nécessaire (lors de la mention d'une information devant rester confidentielle) ;
c) que les enregistrements vidéo seront floutés si nécessaire (lors de la vision d'une information devant rester confidentielle).

3. autorise la diffusion de ces données sous leur forme audio ou vidéo aussi bien que sous leur forme transcrite et anonymée :
a) à des fins de recherche scientifique (mémoire, thèse, article ou exposé scientifique) ☑
b) à des fins d'enseignement universitaire (cours, séminaire) ☑
c) pour diffusion auprès de la communauté des chercheurs (moyennant convention entre équipes) ☑
d) pour diffusion et archivage sur un site dédié à la recherche ou hébergeant des corpus ☑

4. souhaite que la contrainte supplémentaire suivante soit respectée :

Lieu et date:

Signature :

Annexe 2
Entrée du Trésor de la Langue Française Informatisé

SABIR, subst. masc.

A. — Parler composite mêlé d'arabe, d'italien, d'espagnol et de français parlé en Afrique du Nord et dans le Levant. À peine s'il parlait le sabir, ce patois algérien composé de provençal, d'italien, d'arabe, fait de mots bariolés ramassés comme des coquillages tout le long des mers latines (A. DAUDET, Contes lundi, 1873, p. 168). Ce sabir fait de turc, d'arabe, d'espagnol, d'italianismes (...) plutôt que de paroles françaises que parlent tous les marins du Levant (CENDRARS, Bourlinguer, 1948, p. 168).

— En appos. Avec un sourire, elle [une jeune Grecque] s'arrêtait, lui donnait quelque fleur, un brin d'oranger (...) parfois lui disait deux ou trois mots dans un demi-français sabir (LOTI, Matelot, 1893, p. 45).

B. — LING. [P. oppos. à pidgin et à créole dont le système est plus homogène et plus complet] Langue mixte, généralement à usage commercial, née du contact de communautés linguistiques différentes. Les sabirs sont des langues d'appoint, ayant une structure grammaticale mal caractérisée et un lexique pauvre, limité aux besoins qui les ont fait naître et qui assurent leur survie (Ling. 1974).

C. — P. ext.

1. Péj. Langue formée d'éléments hétéroclites, difficilement compréhensible. Synon. fam. charabia. Ou bien l'enseignement du latin sera maintenu (...) ou bien notre langue deviendra une sorte de sabir formé, en proportions inégales, de français, d'anglais, de grec, d'allemand, et toutes sortes d'autres langues (GOURMONT, Esthét. lang. fr., 1899, p. 75).

2. P. anal. « P'pa, et le p'tit Noël (...) y mettra-ti' tet' chose dans mon soulier? » demanda tout à coup Raoul dans son sabir enfantin (COPPÉE, Longues et brèves, 1893, p. 290).

Prononc.: [sy ʀ s]. Étymol. et Hist. 1852 (La langue Sabir, titre d'art. ds l'Algérien, journal des intérêts de l'Algérie, 11 mai d'apr. H. SCHUCHARDT ds Z. rom. Philol. t. 33, p. 457); p. ext. a) 1882 « langue difficilement compréhensible parlée par un étranger qui s'exprime mal » (LOTI, Fleurs ennui, p. 332); b) 1933 ling. synon. de langue franque (franc1*) (MAR. Lex.). Altér. de l'esp. saber « savoir » (v. ce mot) qui servit d'abord à désigner le mélange de fr., d'ar., d'esp. et d'ital. parlé par les Algériens après 1830; cf. Si ti sabir, ti respondir... dans le ballet turc du Bourgeois gentilhomme de MOLIÈRE, IV, 5. Voir H. SCHUCHARDT, ibid., pp. 457-458 et SAIN. Lang. par., pp. 151-153. Bbg. BAL (W.). À propos d'un microsystème de la terminol. ling. fr.: les termes créoles, pidgin, sabir. Zootecnica e vita. 1975, t. 18, n o 1/2, pp. 69-82. **—** QUEM. DDL t. 7.

Annexe 3
Extrait de l'entrée « plurilinguisme » de l'Encyclopaedia universalis

Des systèmes bouleversés : alternances et langues nouvelles

Le modèle du contact interlinguistique appliqué jusqu'ici présuppose que chaque système en présence est distingué sans ambiguïté tant par les locuteurs qui l'emploient que par les linguistes qui le décrivent comme une entité aux limites précises. Un tel modèle est largement redevable au type d'objet que définissent les théories scientifiques : ici l'objet structural de la linguistique moderne ; il est également étayé par la définition de certaines langues comme objet constitutionnel, donc juridique, avec le normativisme que cela entraîne (c'est le cas pour le français), et par les stéréotypes des locuteurs à propos de leurs propres parlers (le français serait une langue, le patois non). Les connaissances plus approfondies que nous avons aujourd'hui de situations linguistiques en voie de changement rapide attirent l'attention sur la part qu'occupent de telles constructions dans notre savoir : en certaines circonstances les langues peuvent devenir des systèmes à bords flous.

L'alternance de langues ou *code-switching* peut illustrer un tel devenir. Considérée autrefois comme une aberration commise par des locuteurs incapables de maintenir séparées leurs différentes langues, l'alternance est aujourd'hui traitée comme une stratégie communicative. Loin d'avoir un statut d'exception, ce mode alterné est attesté dans des parties du monde aussi variées que l'Inde, l'Afrique du Sud, les États-Unis et différentes régions d'Europe. Il ne s'agit plus alors d'éléments empruntés par ignorance du terme adéquat ou comme le seraient des citations, mais du fonctionnement d'un répertoire partagé au sein d'un groupe ou d'une communauté. Dans un tel répertoire, il n'y a pas de phrase complète, et parfois même pas une proposition ou un syntagme autonome qui puisse être attribué à une seule des langues. Les connotations de chacune des langues peuvent ainsi être cumulées et, de surcroît, l'alternance donne lieu à un contraste lui-même porteur de signification. Quel est l'avenir d'un tel répertoire ? Constitue-t-il une étape vers la formation d'une langue différente résultant de l'intrication des langues primitivement présentes, ou bien ne sera-t-il qu'une étape vers l'assimilation d'une des langues au profit d'une autre ? Ces différents cas peuvent sans doute se présenter.

L'alternance se produit généralement à une limite fonctionnelle : à la fin d'une proposition, d'une expression, à l'intérieur de celle-ci ; elle peut également intervenir à des places où l'agencement de la proposition ne la laisse pas prévoir. La théorisation de tels processus reste hypothétique : on préconise une grammaire qui inclut, à un certain niveau d'abstraction, des schèmes communs aux langues en présence, ou bien on admet que toute

alternance est d'abord un procédé pragmatique à but communicatif, qu'elle viole ou non les règles fonctionnelles de l'une ou l'autre langue, qu'elle soit ou non accompagnée d'hésitations, de répétitions ou d'autres disjonctions propres au langage parlé.

De nombreux processus d'affaiblissement ou de perte des distinctions entre systèmes peuvent être décrits dans toute situation sociale limite où un locuteur, ou bien un groupe de locuteurs, est placé dans des circonstances étrangères telles que ses repères langagiers et culturels d'origine ne s'appliquent plus. Certaines de ces situations sont éphémères : celle de l'explorateur de jadis, voire du touriste d'aujourd'hui, ou encore celle de l'étudiant qui apprend une seconde langue ; les transformations que l'on observe restent alors généralement sans lendemain. Sur le plan descriptif, elles peuvent cependant être mises en rapport avec des transformations analogues, durables celles-ci, qui aboutissent à la formation de langues nouvelles dans des conditions de bouleversement de la vie qui sont celles, en général, des migrations. Migrations de la main-d'œuvre d'un pays à un autre, d'un continent à un autre, migration massive vers les villes qui semble aujourd'hui être un phénomène d'échelle mondiale, cas de ces migrations obligées qui accompagnèrent l'esclavage. Dans ce dernier cas, la formation d'un pidgin découle de la nécessité de communiquer où se trouvèrent, face à leurs maîtres, des membres de communautés différentes, cela dans des conditions de grossière inégalité. Un parler commun se développe alors, qui n'est la langue d'origine d'aucune des personnes présentes. Un tel parler puise généralement la majeure part de son matériau lexical dans la seule langue d'intercommunication et de référence commune, celle des maîtres ou des colonisateurs. Il se caractérise par le fait que la forme et les potentialités fonctionnelles de ce matériau sont interprétées selon des schèmes d'emploi des langues d'origine, et davantage encore par des processus de réduction et de simplification dont la description reste incomplète et la théorisation difficile. De tels processus écartent, pour un temps, la plupart des procédés qui contribuent à l'organisation syntaxique et à la redondance (absence générale des désinences, des prépositions, conjonctions, etc.) et limitent l'extension sociale de tels parlers. Il arrive cependant qu'ils deviennent langue commune : ils ont regagné alors une certaine complexité fonctionnelle et peuvent servir à un ensemble étendu de fonctions sociales. C'est le cas aujourd'hui du pidgin bislaman à Vanuato, langue officielle et écrite du pays, mais toujours seconde langue pour les locuteurs. Dans d'autres cas, généralement ceux des communautés d'esclaves aux Antilles, le pidgin devient plus utile pour chacun que sa propre langue d'origine : les enfants l'entendront quotidiennement et grandiront en apprenant à parler dans cette langue. Le pidgin prend alors statut de créole et se transforme encore par la mise en place de nouveaux moyens fonctionnels. Plus récemment, on a pu montrer que la pression permanente des institutions et surtout celle de l'école tendent à faire évoluer le parler de la communauté créole vers la forme normative de la langue dont ce créole tient l'essentiel de son fonds lexical. En

Jamaïque, par exemple, il y a fusion progressive du créole avec l'anglais, fusion qui relie de manière continue un extrême de l'emploi du créole, représenté par les parlers ruraux, à un extrême citadin et cultivé qui se caractérise par l'emploi d'un anglais à norme locale.

[…]

Écrit par Andrée TABOURET-KELLER

Écrit par Pénélope GARDNER-CHLOROS

Annexe 4
Programme Perl d'extraction d'informations sur le sabir

```
###############################################################
# Programme de recherche de sabir -
# Entree : fichier .xml d'une fiche de transcription des Manuscrits de Stendhal
# Sortie : fichier .txt avec les infos sabirs isolees (convertible en Excel, séparateur =
tabulation)
###############################################################

# Lancement programme via console MS Dos (Start -> Run -> cmd)
# cd..
# pgm_sabir.pl (et enter)

#xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx
# Description de l'algorithme
#xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx
# 1) On parcourt le fichier d'entree ligne par ligne.
# 2) Pour chaque ligne
# On regarde s'il y a les infos requises
#    On les ecrit dans le fichier de sortie
# 3) On écrit les infos dans le fichier de sortie .txt.


#XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXX
# Pour imprimer les accents comme dans "Hélène de Jacquelot"
#XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXX

use locale;
use utf8;


#XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXX
# Déclaration des variables et ouverture de fichier
#XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXX
#  s_        scalaire
#  l_        liste
#  h_        hachage
my $s_file1="5896-16-024.xml";      # fichier en entree
# pour l'instant avec une fiche en dur
# apres, il faudra industrialiser le process et elargir a plein de fiches simultanement
my $s_file2="res_sabir.txt";
open(FILEIN,"<", $s_file1);              # ouverture du fichier en lecture
open(FILEOUT,">", $s_file2);             # ouverture du fichier en ecriture
```

```
my $s_line="";      #la ligne courante du fichier
my $s_lineout=""; #la ligne a ecrire dans la fichier de sortie
my $s_reference="";          #nom de la fiche de manuscrit
my $s_transcripteur="";      #le nom du transcripteur -> Transcripteur
my $s_corpus1="";            #le nom du corpus (sous-domaine, entre "domaine" et "document") -> Corpus1
#my $s_corpus2="";           #le nom du corpus_bis le cas echeant (sous-domaine, entre "domaine" et "document") -> Corpus2
#my $s_corpus3="";           # le nom du corpus_ter le cas echeant (sous-domaine, entre "domaine" et "document") -> Corpus3
my $s_date1="";   # le nom de la date de debut de periode probable de redaction ou de la seule date -> Date1
#my $s_date2="";  # le nom de la date de fin de periode probable de redaction le cas echeant -> Date2
my $s_lieu="";      # le nom du lieu de redaction estime ou avere -> Lieu
my $s_tradproposee="";      # la traduction en francais du texte etranger redige par Stendhal -> TradProposee
my $s_textetranger="";      # le texte de Stendhal en etranger -> TextEtranger

#+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
# 1) On parcourt le fichier.xml ligne par ligne.
#+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
while (! eof(FILEIN))        # eof() indique si on a atteint la fin du fichier
{
  $s_line=<FILEIN>;          # Lecture de la ligne courante

  #-------------------------------------------------------------------------------------------------
  # 2) On découpe la ligne via une expression régulière et on recopie l'info interessante
  #    (on lui ajoutera une tabulation lors de l'output)
  #-------------------------------------------------------------------------------------------------
  # On teste grace a une expression reguliere, si on a une info interessante
  # si oui, on recupere l'info interessante
  # l'info est recuperee de la regex via les parenthèses et $1 ou$2.
  # REGEX :
  # La regex vide -> $s_line=~/VIDE/i)
  # Insensible a la casse -> i
  # Pour capturer l'info et la mettre ds une variable $1, $2, etc. -> ( )
  # N'importe quel caractere -> .
  # Le vrai point, il faut l'echapper -> \.
  # Le guillement aussi, il faut l'echapper -> \"
  # Le guillement aussi, il faut l'echapper -> \/
  # On ne decrit pas tout la ligne (donc on se moque du debit de la ligne)
  # mais on dit slmt : si dans ma ligne il y a "nom_fichier=""
  # alors capturer ce qu'il y a apres
  # et on ne s'occupe pas non plus de ce qu'il y a apres la regex
  # pour n'importe quel caractère lettre -> \w
  # pour n'importe quel caractère chiffre -> \d
  # pour n'importe quel caractère d'espace -> \s

  # Info 1 : le nom de la fiche de manuscrit -> Reference
  if ($s_line=~/nom_fichier="([\d\-rv]+)/i)
```

```
  {
            $s_reference=$1;
            print FILEOUT "$s_reference"."\t"; #TEST
  }
  # Info 2 : le nom du transcripteur -> Transcripteur
  if ($s_line=~/nom_transcripteur="([\w\s']+)/i)
  {
            $s_transcripteur=$1;
            print FILEOUT $s_transcripteur."\t"; #TEST
  }
  # Info 3 : le nom du corpus (sous-domaine, entre "domaine" et "document") -> Corpus1
  if ($s_line=~/corpus="([\w_]+)/i)
  {
            $s_corpus1=$1;
            print FILEOUT $s_corpus1."\t"; #TEST
            # Info 4 : le nom du corpus_bis le cas echeant (sous-domaine, entre "domaine" et "document") -> Corpus2
            if ($s_line=~/corpus_bis="([\w_]+)/i)
            {
                        $s_corpus2=$1;
                        print FILEOUT $s_corpus2."\t"; #TEST
               #Info 5 : le nom du corpus_ter le cas echeant (sous-domaine, entre "domaine" et "document") -> Corpus3
        if ($s_line=~/corpus_ter="([\w_]+)/i)
          {
            $s_corpus3=$1;
            print FILEOUT $s_corpus3."\t"; #TEST
          }
        else
            {
                                    print FILEOUT "\t"; #TEST
            }
            }
            else
            {
     print FILEOUT "\t\t"; #TEST
     }
    }
  # Info 6 : le nom de la date de t de periode probable de redaction ou de la seule date -> Date1
  # pour dire "espace" -> \s
  if ($s_line=~/equivalent_debut="([0-9\s]+)/i)
  {
            $s_date1=$1;
            print FILEOUT $s_date1."\t"; #TEST
            # Info 7 : le nom de la date de fin de periode probable de redaction le cas echeant -> Date2
            if ($s_line=~/equivalent_fin="([0-9\s]+)/i)
            {
                        $s_date2=$1;
                        print FILEOUT $s_date2."\t"; #TEST
            }
            else
```

```
		{
			print FILEOUT "\t"; #TEST
		}
	}
	# Info 8 : le nom du lieu de redaction estime ou avere -> Lieu
	if ($s_line=~/lieu_indique="([\w\s-]+)/i)
	{
		$s_lieu=$1;
		print FILEOUT $s_lieu."\t"; #TEST
	}
	# Info 9 : la traduction en francais du texte etranger redige par Stendhal -> TradProposee
	if ($s_line=~/traduction traduction="([^>]+)\">/i) # Tout caractere situe avant le chevron (1 ou n fois) ->[^>]+
	{

		$s_tradproposee=$s_tradproposee." ".$1;
	}
	# Info 10 : le texte de Stendhal en etranger -> TextEtranger
	if ($s_line=~/traduction traduction="[^>]+\">([^>]+)<\/traduction>/i) # Tout caractere situe avant le chevron (1 ou n fois) ->[^>]+
	{

		$s_textetranger=$s_textetranger." ".$1;
	}
}


#+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
# 3) Les textes concatenes sont eux aussi maintenant ecrits dans le fichier
#+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#$s_transcripteur\"\t"$s_corpus1"\t"$s_corpus2"\t"$s_corpus3"\t"$s_date1"\t"$s_date2"\t"$s_lieu"\t"$s_tradproposee"\t"$s_textetranger";
print FILEOUT $s_tradproposee."\t"; #TEST
print FILEOUT $s_textetranger."\t"; #TEST


# Fermeture des fichiers (et vidage des tampons de lecture et d'ecriture)
close(FILEIN);
close(FILEOUT);
```

Annexe 5
Extrait des sabirs recueillis

Refe rence	Trans cripteur	Cor pus1	Cor pus2	Cor pus3	Date1	Date2	Lieu	TradProposee	Textetranger	Contexte
302-1-372v	Cécile Mey nard				18.04 .1801	12.09 .1801	Milan	chevalier servant	cavaliere Servente	On cite ici M.me Nota comme la plus jolie femme de la ville et véritablement elle n'est point mal_ on lui donne 60, 000. l[ivres] de rente, elle a un cavaliere Servente bel homme, et qui dépense beaucoup pour elle, elle est par conséquent inataquable.
302-1-434v	Cécile Mey nard				28.07 .1805	-	Marse ille	J'ai dit au père que le voyage de G. à M. m'avait coûté	Said to the father, that my travel from G[renoble] to M[arseille] had costed to me 6 louis and half.	-
5896-01-136	Hélène de Jac quelot	Corresp ondance _person nelle	Jour naux	Tradu ctions	13.05 .1804	-	Paris	e c'est-à-dire, en examinant si ce personnage dans cette circonstance donnée pouvait e devaitpenser telle chose, et lui donner de la couleur.	; cioè, esaminando se quel tal personaggioin quella data circostanza potea e doveapensare tal cosa, e in quella tal guisa colorarla.	<paragraphe commentaire="Extraits d'Alfieri. A compléter."><ligne>V. 420 </ligne>
5896-07-002	Cécile Meynard	Pensées	Textes sur la littérat ure		10.04 .1803	17.09. 1803	Paris	il ne sera pas mon ami	he shall not be my friend	hier </ligne><ligne>Duverney me reprocha ma causticité, m'en corriger même au </ligne><ligne>risque de paraitre bête, car à cette heure he shall not be my </ligne><ligne><traduction traduction="il ne sera pas mon ami"></traduction>friend. </ligne></paragraphe>
5896-15-011	Hélène de Jacquelo t	Earline			16.02 .1840	20.03. 1840	Rome	du mari	La mere $of the husband $a le diable au corps, femme méchante, et sa grande fille noire, pas bonne.	<paragraphe alignement="gauche"><ligne alignement="gauche" alinea="aucun">La mere <traduction traduction="du mari">of the husband</traduction></ligne><ligne alignement="gauche" alinea="aucun">a le diable au corps,</ligne><ligne alignement="gauche" alinea="aucun">femme méchante, et<douteux> sa </douteux>grande </ligne><ligne alignement="gauche" alinea="aucun">fille noire, pas bonne.</ligne></paragraphe>

Table des illustrations

Sigles et abréviations utilisés

BNF	Bibliothèque nationale de France
CAM	Centre d'analyse des manuscrits
CNRS	Centre national de la recherche scientifique
ITEM	Institut des textes et manuscrits modernes
eKGWB	Edition critique numérique des œuvres complètes et de la correspondance. En allemand : *Digitale Kritische Gesamtausgabe Werke und Briefe.* (*Digitale* se dit aussi *elektronische*)
TAL	Traitement automatique des langues
UMR	Unité mixte de recherche

Glossaire

Allographe	Écrit par une autre main que celle de l'auteur de l'œuvre ou du projet d'œuvre sur laquelle porte l'étude.
Autographe	Ecrit de la main même de l'auteur (et non d'un scribe, d'un copiste ou d'un ami)
Avant-texte	Dossier de genèse « dépouillé », c'est-à-dire dont chacune des pièces a été inventoriée, classée, datée, déchiffrée.
Codicologie	Etude matérielle des manuscrits en tant qu'objets physiques par l'étude des matériaux servant à leur confection et leur utilisation.
Dossier génétique	Ensemble matériel des documents et manuscrits se rapportant à la genèse que nous avons l'intention d'étudier.
Feuillet	Papier manuscrit. Il peut être autographe ou allographe.
Filigrane	Empreinte d'identification d'un papier obtenue en plaçant une forme en cuivre ou en laiton formant un dessin ou une inscription, fixée sur la forme destinée à recevoir la pâte à papier.
Folio	Synonyme de feuillet
Hyperonyme	Terme dont le sens inclut celui d'un ou de plusieurs autres (mot générique) : véhicule est l'hyperonyme de voiture. Voiture est l'hyponyme de véhicule.
Manuscrit définitif	Manuscrit autographe définitif, recopié en fin de rédaction, pour fournir une version lisible au copiste, qui lui produira un document de référence pour l'imprimeur. Egalement appelé Manuscrit (avec une majuscule).
Manuscrit pré-définitif	Manuscrit
Marginale	Toute annotation en marge d'un texte, aussi bien sur un livre publié de Stendhal ou d'un autre auteur que Stendhal lit (Molière, Saint-Simon, etc.) qu'un de ses manuscrits de travail (par exemple les marginales de Lucien Leuwen).
Marginalia	Notes en marge de livres publié ou lu par Stendhal, les siens ou ceux d'autres auteurs. Cette pratique tardive se développe chez Stendhal surtout à la fin des années 1800 et va devenir majoritaire dans l'écriture. Il s'agit donc d'un type particulier de *marginale*. Par exemple, les notes sur un manuscrit d'un projet de roman ne sont pas des marginalia.
Notes de lecture	Notes sur un papier libre et non sur un livre. Ce sont éventuellement des notes de lecture « commentées », dans ce cas Stendhal écrit un petit « h » avant pour signaler qu'il donne son avis.

Mise au net	Texte recopié à partir d'un brouillon, en général par l'auteur (parfois par son copiste)

Table des matières

MOTS-CLÉS : TAL, génétique de textes, généticiens, manuscrits.

RÉSUMÉ

L'apparition de corpus numériques de manuscrits littéraires a enrichi notre patrimoine d'une donnée langagière analysable et traitable automatiquement. La transcription du contenu de ces manuscrits dans un format numérique textuel permet de parcourir en quelques secondes des milliers de mots. L'étude des différentes versions d'une œuvre littéraire donne lieu à des enquêtes fastidieuses de la part des chercheurs en littérature. Se posent alors de nouvelles questions de méthodologie de travail : comment exploiter au mieux l'outil informatique pour assister le chercheur, quelles sont les nouvelles études envisageables grâce aux progrès ? Après un aperçu de l'analyse génétique et d'outils de traitement automatique des langues existants dans le domaine, nous présentons la modélisation sur les manuscrits de Stendhal de trois fonctionnalités à trois niveaux de granularité qui assisteraient les chercheurs en littérature dans leur analyse sur les bibliothèques d'auteurs, le théâtre ou le code-switching.

KEYWORDS: NLP, textual genetic, philologists, manuscripts.

ABSTRACT

The rise of digital corpora of literary manuscripts has added a substantial amount of linguistic data to our heritage that can be analyzed and processed automatically. The transcription of the content of these manuscripts in digital text format allows thousands of words to be examined in a few seconds. Studying different versions of a literary work requires time-consuming examination by researchers in the field of literature. New methodological issues thus arise: how can information technology tools be best put to use to help researchers? What new studies are made possible by this progress? After an overview of textual genetic analysis and of the existing tools for natural language processing in this domain, we will outline the modelling, on Stendhal's manuscripts, of three functions that can aid literary researchers in their analysis of author's libraries, theatrical writing and code-switching.