Familles narratologiques et balisage du roman contemporain

Denise Malrieu

Résumé

Denise MALRIEU
dmalrieu@u-paris10.fr
CNRS-MODYCO PARIS X
Site : infolang.u-paris10.fr/modyco

Familles narratologiques et balisage du roman contemporain

Résumé
L'exposé qui suit s'efforce de définir les dimensions de description du roman dans une perspective qui souhaite joindre profilage des textes et linguistique textuelle pour une caractérisation linguistique de familles narratologiques.
Les dimensions de description privilégiées portent sur le dispositif énonciatif du roman et concernent la désignation des actants principaux, les niveaux de diégèse et leurs attributs, les séquences de discours rapportés, leurs locuteurs et segments introducteurs, les descriptions focalisées.
La méthodologie choisie est celle de la TEI par sa proximité avec XML et les avantages d'interopérabilité. Nous avons enrichi les balises de la TEI concernant le <textClass> et le <profileDesc> et défini un nouveau corps de balises pour décrire les catégories définies plus haut.
Nous donnons ensuite un exemple d'exploitation de textes XMLisés par une analyse contrastive des deux parties correspondant à deux types de narrateur dans *Le Ravissement de Lol V. Stein* de M. Duras.
Nous soulignons enfin les problèmes liés à l'exploitation de balisages assez fins : problèmes des chevauchements de balise pour lesquels la solution n'est pas consensuelle, problème de la lisibilité des textes balisés et des désavantages d'un mélange des différentes couches sémiotiques du balisage, problème du manque d'outils conviviaux pour l'analyse topologique des balises et pour la caractérisation des occurrences par leur chemin dans l'arborescence.

Abstract

The present study tries to define the novels description dimensions with a point of view that joins texts profilage and textual linguistics, with the aim to reach a linguistic characterization of novels narrative families.
The privileged description dimensions relate to the novel's enunciative device, the main actants designation, the diegetic levels and their attributes, the reported speech, their speakers and introducing segments, the focused descriptions.
The chosen methodology is the TEI/XML one, considering her interoperability. We expand the TEI tags, specially the <textClass> and <profileDesc> and define a new tags body in order to describe the previously defined categories.
We then give an example that runs with the defined tags in a contrastive study of the two types of narrator in *Le Ravissement de Lol V. Stein* of M. Duras.
At last, we underline the questions regarding the exploitation of fine tagging : the tags overlap, the readability of tagged texts and the problems generated by the mixing of different semiotic tagging layers; we emphasize the default of convivial query tools for a topologic tags analysis and for a tokens characterization by their paths in the tree.

L'exposé qui va suivre s'inscrit dans une démarche initiée en 2000 sur la caractérisation linguistique des genres écrits, à l'intérieur d'une linguistique textuelle et mettant en jeu la méthodologie TEI de balisage des textes. Il sera centré sur la définition des dimensions nécessaires à la description du

genre romanesque et donc sur la création de balises non définies par la TEI; il fera des propositions concernant le développement de nouvelles modalités de questionnement, devenues nécessaires dès que le balisage des séquences textuelles dépasse en richesse et complexité le simple balisage des . Je terminerai par un exemple d'exploitation de roman balisé à l'intérieur du corpus des oeuvres écrites de Duras, en cours de constitution.

1 – Bref historique de la démarche

Cette démarche part du présupposé que, sur le versant littéraire, les recherches en poétique ou stylistique ne peuvent que s'appuyer sur une analyse linguistique de la matérialité des textes et que, sur le versant linguistique, l'analyse sémantique d'un énoncé implique une linguistique textuelle, qui va définir comment les traits génériques d'un texte vont contraindre l'analyse syntaxique de la phrase. Le genre apparaît comme un concept nécessaire à cette démarche, car le genre est le lieu où s'articulent les contextes d'énonciation et les normes langagières; c'est le lieu où s'explicitent les jeux des différentes sémiotiques qui informent le texte et contraignent les dimensions cognitives de l'interprétation. L'objectif poursuivi consiste à établir un pont entre la démarche de profilage et la linguistique textuelle.

Peut-on caractériser linguistiquement des familles narratives à l'intérieur du roman?

La démarche inductive de profilage de 2540 textes à partir de 250 variables morpho-syntaxiques issues de l'analyseur Cordial[1] (Malrieu et Rastier, 2001), efficace pour différencier les domaines et champs génériques s'est avérée peu probante pour dessiner des familles narratives au sein du roman "sérieux" (le premier facteur n'expliquant pas plus de 30% de la variance).

La démarche inductive de classification des romans rencontrait plusieurs obstacles :

- non disponibilité de corpus *raisonnés* de romans contemporains suffisamment étoffés.
- nécessité de repenser les variables morphosyntaxiques à prendre en compte dans la classification (insuffisance des variables disponibles dans Cordial : nous en avons redéfini environ 300).
- le profilage permet de catégoriser mais non de comprendre les effets du texte: les calculs obtenus sur le texte dans sa globalité effectuent un lissage qui dilue les différences; le roman est par essence un texte composite (par ex le fort taux de 1PS ne peut différencier le roman homodiégétique et le roman hétérodiégétique fortement dialogué). De plus, la désambiguïsation d'un énoncé implique la prise en compte de séquences inférieures au texte, séquences informées par le genre mais qu'il faut caractériser en tant que telles.

D'où l'abandon temporaire de la démarche inductive sur textes entiers pour passer à une représentation qui prenne en compte le genre comme dispositif énonciatif.

2. – Le genre comme dispositif énonciatif :

En reformulant les hypothèses de Bakhtine, on peut avancer que le genre définit des contraintes interprétatives par un réglage du dispositif énonciatif , lié aux rapports sociaux et au dispositif communicationnel; il s'exprime dans des structures textuelles préférentielles et dans la prédilection pour certains types d'énoncés.

Dans le roman, ce dispositif énonciatif est configuré par un projet esthétique (Danon-Boileau, 1982), à la fois dans le mode d'allocution narrateur narrataire, dans la mimésis (les différents discours rapportés, leur poids, leur agencement) (Malrieu 2004), dans le mode de référenciation (dénomination vs désignation qualifiante ou anonymat) (cf plus bas une analyse contrastive des deux types de narrateur dans *le Ravissement de Lol V. Stein* de M. Duras).

La nature de l'univers fictif dépend du projet esthétique de l'auteur. Celui-ci contraint la triangulation des rapports narrateur-lecteur, narrateur-personnages, lecteur-personnages. Le roman peut aménager des scènes énonciatives assez diverses, qui dépendent de la place du narrateur dans le récit, de son ethos (narrateur omniscient ou pas, distancié ou empathique, critique de façon explicite ou indirecte, menant un récit anachronique ou pas).

3 - Les dimensions de description et la méthodologie de balisage:

Il s'agit donc de définir les dimensions de description de ces scènes énonciatives qui peuvent être instables à l'intérieur d'une oeuvre. (On laissera ici de côté le dispositif particulier du roman

[1] http://www.synapse-fr.com

épistolaire)[2]. La méthodologie de balisage choisie est celle de la TEI, car elle présente les avantages d'un standard international lié à XML. Le balisage de la structure arborescente du document est donc celui défini par la TEI; mais nous avons ajouté, en fonction des besoins de description, de nouvelles balises :

- *Les diégèses* : celles-ci sont décrites pour le texte entier dans le <profileDesc> et balisées dans le <body> : nombre de niveaux d'enchâssement de diégèses; nombre de diégèses de chaque niveau; longueur moyenne des diégèses de chaque niveau; chaque diégèse est balisée et décrite par ses attributs : son niveau, type de narrateur (intra- vs extradiégétique; hétéro- vs homodiégétique), l'identité du narrateur et des personnages appartenant à la diégèse, temps du récit de la diégèse, sa longueur.
- *Les focalisations* : descriptions focalisées avec identité du foyer optique.
- *Les registres de la parole* : la parole prononcée (discours direct ou rapporté), la parole intérieure (sous-conversation consciente); leurs marquages explicites sont liés soit à la ponctuation (tiret ou guillemets), soit à des introducteurs du dire (verbes introducteurs ou en incise).

Concernant *les discours rapportés* (DR), on a défini :

- La liste des types de discours rapportés avec un who (identifiant), un speaker éventuel, un whom allocutaire explicite[3] ; les segments (le plus souvent propositions ou incises) introducteurs de discours rapportés.
  (L'automatisation du balisage du DD paraît de prime abord la plus facile ; cependant il est nécessaire d'affecter préalablement le texte à la famille de marquage adéquate (familles que nous avons définies par ailleurs).
  <q type="DD"> désigne le discours direct.
  <q type="DI"> désigne le discours indirect.
  <q type="DDR"> désigne le discours direct rapporté.
  <q type="DRN"> désigne le discours rapporté narrativisé.
  <q type="DIN"> désigne le discours indirect narrativisé.
  <q type="MI"> désigne le monologue intérieur en DD</catDesc
  <q type="MII"> désigne le monologue intérieur en DI.
  <q type="MIN"> désigne le monologue intérieur narrativisé.
  <q type="disc_rapp"> désigne le discussion rapportée résumée.
  <q type="soCalled"> désigne le citation non prise en charge, modalisation autonyme.
  On éclate donc le discours indirect libre en MIN et DRN.

  <ab type="TdP"><desc>description d'un *tour de parole* dans une séquence dialoguée par les attributs facultatifs </desc>:
  who : désigne le locuteur du tour de parole
  whom : par défaut l'interlocuteur
  activity : activités et mouvements du locuteur
  <seg ana="geste"> : gestuelle liée à la parole
  <q type="DD" rend="tiret">contenu des paroles prononcées en discours direct</q>
  Les introducteurs de DR : <seg ana="incise_di"> balise l'incise de dire; <seg ana="int"> balise les introducteurs de DD ou le DI
  <seg ana="incise_pe"> : balise l'incise de discours intérieur (MI).
  L'attribution dans le roman du locuteur d'un tour de parole dans les dialogues doit pouvoir être partiellement automatisé à partir d'une description des familles de marquages du DD selon les périodes et auteurs.
- *Le psycho-récit* : description par le narrateur de la vie mentale et affective du personnage sans reproduction de celle-ci; le codage du psycho-récit est effectué sur la base de présence

[2] Il faut noter que les caractéristiques des œuvres entraînent des exigences différentes concernant le balisage : celui que je propose correspond au dispositif durassien et ne suffirait probablement pas pour décrire le dispositif de N. Sarraute.

[3] Le *who* concerne l'identité stable et unique du personnage locuteur, le speaker concerne le rôle par lequel le locuteur est désigné (ex : Anne Desbaresdes peut être désignée par « la femme », « la mère » etc). Le *whom* allocutaire explicite du tour de parole ne sera dénommé que lorsque le texte le désigne, l'allocutaire par défaut étant l'interlocuteur dans la séquence dialoguée. Dans les exemples qui suivent le key des interlocuteurs n'est pas balisé.

dans la phrase de lexique lié à la cognition ou à la vie affective et émotionnelle du personnage, autre que l'expressivité gestuelle.

- La description des personnages du roman dans <particDesc> : dénomination des personnages et des locuteurs : on déclare les personnages par leur nom propre ("reg"), par le *key*, par leur rôle, on note la d'apparition et de disparition du personnage et /ou de sa dénomination par un nom propre dans le texte. L'établissement des *actants principaux d'une diégèse* se fait par repérage des noms propres humains et des noms communs avec déterminant défini singulier les plus fréquents dans la diégèse. On a ainsi les personnages principaux sans nom propre et les actants non humains principaux (qu'on peut aussi identifier par calcul des spécificités, par ex. dans *Moderato Cantabile* : la fleur de magnolia, la musique, la mer).

Un exemple de profileDesc pour *Moderato cantabile*:

```
<profileDesc>
        <textClass>
            <domaines>littérature</domaines>
            <gn>ro.ps</gn>
        </textClass>
        <totDieg nbNiv="2"/>
        <diegNiv2 nb="1" surf="?"/>
        <diegNiv1 id="1" narr="extra_hétérodiég" narTemp="pas">
            <particDesc><personae key="ad ch pa enf mg in"/> </particDesc>
            </diegNiv1>
        <diegNiv2 id="2" narr="intra_hétérodiég" narTemp="pas pres"><particDesc>
        <personae key= "Fas Has"/>
            <particDesc><personae key="pa Fas Has"/></particDesc></diegNiv2
        <seqNiv2 id="2" nb="?" lm="?"/>< catDesc>nombre de séquences de la diégèse "2"
de niveau 2; longueur moyenne de ces séquences</catDesc>
        </totDieg>
```

Un extrait de texte balisé :

```
<ab locus="appartement_de_mg">
<p><q type="DD" rend="tiret">- Veux-tu lire ce qu'il y a d'écrit au-dessus de ta partition? <seg
ana="incise_di"> demanda la dame. </seg></q> </p>
<p><q type="DD" rend="tiret">- Moderato cantabile<seg ana="incise_di">dit l'enfant</seg>.</q>
</p>
<p><seg ana="geste">La dame ponctua cette réponse d'un coup de crayon sur le clavier.</seg>
<seg ana="geste">L'enfant resta immobile, la tête tournée vers la partition.</seg> </p>
<p><q type="DD" rend="tiret">- Et qu'est-ce que ça veut dire, moderato cantabile?</q> </p>
<p><q type="DD" rend="tiret">- Je ne sais pas.</q> </p>
<p><seg ana="geste">Une femme, assise à trois mètres de là, soupira.</seg></p>
<p><q type="DD" rend="tiret">- Tu es sûr de ne pas savoir ce que ça veut dire, moderato
cantabile? <seg ana="incise_di"> reprit la dame. </seg></q> </p>
<p>L'enfant ne répondit pas. <seg ana="geste"><PR>La dame poussa un cri d'impuissance
étouffé,</PR> tout en frappant de nouveau le clavier de son crayon.</seg> <seg ana="geste">Pas
un cil de l'enfant ne bougea.</seg><seg ana="geste"> La dame se retourna.</seg> </p>
<p><q type="DD" rend="tiret">- Madame Desbaresdes, quelle tête vous avez là<seg
ana="incise_di">, dit-elle</seg>.</q></p>
```

Balisage des diégèses dans le <body>:

```
<diegNiv2><p><q type="DD" rend="tiret">- Ce cri était si fort que vraiment il est bien naturel que
l'on cherche à savoir. J'aurais pu difficilement éviter de le faire, voyez-vous.</q> </p></diegNiv2>
<p>Elle but son vin, le troisième verre.</p>
<diegNiv2><p><q type="DD" rend="tiret">- Ce que je sais, c'est qu'il lui a tiré une balle dans le
cœur.</q> </p></diegNiv2>
<p>Deux clients entrèrent. Ils reconnurent cette femme au comptoir, s'étonnèrent.</p>
<diegNiv2><p><q type="DD" rend="tiret">- Et, évidemment, on ne peut pas savoir pourquoi ?</q>
</p></diegNiv2>
```

4 – **L'exploitation du corpus XMLisé**

L'exploitation des textes balisés reste malheureusement encore très peu ergonomique et conviviale pour le public littéraire ou linguistique et bon nombre de fonctionnalités souhaitables ne sont guère accessibles pour le moment.

- **On définit d'abord l'espace de recherche** : corpus entier *vs* sous ensemble de textes défini par un (ou +) trait(s) balisé(s) dans l'en-tête *vs* un texte.
  Ex : *l'étude des modaux dans le roman* implique de prendre en compte d'un côté (modaux) type de verbes, temps et personnes verbaux; de l'autre (corpus) de prendre en compte le champ générique et le genre (balisé dans l'en-tête dans <textClass>) et dans le roman i) le temps du récit de la diégèse en cours; ii) la séquence en cours : type de DR *vs* discours narratorial.

- **Analyses statistiques contrastives** :
  On peut donner quelques exemples d'exploitations pratiquées sur le roman :
  - type de narrateur et surface relative des différents DR (Malrieu, 2004)
  - les données statistiques morpho-syntaxiques peuvent diagnostiquer certains traits narratologiques : l'étiquetage des verbes du seul discours narratorial permet de voir s'il s'agit d'un récit au passé ou au présent; l'examen sur les différentes div de ce discours permet de voir si ce temps du récit est stable ou pas. De même le poids de la 1S dans ce discours narratorial permet de dire s'il s'agit d'un récit homodiégétique.
  - Les modes de désignation des personnages: dénomination vs désignations autres selon les séquences textuelles.
  - Sur un corpus important de romans, et dans une optique plus socio-historique, on peut envisager toutes sortes d'exploitation statistique de données synchroniques ou diachroniques, sur les genres, les actants, les lieux, les moments privilégiés des diégèses, etc.

- **L'analyse topologique des balises**: ce genre d'analyse n'est pas pratiqué sur l'écrit, (il l'est davantage sur les corpus oraux), il présente un grand intérêt et demanderait le développement de fonctionnalités spécifiques: les cooccurrences, positions respectives, enchaînements, rythmes, répétitions de balises fournissent des informations précieuses sur le fonctionnement du texte et peuvent donner lieu à des visualisations graphiques résumant la répartition d'un phénomène ou d'une classe de phénomènes dans le texte: ex dans le roman, les enchaînements gestuelle/psycho-récit; psycho-récit/monologue intérieur/ DD; sans parler des rythmes prosodiques.

- **Concordanciers enrichis** : le balisage des séquences textuelles permet un saut qualitatif dans la caractérisation des occurrences d'un phénomène étudié : en effet, on ne dispose plus seulement du contexte d'une occurrence dans une fenêtre de taille *n*, mais on doit pouvoir caractériser chaque occurrence par l'information sur ses contextes ascendants ou chemins dans l'arborescence: ex : occurrences de *on* sujet de telle catégorie de verbe à tel temps, selon les contextes: proposition principale ou relative ou conditionnelle; phrase interrogative vs déclarative; type de discours rapporté vs narratorial; paragraphe d'introduction de chapitre, etc;
  ex : densité des *on* selon les chapitres du roman et leurs types de séquences dominantes.

- **Aide à la désambiguïsation par la prise en compte du contexte balisé**:
  On peut considérer que la prise en compte de l'arborescence des unités supraphrastiques pour l'aide à la désambiguïsation des acceptions (ou des valeurs des temps verbaux par exemple) est un domaine encore en friche : le problème du coût de constitution des corpus finement enrichis fait pencher la balance vers les traitements statistiques lourds sur corpus pauvrement annotés. Mais il n'est pas évident que le coût final de ce dernier choix soit moindre, et la mutualisation de corpus richement annotés permettrait des exploitations illimitées dans les différentes disciplines des sciences du langage ou littéraires.
  ex : les modes de *résolution de la référence anaphorique* diffèrent selon que l'on est dans une séquence dialoguée ou dans le discours narratorial.
  ex : *valeur des temps verbaux* : si on est dans un genre narratif, si le temps du récit de la diégèse en cours = passé, si le conditionnel est dans un DI, alors le conditionnel a de fortes chances d'être un futur du passé.

5 - Un exemple d'analyse comparée des narrateurs à l'intérieur d'un œuvre de M. Duras

Les ruptures narratoriales à l'intérieur d'un œuvre ne sont pas rares chez Duras. Nous allons essayer de qualifier linguistiquement les deux configurations successives à l'intérieur du *Ravissement de Lol V. Stein* [4]: d'abord narrateur intra-hétérodiégétique, puis intra-homodiégétique. Ces deux configurations n'autorisent pas le même psycho-récit (Cohn). L'intradiégéticité met en scène un narrateur qui peut à la fois affirmer son non savoir sur le passé de Lol et impliquer le lecteur dans ses interrogations et s'adonner à un psycho-récit très empathique et projectif qui dépasse largement la parole intérieure ou le savoir du personnage sur lui-même. Le passage au récit homodiégétique affirme la thèse épistémique de Duras : il n'y a de connaissance de l'autre que dans la relation : le récit de cette deuxième partie favorise l'emprise du lecteur en combinant le mimétisme (récit au présent, dialogues en discours direct), et le psycho-récit du vécu du narrateur dans ses réactions à Lol, les deux étant intimement mêlés.
Pour comparer les deux parties, nous avons balisé selon la méthodologie de la TEI-XML[5] les différents discours rapportés, les séquences narratoriales (intradiégétiques homo- vs hétérodiégétiques) ainsi que le psycho-récit[6].
Les variables utilisent les sorties de l'analyseur CORDIAL.

Résultats :
Poids des différents discours dans les deux parties
Par rapport à la surface totale en nombre de mots de chaque partie, le récit évènementiel (ce qui n'est ni discours rapportés (DR), ni psycho-récit (PR)) occupe des surfaces similaires (environ 40%). Le PR occupe une place plus importante dans le partie hétérodiégétique (38,5% contre 22%), le credo du narrateur sur Lol. La partie homodiégétique comporte davantage de discours direct (DD : 23% de la surface contre 2%) par contre les DR narrativisés sont moins importants (DRN 40%, MIN 15% et le DI 20% dans l'hétérodiégétique[7] n'atteignent pas 10% dans la partie Ho).
La partie homodiégétique connaît une configuration déictique cohérente : présentification du récit, dont le poids du DD est l'expression, récit au présent, marques déictiques spatiales (immédiateté du corporel et du visuel, qui embarque le lecteur dans un espace partagé).

Les temps verbaux :
Comme le montre le Tableau 1, l'homodiégétique est massivement dans le présent[8], le passé (IMP, PS et PQP) est plus important dans le récit hétérodiégétique (récit Hé).

	PR-Ho	PR-Hé
Présent indic	70,2	41,4
Imparfait	5,7	21,9
PS	1,5	11,4
PC	10,5	5,7
Futur	4,2	0,2
Conditionnel	3,4	3,6
PQP	0,5	7
Impératif	0,8	0,7

Tableau 1 : Les temps dans les deux psycho-récits

Les propriétés de la phrase : longueur et ponctuations faibles
La distinction du récit évènementiel (récit tout court) et du PR permet d'analyser en quoi l'expressivité des affects et l'idéalisation dans le PR sont corrélés avec des rythmes de phrase différents du récit et d'explorer si ces derniers diffèrent dans l'homo- et l'hétérodiégétique.
Nous avons observé *la répartition des longueurs de la phrase* : le PR, plus nettement le PR Hé, connaît des phrases plus longues que le récit, ce qui confirmerait le caractère plus élaboré du PR dans la partie Hé (plus fort % de substantifs et adjectifs/ mots signifiants, de noms abstraits, de pronoms relatifs); l'hétérodiégétique (récit comme PR) connaît des phrases plus longues que

[4] Les résultats plus complets de cette étude sont accessibles sur le site : infolang.u-paris10.fr/modyco
[5] http://www.tei-c.org/P4X/
[6] Est balisé comme PR tout passage faisant allusion à la vie psychique, émotionnelle du personnage.
[7] DRN = discours rapporté narrativisé, MIN = monologue intérieur narrativisé, DI = discours indirect, Ho = homodiégétique, Hé = hétérodiégétique; 1S = première personne du singulier.
[8] Ce qui le distingue du récit homodiégétique au passé comme *Le rivage des Syrtes* par exemple.

l'homodiégétique. On observe davantage de phrases très brèves dans le récit Ho (10,4% contre 2%). Les phrases ≤ 7mots représentent 18,5% dans le PR Hé contre 30,6% dans le PR Ho.

	PR-Ho	PR-Hé	Réc-Ho	Réc-Hé	Hiérarchie
Quartile 1	6	8	5	7	PR Hé > autres
Médiane	12	14	8	12	PR Hé > autres
Quartile 3	21	24	13	18	PR > Récit

Tableau 2 – Répartition des longueurs des phrases en nombre de mots

Nombre de virgules et longueur des phrases :

a) *récit homo vs hétérodiégétique* :

nombre de virgules : le récit Ho comporte davantage de phrases sans virgule (63,3% contre 52,8%), le récit Hé plus de phrases de 2 à 3 virgules.

b) *PR homo- vs hétérodiégétique* :

Nombre de virgules : le PR Ho, qui connaît des phrases plus courtes, comporte davantage de phrases sans virgules, moins de phrases à 2 ou plus de 4 virgules, sans exclure l'existence de phrases de 8 à 18 virgules. Les phrases à plus de 5 virgules représentent 4,5% dans le PR Ho contre 6,5% dans le PR Hé.

Les différences entre homo- et hétérodiégéticité montrent entre autres que l'homodiégétique est dominé par des phrases brèves, il est moins dans la représentation au passé d'un point de vue externe, plus dans la vivacité des évènements au présent. L'homodiégéticité induit une autre forme de psycho-récit : elle autorise moins l'intrusion du discours narratorial de l'auteur et le psycho-récit s'inscrit par petites touches, par phrases brèves à l'intérieur du récit évènementiel ou des dialogues. Cependant elle n'interdit pas la phrase fortement rythmée, exprimant l'exacerbation des affects (cf. l'apparition de la virgule dans des phrases plus courtes et le plus grand nombre de virgules dans les phrases longues dans le psycho-récit homodiégétique).

Les indices thématiques : fréquence comparée des lexèmes

La comparaison des lexiques des différentes parties vient confirmer ces différences[9].

Les calculs suivants portent sur les lemmes et non sur les formes. Le % d'hapax /au nombre de lemmes différents de chaque récit ou PR montre davantage d'hapax dans les PR (60% contre 52 et 57% dans le récit homo et le récit hétéro). On peut donc dire que le psycho-récit est le lieu de la richesse lexicale, ce qui n'a rien de surprenant, vu son affinité avec la prose poétique.

Les lemmes statistiquement plus fréquents dans les PR par rapport aux récits correspondent à un lexique ayant trait aux affects (*absence, aimer, amour, cœur, craindre, douleur, dieu, larme, idée, ignorer, inconnu, infini, joie, mensonge, mentir, mort, nommer, ordre, penser, peur, rêve, souffrance, souffrir, souvenir, souhaiter, surprendre, tristesse*) mais aussi l'adjectif démonstratif singulier, certains connecteurs (*alors que, mais*).

L'examen du lexique plus fréquent dans le psycho-récit Hé / total indique un déficit pour les pronoms personnels 1S, et une surreprésentation des adjectifs démonstratifs et interrogatifs, de l'exclamation, qui marquent l'implication du narrateur, des privatifs (*ignorance, immobilité, immuable, impossible, inconnu, inconsolable, infini, infirme, inimportance, vain*), des références au monde distal (*infini, monde, univers, murer, terre, dieu, pénétrer, nom, nommer, avenir, à jamais*), des métaphores (*aurore, cendre, port, hiver, lumière, été, jour, nuit, navire, blancheur d'os, souffle, trou, mot-trou, mot-absence, gong* etc).

Les spécificités du lexique du psycho-récit Ho par rapport au lexique total (récit + PR) concernent, à côté du pronom personnel ou possessif 1S, les pronoms quantificateurs (*tout* et *rien*) ou les adverbes évaluatifs (*trop, si*), le lexique affectif (*larme, mentir, mensonge, souhaiter, joie, amour, peur, épouvante, rêver, sentiment, victoire, cœur, mort, souffrir, aimer, apaiser, accueillir, broyer, confiance, détruire, douceur, effrayer, horreur* etc), le lexique épistémique (*version, comprendre, ignorer, sens, savoir, se tromper*).

Conclusion

S'il est bien clair que le *literary computing* et plus précisément la méthodologie de balisage liée à XML et à la TEI autorisent maintenant des analyses contrastives fines des textes, il me semble que les

[9] La comparaison des fréquences des lemmes des différentes parties est faite par calcul de l'écart réduit selon la formule Z = (k-fp)/RACINE(fpq) où k = la fréquence du lemme dans la partie du texte étudiée, f = la fréquence du lemme dans le corpus de référence, p = le % total de lemmes du texte étudié par rapport au total de lemmes du corpus de référence, q = 1-p.

obstacles à surmonter pour convaincre la communauté des littéraires (qui est loin de partager ce point de vue dans sa majorité) sont de quatre ordres :

- arriver à constituer des corpus balisés suffisamment volumineux et mutualiser ces ressources : pour cela il faut surmonter le goulot d'étranglement concernant le passage au format XML. Si le balisage de la structure logique du texte est évidemment d'un intérêt très limité pour l'analyse des textes littéraires, il paraît possible et nécessaire de développer des outils de balisage automatique modulables selon les genres textuels. Par exemple, le balisage de la valeur sémantique de l'italique pourrait être automatisé car son usage est réglé à l'intérieur d'un genre, de même le balisage du discours direct à l'intérieur des familles chronologiques de son marquage typographique; de même le balisage des entités nommées et des actants principaux dans le roman.
- Un travail collectif de représentation des connaissances liées à un balisage plus fin que la structure logique s'avère nécessaire; une mutualisation des représentations des balises gagnerait aussi à se définir selon les grands genres de textes (poésie, roman, articles de presse, etc).
- L'intégration des étiquetages morpho-syntaxiques et syntaxiques au sein d'un corpus XMLisé n'est pas non plus aisée pour l'instant. Le problème d'une représentation ergonomique du balisage se pose aussi et la difficulté à lire les corpus lourdement "farcis" pousserait à une séparation en portée des différentes couches de balises (structure logique, séquences textuelles autres, balisage morpho-syntaxique, balisage prosodique etc.).
- Enfin l'exploitation des textes balisés est pour l'instant loin d'être conviviale pour le linguiste ou littéraire non programmeur, et il semble urgent de mettre à la disposition de ces communautés des outils de questionnement qui autorisent aussi bien la qualification de toute occurrence par son chemin dans l'arborescence du texte, que l'analyse topologique de la répartition des balises dans le texte, que la visualisation des occurrences d'un phénomène au sein du texte source.

Références

Bakhtine, M. (1984 [1952-53]), *Esthétique de la création verbale*, Paris, Gallimard.

Benveniste, E. (1966). *Problèmes de linguistique générale*, Paris, Gallimard.

Biber, D., Conrad, S., Reppen, R. (1998), *Corpus Linguistics: Investigating Language structure and Use*, Cambridge University Press.

Bird, S. & Liberman, M. (2001). A formal framework for linguistic annotation, *Speech Communication*, 33, 23-60.

Bouquet, S. (éd.) (2004). "Les genres de la parole", *Langages, n° 153*.

Bronckart, J.P. (1996). *Activité langagière, textes et discours*, Lausanne, Delachaux et Niestlé.

Cohn, D. (1981). *La transparence intérieure*. Paris, Le Seuil.

Declerck, R. (2003). "How to manipulate tenses to express a character's point of view", *Journal of Literary Linguistics*, 85-112.

De Mattia, M., Joly, A. (2001). *De la syntaxe à la narratologie*, Paris, Ophrys.

Fillietaz, L. et Grobet, A. (1999) "L'hétérogénéité compositionnelle du discours : quelques remarques préliminaires, *Cahiers de linguistique française*, n° 21, 213-260.

Genette, G. (1972), *Figures III*, Paris, Le Seuil.

Genette, G. (1983), *Nouveau discours du récit*, Paris, Le Seuil.

Genette, G. (éd.) (1986). *Théorie des genres*, Paris, Le Seuil.

Habert, B. et coll. (2000). "Profilage de textes : cadre de travail et expérience", *Actes des 5èmes Journées JADT*.

Hamburger, K. (1986, (trad.), [1957]). *Logique des genres littéraires*, Paris, Le Seuil.

Henrot, G. (2000), *L'usage de la forme. Essai sur* les Fruits d'or *de Nathalie Sarraute*, Biblioteca Francese, Unipress, Padova.

Kuyumcuyan, A. (2002), Diction et mention. Pour une pragmatique du discours narratif, Berne, Peter Lang.

Langue française (2001). "La Parole intérieure." n° 132.

Lintvelt, J. (1989). *Essai de typologie narrative. Le "point de vue".Théorie et analyse*. Corti. Paris.

Lips, M. (1926), *Le Style indirect libre*. Payot. Paris.

Maingueneau, D. (1986). *Eléments de linguistique pour le texte littéraire*, Paris : Bordas.

Maingueneau, D. (2000). Instances frontières et angélisme narratif, *Langue française*, 128; pp. 74-95.
Malrieu, D. & Rastier, F. (2001). "Genres et variations morpho-syntaxiques *", T.A.L., 42, 2*, 547-577.
Malrieu D. (2002). "Stylistique et Statistique textuelle: À partir de l'article de C. Muller sur les pronoms de dialogue", JADT 2002, 6èmes Journées internationales d'analyse des données textuelles, St-Malô, 13-15 mars 2002.
Malrieu, D. (2004). Linguistique de corpus, genres textuels, temps et personnes, *Langages*, *153*, 73-85.
Malrieu, D. "Discours rapportés et typologie des narrateurs dans le genre romanesque", Actes du colloque Ci-Dit de Cadiz "Dans la jungle des discours", 11-14 mars 2004, (à paraître chez l'Harmattan).
Malrieu, D. (2006). Type de narrateur et place du lecteur dans *Le ravissement de Lol V. Stein* (infolang.u-paris10.fr/modyco)
Marnette, S. (2002). "Étudier les pensées rapportées en français parlé: Mission impossible?". *Faits de Langues, 19*, p 211-20.
Marnette, S. (2002). "Aux frontières du discours rapporté". *Revue Romane*, *37.1*, p 3-30.
Marnette, S. (2001). "The French Théories de l'Énonciation and the Study of Speech and Thought Presentation". *Language and Literature*, *10.3,* p 261-80.
Mellet, S. & Vuillaume, M. (éds.) (2000). *Le style indirect libre et ses contextes*. Rodopi. Amsterdam-Atlanta.
Newman A. S. (1976). *Une poésie des discours. Essai sur les romans de Nathalie Sarraute*, Genève : Droz.
Philippe, G. (1995). "Pour une étude linguistique du discours intérieur dans "Les Chemins de la liberté" : le problème des modalités du discours rapporté", *Ritm, 11*, 119-155.
Philippe, G. (ed.) (2000). L'ancrage énonciatif des récits de fiction, *Langue française*, 128.
Rabatel, A., 1998, *La construction textuelle du point de vue*, Lausanne, Paris, Delachaux & Niestlé
Rastier, F. (2001). *Arts et sciences du texte*. Paris, PUF.
Rivara, R. (2000). *La langue du récit*, Paris, L'Harmattan.
Rosier, L. (1999). *Le discours rapporté. Histoire, théories, pratiques.* Paris, Bruxelles, Duculot
TEI, http://www.tei-c.org
TEI P5, http://www.tei-c.org/P5/