

Profils linguistiques et structuration textuelle

Mathilde Dargnat

LPL – 29 avenue Robert Schuman – 13 621 Aix-en-Provence – France

Abstract

By analyzing a chronological sequence (1968-1998) of five plays by the Quebecois writer Michel Tremblay, I explain how the author, when representing the vernacular, adds a meta-literary dimension relevant to text structure to the socio-linguistic dimension. I make use of two simple statistical tests to explore the linguistic profile of characters, as it emerges from an annotated corpus. This study shows that, by connecting narratology issues with such profiles, one can shed light on the question of how author styles are situated in the history of literary practices.

Résumé

En m'appuyant sur l'analyse d'une série chronologique (1968-1998) de cinq pièces de théâtre de l'écrivain québécois Michel Tremblay, j'expose comment la représentation de l'oral populaire construite par l'auteur s'enrichit, par ajout au marquage sociolinguistique, d'une dimension métalittéraire de marquage de la structuration textuelle. L'étude repose sur deux tests statistiques simples portant sur le profil linguistique des différents personnages, constitué à partir d'un corpus balisé. Elle montre que l'articulation des problématiques de type narratologique et des profils linguistiques statistiquement identifiables éclaire la question du positionnement des styles d'auteur dans l'histoire des pratiques littéraires.

Mots-clés : analyse de corpus assistée par ordinateur, codage xml, tests statistiques non paramétriques, stylistique, textes dramatiques, variation linguistique.

1. Problématique générale et hypothèses

Comment lisons-nous d'ordinaire ? Vite, trop vite, en aperçu. Mais supposons-nous doués d'une intelligence rigoureuse et informés de mieux en mieux de tout ce que la lecture courante laisse échapper : que de petites perceptions, que de liaisons d'abord insensibles apparaîtraient¹.

1.1. Problématique générale

Cette étude s'insère dans le cadre d'un travail plus large (Dargnat, 2006) sur l'évolution de la représentation de l'oralité populaire québécoise (désormais OPQ)² dans un corpus de textes

¹ Y. Bélaval, préface à A. Becco, Du Simple selon G. W. Leibniz. Discours de Métaphysique et Monadologie. Étude comparative critique des propriétés de la substance appuyée sur l'opération informatique « MONADO 74 ». Vrin/CNRS, 1975, p. VIII.

² La pratique linguistique visée a un temps été appelée « joual ». Le terme joual est une déformation phonétique du mot cheval et, par métonymie, a fini par désigner la manière de parler elle-même. Il n'est pas possible donner une seule définition de ce terme, car, empreint d'idéologie, il est fortement connoté symboliquement dans la société et la littérature québécoises (voir Daoust, 1983 et Dargnat, 2006, tome 2, p. 54-64 pour une compilation des définitions).

dramatiques de l'écrivain Michel Tremblay (sous-corpus *tremblay*)³, dont les productions sont assez représentatives des différents courants esthétiques au Québec depuis les années soixante. J'ai d'abord été amenée à définir l'objet OPQ dans ses composantes linguistiques majeures (en particulier phonétique, syntaxique, lexicale et discursive) en m'appuyant sur les études existantes et en constituant un sous-corpus représentatif (désormais *frcapop*), à partir des corpus de référence *Sankoff-Cedergren* et *Montréal 84*⁴. La liste de traits relevés dans lesdites études a été validée dans le sous-corpus *frcapop*, afin d'obtenir une liste plus restreinte de variables linguistiques jugées caractéristiques de l'objet, et de proposer ainsi une sorte de pierre de touche pour la comparaison entre une pratique linguistique attestée et les textes de théâtre. L'enjeu est ici la nature plus ou moins réaliste des saillances linguistiques⁵ présentes dans les textes. L'objectif principal est de montrer que l'usage que fait Michel Tremblay des marques d'OPQ dans ses textes change au fil des années, et de donner une interprétation stylistique de ce changement.

1.2. Hypothèses

Ce travail repose sur deux hypothèses principales :

- Les saillances linguistiques dans les textes de théâtre renvoient à des observables linguistiques et sont donc des cautions du réalisme langagier des personnages ;
- La distribution de ces marques littéraires d'OPQ évolue au fil des textes et des années. Plus précisément, leur fonction de distinction des personnages s'enrichit d'une fonction de marquage de la structure énonciative des textes en question.

Le présent article prend pour acquise et validée la première hypothèse et expose le raisonnement et les analyses qui conduisent à valider et expliquer la seconde, notamment au moyen de la catégorisation statistique des profils linguistiques. Nous ne traitons donc ici que du sous-corpus de textes dramatiques, et nous n'illustrons l'analyse générale qu'au moyen de deux pièces, dont la mise en perspective est très significative⁶. Il est évident que parler d'*évolution* des fonctions des marques d'OPQ, et plus généralement d'*évolution* stylistique, suppose que l'on considère les trois autres pièces du corpus dramatique, dont les études corroborent la dynamique mise en évidence dans la comparaison des deux textes ici analysés.

³ Ce sous-corpus se compose de *Les Belles-soeurs* (1968), *Bonjour, là, bonjour* (1974), *L'Impromptu d'Outremont* (1980), *Le Vrai Monde ?* (1984) et *Encore une fois, si vous permettez* (1998).

⁴ Le corpus Sankoff-Cedergren a été fait par Gillian Sankoff et Henrietta Cedergren en 1971 et le corpus Montréal 84 a été fait par Pierrette Thibault et Diane Vincent en 1984. Voir Thibault & Vincent (1990).

⁵ Par saillance linguistique, j'entends toute perception d'un écart par rapport à la norme écrite. Ce sont, par exemple, les néographies phonétisantes, les particules discursives caractéristiques du français parlé au Québec, les structures syntaxiques de l'oral comme les relatives non standard et les dislocations, certaines constructions interrogatives et exclamatives au moyen de la particule « -tu, » ou encore l'usage des sacres et des québécismes.

⁶ Pour les résultats concernant les autres pièces, les schémas de leur structure énonciative et une discussion plus approfondie, voir ma thèse en ligne (Dargnat, 2006) et les annexes correspondantes : <http://mathilde.dargnat.free.fr>, vol. I, chap. 6 et annexes.

2. Traitement logiciel et statistique des données

2.1. Exploitation logicielle du corpus : intérêts et limites

Pour travailler de manière structurée et systématique sur le matériau linguistique attribué aux personnages de chaque pièce constitutive de *tremblay*, j'ai balisé les textes et constitué une base de données exploitable par le logiciel *Webllex* (Heiden, 2002). La structuration de l'information textuelle a été faite en XML en suivant les indications de la TEI⁷ (Burnard & Sperberg-McQueen, 2002, chap. 10 ; Bernet, 2001). Il en résulte une base dans laquelle on peut travailler en sélectionnant le discours des personnages, tout en conservant les didascalies et le paratexte dans l'édition, pour comparer aussi bien les personnages entre eux, les sections textuelles entre elles (actes, scènes, etc.) que les pièces entre elles. Cela a notamment permis d'établir le vocabulaire complet et d'en extraire manuellement certaines des particularités littéraires de l'OPQ de manière quantitative, comme par exemple les apostrophes devant consonnes, les « néographies phonétisantes » (Anis, 1999), le paradigme oral des pronoms personnels, les régionalismes, les sacres, etc. Ce décompte est l'étape préliminaire de la catégorisation statistique des profils linguistiques des personnages telle qu'elle est présentée dans la section 2.2.

Cependant, le traitement d'encodage et l'exploitation logicielle demeurent insuffisants pour mener à bien l'analyse stylistique. Cela pour deux raisons principales et indépendantes. Premièrement, seule la structure textuelle a été encodée, l'unité la plus large étant le texte lui-même et l'unité la plus petite la réplique. Ce qui veut dire qu'aucun codage proprement linguistique n'a été fait⁸, alors que je souhaitais travailler sur des phénomènes de types phonographique, lexical, syntaxique et pragmatique. L'aspect quasiment brut du corpus, non étiqueté, rend très difficile voire impossible la récupération systématique par une requête simple des phénomènes constructionnels sans marque spécifique ou sans marque du tout (par exemple les dislocations, l'absence de déterminant, les relatives pléonastiques, les parataxes, etc.). La liste des variables retenues pour l'analyse statistique est donc contingentée dès le départ par les outils d'observation déployés. Deuxièmement, le logiciel utilisé, qui a été choisi pour sa souplesse de structuration et sa disponibilité sur le web, ne propose pas à proprement parler de tests statistiques⁹. L'encodage et l'exploitation logicielle n'ont donc été qu'une étape

⁷ Je remercie Serge Heiden et Charles Bernet de leurs conseils et de leur aide dans cette tâche, laborieuse mais nécessaire.

⁸ On aurait pu envisager pour cela d'utiliser un étiqueteur comme *TreeTagger* (<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger>), mais, compte tenu de ce que les formes linguistiques sont non standard et non systématiques dans les textes de théâtre, cela n'aurait pas été très rentable. Il aurait fallu faire de très nombreuses vérifications et, par exemple, redéfinir un dictionnaire de formes au fur et à mesure du codage. Le codage morpho-syntaxique du sous-corpus linguistique *fricapop* pose à ce niveau-là moins de problèmes, puisque les normes de transcription sont explicites et relativement bien observées par les transcripteurs.

⁹ Il serait faux de dire que ce logiciel ne comporte pas d'outil de mesures contrastives, puisqu'il propose un calcul des spécificités du vocabulaire ou d'une chaîne définie par une expression régulière, selon la syntaxe d'interrogation CQP (Heiden, 2002, chap. 6) – calcul qui repose sur les études de P. Lafon (1984). J'ai effectué quelques tests ponctuels, mais je souhaitais avant tout utiliser des tests statistiques de significativité, comme ceux ici présentés, et pas seulement des outils de représentation. Il serait cependant intéressant et bienvenu, comme le suggère un lecteur anonyme, que je remercie pour sa remarque, de regarder si une analyse des spécificités pour chaque personnage révélerait des associations particulières entre les différents types de marques d'OPQ et les personnages. Cette étude pourra faire l'objet d'un développement ultérieur, au même titre qu'une analyse factorielle et qu'une analyse en constituant principal.

dans le raisonnement et l'analyse de la problématique, mais une étape importante pour la qualité des relevés quantitatifs à la base de l'échantillonnage maintenant présenté.

2.2. Profils linguistiques des personnages et tests non paramétriques

Pour savoir si la représentation de l'OPQ a changé, deux critères se combinent : celui des marques elles-mêmes (leur nature et leur présence/absence) et celui de leur distribution textuelle (niveau énonciatif concerné et différenciation des personnages).

2.2.1. Variables linguistiques retenues

L'idéal aurait été de travailler sur l'ensemble des marques représentatives de l'OPQ, mais il n'était pas possible ou tout simplement pas pertinent de le faire. En effet, en plus des raisons évoquées ci-dessus à propos de l'interrogation de la base dans *Weblex*, la prise en compte de certains phénomènes trop rares n'était pas pertinente pour une analyse comparative basée sur des tests statistiques. Seules douze variables ont été conservées, compte tenu de leur caractère représentatif de l'OPQ dans différentes composantes de la description linguistique, de leur observabilité dans le corpus et de la possibilité de les compter automatiquement. Cela a parfois conduit à regrouper les phénomènes sous des classes génériques, comme c'est le cas pour les variables lexicales, et ce pour des raisons d'abord quantitatives.

1. Phonographie

- a. « pis »
- b. « ben »
- c. « toé », « moé »
- d. ouverture de « è » en « a » devant « r » (ex. farmer, narveuse)
- e. apostrophes devant consonnes (ex. v'nir, à'voulait)

2. Morpho-syntaxe

- f. « à » et « dans » + N (ex. à'soir, dans'cuisine)
- g. particule « ça fait que »
- h. particule interrogative et exclamative « tu » (ex. c'est-tu beau ! A' vient-tu ?)
- i. négation en « pas » sans « ne » (j'en veux pas)

3. Lexique

- j. sacres (ex. crise, câlisse)
- k. québécoisismes (ex. garrocher, magané)
- l. anglicismes (ex. pinottes, grocerie)

Variables linguistiques retenues pour les profils linguistiques

2.2.2. Tests non paramétriques : Welch et Fisher

Une fois le profil OPQ-type déterminé en fonction de ces douze variables, j'ai voulu tester si l'analyse de leur distribution dans les textes mettait en évidence des profils linguistiques différenciés pour chaque personnage. Autrement dit, l'hypothèse testée a été la suivante :

Hypothèse

Est-ce que la différence de distribution des variables observée entre les différents personnages, à l'intérieur de chaque pièce, est significative (non due au hasard) ?

J'ai utilisé deux tests non paramétriques, le test de Welch, qui mesure une différence de moyennes, et le test de Fisher, qui, comme celui du Khi carré, mesure une corrélation – en l'occurrence une corrélation entre l'identité du personnage et les proportions de marques

d'OPQ. Les deux tests sont censés donner des résultats cohérents. Il s'agit de tests très simples¹⁰, qui correspondent à des situations d'analyse très courantes en stylistique, tant du point de vue de la poétique que de la sociolinguistique¹¹ (cf. Muller, 1992), et qui permettent déjà d'aboutir à des conclusions intéressantes. Outre leur pertinence par rapport à l'hypothèse, ces deux tests ont été choisis en fonction du type de données recueillies. Ces données sont proportionnelles, leur distribution n'est pas symétrique, leurs variances ne sont pas homogènes et leur taille est par moment insuffisante pour accepter l'approximation par une distribution normale. Dans ces cas-là, les tests de significativité non paramétriques présentent plus de garanties. D'autre part, le Fisher a été préféré au Khi 2, car il est plus précis, plus sûr, et convient mieux aux données à faible score, ce qui est le cas ici. Pour des discussions sur ces deux questions (non paramétrique et différence entre Khi carré et Fisher), voir entre autres Rietveld & van Hout (2005, p. 125), Manning & Schütze (2002, p. 50-54), Oakes (1998, p. 10-16) et Ramousse et al. (1996, chap. 3).

2.2.3. Présentation et lecture des résultats

Les calculs ont été faits avec le logiciel statistique *R* (Venables & Ripley, 2002) et ont consisté à tester l'hypothèse mentionnée ci-dessus par paires de personnages, à l'intérieur de chaque pièce¹². Les résultats sont présentés sous forme de tableaux à double entrée, dans lesquels l'intersection d'un personnage avec un autre est notée 1, lorsque la différence a été jugée statistiquement significative, et non notée, lorsqu'elle n'a pas été jugée significative. Le seuil de significativité choisi est 5%, et les résultats obtenus, convergents pour les deux tests, sont conservés avec un seuil de 1%. Je ne commente ici que deux tableaux, qui correspondent respectivement aux pièces *L'Impromptu d'Outremont* (IO data) et *Encore une fois, si vous permettez* (EF data). Je ne reproduis qu'un seul tableau par pièce puisque les résultats, en termes de significativité (1 ou rien), sont identiques pour les deux tests. La fréquence des marques d'OPQ est également indiquée (par rapport à la taille totale du discours de chaque personnage), c'est elle qui permet de déterminer les pôles /+OPQ/ et /-OPQ/.

ANALYSE DE L'IMPROMPTU D'OUTREMONT --- seuil = 5%

Test de Welch et test de Fisher

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]		Fernande	Lorraine	Lucille	Yvette
[2,]	Fernande		1	1	1
[3,]	Lorraine	1		1	1
[4,]	Lucille	1	1		
[5,]	Yvette	1	1		

Fréquences pour IO (total des occurrences des 12 variables / taille totale de chaque personnage, en nombre de mots).

¹⁰ L'analyse statistique pour regrouper ces douze variables, sur le modèle d'une analyse factorielle pour des données non paramétriques, est en cours, mais je n'en parlerai pas ici.

¹¹ Le terme *style* est en effet utilisé pour des phénomènes littéraires et pour des phénomènes sociolinguistiques. Dans les deux cas, c'est le caractère significatif ou non d'une variation linguistique qui est en jeu, seuls les critères de variation et leur interprétation diffèrent selon le point de vue adopté. Cf. Lefebvre (1983) et Rastier (2001) pour différentes facettes définitoires de ce terme.

¹² Je remercie Jacques Jayez, qui m'a aidée pour la programmation dans *R*.

Fernande : 0.01090093 (taille totale = 6 238)

Lorraine : 0.04473928 (taille totale = 3 241)

Lucille : 0.02382926 (taille totale = 4 826)

Yvette : 0.01764095 (taille totale = 2 891)

Dans cette pièce, on peut déterminer trois groupes de personnages. Fernande, qui correspond au pôle /-OPQ/, Lorraine, l'extrême inverse, qui correspond au pôle /+OPQ/, et au milieu, Yvette et Lucille, qui sont associées à des profils linguistiques intermédiaires.

ANALYSE DE ENCORE UNE FOIS SI VOUS PERMETTEZ --- seuil = 5%

Test de Welch et test de Fisher

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]		Nana	Le Narrateur	N1	N2
[2,]	Nana		1	1	
[3,]	Le Narrateur	1		1	1
[4,]	N1	1	1		1
[5,]	N2		1	1	

Fréquences pour EF (total des occurrences des 12 variables / taille totale de chaque personnage, en nombre de mots).

Nana : 0.1026456 (taille totale = 11 642)

Le Narrateur : 0.08095029 (taille totale = 4 546)

N1 : 0.0009191176 (taille totale = 1 088)

N2 : 0.1061307 (taille totale = 3 458)

Cette deuxième pièce est plus complexe dans sa structure. Il n'y a en réalité que deux personnages déclarés : Nana et Le Narrateur. N1 et N2 renvoient à la décomposition du personnage Le Narrateur, en fonction de son interlocuteur dans le texte : le public pour N1 et Nana pour N2. Il n'y a donc textuellement que deux personnages, mais statistiquement, cela correspond à trois profils linguistiques différents : d'abord Nana et N2, qui se présentent comme le pôle /+OPQ/, ensuite N1, qui correspond au pôle /-OPQ/, et enfin Le Narrateur, ce collage de N1 et N2, dont le profil est hybride.

D'un point de vue purement typologique, l'analyse pourrait s'arrêter là, puisque les variables retenues ont permis de mettre en évidence l'existence de plusieurs profils langagiers correspondant à un personnage ou à un groupe de personnages. Cependant, le cadre stylistique général de la recherche et l'hypothèse d'une évolution de la fonction des marques d'OPQ dans les textes demandent que l'on interprète ces résultats aussi d'un point de vue qualitatif.

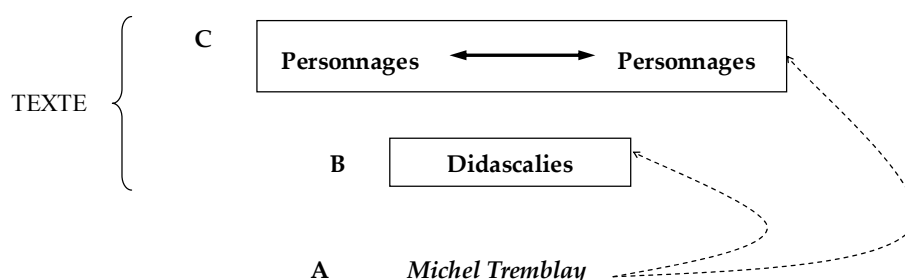
3. Interprétation des résultats

3.1. Répartition des marques d'OPQ et valeur stylistique

Les deux pièces choisies pour les besoins de la démonstration montrent relativement bien le changement de la fonction des profils linguistiques dans l'œuvre de Michel Tremblay, et on pourrait dire dans le théâtre québécois depuis les années soixante. L'analyse complète des cinq pièces constitutives du sous-corpus *tremblay* montre que cette évolution n'est ni linéaire ni radicale, mais je m'arrêterai essentiellement sur la fonction sociolinguistique et énonciative standard des profils (3.1.1) et sur leur fonction métatextuelle (3.1.2). Je rappelle ici que les

profils ont été calculés sur la base du matériau verbal attribué aux personnages et que, même si elles participent à la structuration textuelle, il n'a pas été question de prendre en compte les didascalies, qui sont, de manière très commune, rédigées sans marques d'OPQ¹³. Pour rendre compte de ces deux fonctions, on peut schématiser la structure énonciative du texte sous la forme d'un « feuilleté », à l'instar de ce proposent Molinié (1989, 1998) et Stolz (1999), et dans un autre domaine Haller & Schneuwly (1996), ou encore sous la forme d'une « pile narrative », si l'on suit Ryan (1991, chap. 9). Dans tous les cas, ce qui importe, c'est la prise en compte d'une hiérarchie des différents niveaux et des différents actants de l'énonciation textuelle, en particulier littéraire.

3.1.1. Fonction sociolinguistique et énonciative standard des profils linguistiques



Légende

A : l'écrivain

↔ relation locuteur/allocutaire

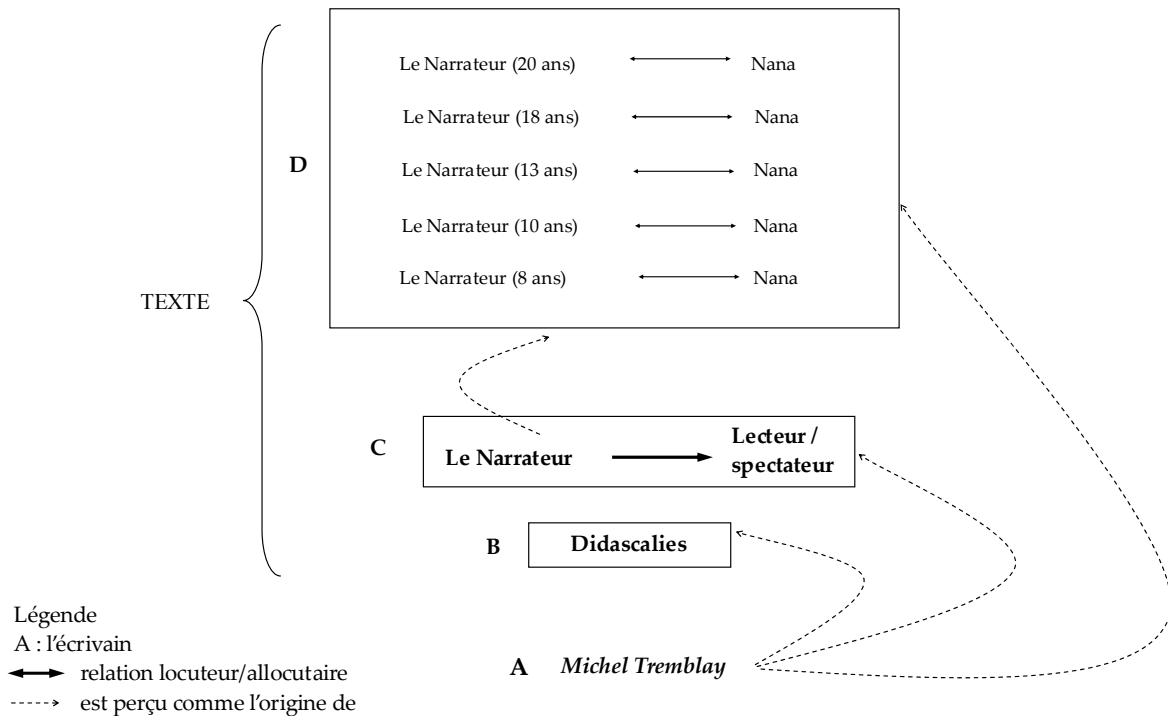
-----> est perçu comme l'origine de

Structure textuelle (énonciative) de L'Impromptu d'Outremont (1980)

Dans ce texte, l'usage de profils linguistiques différenciés joue à un seul niveau énonciatif (niveau C), celui des échanges entre personnages d'une même dimension spatio-temporelle. Du point de vue fictionnel, les quatre personnages sont censés être originaires du même milieu, la bourgeoisie francophone montréalaise des années quatre-vingt. On pourrait donc s'attendre à ce qu'il y ait très peu de marques d'OPQ. La différenciation linguistique manifeste entre Fernande (/ -OPQ /) et Lorraine (/ +OPQ /) est d'ordre sociolinguistique. La première est présentée comme garante des valeurs familiales et culturelles traditionnelles, qualifiée de « vierge de la langue française », et la seconde, mariée à un immigré italien qui parle anglais, habite dans l'Est montréalais (quartiers populaires), est la « quétaine » (ringarde) de service, qui parle mal, jure et crie, et a en quelque sorte trahi ses origines. Les deux autres personnages, Yvette et Lucille, aux profils statistiquement intermédiaires, sont aussi des intermédiaires au niveau narratif, puisque, dans les rixes verbales qui opposent la « pincée » et la « traître », elles défendent tantôt l'une tantôt l'autre et prônent un certain consensus quant aux sujets polémiques abordés.

¹³ La distinction profil linguistique des didascalies (/ -OPQ /) et discours des personnages (/ +OPQ /) contribue à la structuration linguistique du texte dramatique, mais n'apporte rien de très nouveau. Ce serait plutôt l'inverse qui serait stylistiquement pertinent.

3.1.2. Fonction métatextuelle des profils linguistiques



Structure textuelle (énonciative) de Encore une fois, si vous permettez (1998)

La structure de ce texte est moins canonique, puisqu'il y a une strate fictionnelle (et énonciative) supplémentaire. Le Narrateur, qui est lui-même un personnage, joue sur deux registres et constitue le pivot de deux mondes possibles : celui où, adulte, dans une posture de dramaturge pris en pleine action de création, il semble s'adresser au public de la représentation et où il parle de Nana à la troisième personne (niveau C dans le schéma) ; celui où, présenté de manière chronologique comme un enfant, un adolescent et un jeune adulte (de 8 à 20 ans), il s'adresse directement à Nana et ne rompt pas le pacte de vraisemblance (niveau D). D'un point de vue linguistique, le premier correspond à N1, c'est-à-dire au pôle /-OPQ/ et le second, tout comme Nana, correspond au pôle /+OPQ/. La mise en regard des profils linguistiques statistiquement identifiés avec l'organisation hiérarchisée du discours que constitue le texte dramatique soulève deux questions importantes, qui touchent à deux problématiques centrales de la théorie littéraire. D'une part, c'est le marquage linguistique de la structure textuelle qui est en jeu dans ce texte. La perception de deux profils linguistiques au sein d'un même personnage va de pair avec la perception de la « métalepse narrative » (Genette, 2004 ; Pier & Schaeffer, 2005) que ce personnage opère en faisant tomber ce qu'on appelle au théâtre le « quatrième mur »¹⁴. Les marques d'OPQ servent alors d'indices linguistiques de la fiction, ce qui n'était pas le cas dans *L'Impromptu d'Outremont*. D'autre part, l'hybridité linguistique du personnage Le Narrateur pris dans sa globalité renvoie aussi à une hybridité ontologique : comment définir l'identité d'un entre-deux-mondes-possibles ? Quel référent, même fictionnel, convient-il de lui attribuer ?

¹⁴ La notion de quatrième mur est généralement attribuée à A. Antoine, mais c'est à B. Brecht que l'on doit plus spécifiquement l'idée d'une chute du quatrième mur, dans la théorie du théâtre épique.

3.2. Conclusion

La mise en évidence de profils linguistiques statistiquement identifiables et leur interprétation narratologique a montré une différence du traitement des marques d'OPQ entre 1980 et 1998. Ce passage d'une fonction sociale, qui cautionne le réalisme littéraire, à une fonction affective (relation de souvenirs d'enfance), qui conduit au métalittéraire, tel qu'il a été observé dans la comparaison des deux textes analysés, et tel qu'il peut l'être d'une manière générale dans le sous-corpus *tremblay*, correspond également au changement des fonctions littéraires des marques d'oralité populaire observable dans la littérature québécoise, et à la dynamique générique du théâtre québécois (cf. Nepveu, 1999 ; Dion et al., 2001). D'un point de vue méthodologique, le codage et l'analyse linguistique et statistique du corpus littéraire m'ont servi à la fois de support et d'outil heuristique pour développer une analyse stylistique.

Références

- Anis J. (1999). Chats et usages graphiques. In Anis J. (ed.), *Internet, communication et langue française*. Hermès, pages 71-90.
- Bernet C. (2001). Formatage et finalisation de corpus : le corpus Théâtre. *Programmes fédératifs de l'ILF*. CNRS, pages 92-96.
- Burnard L. and Sperberg-McQueen C. M. (eds) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, XML version. Url: <http://www.tei-c.org>. Voir en particulier le chapitre 10 « Base Tag Set for drama ». Url : <http://www.tei-c.org/P4X/DR.html>
- Daoust P. (1983). *Jugements sur le joul à la lumière de la linguistique et de la sociolinguistique (1959-1975)*, thèse de l'Université de Montréal.
- Dargnat M. (2006). *L'oral comme fiction*. Thèse de l'Université de Provence et Ph. D. de l'Université de Montréal.
- Dion R. et al. (eds). (2001). *Enjeux des genres dans les écritures contemporaines*. Nota Bene.
- Genette G. (2004). *Métalepses. De la figure à la fiction*. Éditions du Seuil.
- Haller S. and Schneuwly B. (1996). Feuilleté énonciatif et mise en bouche. *AILE*, 8 : 129-151. Url : <http://aile.revues.org/document1240.html>.
- Heiden S. (2002). *Weblex, Manuel utilisateur, version 4.1*. CNRS/ENS-LSH.
- Manning C. D. and Schütze H. (2002 [1999]). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.
- Molinié G. (1989). *La Stylistique*. Presses Universitaires de France (Que sais-je ?).
- Molinié G. (1998). *Sémiostylistique, L'effet de l'art*. Presses Universitaires de France.
- Muller C. (1992 [1973]). *Initiation aux méthodes de la statistique linguistique*. Champion.
- Nepveu P. (1999 [1988]). *L'écologie du réel*. Boréal.
- Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.
- Pier J. and Schaeffer J.-M. (eds) (2005). *Métalepses : entorse au pacte de la représentation*. Presses de l'EHESS.
- R (2002). Logiciel d'analyse statistique. Auteurs : W. N. Venables & Brian D. Ripley. Url : <http://cran.r-project.org/>.
- Ramousse R. et al. (1996). *Introduction aux statistiques*. Url : <http://cons-dev.org/elearning/stat/St2b.html>
- Rastier F. (2001). Vers une linguistique des styles. *L'information grammaticale*, 89 : 3-6.

- Rietveld T. and van Hout R. (2005). *Statistics in Language Research : Analysis of Variance*, Mouton de Gruyter.
- Ryan M.-L. (1991). *Possible Worlds. Artificial Intelligence and Narrative Theory*. University of Indiana Press.
- Stolz C. (1999). *Initiation à la stylistique*. Ellipses.
- Thibault P. and Vincent D. (1990). *Un corpus de français parlé. Montréal 84 : historique, méthodes et perspectives de recherche*. Presses de l'Université Laval.
- Venables W. N. and Ripley B. D. (2002). *Modern Applied Statistics with S*. Springer.