À propos de quelques n-grammes significatifs d'un corpus poétique du XIXe siècle

Dominique Legallois

A PROPOS DE QUELQUES n-GRAMMES SIGNIFICATIFS D'UN CORPUS POÉTIQUE DU XIXe SIÈCLE

Dominique LEGALLOIS

La notion de n-gramme, qui sera au cœur de cette étude, se montre particulièrement fertile pour mettre en évidence quelques-uns des aspects – esthétiques et linguistiques – d'une collection de textes du XIX[e] siècle appartenant au genre poétique. Nous entendons par n-grammes, les segments répétés – les suites de mots – détectés automatiquement à partir de logiciels ad hoc. Certains de ces n-grammes ont une fonction discriminante au regard d'autres corpus composés de textes de la même période, mais appartenant à des genres littéraires différents (roman et correspondance). La technique employée et expliquée tout au long de ce travail est relativement simple, facilement reproductible, vérifiable et, bien sûr, perfectible par quiconque. Il s'agit donc d'approcher ici le discours poétique de façon globale, par une méthode quantitative, dont les résultats permettent de mieux saisir certaines dimensions d'un discours littéraire particulier. Ces dimensions sont au nombre de quatre : la dimension fréquentielle, la dimension morpho-syntaxique, la dimension esthétique, la dimension tonale (ou prosodique, en un sens particulier). Ces dimensions sont précisées à l'aide de certaines notions forgées dans le cadre de la linguistique contextualiste[1], perspective qui vise à démontrer le rôle fondamental de la phraséologie dans la production discursive. Il est important de signaler que notre étude n'est pas lexicométrique dans le sens où elle ne propose nullement la cartographie d'un corpus ou d'un genre (par analyse factorielle, par exemple) ; elle n'est pas, non plus, « interprétative », car l'objectif se limite à la « simple » observation, qui bien sûr, rencontrera certains discours critiques ; il s'agira alors de montrer la solidarité entre les données et ces interprétations. La motivation principale de l'étude est l'identification de « portes » donnant accès à des spécificités globales du corpus. Notre position reste donc celle du linguiste, utilisant parfois les outils que lui offre une conception étendue de la phraséologie, et non celle du lexicométricien apportant avec lui un bagage statistique important, ni celle du poéticien, spécialiste de l'interprétation d'une certaine esthétique.

1. CORPUS, OUTILS ET MÉTHODE

Les corpus que nous avons travaillés sont les collections proposées par le CNRTL[2].
Ainsi, pour composer nos corpus, avons-nous téléchargé :

- l'intégralité des textes poétiques proposés pour la période 1800-1900, à savoir 44 œuvres, 28 auteurs différents (parmi lesquels : Hugo, Leconte de Lisle, Banville, Flaubert, Quinet, Musset, Béranger, Lamartine, Baour-Lormian, etc.). Nous avons ajouté à la liste Verlaine (*Poèmes Saturnien -Fêtes Galantes Bonne Chanson-Romances Sans Paroles, Sagesse, Jadis Et Naguère*), Baudelaire (*Les Fleurs du Mal*) et Rimbaud (*Les Illuminations*). Soit un corpus poétique de 1363057 mots.

[1] Pour une vue d'ensemble, cf. D.Legallois (2008).
[2] http://www.cnrtl.fr/ : centre national des ressources textuelles et linguistiques.

- L'intégralité des textes romanesques proposés pour la période 1800-1900, à savoir 51 œuvres, 35 auteurs (parmi lesquels : Hugo, Zola, Chateaubriand, Soulié, Sue, Nodier, Verne, Loti, etc.), soient 4263670 mots.
- L'intégralité des textes de correspondance proposés pour la période 1800-1900, à savoir 6 œuvres, 4 auteurs différents (Hugo, Lamennais, Guérin, Du Camp), pour un total de 994118 mots. Il va de soi que ce corpus est moins « compétitif » que les deux autres (peu d'œuvres et d'auteurs). Il nous permet néanmoins, grâce à sa spécificité générique, d'identifier les unités discriminantes du corpus Poétique.

Le choix effectué est délibérément pragmatique ; il s'agissait de télécharger des textes numérisés et disponibles, sans que nous ayons à nous poser la question des critères de sélection (mis à part la période). Les corpus, précisons-le, ne visent aucunement à la représentativité des genres poétique, littéraire et épistolaire. Le genre n'est pas l'objet de cette étude. Celle-ci, en effet, est motivée par la *généricité*, entendue selon la conception de J.M. Schaeffer (1986) : un facteur productif de la constitution de la textualité, par un jeu de répétitions, d'imitations, d'emprunts, de transformations de formes et de thèmes, jeu opéré par un texte par rapport à d'autres textes. Autrement dit, quelle que soit la qualité représentationnelle de nos corpus, celle-ci est secondaire puisque l'objet est d'observer le travail de la généricité dans les textes, et non la prescription générique – même si les discours critiques auxquels nous ferons parfois référence se fondent sur la notion de genre.

La méthode d'analyse adoptée est celle de l'identification des n-grammes : nous définirons la notion de *n-gramme* en paraphrasant Lafon et Salem (1983 : 163) : est un n-gramme, toute suite de n formes graphiques non séparées par un délimiteur (signes de ponctuation usuels) qui se répète au moins deux fois. Le terme technique de *n-gramme* a deux avantages : en premier lieu, la variable *n* permet de symboliser la composition de ces segments - segment de deux mots (2-grammes), de trois mots (3-grammes), etc. Mais surtout, *n-gramme* recouvre le flou terminologique en vigueur, particulièrement dans le domaine anglo-saxon, car la linguistique française s'est peu intéressée au phénomène[3] : *lexical bundles* (Biber *et al.*, 1999), *chains* (Stubbs & Barth, 2003), *clusters* (Scott, 1996), *recurrent word-combinations* (Altenberg, 1998).

Pour cette étude, le recensement des n-grammes a été effectué avec le logiciel *Collocates*[4] qui permet l'extraction automatique de ces unités, sans qu'il ait été nécessaire d'avoir indiqué un mot cible particulier. Nous avons donc recherché les 2, 3 et 4-grammes qui se répétaient au moins 50 fois. Seuls les 100 premiers n-grammes de chaque corpus ont été pris en compte.

Précisons que les segments repérés ne sont pas des formes lemmatisées.. En effet, dans le corpus Poétique où la versification est une contrainte essentielle, la composition syllabique des unités est fondamentale et prime sur les dimensions lemmatiques. Ainsi, les unités amalgamées (*du / de le*), ou encore les mots dont le nombre de syllabes peut être différent de leurs lemmes (*chanterons / chanter*) constituent des éléments dont la

[3] A l'exception notoire de Lafon et Salem (1983).

[4] Conçu par M. Barlow en 2004.

dimension doit répondre à celle d'un tout, le vers. Aussi, le 3-gramme *au fond de*, par exemple, est-il avant tout, dans le processus de composition poétique, un trisyllabique.

Comme indiqué plus haut, ce travail s'inspire d'une conception étendue de la phraséologie, qui tend à dépasser le recensement des unités généralement et traditionnellement reconnues comme figées (expressions idiomatiques, collocations). Dans ce cadre, le discours est considéré comme étant en partie constitué de groupes lexico-grammaticaux, non pas compositionnellement produits, mais convoqués en blocs. Cette conception, véritable gageure face à la nature a priori ouverte de la composition poétique, a été élaborée principalement en Grande-Bretagne, autour des travaux de John Sinclair. Nous ferons explicitement référence aux concepts « phraséologiques » proposés par cette école dite « contextualiste », mais sans les développer[5].

2. LES DONNÉES

Les 2-grammes ont été extraits sur la base du t-score ; l'analyse se fonde à la fois sur les 100 premiers 2-grammes de chaque corpus, et sur les *2-grammes clés* – dont nous donnons la liste ci-dessous, par ordre de fréquence des occurrences - , c'est-à-dire les séquences propres (non partagées) à chaque corpus.

2-GRAMMES CLÉS DE POÉSIE :
{*comme un, qu' un, la terre, la nuit, l'ombre, au fond, l'air, l'amour, le ciel, la mort, mon cœur, l'on, comme une, la mer, l'âme, l'eau, dans son, le monde, sous les, l'heure, le vent, les yeux, un jour, sous le, du ciel, à son, dans sa, à travers, le jour, le soleil, qu' une, sur son, qui se, ses yeux, l'œil, par les, au milieu, du monde, la tête, où l', par la, les dieux, autour de, jusqu'au, les cieux, de mes*}

2-GRAMMES CLÉS DE ROMAN :
{*et de, à n', elle à, en des, les pas, et il, à pas, que les, à ce, de ce, il avait, qui il, on la, pour le, pour la, à elle, qu' à, il se, à une, et se, plus et, l'est, la plus, pas à, de lui, et des, pas les, il était, et s', et qui, à de, dans de, vous le, à se, je l', ne l', et sur, est le, et son, en un, s'était, de tout, un des, à s', une de, et qu'*}

2-GRAMMES CLÉS DE CORRESPONDANCE :
{*m'a, victor hugo, vous avez, vous êtes, pour moi, n'ai, tout ce, qui est, à madame, a été, paul meurice, il me, de votre, à paul, je crois, que nous, est une, ne sais, j' espère, à paris, j ampère, je t', à m, hauteville-house, mon cher, nous avons, vous me, j' y, à auguste, qui me, de ma, n' en, j'en, le plus, auguste vacquerie, si vous, mais je, j'avais, la France, quelque chose, a fait, pour vous*}.

Pour les 3-4 grammes (et au-delà), *Collocates* n'offre pas les services du t-score, aussi nous avons normalisé les fréquences absolues selon le nombre de mots de chaque corpus. La place nous manque pour donner l'intégralité des résultats ; nous indiquons seulement la liste des 50 premiers 3-grammes de chaque corpus avec leur fréquence pondérée.

Poétique		**Roman**		**Correspondance**	
n' est pas	*405*	*n' est pas*	*464*	*il y a*	*1089*
c' est un	*275*	*il y a*	*463*	*je n' ai*	*671*
ce n' est	*271*	*il n' y*	*415*	*n' est pas*	*566*
c' est le	*266*	*ce qu' il*	*293*	*que j' ai*	*520*
dans l' ombre	*252*	*il y avait*	*286*	*que je vous*	*464*

[5] Pour une réflexion sur les rapports entre poésie et phraséologie, cf. l'article de I. MacKenzie, 2003.

c' est la	*247*	*je n' ai*	*276*	*je ne sais*	*433*
je n' ai	*219*	*ce n' est*	*276*	*il n' y*	*424*
c' est l'	*215*	*qu' il avait*	*240*	*à paul meurice*	*410*
dans la nuit	*207*	*n' était pas*	*237*	*h – h*	*390*
n' est plus	*196*	*que j' ai*	*227*	*c' est un*	*385*
au fond de	*183*	*n' y a*	*216*	*ce n' est*	*374*
de la terre	*182*	*c' est un*	*199*	*ce que vous*	*361*
j' ai vu	*182*	*qu' il ne*	*196*	*ce que je*	*358*
et l' on	*177*	*qu' il n'*	*195*	*c' est une*	*334*
que j' ai	*176*	*au milieu de*	*194*	*à auguste vacquerie*	*319*
dans l' air	*172*	*qu' il y*	*187*	*j- j ampère*	*317*
et c' est	*170*	*que je ne*	*187*	*n' y a*	*313*
que l' on	*158*	*tout à coup*	*180*	*tout ce que*	*305*
tout à coup	*158*	*et de la*	*175*	*que je ne*	*304*
de la nuit	*153*	*n' avait pas*	*173*	*ce qu' il*	*297*
au fond des	*146*	*que l' on*	*173*	*en ce moment*	*288*
n' a pas	*142*	*je ne sais*	*170*	*et je vous*	*285*
sur la terre	*137*	*de la vie*	*162*	*madame victor hugo*	*276*
il n' est	*134*	*n' a pas*	*158*	*c' est la*	*271*
qu' il est	*131*	*c' est une*	*153*	*c' est le*	*256*
de la vie	*131*	*que c' est*	*144*	*à madame victor*	*251*
à la fois	*129*	*et qu' il*	*142*	*qu' il y*	*247*
où l' on	*128*	*ce que je*	*141*	*je ne puis*	*238*
de la mer	*128*	*c' était un*	*136*	*n' ai pas*	*230*
à l' heure	*127*	*il n' avait*	*134*	*que vous avez*	*225*
ce qu' il	*127*	*c' est le*	*133*	*je vous aime*	*218*
et de l'	*125*	*n' est-ce pas*	*128*	*je vous envoie*	*214*
et de la	*125*	*n' y avait*	*125*	*je vous prie*	*212*
au milieu des	*124*	*tout ce qui*	*125*	*je vous remercie*	*208*
je ne sais	*120*	*tout le monde*	*124*	*j' ai vu*	*208*
l' heure où	*117*	*c' est que*	*123*	*ce que j'*	*208*
de l' ombre	*109*	*c' est la*	*122*	*que c' est*	*206*
la terre et	*107*	*de ne pas*	*122*	*je vous serre*	*205*
c' est moi	*106*	*de l' homme*	*121*	*que je suis*	*204*
à travers les	*106*	*ce qu' elle*	*121*	*je ne vous*	*203*
et qu' il	*105*	*que je n'*	*120*	*tout ce qui*	*201*
de l' homme	*105*	*de l' autre*	*120*	*il me semble*	*201*
ainsi qu' un	*103*	*de tous les*	*118*	*je vous ai*	*200*
au fond du	*101*	*et de l'*	*117*	*j ampère à*	*199*
c' est là	*101*	*à la fois*	*117*	*je l' ai*	*197*
n' est point	*100*	*où l' on*	*116*	*paul meurice à*	*196*
n' est-ce pas	*98*	*qu' elle avait*	*116*	*meurice à paul*	*193*
il y a	*95*	*n' ai pas*	*114*	*que vous me*	*186*
et j' ai	*95*	*à l' heure*	*113*	*c' est que*	*185*
de l' eau	*94*	*qu' il était*	*112*	*je me suis*	*185*

- Tableau des 50 premiers 3-grammes pour chaque corpus

4-grammes

Poétique ne possède que 14 4-grammes, dont voici la liste :{*ce n'est pas, à l'heure où, je n'ai pas, c'est moi qui, il n' y a, il n'est pas, c'est l'heure, au fond de la, n'est qu'un, au fond de l', au fond d'un, ce n'est point, ce n'est plus, de la terre et}*

Pour Roman et Correspondance, nous ne donnons que les 15 premiers 4-grammes

Roman :{*il n' y a, ce n' est pas, il n' y avait, je n' ai pas, qu' il y a, n' y a pas, pour la première fois, tout à l' heure, il n' est pas, tout ce qu' il, qu' il n' y, il y a des, de temps en temps, ce qu' il y, ce n' était pas}*
Correspondance :{*il n' y a, à madame victor hugo, je n' ai pas, ce n' est pas, à paul meurice à, paul meurice à paul, meurice à paul meurice, qu' il y a, j- j ampère à, de j- j ampère, à auguste vacquerie à, vacquerie à auguste vacquerie, auguste vacquerie à auguste, il me semble que, ampère de j- j}*

Commentaire sur les données

Plusieurs remarques s'imposent. Nous nous sommes demandé d'abord quelle serait la variation des résultats sans la présence dans Poétique de *La légende des Siècles* de V. Hugo : l'œuvre est « quantitativement » importante et les 3-grammes laissent apparaître des motifs hugoliens (*dans l'ombre, dans la nuit*) identifiés (au niveau des simples lexèmes et non au niveau des n-grammes) par E. Brunet (1988). *Ombre* est ainsi reconnu par le lexicométricien comme un substantif excédentaire[6]. Un examen montre que Hugo, plutôt que d'imposer « ses » n-grammes, renforce la distribution générale de l'ensemble. La seule différence notoire est que, sans Hugo, *dans la nuit*[7] supplante *dans l'ombre* à la première place des 3-grammes avec nom. Les analyses portent donc sur le corpus complet.

Par ailleurs, et comme il sera illustré dans les analyses, le seuil de 3 permet des observations plus fines : les 3-grammes précisent la construction des unités par rapport au seuil 2, et, pour Poétique, ils favorisent une saisie infiniment plus riche que le seuil 4. On observe par ailleurs que certains n-grammes sont autonomes (par ex. *dans l'air, je vous aime,* etc.) : ils renvoient à des cadres, des objets, ou des procès précis. D'autres n'ont pas le même degré d'autonomie (par ex. *et de la, la terre et*). Cette différence sera exploitée dans notre analyse sur la syntacticité des corpus.

Le corpus Correspondance est non seulement constitué de peu d'œuvres, mais présente également un problème certain. On peut ainsi se demander si *à paul meurice* (par exemple), est un véritable 3-grammes : [prénom + nom] forme une seule entité. De même, et pire encore, *j- j ampère* n'est sûrement pas un n-gramme. Néanmoins, pour des raisons techniques, nous avons choisi de conserver ces occurrences.

3. ANALYSES

Comme nous l'avons dit précédemment, les analyses que nous proposons ne peuvent être que partielles (vu le nombre de données). Nous segmentons cette partie selon les quatre dimensions dont nous allons rendre compte.

3.1. La dimension fréquentielle (phraséologie et idiomaticité)

Il s'agit ici d'examiner la fréquence globale des n-grammes dans les corpus, et non la fréquence de tel n-gramme dans les trois corpus. Cette fréquence est calculée selon la moyenne de la fréquence des 100 premiers 3-grammes. On obtient ainsi pour Poétique : 156 ; pour Roman : 180 ; pour Correspondance : 305. Les résultats confortent la représentation généralement en vigueur : le texte poétique est beaucoup moins contraint par la présence d'unités préformées. Il est donc linguistiquement plus « créatif ».

[6] M. Milner consacre d'ailleurs un chapitre de son ouvrage (2005) à la thématique de l'ombre chez Hugo.
[7] *Dans la nuit* est également très présent chez Leconte de Lisle

Toutefois, la notion de créativité, essentiellement esthétique, reste relativement impressionniste pour la linguistique. Aussi préférons-nous parler pour Poétique d'une composition plus *ouverte*, car moins contrainte par la présence d'unités prédéfinies. La notion d'*ouverture* fait ici écho à l'un des deux principes - énoncés par J. Sinclair – à l'œuvre dans la production textuelle. Pour l'auteur, en effet, l'énonciation est « régie » par un *principe phraséologique* : l'énonciateur a à sa disposition un grand nombre de *phrases* (au sens anglais du terme) qui constituent des unités préformées, même si ces unités paraissent analysables en segments. *A contrario*, l'énonciation textuelle peut être « régie » par le *principe du libre choix* :

> "*this is a way of seeing language as the result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness"* (Sinclair, 1991: 109- 110).

Les deux principes sont évidemment les deux pôles opposés d'un continuum. Poétique, à travers ses 3-grammes montre donc une sensibilité moindre au principe phraséologique. Il est en effet assez évident que Correspondance, de loin le corpus le plus « figé », est davantage normé par des routines interactionnelles (termes d'adresse, phrases performatives, par exemple). Ainsi, les n-grammes nous renseignent sur la composition du texte : composition, au sens de processus de textualisation ; composition, au sens d'un ensemble d'ingrédients contribuant à la *texture* des discours, c'est-à-dire sa consistance en termes d'unités « prévisibles ». Cette texture est donc plus souple pour Poétique puisque le nombre de n-grammes non partagés est légèrement plus élevé. Cependant, si la nature phraséologique de Poétique est moins élaborée (et encore une fois, conformément aux attentes), on ne peut pas dire pour autant que le corpus soit moins « idiomatique ». Si l'idiomaticité se définit par les caractéristiques appartenant en propre à un objet, on peut alors apprécier l'idiomaticité des corpus en fonction du nombre des n-grammes clés :

	Poésie	Roman	Correspondance
2-grammes	46	46	39
3-grammes	45	34	32

- Nombre de 2-3 grammes clés (100 premières occurrences)

Le tableau montre que Poésie n'est pas moins idiomatique, au contraire, que les deux autres corpus. C'est cette idiomaticité, liée à la généricité, qui sera analysée dans la suite de l'article.

3.2. Particularités morphosyntaxiques et sémantiques

Il convient à présent de préciser la nature des n-grammes. En cela, l'observation des 2-grammes clés de Roman offre des éléments intéressants : absence de morphèmes lexicaux, mais un nombre conséquent de conjonctions ou de prépositions incolores. Comparé à Poésie (mais aussi à Correspondance), Roman manifeste une forte densité d'outils relationnels déterminant la complexité syntaxique du corpus. Les liens grammaticaux y sont marqués, et immédiatement observables, dans les 2-grammes non autonomes, véritables chevilles du texte. Au contraire, Poétique comprend des 2-

grammes constitutionnellement complets et autonomes. Ce point nous parait être à rapprocher d'une remarque récente de L. Victor :

> *Il y a certainement un lien structurel entre genre littéraire et traitement par l'écriture littéraire de la composante syntaxique. Dans la prose, narrative ou non, la grammaire paraît fortement sollicitée, elle peut l'être même jusqu'à une sorte d'hypertrophie, comme on peut le voir aisément dans Proust, Céline, ou Claude Simon, par des formes différentes.* […]*Par nature, si on peut dire, la poésie n'est pas un genre hypersyntaxique*.

En comparant des poèmes de Ronsard, Baudelaire et Jaccottet, L. Vicor perçoit une évolution vers une asyntacticité du genre poétique (un minimalisme syntaxique) – ce qui ne veut pas dire dysgrammaticalité. Limitée à une période, notre analyse n'a pas pour objectif de confirmer cette observation (qui correspond, là encore, sans doute, à l'intuition), mais de montrer qu'une syntacticité avec des variations de densité, constitue un critère discriminant de la généricité.

L'observation de cette propriété ne peut porter véritablement que sur les 2-grammes. On voit mal comment les suites de trois mots seraient majoritairement « dispensés » de relateurs ou de subordonnants. Néanmoins, là encore, les subordonnants et les conjonctions de coordination dans Poétique sont largement minoritaires par rapport à Roman, ce qui atteste la dimension hyposyntaxique supérieure de Poétique.

Au niveau sémantique, les 2-grammes « lexicaux » de Poétique reflètent les thématiques identifiées par la critique du Romantisme : éléments naturels et spirituels (*la terre, la nuit, l'ombre, l'air, l'amour, le ciel, la mort,* etc.), dont les noms déterminés par l'article défini singulier désignent des singletons, et non des entités pluriels. Mais il est plus intéressant d'observer ces noms dans les suites à trois éléments, puisque celles-ci nous renseignent sur le profil syntaxique et conceptuel de ces items et de leurs référents. La caractéristique fondamentale est alors la présence de la préposition *dans*, dans les 100 premiers 3-grammes, avec une fréquence très importante pour certaines suites : *dans l'ombre, dans la nuit, dans l'air, dans les bois, dans les cieux, dans le ciel, dans les airs*, alors qu'aucune occurrence de n-gramme avec *dans* n'apparaît dans Roman (la première n'apparaît qu'en 167^{e} position, avec une fréquence modeste de 61 : *dans le monde*, suivi de loin par *dans la chambre, dans la vie, dans la maison, dans le cœur*)[8].

Il est donc nécessaire d'interpréter non pas la seule spécificité lexicale, qui est de toute façon connue de la critique, mais la construction de ces items : une intériorité construite par *dans*, une intériorité des éléments naturels qui constitue en elle-même une thématique. Mais il y a plus, car cette intériorité – qui mériterait bien sûr un long développement, et une définition plus précise- nous semble être complémentaire de la présence importante de la locution prépositionnelle *au fond de/des/du/d' :*

Vénus l'a blessé soudain des mêmes traits / dont elle abuse, au fond des antiques forêts (Moréas)

– qui apparaît d'ailleurs trois fois (*au fond de la / au fond de l'/ au fond d'un*) parmi les quatorze 4-grammes de Poétique. Il est inutile de souligner que les items grammaticaux

[8] Où l'on voit se dessiner avec *chambre* et *maison* une intériorité domestique. Correspondance est encore plus modeste quant à la présence de *dans* dans les n-grammes.

contribuent, au même titre que les lexèmes, à la construction et aux déploiements thématiques : ici, la profondeur, cause d'insondabilité – thème poétique récurrent du XIX[e] siècle, qui sera discuté par le critique J.P. Richard, dans son livre *Poésie et Profondeur*.

Un autre 3-grammes se démarque selon nous, qui demande un effort d'investigation supplémentaire : *on ne sait* ; sa fréquence pondérée (67) n'est peut-être pas spectaculaire, mais elle est nettement supérieure à celle observée dans Roman et Correspondance. Au regard des concordances, *on ne sait* est très lié à la généricité poétique. Peut-être même peut-on le considérer comme révélateur d'une thématique. En effet, *on ne sait* est à la fois une forme « phrastique » (sujet + verbe), mais aussi, et surtout ici, une locution à noyau verbal faisant partie, selon Wilmet, des quantifiants-caractérisants stricts synthétiques *qui se contentent d'additionner à la quantification de base une information caractérisante vague* (Wilmet, 1999 :234) :

Dans les lointains mourants, on ne sait quel flot bleu passe, et traverse encore l'insondable océan de verdure sonore (Banville)
Ce fut dans on ne sait quel ravin inconnu / Que Tiphaine atteignit le pauvre enfant farouche (Hugo)

60 % des emplois de *on ne sait* sont consacrés à cette fonction, pour seulement 31% des emplois dans Roman et 15% dans Correspondance. A l'intériorité et à la profondeur s'ajoute le thème de l'indétermination, thème complémentaire si l'on considère de plus, que l'intériorité porte majoritairement sur des motifs « obscurs » (*ombre* et *nuit*) – sources d'indétermination par excellence. La complémentarité reflète le rapport entre la modalité du voir et celle du savoir : la difficulté de voir et de distinguer à travers des espaces, et la difficulté d'identifier – donc de nommer – des objets. Outre le lyrisme, le sublime est une qualité d'expérience propre à la poésie qui, « *remet en cause la distinction entre un sujet percevant et un objet perçu qui fonde le principe de représentation »* (Y. Le Scanff 2007 :160).

On aurait là, phénoménologiquement, l'érosion de la représentation qui se voit abolie par l'indétermination de l'objet par un « sujet » qui ne peut, par conséquent, et en retour, se constituer comme tel.

Sur un terrain plus linguistique, Poétique se distingue par la nature des locutions prépositionnelles, adverbiales, ou conjonctives « prépositionnelles » : elles ont un emploi « spatial » (déterminé, bien sûr, par leurs compléments) beaucoup plus affirmé que dans Roman et Correspondance :

Poétique : *au fond de, au fond des, à l'heure, au milieu des, à travers les, au fond du, au fond d', au milieu de, au bord de, jusqu' à la, sur le bord*,
Roman : *au milieu de, à l'heure, au milieu des, au fond de, en ce moment, en même temps, au moment où, jusqu' à la, au bout de,*
Correspondance : *en même temps, au milieu de.*

Les locutions temporelles sont sous-employées dans Poétique. Par ailleurs la notion de *cadre collocationnel* proposée par Renouf et Sinclair (1991- *collocational framework*) nous semble pertinente pour, cette fois-ci, caractériser la sous-spécificité de ces locutions. Les cadres collocationnels sont des combinaisons discontinues d'items grammaticaux qui entourent ou « encerclent » des mots lexicaux. Ces combinaisons

sont fonctionnelles, dans le sens où elles reflètent une sémantique particulière. La structure n'est donc pas considérée comme hiérarchique et compositionnelle, mais comme plate et globale ; c'est une matrice productive, relevant de la phraséologie étendue. Le cas qui nous intéresse ici est la forme < au * de >[9] observée non plus parmi les cent premières occurrences mais parmi tous les 3-grammes. Ce cadre collocationnel est particulièrement productif dans Roman[10] : {*milieu, dessus, fond, bout, bord, pied, dessous, devant, delà, moment, moyen, moment, coin, bas, sein}*. Peu productif pour Correspondance {*milieu, dessus, fond*}, à peine plus pour Poétique {*fond, milieu, bord, sein, dessus, bout*} dans un registre largement « spatial ». Si le cadre < au * de > est peu productif pour Poésie – alors même qu'il détermine des 3-grammes fréquents, on peut conclure à une certaine spécialisation, qui serait propre là encore à l'idiomaticité générique de Poétique.

Pour clore cette partie, terminons par deux observations dont nous esquissons à peine les conséquences. D'abord la présence dans Poétique du 3-grammes discriminant *n'est plus* pourrait être l'indice d'une thématique de la « révolution », c'est-à-dire de ce qui est résolument de l'ordre du passé, regretté ou non.

De nos temps l'épopée n'est plus la propriété d'un peuple à l'exclusion d'un autre (Quinet)
Tout le vieux fer romain n'est plus que de la rouille (Hugo)

Enfin, on constatera dans Poétique l'évitement de la forme thétique *il y a* marqueur de thématicité (au sens informationnel), surreprésentée dans Roman et surtout dans Correspondance (par rapport à Poétique). Cet évitement est si fort qu'il conviendrait – ce que nous ne pouvons faire ici – d'en déterminer les raisons.

3.3. Particularités esthétiques : indices de littérarité poétique.

Généricité et idiomaticité sont étroitement liées à des facteurs esthétiques, parmi lesquels les outils de la comparaison figurée. On constate dans Poétique l'emploi discriminant des 2 -grammes outils de comparaison : *comme un / comme une* et du 3-grammes *ainsi qu'un*. *Comme* suivi de l'article indéfini est suivi de façon privilégiée par des noms quasi-systématiquement expansés, regroupés ici en classes : {*homme, dieu, enfant, femme*}, {*oiseau, aigle*}, {*flot, torrent*} ainsi que par *fleur* et *ombre*, alors qu'*ainsi qu'un* n'a pas de collocatif marqué. Poétique est donc un corpus qui manifeste clairement sa structure esthétique figurée. A nouveau, il s'agirait d'effectuer une analyse diachronique (en comparant des siècles différents) pour établir l'évolution des profils de généricité portant sur la figurativité. A noter que les métaphores, qui par définition sont délestées de marqueurs, échappent totalement à l'observation des n-grammes.

Les éléments de littérarité poétique sont également lexicaux. Nulle surprise, certaines formes sont reconnues comme intrinsèquement « poétiques » : *dans les cieux, dans les airs*. Cependant, une certaine prudence interprétative est de règle pour mettre en évidence les formes témoignant de la littérarité poétique. Par exemple, l'unité *dans les cieux* peu paraître discriminante dans la mesure où elle n'apparaît pas comme n-gramme

9 * vaut pour un nom (généralement).
10 Même s'il faut prendre en compte la dimension de ce corpus par rapport aux autres.

de fréquence égale ou supérieure à 50[11] de Roman et de Correspondance. En fait, ce n'est pas *dans les cieux* qui est discriminant mais *cieux*, qui est privilégié – qu'on nous permette cette évidence – par le discours poétique. L'unité *dans les cieux* correspond dans Poétique à 12 % des emplois de *cieux*, et à 14% des emplois du même lexème dans Roman. L'égalité est presque parfaite si l'on raisonne au niveau de la séquence répétée.

En revanche, *dans la nuit* est incontestablement discriminant puisque l'unité correspond à 15 % de tous les emplois de *nuit* dans Poétique, pour seulement 5% dans Roman. Elle est doublement discriminante puisque *nuit* est suremployé dans Poésie par rapport à Roman. Contrairement à *dans les cieux*, cette propriété « littéraire » de *dans la nuit* ne correspond pas à l'intuition du lecteur « moyen ».

On retiendra aussi le 4-grammes *à l'heure où* « d'essence » poétique ; l'occurrence est absente du relevé des n-grammes des autres corpus. Mais il convient de ne pas limiter la dimension phraséologique au seul item, et de considérer les n-grammes comme des points d'accès, que peuvent prolonger l'observation de concordances. A la lumière des emplois, on remarque que *à l'heure où* a des exigences dans Poétique qu'il n'a pas dans Roman (où la forme apparaît seulement 29 fois, donc en deçà du seuil de la fréquence 50) : il est suivi quasi-systématiquement d'une proposition évoquant le retour d'un phénomène naturel. Cette distribution privilégiée reçoit le nom de *préférence sémantique*[12] dans l'approche de la linguistique contextualiste. Par exemple, Roman privilégie des actions humaines itératives ou non[13] :

Car d'ordinaire Louise ne manquait pas, à l'heure où elle se couchait, d'ouvrir sa fenêtre (Champfleury)
Madame Paturot jouirait, à l'heure où j'écris, d'une statue (Reybaud)

Tandis que Poétique privilégie nettement les cycles naturels :

A l'heure où l'horison lentement se colore des rayons du soleil (Michaud)
Jusqu' à l'heure où le soir vint obscurcir les cieux (Baour-Lormian)

A l'heure où est donc une véritable machine à clichés poétiques.

3.4. Prosodie pragmatique

La notion de prosodie sémantique a été proposée et illustrée par B. Louw (1993), mais inspirée par J. Sinclair. Cette notion, plus métaphorique dans le contexte français que dans le contexte britannique, reste redevable à la conception de Firth du système phonologique prosodiquement organisé : certains traits ne sont pas inhérents à des unités discrètes mais se distribuent sur plusieurs unités (suprasegmentalité). Au niveau sémantique, la prosodie sémantique d'une unité se caractérise par

« *the constituant aura of meaning with which a form is imbued by its collocates* » (Louw, 1993 : 157).

Dans une conception plus récente, J. Sinclair (2004) montre que la prosodie sémantique est en fait une prosodie *pragmatique* dans la mesure où ce qui est exprimé par ces

11 *Dans les cieux* apparaît seulement 17 fois dans Roman.
12 La notion est développée dans Stubbs (2002 : 65).
13 Même si, évidemment, les actions humaines cycliques peuvent être en partie déterminées par les cycles naturels.

formes relève de l'attitude énonciative du locuteur[14]. On relève ainsi pour toute langue des constructions plus ou moins saturées lexicalement qui sont étroitement associées à des postures énonciatives telles que l'indignation, l'incrédulité (*toi, participer au Marathon ?*), etc. La notion d'une prosodie pragmatique se justifie par le fait que la valeur pragmatique est véhiculée par la construction (par exemple [Pronom tonique + proposition infinitive]), indépendamment de la variation lexicale. Dans l'analyse poétique, la prosodie pragmatique est à rapprocher de ce que D. Combe (1992) appelle l'*impression affective pré-réflexive*, qui relève de la tonalité mais aussi de l'éthos de l'énonciateur. Une forme semble particulièrement spécialisée dans l'expression de l'éthos : la suite *je suis le*. Ce 3-grammes peut sembler d'une fréquence anodine (fréquence pondérée de 68) ; il est pourtant discriminant puisque il est absent de la totalité des n-grammes des corpus Roman et Correspondance (et non seulement des cent premiers).

Dominée par le lyrisme, la poésie romantique ne cesse de mettre en scène des modes énonciatifs dans lesquels s'expriment des rapports intimes avec des unités transcendantes. Ainsi, comme l'écrit Y. Vadé :

Aucun poète n'a écrit la formule folle « je suis tout ». Mais presque tous ont écrit comme si le poète en tant que tel entretenait des relations privilégiées avec l'univers, ou ce qu'ils aiment à nommer la Nature, l'infini, l'inconnu, Dieu. (1996 : 25).

La « formule » *je suis le*, parce qu'elle est en quelque sorte imbibée de ces différents emplois, est consacrée à l'expression du lyrisme – cette valeur « suprasegmentale » qui détermine dans le discours un mode d'être face à l'univers. Aussi, dans notre corpus, *je suis le* est quasi systématiquement consacrée à ce que l'on pourrait appeler la *grandiloquence*. Grandeur du discours, mais plus encore grandeur de l'énonciation et de ces paramètres : sujet et circonstances. *Je suis le* construit un éthos lyrique du sujet énonciateur ; quelques exemples :

Je suis le coeur de la vertu, Je suis l'âme de la sagesse (Verlaine)
Je suis la plaie et le couteau ! Je suis le soufflet et la joue ! (Baudelaire)
Je suis le pleur qui toujours coule ; je suis le soupir qui toujours recommence (Quinet)

Ces trois citations sont typiques de beaucoup d'occurrences, dans le sens où est manifeste l'amplification construite par la redondance de la formule. Cette inflexion est également secondée par un 4-grammes, moins « spécialisé » cependant : *c'est moi qui*

Roi, c'est moi qui suis ma cage, et c'est moi qui suis ma clé (Hugo)
C'est moi qui suis l'amas des yeux et des rayons (Lamartine)

Ce 4-grammes est présent dans Roman, mais, évidemment, sans sa valeur prosodique lyrique, ce qui démontre la généricité de la valeur, qui, d'ailleurs, impose un régime sémantique particulier au nom complétant le n-gramme. En effet, il nous semble exclu de parler d'emploi métaphorique du nom, tant lyrisme et sublime transcendent les catégories interprétatives habituelles.

La notion de prosodie pragmatique, travaillée par la technique des n-grammes, nous semble une notion féconde pour établir une sémantique tonale, affective, de certaines formes génériquement situées.

[14] Cf. Legallois (2008)

Conclusion

Ce travail constitue donc une première approche de la généricité poétique par le biais des n-grammes, généricité entendue non pas comme une transcendance qui contraint la production, mais comme un ensemble de relations intertextuelles. L'ambition était de considérer les n-grammes comme des entrées possibles dans la globalité d'un corpus, afin d'en déterminer quelques unes des caractéristiques. L'analyse a été nécessairement partielle, focalisée sur certaines unités plutôt que sur d'autres[15].

Les n-grammes étant des segments répétés, nous avons fondé notre étude sur l'idée d'un principe phraséologique général en œuvre dans tout discours. Ce principe se décline sur plusieurs dimensions, avec des degrés très variables. Les notions comme celles de cadres collocationnels, de prosodie pragmatique, de préférence sémantique, d'idiomaticité - élaborées dans un autre domaine que l'analyse critique - nous ont ainsi parues utiles pour préciser certaines tendance génériques.

Nous terminons cet article par un relevé des 3-grammes de Poétique (discriminants ou non) présents dans un célèbre passage des *Misérables* qui décrit les éléments de la scène où Cosette est envoyée chercher de l'eau :

L'obscurité est vertigineuse. Il faut à l'homme de la clarté. Quiconque s'enfonce dans le contraire du jour se sent le cœur serré. Quand l'œil voit noir, l'esprit voit trouble. Dans l'éclipse, **dans la nuit**, dans l'opacité fuligineuse, **il y a** de l'anxiété, même pour les plus forts. […] Une réalité chimérique apparaît dans la profondeur indistincte. L'inconcevable s'ébauche à quelques pas de vous avec une netteté spectrale. On voit flotter, **dans l'espace** ou dans son propre cerveau, **on ne sait** quoi de vague et d'insaisissable comme les rêves des fleurs endormies. **Il y a** des attitudes farouches sur l'horizon. On aspire les effluves du grand vide noir […]. On éprouve quelque chose de hideux comme si l'âme s'amalgamait **à l'ombre**.

Comme on le remarque, ce passage en prose « emprunte » un nombre appréciable de 3-grammes poétiques. Il serait ainsi intéressant d'examiner les relations intertextuelles des n-grammes dans les textes génériquement différents, mais reliés thématiquement, afin d'élaborer des « sémantiques interprétatives » fondées sur la matérialité des discours. Les n-grammes autonomes ne sont-ils pas, en quelque sorte, des morphèmes textuels, c'est-à-dire des unités minimales chargées des valeurs sémantiques transitant de discours en discours ?

Dominique LEGALLOIS
Université de Caen
CRISCO (EA 4255)

[15] Elle pourrait être poursuivie à un autre niveau : par exemple, en examinant la position des n-grammes dans le vers (par exemple *dans la nuit*) et en mesurant la constance ou la variation de cette place.