Caractérisation quantitative de textes. Application à l'oral représenté, en diachronie

Bénédicte Pincemin, Denise Malrieu

In *Driss ABLALI, Sémir BADIR, Dominique DUCARD (dir)*
Documents, textes, œuvres -Perspectives sémiotiques
Presses Universitaires de Rennes, 2014, p. 43-56.

Caractérisation quantitative de textes. Application à l'oral représenté, en diachronie.

Bénédicte PINCEMIN et Denise MALRIEU

Résumé
La caractérisation des textes d'un corpus peut être basée sur des jeux de mesures linguistiques ou stylistiques comme la longueur moyenne des phrases, la proportion des différentes catégories grammaticales, etc. La question abordée ici est celle du choix des mesures à utiliser, en étant attentif à leur cohérence d'ensemble et à leur interprétativité. Une dizaine de modèles de mesure génériques sont proposés, permettant de rapporter les caractérisations textuelles à des principes linguistiques descriptifs fondamentaux comme l'organisation paradigmatique et syntagmatique de la langue, la linéarité du texte, les liens de dépendance syntaxique. Ces modèles de mesure ont été implémentés dans le logiciel libre de textométrie TXM, et expérimentés sur un corpus de quatre textes littéraires français, finement annotés au niveau du discours représenté (discours direct, indirect, parole intérieure, etc.).

Mots-clés
linguistique de corpus, textualité, sémantique interprétative, textométrie, stylométrie, discours représenté.

1. Introduction

La linguistique textuelle suivant une approche différentielle trouve un important terrain d'observation dans l'analyse statistique de corpus de textes. En particulier, chaque texte peut être représenté par un certain nombre de mesures sur différents paliers (morphèmes et mots, phrases, paragraphes) à partir desquelles le situer par rapport aux autres : par exemple, sa proportion de noms par rapport aux verbes, celle des différents temps verbaux, le rythme des ponctuations, la longueur moyenne des paragraphes... Cette communication propose des repères méthodologiques et des outils pour la construction de tels jeux de mesures, en veillant à leur interprétabilité linguistique et à leur cohérence.

2. Contexte

2.1. Une expérimentation en corpus sur les genres et les variations morphosyntaxiques

Pour la sémantique interprétative, les paliers de description local (lexical, morphosyntaxique) et global (textuel, intertextuel) sont en interaction forte et multiple. Au début des années 2000, Rastier et son équipe expérimentent en corpus cette interaction local/global. Le travail de recherche porte sur 251 mesures produites par l'analyseur Cordial (telles que « nombre moyen de mots par phrase » ou « % de noms communs par rapport aux noms (communs ou propres) ») sur un corpus de 2541 ouvrages français. Les résultats confirment empiriquement qu'une typologie globale de ces textes peut se construire sur la base de faisceaux d'indices linguistiques locaux (Malrieu & Rastier, 2001).

2.2. Discussion des mesures fournies par le logiciel Cordial

Le jeu de mesures obtenu par le logiciel Cordial n'est cependant pas pleinement satisfaisant ici, car il n'a pas été conçu pour ce type d'expérimentation, utilisant toutes les mesures dans leur ensemble. Il a été constitué progressivement en réponse à des demandes d'utilisateurs dispersées, fournissant aux uns et aux autres les mesures les intéressant. Il se présente de ce fait comme une compilation foisonnante, organisée en grandes catégories hétérogènes : Totaux, Moyennes, Phrases, Ponctuations, Morphologies, Usage, Noms,... Certaines mesures, disséminées dans plusieurs de ces catégories ou non, s'avèrent redondantes. Par exemple, pour les articles définis :

- % de déterminants par rapport à l'ensemble des mots
- % d'articles définis par rapport à l'ensemble des mots
- % d'article définis par rapport à l'ensemble des déterminants

- % d'articles définis par rapport à l'ensemble des articles

On déploie ainsi toutes sortes de combinatoires qui démultiplient artificiellement les mesures, emboîtements de catégories comme emboîtements de paliers de réalisation (caractères, mots, propositions, phrases, paragraphe, texte).

Cette profusion de mesures ne garantit pas non plus l'absence de lacunes : comme on n'a pas de conception d'ensemble du jeu de mesures, on manque de repères pour s'assurer d'avoir parcouru méthodiquement des dimensions de caractérisation textuelle.

Enfin, nous examinons aussi les mesures sous l'angle de leur interprétation. Par exemple, le « % de verbes au subjonctif présent par rapport à l'ensemble des verbes conjugués » n'a pas une pertinence claire, si l'on considère que le subjonctif est sans doute davantage appelé par l'emploi de certaines conjonctions que positionné dans un équilibre de temps et modes verbaux.

2.3. Autres travaux basés sur des jeux de mesures

Des travaux antérieurs ont également observé le lien entre indices locaux et caractérisation globale des textes, sur la base d'autres jeux de mesures.

Biber (1988) construit son jeu de mesures en dressant un inventaire des mesures évoquées dans la littérature scientifique sur la caractérisation textuelle, dans la limite des mesures réalisables sur la base d'une annotation semi-automatique du corpus. S'il élabore un ordre pour lister ses variables, il ne cherche pas à les décrire de façon unifiée ni à les structurer de façon à mettre en évidence le parcours méthodique d'un espace de description.

Dans le contexte du début des années 1980, Bronckart *et al*. (1985) ne disposent pas encore de logiciels d'étiquetage du type de Cordial, et élaborent leurs mesures en exploitant une annotation pragmatiquement parcimonieuse. Leur réflexion dégage cependant plusieurs propositions très intéressantes pour la conception de mesures. Ainsi, pour obtenir des grandeurs relatives, ils remarquent que le caractère linguistique des données appelle deux référentiels différents : le nombre de verbes doit méthodiquement servir d'aune au décompte des traits verbaux (ex. temps) ou des différents types de structures liées aux verbes (ex. propositions) ; en revanche c'est au nombre de mots que l'on rapporte le décompte général des lexèmes ou celui des organiseurs argumentatifs, au sens où ces éléments sont considérés sous l'angle de leur occupation de la surface textuelle. Bronckart *et al*. introduisent également des concepts originaux pour modéliser leur perception de certaines structures linguistiques, par exemple en définissant la densité syntagmatique comme le rapport entre le nombre de noms noyaux (Ny) et celui de qualifiants (Q). Cette manière linguistique de concevoir des mesures a nourri notre réflexion.

3. Proposition théorique : repères pour la conception de mesures textuelles

3.1. Objectif

Nous nous plaçons dans le cadre d'une analyse sur corpus de textes intégraux, annotés morphosyntaxiquement, et éventuellement structurés (typiquement avec un codage XML, permettant par exemple de désigner le titre, les paragraphes, ou d'autres éléments de structure du texte). Il s'agit de se baser sur des décomptes d'étiquettes pour obtenir une caractérisation textuelle pertinente. Nous supposons avoir toute liberté pour choisir les étiquettes parmi celles disponibles et pour combiner les décomptes en mesures.

Du fait de la multiplicité des points de vue possibles sur le texte, l'objectif n'est pas de définir une liste (universelle) de mesures, mais de dégager des repères méthodologiques pour la conception de mesures linguistiquement motivées. Ces repères ont pris la forme de modèles de mesure génériques, utilisables comme des « moules » pour la construction de mesures appropriées à un corpus et une problématique.

Pour les exemples nous travaillons sur un corpus français, mais les principes mobilisés relèvent de la linguistique générale, et sont sans doute au moins en partie transposables à d'autres langues.

3.2. Mesures liées au déploiement du texte

Ampleur

Une Ampleur est une mesure de longueur : c'est l'Ampleur que l'on utilise par exemple pour demander la taille du texte. On peut choisir l'unité de décompte. On s'attache à choisir une unité s'approchant au mieux du ressenti du lecteur face au texte : pour le texte ou pour la phrase, l'impression de volume globale pourra dans certains cas être d'abord liée au nombre de mots ; et si l'on opte pour ces deux mesures, on pourrait se dispenser de calculer la longueur du texte en phrases, moins parlante et déjà indirectement

représentée par les deux mesures précédentes. Les unités de mesure possibles sont nombreuses : s'il y a les classiques paliers syntagmatiques (caractère, mot, phrase, paragraphe), on a aussi toutes les unités correspondant à un découpage rythmique (vers), une subdivision (éléments d'une liste, tours de parole), une structuration (entrée dans un lexique).

Profondeur

La Profondeur veut rendre compte de l'organisation éventuellement hiérarchique de structures textuelles et de leurs inclusions plus ou moins développées. Nous proposons de décompter le nombre maximum d'emboîtements observés pour une structure ou un ensemble de structures donnés (d'autres implémentations restant envisageables). Par exemple, on peut ainsi caractériser un texte par le nombre de niveaux de sections utilisés, ou la présence de listes emboîtées. Exemple concret, dans la *Quête du Graal*, on décompte jusqu'à 3 niveaux d'imbrication dans le discours représenté (le locuteur cite un passage de la Bible, dans lequel un des personnages prend la parole).

3.3. Mesures paradigmatiques

Proportion

On mesure ici les proportions prises par différents cas, compris comme représentant différents choix linguistiques en alternative ; autrement dit, pour un paradigme donné, on mesure la répartition observée des différents cas. Cette mesure est typiquement mobilisée sur des sous-catégories morphosyntaxiques par rapport à leur catégorie (ex. proportion de possessifs parmi les déterminants) ou sur des traits morphosyntaxiques par rapport à une ou plusieurs catégories qui les portent (ex. proportion de verbes qui sont à l'imparfait). Ceci étant, l'approche interprétative invite à bien penser la construction des paradigmes utilisés, quitte éventuellement à s'écarter des groupements « évidents » suivant les désignations traditionnelles.

Si l'on considère le système des classes grammaticales (Nom, Verbe, etc.), l'interprétation de la proportion de chaque catégorie par rapport l'ensemble ne rend pas compte des dépendances induites par la syntaxe : par exemple, la proportion d'adjectifs ou de déterminants dans un texte ne s'interprète pleinement qu'en lien avec celle des noms. On préférera donc sur ces catégories des mesures rendant mieux compte de ces dépendances (cf. *Ratio*, plus loin). Autre exemple de prise de recul par rapport aux paradigmes descriptifs traditionnels, veut-on penser le pronom relatif comme un cas de pronom (en concurrence paradigmatique avec les pronoms personnels, possessifs, indéfinis), ou plutôt mesurer sa présence comme en alternative avec les différentes formes de subordination, ou d'expansion nominale ? Et pour les ponctuations : prendra-t-on les ponctuations dans leur ensemble pour évaluer leur équilibre respectif, ou voudra-t-on plutôt d'abord rendre compte des proportions de ponctuations fortes, semi-fortes et faibles, puis ensuite rapporter l'usage du point d'interrogation aux seules ponctuations fortes ? Quant au nom, peut-être devrait-il être d'emblée mis en système avec le pronom (pronoms personnels, possessifs, démonstratifs), en constituant ainsi une classe de « quasi-noms » à l'intérieur de laquelle étudier l'équilibre des différentes catégories. Bref, la mise en œuvre des mesures de Proportion est vraiment l'occasion de questionner la représentation du fonctionnement linguistique mobilisée.

Diversité

Pour certains paradigmes, peut nous intéresser aussi le caractère plus ou moins riche de leur utilisation : se cantonne-t-on à l'usage d'un seul cas, de quelques-uns, ou exploite-t-on un large ensemble des possibilités linguistiques ? Nous proposons d'évaluer cette diversité de réalisations d'un paradigme par le décompte des différents cas attestés : par exemple, combien de formes différentes de pronom relatif sont utilisées dans le texte. Il peut certes y avoir un biais si les textes à caractériser sont de longueurs très différentes, la Diversité pouvant être bridée sur un texte court. Ceci étant, cette mesure rend compte *a minima* d'un inventaire des formes attestées et mobilisées, et leur simple effectif peut déjà exprimer un mode parcimonieux ou gourmand de recours à la langue.

3.4. Mesures syntagmatiques

Ratio

Cette mesure considère une dépendance syntaxique et mesure un pro-rata entre un élément pôle et un élément lié. Par exemple, sachant la dépendance de l'adjectif qualificatif au nom, il est parlant d'évaluer le nombre d'adjectifs par rapport à celui de noms communs (ou quasi-noms définis ci-dessus). De même, une évaluation de la valence des verbes utilisés pourrait être approximée par un Ratio entre les quasi-noms (noms et pronoms) et les verbes.

Dépendance

On souhaite ici décrire des figements ou des affinités contextuelles, et mesurer l'autonomie d'emploi de composants par rapport à une construction qui les rassemble. Un calcul simple consisterait à rapporter la fréquence du composé à la fréquence du composant le moins fréquent : une valeur de 1 caractériserait la dépendance totale d'au moins l'un des composants à l'expression composée ; une valeur proche de zéro dénoterait une faible force de liaison pour l'ensemble des composants, au regard de leurs emplois dans le corpus. Cette mesure servirait notamment à caractériser des enchaînements de structures textuelles ou de catégories descriptives.

Taux de présence linéaire

Pour certains phénomènes, diffus sur de multiples catégories (par exemple, les marques de première personne du singulier) ou perçus comme relativement autonomes ou polyvalents (adverbes, prépositions, conjonctions), on voudra mesurer leur occupation dans le volume du texte. La formule proposée consiste à rapporter la fréquence de l'élément considéré à l'Ampleur du texte : si l'unité considérée est le mot, cela correspond à un classique calcul de fréquence relative. Un paramètre d'échelle, multiplicateur, peut être ajouté pour obtenir des grandeurs plus lisibles. Une échelle de mille permet de lire le résultat comme le nombre moyen d'apparition du phénomène pour mille mots.

Mathématiquement, le calcul du Taux de présence linéaire peut dans certains cas prendre la même forme qu'un calcul de Proportions : ainsi, le taux de présence linéaire des adverbes pour l'unité mot fait appel au même calcul qu'une proportion des adverbes par rapport à l'ensemble des catégories grammaticales. D'autres rapprochements de la sorte pourraient être faits avec d'autres mesures : plusieurs mesures ici proposées ne font appel qu'à un simple rapport entre deux décomptes de fréquences (Proportions, Ratio, Taux de présence linéaire). Les principes de définition des types de mesure ne sont donc pas mathématiques, mais linguistiques : notre recherche ne vise pas ici la mise au point de formules, mais la mise en évidence de fonctionnements linguistiques pouvant se traduire quantitativement et donner lieu à des mesures interprétables. D'une mesure à l'autre, les formules peuvent être identiques, mais l'interprétation est différente. Or l'interprétation est déterminante, puisque c'est elle qui définit le choix des opérandes. C'est le rapport à l'interprétation linguistique qui donne toute sa consistance à notre proposition.

Cadence

La caractérisation d'un texte peut aussi passer par des mesures de rythme. Pour un phénomène régulièrement présent, le Taux de présence linéaire rend compte de l'espacement moyen, plus ou moins grand, entre deux occurrences. Mais cette mesure serait insatisfaisante pour rendre compte de phénomènes apparaissant épisodiquement ou pour des phénomènes concentrés sur une partie du texte. La Cadence se focalise sur les passages où le phénomène apparaît : elle calcule la distance moyenne entre une occurrence du phénomène et l'occurrence la plus proche. L'unité pour mesurer la distance est typiquement le nombre de mots, mais selon les cas cela peut aussi être le nombre de phrases, de paragraphes, de tours de parole, etc. Ainsi, on peut souhaiter calculer la Cadence en mots de la virgule, et la Cadence en phrases des points d'interrogation.

3.5. Pour toutes les mesures : *domaine*, *multiplicité* et *synthèse*

Une mesure ne s'applique pas nécessairement à tout le texte : par exemple, si l'on veut caractériser un texte par la longueur de son titre, on calculera une Ampleur avec pour *domaine* d'application du calcul le titre. Le *domaine* peut être un moyen de contextualiser une mesure : par exemple, la proportion d'usage du présent dans les passages au discours direct. Pour toutes les mesures que nous avons définies, le paramètre de domaine est disponible et permet de définir sur quelle portée la mesure est appliquée au texte.

Par ailleurs, tout domaine peut être vu soit comme un tout (malgré d'éventuelles discontinuités), soit comme la réunion de plusieurs sous-parties (s'enchaînant directement ou éparses au fil du texte). Dans ce second cas, nous disons que le domaine est vu dans sa *multiplicité*. La mesure est alors effectuée sur chacune des parties du domaine. On obtient une série de valeurs, et c'est l'application d'une *synthèse* qui permet de réduire la pluralité des résultats pour les parties à une seule valeur pour le texte. Par exemple, considérons l'Ampleur du texte en mots, en choisissant comme domaine l'ensemble de ses phrases. Si le domaine n'est pas vu dans sa multiplicité, alors le résultat obtenu est le nombre de mots du texte (à l'exception des mots qui ne seraient pas dans des phrases). Si le domaine est vu dans sa multiplicité, alors

on calcule la longueur de chacune des phrases en nombre de mots, puis on applique une synthèse, par exemple un calcul de moyenne : le résultat de la mesure est alors la longueur moyenne (en mots) des phrases du texte.
Une dizaine de formes de synthèses sont envisageables, comme : la moyenne, l'écart-type, le minimum, le maximum, la médiane, les premier et troisième quartiles, et des notions originales comme l'intensité (moyenne sur les valeurs non nulles) ou la diffusion (proportion de parties où la valeur est non nulle).
La combinaison des différents types de mesure, des différentes définitions de domaines, et des formes de synthèse, permet de traduire et interpréter dans notre modèle un très grand nombre de mesures existantes, comme de concevoir des mesures adaptées à de nouveaux contextes.

3.6. Discussion : mesures et pré-interprétation

Cette démarche de mise au point de mesures engage le chercheur dans une modélisation des données textuelles et dans une représentation de ses points d'attention et de ses attentes. Plutôt que de prendre le texte tel quel, dans l'inventaire brut de ses mots ou de ses étiquettes, celui-ci se retrouve médiatisé par une série de mesures choisies. Mais alors, les mesures ne plaquent-elles pas certains *a priori* qui limiteraient les découvertes d'observables inattendues ?
Au moins deux éléments de réponse peuvent être proposés. Tout d'abord, il n'y a pas de représentation neutre du texte : la manière même de le découper en mots, ou le choix du jeu d'étiquettes, définissent non moins des points de vue. Autrement dit, le chercheur est forcément engagé par sa façon d'envisager ses données. En proposant des principes méthodologiques de conception de mesures, le présent travail entend justement donner des repères pour éviter des représentations arbitraires et s'efforcer de rester proche du texte et de son fonctionnement. Comme l'intérêt des observations est lié à la pertinence de la représentation de travail, le fait de définir cette représentation sur la base de principes linguistiques est un appui pour contrôler sa pertinence.
Par ailleurs, les résultats obtenus à l'aide des mesures sont non seulement les valeurs calculées pour chacune des mesures, mais aussi les résultats issus de nouveaux traitements applicables au tableau des mesures. Ainsi, des observables non directement prévus peuvent apparaître par la convergence ou la recomposition de mesures initiales. Ceci étant, si nous avons avancé sur la question de la définition individuelle de mesures, celle de la cohérence d'ensemble et de l'équilibre des jeux de mesure n'est encore que partiellement traitée (par le fait que les types de mesure donnent chacun une manière unifiée de définir différentes mesures).

4. Première mise en œuvre

4.1. Implémentation des mesures dans le logiciel TXM

La formalisation des différents types de mesure a permis de spécifier un nouveau composant *Mesures* dans la plateforme ouverte TXM (http://textometrie.ens-lyon.fr/). TXM est un logiciel pour l'analyse textométrique de corpus, conçu pour traiter les corpus enrichis, structurés et annotés.
Les premières utilisations de ce composant Mesures ont vraiment montré l'intérêt de disposer de l'environnement de toute l'application de textométrie pour la mise au point et l'interprétation des mesures. En effet, on dispose de tous les moyens de repérage et d'observation en contexte de ce que l'on compte, ce qui permet en amont d'affiner la définition des paramètres de la mesure, et en aval de contrôler l'interprétation en consultant le texte dans le mode de présentation le plus adapté.

4.2. Problématique linguistique : analyse des discours représentés dans les textes fictionnels

Dans la perspective d'une caractérisation des textes fictionnels sur la composante dialogique (au sens de Rastier), nous avons choisi comme domaine les discours représentés (*DR*). Nos variables sont le genre textuel et les différents types de DR. Le questionnement porte sur la présence relative des différents DR, leurs enchaînements et emboîtements, leur mode d'introduction et de marquage, la richesse des verbes introducteurs. Ces mesures visent à caractériser des types de narrateurs différents (Malrieu, 2006).

4.3. Corpus

Pour cette première expérimentation nous avons choisi un corpus de quatre textes fictionnels contrastés par le genre et en diachronie :
– Duras, *L'Amant de la Chine du Nord* (XX^{e}, roman), désigné par la suite ACN
– Stendhal, *La chartreuse de parme*, livre 1 (XIX^{e}, roman)
– Mme d'Aulnoy, *Le Pigeon et la Colombe* (XVIII^{e}, conte)

– Gripari, *Les contes de la rue Broca* (XX^{e}, contes)

Les textes ont été balisés en XML-TEI : un en-tête (<header>) rassemble les informations sur le texte (date, auteur, etc.) ; dans le texte sont balisés les divisions ( ) (chapitres, contes), les discours directs (DD) et indirects (DI), la parole intérieure (MI) au discours direct ou narrativisée (MIN) (codage avec l'élément <q> et différentes valeurs pour la propriété *type*), les incises de dire et de parole intérieure ainsi que les segments introducteurs de ces discours (<seg> avec différentes valeurs de l'attribut *ana*), les adresses, les DN (discours narrativisé), le psycho-récit (PR), les lettres, les citations. L'import dans TXM a permis d'ajouter automatiquement sur chaque mot sa catégorie grammaticale et son lemme.

4.4. Observations réalisées et présentation de quelques résultats

Mesures implémentées :

● Caractérisation générale du texte :
– Longueur (Ampleur)
– Proportion de phrases sans verbe conjugué
● Les DR dans le texte:
– Taux de présence linéaire de chaque type de DR
– Longueur de chaque type de DR (quartiles) (Ampleur)
● Propriétés du DD
– Taux de présence linéaire des ponctuations fortes
– Proportion de points, points d'interrogation et d'exclamation dans les ponctuations fortes du DD
– Proportion de phrases sans verbe conjugué
● L'insertion des DR dans le texte
– Proportion de DD typographiquement marqués avec un tiret, avec des guillemets
– Proportion de DD (resp. MI, MIN) avec incise ou/et avec verbe introducteur
– Diversité en lemmes des verbes des incises, et des verbes introducteurs pour chaque type de DR
● Les rapports entre DR
– Inclusions : proportions de DI dans du DD et de DD contenant du DI, proportion de MI(N) dans du PR, proportion de DN dans du PR ;
– Enchaînements : proportion de DD (resp. DI, MI, MIN) précédé (resp. suivi) immédiatement (resp. à moins de 10 mots) d'un PR.

Présence des DR

Mesures	***ACN***	***BROCA***	***CHARTREUSE1***	***PIGEON***
Longueur du texte (en mots)	63 418	22 878	104 626	21 058
Taux de présence linéaire de DD	**36,1 %**	**41,0 %**	32,5 %	31,0 %
Taux de présence linéaire de DI	9,9 %	0,8 %	5,0 %	8,6 %
Taux de présence linéaire de MI	**0,03 %**	3,4 %	10,3 %	4,4 %
Taux de présence linéaire de MIN	0,4 %	0,1 %	0,8 %	0,9 %
Taux de présence linéaire de DN	0,3 %	2,0 %	**4,3 %**	**5,3 %**
Taux de présence linéaire de PR	15,0 %	1,1 %	6,6 %	17,2 %

Tableau 1 : Longueur des textes et taux de présence linéaire des DR

Les deux textes contemporains accordent une place prépondérante au DD alors que le récit d'évènements de parole par le narrateur (DN) est plus présent dans la *Charteuse* et le *Pigeon*.

Le MI est quasi absent dans le roman scénario de film alors qu'il est présent dans les trois autres. Le taux de présence du DI varie de 1% (*Broca*) à 10% (*L'ACN*), en rapport avec son mode d'introduction : par le narrateur *vs* à l'intérieur d'un DD.

Richesse lexicale des verbes introducteurs de DR

L'analyse de la richesse lexicale des verbes introducteurs se heurte pour l'instant à deux difficultés. Tout d'abord, il y a les erreurs de reconnaissance des verbes par l'étiqueteur (Treetagger), en particulier l'absence de distinction entre auxilaires et verbes. Mais surtout, le segment introducteur du DR peut comporter plusieurs verbes dont certains ne sont pas introducteurs (*commencer, éloigner, vivre* etc.) : il faudrait donc baliser le verbe introducteur proprement dit.

Diversité en lemmes	*ACN*	*BROCA*	*CHARTREUSE1*	*PIGEON*
Verbes introducteurs	80	101	149	58
Verbes des incises de dire ou de pensée	5	11	110	24

Tableau 2 : Diversité en lemmes des verbes introducteurs et des verbes d'incise

On constate que la diversité lexicale des verbes introducteurs est faible, avec une chute diachronique vertigineuse : les verbes d'incise de *dire* ou de pensée passent de 110 dans la Chartreuse à 5 dans *l'ACN* et 11 dans *Broca*. La grande différence entre deux textes comparables quant à leur volume, *la Chartreuse* et *l'ACN*, n'est pas sans rapport avec le type de narrateur : dans *la Charteuse*, le narrateur qualifie la parole prononcée, dans *l'ACN* la parole est donnée de façon brute, c'est le psycho-récit qui en délivre la couleur. Cela est confirmé par l'analyse des enchaînements psycho-récit/type de DR.

Positions respectives des DR et enchaînements

Ici apparaît clairement l'intérêt de l'analyse des arborescences textuelles et non plus des distances linéaires, car les rapports d'inclusion des DD et DI mettent en évidence une disparité des dispositifs énonciatifs de *l'ACN* par rapport aux autres œuvres. Dans *l'ACN* seuls 14% des DI sont inclus dans un DD contre 54% dans *Broca* et 34 et 23% pour *La Charteuse* et *le Pigeon*. Une analyse des positions respectives des DD et DI montrerait des dispositifs différents : alternance de DD et DI dans une même scène énonciative *vs* inclusion du DI dans un DR.

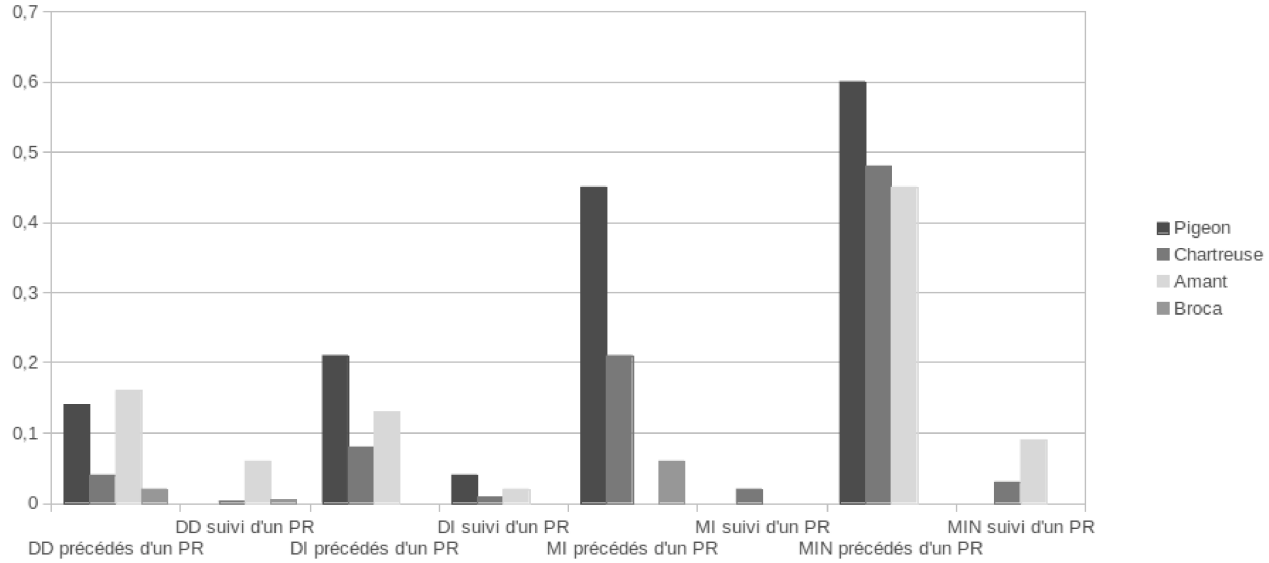

Figure 1 : Pour les différents types de DR, proportion de leurs attestations précédées ou suivies à moins de 10 mots par un PR

Les positions respectives des DR et du PR (figure 1) sont aussi instructives : s'il est prévisible que le MI et le MIN soient fréquemment précédés d'un PR (de 22 à 45% dans les textes où ils existent, le MIN étant inclus à 8% dans le PR dans *l'ACN*), on peut noter, en ce qui concerne les positions relatives du DD et du PR, un fort contraste entre d'un côté *l'ACN* et *le Pigeon*, de l'autre *Broca* et la *Chartreuse* : pour les premiers 16 et 14% des DD sont précédés à moins de 10 mots d'un PR contre 2 et 4% pour les autres. On retrouve les mêmes différences entre textes pour le DI précédé d'un PR à moins de 10 mots : 14 et 21% contre 0 et 8%. On voit se confirmer la différence de dispositif entre *La Chartreuse* et *l'ACN* : dans ce dernier le MI est absent et le PR fonctionne avec le DD, alors que dans le premier, le PR est fortement lié au MI et MIN.

4.5. Discussion : limites et ouvertures

Cette première expérimentation repose sur un corpus très finement annoté ; en contre-partie, il n'y avait qu'un seul texte pour représenter un genre (conte ou roman) à une période (siècle), si bien que les observations restent très liées aux textes considérés et ne peuvent être généralisées dans l'immédiat.

Nous avons vu que la qualité de l'analyseur morphosyntaxique utilisé pour l'étiquetage lexical a évidemment une incidence directe sur les mesures l'utilisant. Ici, la plupart des mesures étaient supralexicales et se basaient sur le codage fiable des passages de DR, mais il est rare de disposer de

corpus avec une telle richesse d'annotation, si bien que l'étiquetage morphosyntaxique prend plus souvent une place centrale dans l'analyse. Ceci étant, l'intégration du calcul de mesures dans un environnement de textométrie aide à identifier les erreurs, qualitativement (retour au texte, observation en contexte) et quantitativement (effectifs en jeu), pour les prendre en compte au moment de l'interprétation des mesures.
Enfin, notons que les mesures rencontrent des limites lorsqu'il existe de grosses disparités de taille des documents et des fréquences très faibles dans certains textes du corpus.
Les chiffres produits ont été ici directement considérés pour l'analyse ; pour un corpus de textes plus nombreux ou pour mettre en évidence des corrélations entre mesures, la production de tableaux chiffrés pourrait n'être qu'une étape intermédiaire avant des procédures d'analyse des données, aidant à dégager des dimensions de synthèse pour décrire la diversité interne d'un corpus.

5. Conclusion

Face à la multiplicité de mesures envisageables pour caractériser des textes en corpus, nous avons effectué un travail d'abstraction pour rapporter cette diversité à un petit nombre de types de mesure linguistiquement motivés. Nous avons proposé deux types de mesures liés à l'étendue textuelle, deux types de mesures paradigmatiques pour décrire la variation linguistique et les effets de préférence, et quatre types de mesures syntagmatiques rendant compte des dépendances et des rythmes. Ces types de mesures ne se définissent pas d'abord par des formules mathématiques, mais par une interprétation linguistique, déterminante pour l'instanciation des différents éléments de la formule.
Le contexte de mise au point de ces types de mesures relève de la linguistique de corpus : étude différentielle des genres textuels et de leur évolution diachronique, sous l'angle des formes de discours représenté. L'illustration développée dans cet article montre que les mesures offrent de nouveaux observables ajustés à la problématique de recherche, qui pourraient également intéresser l'analyse littéraire et stylistique.

Remerciements

Cette recherche n'aurait pu être menée sans le concours de Serge Heiden et Matthieu Decorde, de l'équipe TXM (Lyon, ICAR), qui se sont impliqués dans la conception informatique et l'implémentation d'un module *Mesures* au sein du logiciel TXM. La dette est non seulement technique, mais aussi scientifique, puisque pour être ainsi concrétisées les idées doivent être discutées et précisées.

Bibliographie

BIBER D., *Variation across speech and writing*, Cambridge University Press, 1988.
BRONCKART J.-P., BAIN D., SCHNEUWLY B., DAVAUD C., PASQUIER A., *Le fonctionnement des discours – Un modèle psychologique et une méthode d'analyse*, Delachaux & Niestlé, 1985.
MALRIEU D., RASTIER F., « Genres et variations morphosyntaxiques », *Traitement automatique des langues*, 42 (2), 2001, p. 547-577.
MALRIEU D., « Discours rapportés et typologie des narrateurs dans le genre romanesque », Actes du Colloque Ci-Dit de Cadiz, 11-13 mars 2004, J. M. LOPEZ-MUÑOZ, S. MARNETTE et L. ROSIER (dir), *Dans la jungle des discours : genres de discours et discours rapporté*, Cadix, Presses de l'Université de Cadix, 2006.

Présentation des auteurs

Bénédicte Pincemin est chargée de recherche CNRS en linguistique au laboratoire ICAR (UMR 5191, Lyon) (http://icar.univ-lyon2.fr/membres/bpincemin/). Ses travaux portent sur la sémantique des textes, selon l'approche initiée par François Rastier (sémantique interprétative, différentielle, unifiée). Elle s'intéresse en particulier à l'instrumentation de la lecture et de l'analyse de grands corpus de textes intégraux. Elle développe et expérimente actuellement ses propositions dans le contexte de la plateforme ouverte TXM, logiciel de textométrie (analyse statistique de données textuelles) (http://textometrie.ens-lyon.fr/).
Pincemin Bénédicte (2012) - « Sémantique interprétative et textométrie », Texto!, XVII, 3, numéro coordonné par Christophe Cusimano. En ligne : http://www.revue-texto.net/index.php?id=3049.
Denise Malrieu a développé l'application de la méthodologie de balisage XML-TEI de textes littéraires dans l'optique d'une sémantique textuelle prenant en compte les genres, plus particulièrement la composante dialogique des textes. Son travail actuel porte sur l'articulation de la linguistique de la langue et de la linguistique de la parole dans l'interprétation sémantique du verbe dire (Lambert-Lucas, 2012).