Cédille
revista de estudios franceses

ISSN: 1699-4949

nº 19 (primavera de 2021)

Monografías 12

Linguistique textuelle et graphématique du français en diachronie
Elena Llamas-Pombo & Diana Gatea, editoras científicas

Approche quantitative des marques graphiques et lexicales de l'oral représenté à travers les corpus *BFM* et *BVH*

Alexei LAVRENTIEV
École Normale Supérieure de Lyon – CNRS
alexei.lavrentev@ens-lyon.fr
https://orcid.org/0000-0001-8306-3653

Resumen

En este artículo, presentamos una experiencia de uso de corpus digitales –la *Base de français médiéval* y las *Bibliothèques virtuelles humanistes* – para el estudio de la señalización gráfica y léxica del oral representado y del discurso citado. Según los métodos de edición y el estado del marcado, se pueden aplicar diferentes búsquedas al conjunto de los corpus o a algunas de sus partes. Los indicadores estudiados incluyen la frecuencia relativa del verbo *dire* ('decir'), la frecuencia de los diferentes signos de puntuación empleados en las fronteras entre elementos que constituyen episodios de oral representado y, por último, las marcas de cita de fuentes escritas.

Palabras clave: oral representado, puntuación, diacronía de la lengua francesa, anotación, filología digital.

Résumé

L'article présente une expérience d'exploitation de corpus numériques, la *Base de français médiéval* et les *Bibliothèques virtuelles humanistes*, pour l'étude de la signalisation graphique et lexicale de l'oral représenté et des citations. Selon les méthodes d'édition et l'état de balisage, différentes requêtes peuvent être appliquées à l'ensemble des corpus ou à leurs parties. Les indicateurs étudiés incluent la fréquence relative du verbe *dire*, la fréquence des différentes marques de ponctuation utilisées aux frontières des éléments qui constituent les épisodes d'oral représenté et les marques de citation de sources écrites.

Mots-clés : oral représenté, ponctuation, diachronie de la langue française, annotation, philologie numérique.

Abstract

This paper presents an experience of using digital text corpora, *Base de français médiéval* and *Bibliothèques virtuelles humanistes*, in a research on graphical and lexical markers of

* Artículo recibido el 15/09/2020, aceptado el 30/03/2021.

represented spoken utterances and citations of written sources. Depending on editing practices and on the state of markup, different queries may be applied to the whole corpora or to some parts of them. The indicators examined in the paper include the relative frequency of the verb *dire* ('to say'), the frequency of different punctuation marks used at the borders of the structural elements of represented speech, and marks of written sources citations.
Keywords: represented speech, punctuation, diachrony of the French language, annotation, digital philology.

1. Introduction

Dans cet article, nous visons à explorer les possibilités que les corpus numériques disponibles à ce jour peuvent offrir pour l'étude de la signalisation de l'oral ou discours représenté sur une échelle diachronique large, grâce à l'analyse quantitative des marques graphiques et lexicales qui s'y trouvent.

Les grands corpus diachroniques du français ancien, tels que la *Base de français médiéval* (BFM) ou la *Base textuelle du moyen français* sont malheureusement inutilisables pour les recherches sur la ponctuation et les systèmes graphiques, car ils sont constitués, pour la grande majorité des textes, à partir d'éditions scientifiques où la ponctuation est entièrement refaite selon les normes de la typographie moderne. D'un autre côté, les études sur la ponctuation de l'oral représenté menées sur les manuscrits sont rares et leur corpus est restreint en raison de la difficulté du dépouillement (Marnette, 2006 ; Llamas-Pombo, 2010).

La normalisation de la ponctuation présente toutefois certains avantages pour l'analyse linguistique. En l'occurrence, grâce aux guillemets introduits par les éditeurs, il est possible de repérer et baliser d'une manière quasi-automatique les passages au discours direct. Ce procédé a déjà permis de réaliser une série d'études sur le fonctionnement du discours rapporté et de ses marqueurs sur un corpus de plusieurs millions de mots (Guillot-Barbance *et al.*, 2017, 2018), mais la ponctuation n'a pas fait partie des phénomènes analysés.

Le développement récent d'éditions d'un nouveau type, « nativement numériques », constitue une avancée déterminante dans la qualité de données pour la recherche. Ces éditions commencent depuis quelques années à enrichir le corpus de la BFM. Grâce à la transcription « multi-facette », elles permettent à la fois de préserver l'information sur la ponctuation, la segmentation graphique et l'usage des majuscules de l'original et d'ajouter une couche normalisée facilitant la lecture du texte et l'application des outils de TAL (traitement automatique du langage). À l'édition de la *Queste del saint Graal* (2019), première dans son genre, s'ajoutent désormais la *Vie de saint Alexis* (2018), l'*Image du monde* en prose de Gossouin de Metz (2018) et les *Quinze joies de mariage* (2018). Malgré ses dimensions restreintes, ce corpus de nouvelles éditions offre des possibilités inédites pour l'analyse quantitative de la ponctuation

médiévale. Toutes les occurrences du discours direct y sont soigneusement balisées, ce qui permet d'extraire et d'analyser l'ensemble des marques graphiques et lexicales.

Pour le XVI^e^ siècle, le corpus *Epistemon* des *Bibliothèques virtuelles humanistes* offre des données fiables sur les graphies et la ponctuation grâce au principe de la transcription patrimoniale, très fidèle aux documents sources. En revanche, l'analyse linguistique de ces données demande un certain effort, car l'absence de normalisation des graphies et de la ponctuation ne permet pas d'appliquer d'une manière efficace les outils de TAL. Ainsi, le repérage automatique du discours direct basé sur la ponctuation ne peut fonctionner que dans les textes où les guillemets ou d'autres marques graphiques spéciales sont présents dans les originaux. Néanmoins, grâce aux outils d'analyse de corpus, comme la plateforme TXM (Heiden *et al.*, 2010), il est possible de rechercher l'ensemble des marques susceptibles de signaler le discours direct (*verba dicendi*, interjections, etc.), d'annoter les occurrences pertinentes et d'appliquer ensuite des méthodes d'analyse qualitative et quantitative.

Dans cet article nous présentons d'abord les corpus étudiés et la manière dont nous les avons utilisés. Les données quantitatives concernant la ponctuation du discours direct et des citations sont ensuite analysées. Même si les données analysées sont loin d'être représentatives de l'époque étudiée dans son ensemble, elles permettent de formuler et vérifier un certain nombre d'hypothèses et d'ouvrir de nouvelles pistes de réflexion.

2. Corpus étudiés : leurs avantages et limitations

2.1. Base de français médiéval et ses éditions « nativement numériques »

La *Base de français médiéval* est un ensemble de corpus de textes datés du IX^e^ au XV^e^ siècle régulièrement complété et enrichi d'annotations linguistiques. À ce jour, le corpus principal BFM2019 est composé de 170 textes intégraux, soit près de 4,7 millions de mots au total. La grande majorité des textes de la BFM sont des éditions scientifiques numérisées. Ce choix assumé à la fondation de la *Base* a permis de constituer un corpus de taille importante en relativement peu de temps. Grâce à la sélection d'éditions « bédiéristes », les données de la BFM sont suffisamment fiables pour la plupart des recherches linguistiques (Lavrentiev, 2007). La ponctuation interprétative moderne introduite par les éditeurs facilite la compréhension des textes et leur traitement par les outils automatiques. Il existe cependant des questions de recherche pour lesquelles les données des sources primaires manuscrites ou imprimées sont indispensables. C'est pourquoi le besoin d'éditions plus proches de ces sources s'est fait d'année en année plus évident. Les technologies numériques permettent à présent de résoudre le dilemme entre lisibilité et fidélité aux sources primaires grâce à la représentation « multi-facette » qui donne au lecteur-utilisateur la possibilité d'adapter la présentation du texte à ses besoins et intérêts. Ces éditions permettent ainsi de répondre à des

attentes multiples et rendent possible une approche quantitative et diachronique des marques de l'oral représenté.

Inspirée du projet *Charrette*, l'édition numérique de la *Queste del saint Graal* (2019) a été la première expérience de ce nouveau type d'édition dans le cadre de la BFM. Initié en 1999, ce projet continue d'évoluer et de s'enrichir périodiquement jusqu'à présent. Sa dernière révision date de 2019, mais elle concerne surtout l'habillage graphique du site et les fichiers PDF, les transcriptions et les annotations sont stables depuis 2007, à l'exception de la correction de quelques coquilles. Les principes méthodologiques de cette édition sont exposés dans son introduction et dans un chapitre du *Manuel de la philologie de l'édition* (*cf.* Marchello-Nizia, Lavrentiev et Guillot-Barbance, 2015). Ils sont également valables pour les autres éditions originales de la BFM que nous présenterons plus bas.

Le lecteur peut accéder à trois types de transcription : « normalisée », « diplomatique » et « facsimilaire ». La transcription normalisée est la plus proche des pratiques traditionnelles des éditions critiques de textes en français médiéval. Toutefois, la fidélité au manuscrit y est préservée dans la mesure du possible : les interventions éditoriales se limitent à la ponctuation du discours direct, aux majuscules dans les noms propres, aux lettres « ramistes », à la segmentation des mots, à la résolution des abréviations et à la correction, toujours signalée, des erreurs manifestes, comme des mots et syntagmes répétés. La transcription diplomatique est encore moins interventionniste : les majuscules et les minuscules du manuscrit sont toujours préservées et les résolutions des abréviations sont signalées par les italiques. Les marques de ponctuation sont représentées par des caractères modernes (le point, la barre oblique, le point d'interrogation et la séquence « point + accent aigu » pour le *punctus elevatus*), mais les principales distinctions sont ainsi conservées et aucune marque de ponctuation n'est ajoutée ou supprimée par rapport au manuscrit. La transcription facsimilaire, enfin, est plus précise encore dans la reproduction des graphies du manuscrit, jusqu'à la distinction des variantes de lettres (allographes), telles que le *s* « long » ou le *r* « rond », les marques d'abréviation et de ponctuation médiévales. La transcription facsimilaire est la plus complexe à réaliser et peut nécessiter, pour sa visualisation, l'installation d'une police de caractères compatible MUFI (*Medieval Unicode Font Initiative*). Ce type de transcription est intéressant pour des recherches paléographiques ou pour l'analyse linguistique très fine des graphies médiévales, mais la transcription diplomatique suffit pour des études linguistiques de la ponctuation, puisque les distinctions principales y sont préservées.

L'édition de la *Queste del saint Graal* fait bien évidemment partie de notre corpus d'étude. Son volume est assez important (118 488 occurrences de mots et de ponctuations), dont la moitié (50,2% des occurrences) se trouve dans le discours direct des personnages. Les prises de parole vont de quelques mots à de longs monologues qui s'étalent sur plusieurs paragraphes et peuvent contenir jusqu'à deux niveaux de discours

direct imbriqué, comme c'est le cas dans la colonne 167c où un chevalier raconte à Galaad une histoire dans laquelle Josephé, fils de Joseph d'Arimathie selon le roman, parle au roi Evalach et prédit ce que ce dernier dira au moment où il sera en danger de mort. La *Queste del saint Graal* offre donc de nombreux exemples pour l'étude de la ponctuation du discours direct dans son manuscrit de base (Lyon, Bibliothèque municipale, Palais des arts 77, milieu du XIII[e] siècle), mais il est loin de représenter toute la diversité des pratiques médiévales.

La *Vie de saint Alexis* est un texte beaucoup plus court (5 536 occurrences), mais il joue un rôle particulièrement important pour l'étude de l'histoire de la langue française, puisqu'il s'agit de l'une des plus anciennes œuvres qui nous soit parvenue. Le manuscrit Hildesheim, St. Godehardi, qui a servi de base à l'édition de T. Rainsford et C. Marchello-Nizia (2018), comme à beaucoup d'autres, est également l'un des plus anciens (ca. 1120). L'édition est dotée d'un riche apparat critique (135 notes) et présente un certain nombre de choix de lecture originaux. Malgré de nombreuses études qui portent sur le texte et le manuscrit, la ponctuation du discours direct n'y a pas attiré jusqu'à présent l'attention des chercheurs. La part du discours direct est moins importante dans ce texte que dans la *Queste del saint Graal*, mais elle représente tout de même 33% des occurrences-mots.

Une autre édition numérique « multi-facette » de la BFM est celle de l'*Image du monde* en prose attribuée à Gossouin de Metz. Elle est réalisée par N. Kanaoka à partir du manuscrit BnF, fr. 574 (XIV[e] siècle). Malheureusement, ce texte « scientifique » n'offre que très peu d'occurrences du discours direct (0,4% des occurrences, soit 230 sur 51 066, réparties dans sept prises de paroles ou citations « déguisées »). Malgré ce nombre limité d'occurrences, l'*Image du monde* fournit quelques exemples intéressants pour l'étude de la ponctuation du discours direct.

L'édition des *Quinze joies de mariage* d'après le manuscrit Rouen, Bibliothèque municipale, 1052 (milieu du XV[e] siècle), a été réalisée grâce à la collaboration entre les Presses universitaires de Rouen et du Havre et l'équipe de la BFM. Dans le cadre d'une convention, la BFM a reçu les fichiers qui ont servi à la publication parue en 2009 (édition établie par M. Guéret-Laferté, S. Louis et C. Mora). Le corps du texte a été adapté aux normes des éditions multi-facette de la BFM et son apparat critique complété par N. Kanaoka. Comme dans la *Vie de saint Alexis*, le discours direct représente près de 33% du volume du texte, soit 13 291 occurrences sur 39 564.

Les transcriptions normalisées des éditions que nous venons de présenter sont intégrées au corpus BFM2019 accessible à tous. Toutefois, à l'exception de la *Queste del saint Graal*, la consultation des différentes facettes de la transcription est à l'heure de l'écriture de cet article expérimentale. L'accès à la lecture synoptique des différentes transcriptions et des images des manuscrits, ainsi qu'au moteur de recherche en ligne, se fait sur demande auprès du gestionnaire de la BFM.

Bien entendu, les données provenant des quatre éditions multi-facette de la BFM sont insuffisantes pour tirer des conclusions sur la ponctuation du discours direct dans les manuscrits français médiévaux en général, mais elles fournissent des exemples assez diversifiés sur plusieurs plans et permettent de voir le potentiel de ce genre d'édition pour les recherches linguistiques.

Le corpus général BFM2019 sera utilisé, lorsque ce sera possible, pour mettre en perspective les données obtenues sur le corpus restreint.

2.2. Corpus Epistemon des Bibliothèques virtuelles humanistes

Les *Bibliothèques virtuelles humanistes* (BVH) sont devenues au fil des années une ressource incontournable pour l'étude de la littérature française du XVI^e^ siècle. C'est le corpus *Epistemon* constitué de transcriptions « patrimoniales » des imprimés de l'époque et de certains manuscrits d'auteurs (Rabelais, Montaigne, etc.) qui nous intéresse pour cette étude, grâce à ses transcriptions XML-TEI, qui peuvent être analysées par des outils textométriques. Le corpus peut être interrogé en ligne via un portail TXM[1], le même logiciel que celui de la *Base de français médiéval*. Nous avons toutefois travaillé avec une version plus récente du corpus, que nous avons obtenue auprès de l'équipe des BVH et que nous avons traitée avec l'application TXM pour ordinateur personnel. Le corpus que nous avons préparé pourra dans l'avenir être intégré au portail TXM des BVH.

L'ensemble du corpus *Epistemon* que nous avons étudié en 2019 est composé de 153 unités textuelles (soit 89 manuscrits et 64 imprimés, près de 3 900 000 occurrences au total). Un cas particulier est constitué par « l'exemplaire de Bordeaux » des *Essais* de Montaigne (1588), un livre imprimé portant de nombreuses annotations manuscrites de l'auteur.

Le principe de transcription patrimoniale, très fidèle aux documents sources, a présidé aux choix d'encodage des BVH (Dufournaud *et al.*, 2017). Même les réclames, les titres courants et les numéros de page sont reproduits. Les ponctuations sont évidemment respectées, ainsi que les majuscules et les minuscules. Les résolutions d'abréviations et certaines régularisations graphiques (notamment pour les noms propres) sont proposées en complément grâce à la balise TEI <choice>. En revanche, aucune régularisation n'est proposée pour la ponctuation, ce qui rend difficile le repérage automatique du discours direct, lorsqu'il n'est pas indiqué par une ponctuation spécifique dans les sources.

Certains imprimés du XVI^e^ siècle utilisent les italiques ou les guillemets dans les marges pour marquer les citations. Nous avons analysé ces usages, qui sont toutefois rares dans le corpus et ne concernent pas a priori le discours direct de type « oral représenté » (voir *supra*). Par ailleurs, de nombreux textes du corpus *Epistemon* ne semblent pas contenir de discours direct (par exemple, des traités ou certaines poésies) et ne sont

[1] https://txm.bvh.univ-tours.fr/txm

donc pas pertinents pour notre analyse. Pour ces raisons, après avoir effectué quelques sondages sur l'ensemble du corpus, nous avons sélectionné pour une analyse approfondie un sous-corpus de cinq romans dans lesquels nous avons annoté manuellement les occurrences du discours direct signalé par le verbe *dire*. Il s'agit des romans suivants (nous indiquons pour chaque texte le lieu d'impression, le nom de l'imprimeur ou du libraire, l'année d'impression, le lieu de conservation de l'exemplaire transcrit et le nombre de mots et de ponctuations) :

1. *Pantagruel* de François Rabelais (Lyon, Claude Nourry, 1532, BnF, 43 138 occ.) ;
2. *Le Tiers livre* de François Rabelais (Paris, Chrétien Wechel, 1546, Bibliothèque municipale de Tours, 64 733 occ.) ;
3. *Alector* de Barthélemy Aneau (Lyon, Pierre Fradin, 1560, Bibliothèque de l'Université de Virginie, 79 993 occ.) ;
4. *Nouvelles recreations et joyeux devis* de Bonaventure Des Périers (Lyon, Guillaume Rouillé, 1561, Bibliothèque municipale de Lyon, 84 212 occ.) ;
5. *Le Tableau des riches inventions…* de Francesco Colonna « exposé » en français par Béroalde de Verville (Paris, Matthieu Guillemot, 1600, bibliothèque du CESR à Tours, 163 162 occ.).

L'ensemble des cinq romans constitue un corpus de près de 435 000 occurrences, soit près de 11% du volume total d'*Epistemon*.

3. Annotation de l'oral représenté

Comme on vient de le voir, l'analyse quantitative du marquage de l'OR suppose avant tout la restitution sous forme numérique des marques graphiques présentes dans la source primaire, mais elle implique également qu'on identifie les éléments lexicaux et les structures syntaxiques pertinents pour l'analyse. L'annotation du corpus est donc indispensable afin de quantifier précisément les marques graphiques et lexicales qui accompagnent l'oral représenté dans les textes. Pour cette annotation, nous nous basons sur la notion d'épisode d'oral représenté (désormais abrégé OR) définie par Christiane Marchello-Nizia (2012) et précisée par Céline Guillot-Barbance *et al.* (2018). Toutefois, nous ne nous limitons plus strictement au discours direct, comme c'était le cas des études précédentes, mais annotons toutes les occurrences qui relèvent de l'OR au sens large.

Nous entendons par un épisode d'OR une séquence discursive se donnant comme une restitution à l'écrit de paroles prononcées oralement. Les épisodes d'OR sont composés de prises de paroles (désormais abrégé PP) : soit d'une seule (en cas de monologue) soit de plusieurs (en cas de dialogues entre plusieurs personnages). Les éléments qui précisent les limites externes et les articulations internes des épisodes d'OR, soit l'annonce, le rappel et l'incise, en font partie intégrante. L'annonce et l'incise se caractérisent par la présence d'un verbe de parole ; il est donc relativement facile

de les repérer dans les concordances pour ensuite annoter les épisodes d'OR auxquels ils participent. D'autres éléments, comme les termes d'adresse (*dame, mes sire*, etc.) ou les interjections, souvent situés au début de prise de parole, peuvent également servir à repérer les épisodes d'OR dans de grandes masses de données.

Dans la réalité les épisodes d'OR se présentent souvent comme un mélange de discours direct et indirect, comme dans le petit dialogue de l'*Image du monde* cité dans l'exemple suivant[2] :

> (1) Dont aucu*n*s || porroit **demander. pu||is** que eles uienne*n*t || toutes de la mer com||ment ce est que yaue || douce en uient**. A** ce || **respont** .i. des aucteur*s* || que liau qui a son || cours par la douce t*er*re || est douce . (Gossouin 2018 : 79c).

Nous distinguons ici deux annonces, une prise de parole au discours direct et un discours indirect. Le début du discours direct (après l'annonce) est marqué par une ponctuation faible (point + minuscule). Sa fin, qui correspond aussi au changement de locuteur et au début de l'annonce du discours indirect, prend une ponctuation forte. S'il n'y avait pas de subordonnée causale au début de la prise de parole, il aurait été difficile de dire s'il s'agit d'un discours direct ou indirect.

Dans certains textes du corpus *Epistemon*, nous avons des épisodes d'OR où le mélange entre le discours direct et indirect des personnages avec les énoncés où l'auteur s'adresse au lecteur est encore plus complexe. L'exemple suivant est tiré des *Nouvelles recreations* de Bonaventure Des Périers :

> (2) Je loueroys beaucoup plus celuy de || nostre temps, qui ha esté si plaisant en sa vie, que par || une antonomasie on l'ha appellé le Plaisantin : Chose || qui luy estoit si naturelle & si propre, qu'à l'heure mes||me de la mort, combien que tous ceux qui y estoyent || le regrettassent : si ne peurent ilz jamais se fascher : tant || il mourut plaisamment. On luy avoit mis son lit au || long du feu sus le plastre du foyer, pour estre plus chau||dement. Et quand on luy **demandoit, Or** ça mon amy, || ou vous tient il **? Il respondoit** tout foiblement, n'ayant || plus que le cueur & la langue**, Il** me tient**, dit il,** entre le || banc & le feu**, qui** estoit à dire, qu'il se portoit mal de || toute la personne. Quand ce fut à luy bailler l'extreme || unction, il avoit retiré ses piedz à cartier tous en un || monceau. Et le prestre **disoit, Je** ne say ou sont ses || piedz **: Eh** regardez**, dit il,** au bout de mes jambes, vous || les trouverez**. Et** mon amy ne vous amusez point à rail-||ler **luy disoit on :** recommandez vous à Dieu **: Et** qui y || va **? dit il : Mon** amy, vous irez aujourd'huy si Dieu plaist**. || Je** voudrois bien estre asseuré**, disoit il,** d'y pouvoir || estre demain pour tout le

[2] Pour tous les exemples de la BFM nous adoptons la transcription diplomatique. Pour les exemples d'*Epistemon*, nous reproduisons la transcription patrimoniale dans sa version régularisée non corrigée. Dans les deux cas, nous représentons un saut de ligne par deux barres verticales. Nous mettons en gras les signes de ponctuation marquant les éléments d'épisodes d'OR, ainsi que les premiers mots qui les suivent, les verbes de parole et les incises.

> jour**. Recommandez** vous à || luy, & vous y serez en huy**. Et** bien**, disoit il,** mais que || j'y soys, je feray mes recommandations moy-||mesmes**. Que** voulez vous de plus naif || que celà ? quelle plus grande felici-||té ? (Des Périers 1561 : 9).

Ce passage contient plusieurs prises de parole associées à des annonces ou des incises. La ponctuation forte est généralement utilisée au début et à la fin des prises de parole, mais les signes de ponctuation sont assez variables : on trouve aussi bien des virgules que des points et des deux-points. Les incises sont généralement marquées par des ponctuations faibles (virgules ou deux-points), mais dans un cas aucune ponctuation ne sépare l'incise du discours direct qui la précède. Le passage se termine par des questions rhétoriques adressées au lecteur.

Un dernier exemple est tiré du *Tableau des riches inventions* exposé par François Béroalde de Belleville :

> (3) Je **respondis,** || **je** mettrois peine de m'avancer à la servir selon le merite de sa pudicité & l'honneur || que je dois à vostre respect **: Adonc** elle me **dit ; Mais** encores Poliphile luy estes || vous autant affectionné que vous feignez estre son serviteur **? Je** luy **replique Ma-||dame**, je vous proteste que ma vie ne m'est point tant aggreable que mon beau || sujet, aussi je luy ay tant voué d'amitié que si l'extremité d'amour se peut estimer || on la trouvera en moy pour son occasion**. Où** est-ce doncques **dit-elle** que vous || avez abandonné cét object tant extresmement aymé**. Je respondis** que je ne sça-||vois, & mesme je luy **dis** ainsi**, je** ne sçay en quel lieu je suis ny quelle avanture || me conduit **: Lors** en se sous-riant elle me **dit, que** donneriez vous à qui vous fe-||roit recouvrer vostre maistresse ? Ne vous donnez plus de soucy, faites bonne || chere & n'affligés plus vostre coeur, vous la trouverez bien tost**. Avec** tels devis les || Nymphes se bagnerent & moy avec elles : (Colonna 1600 : 27).

Ici l'auteur dialogue avec une nymphe, et son discours se confond par moment avec le récit à la première personne. La ponctuation semble très instable : on trouve des ponctuations faibles et fortes au début des prises de parole (dans un cas on a même la majuscule seule). La seule incise qui figure dans le passage n'est pas ponctuée, mais on trouve plus haut sur la même page du texte une incise placée entre parenthèses.

Compte tenu des différences importantes entre les données provenant des éditions de la BFM et du corpus *Epistemon*, nous avons appliqué des stratégies d'annotation différentes à ces deux corpus. Nous avons toutefois veillé à ce que ces annotations produisent *in fine* des indicateurs comparables.

Pour les textes de la BFM, où l'ensemble des passages au discours direct a été balisé, nous avons extrait les premiers mots de ces passages, ainsi que les premiers mots qui suivent leur fin et analysé la ponctuation qui les précède grâce à la transcription diplomatique.

Ces extractions automatiques ont été faites à l'aide de la commande Index de TXM avec les requêtes suivantes :

Élément recherché	Requête CQL[3]
Premier mot du discours direct et la ponctuation éventuelle qui le précède	[word="\p{P}+"]? <q> [word="\p{P}+"]? [word!="\p{P}+"]
Ponctuation éventuelle de fin de discours direct et le premier mot suivant	[word="\p{P}+"]* </q> [word="\p{P}+"]? [word!="\p{P}+"]

Tableau 1. Requêtes pour l'extraction des occurrences à annoter (début et fin du discours direct).

La « borne » du discours direct est indiquée par <q> (début) ou par </q> (fin). Les contraintes portant sur les occurrences sont données entre crochets. À chaque fois, il y a deux occurrences facultatives de ponctuations (\p{P} désigne n'importe quel caractère de la classe « ponctuation » du standard Unicode, « word » correspond à la forme de l'occurrence dans la transcription normalisée) et d'une occurrence d'un mot (défini comme toute occurrence sauf ponctuation).

La commande Index de TXM génère un tableau où la première colonne correspond à la forme des occurrences qui répondent à la requête et la deuxième, à leur nombre dans le corpus interrogé. Il est possible de sélectionner une ou plusieurs propriétés d'affiche pour la forme, c'est-à-dire « dipl » (la transcription diplomatique) dans notre cas. Ce tableau peut être exporté et édité ensuite avec un logiciel tableur, par exemple *Microsoft Excel* ou *LibreOffice Calc*. C'est ce dernier que nous avons utilisé pour l'annotation.

La grille d'annotation appliquée a été extrêmement simple : « 1 Forte » pour une ponctuation forte (une majuscule précédée ou non d'un signe de ponctuation), « 2 faible » (un signe de ponctuation suivi d'une minuscule), « 3 aucune » (minuscule sans ponctuation) et « 4 n/a ». La dernière étiquette a été utilisée dans les cas de « fausses bornes ». En effet, les contraintes du langage XML imposent que les balises <q> soient fermées à chaque fin de paragraphe ou de strophe, même si le discours direct continue.

Un extrait du tableau annoté est présenté ci-dessous :

[3] CQL (Corpus Query Language) est un langage de requête du moteur de recherche CQP (http://cwb.sourceforge.net) utilisé par TXM. Le *Manuel utilisateur TXM 0.7* présente, dans son chapitre 12, les principaux éléments de ce langage (Heiden *et al.,* 2015 : 122-131).

Dipl	annot	F
, e	2 faible	3
, certes	2 faible	2
, mercit	2 faible	2
. {F}ilz alexis	1 Forte	2
, an	2 faible	1
, as	2 faible	1
, dama	2 faible	1
, de	2 faible	1
, filz	2 faible	1

Tableau 2. Extrait des occurrences annotées (début du discours direct dans la *Vie de saint Alexis*).

Une fois les annotations terminées, les décomptes peuvent être réalisés très facilement grâce à l'outil « Table de pilote » (équivalent de « Table d'analyse croisée dynamique » sous Excel).

Pour les textes du corpus *Epistemon,* l'extraction automatique de l'ensemble des passages au discours direct étant impossible, nous avons procédé à un sondage des occurrences du verbe *dire*, le plus fréquent des verbes de parole. Selon nos estimations, sa fréquence est de 8 à 9 supérieure à celle des verbes *respondre* et *demander* dans le corpus *Epistemon*.

La requête CQL suivante a été utilisée pour obtenir une concordance des différentes formes du verbe *dire* :

```
[word="dise?n?t?|disiez|dis(i?ons|o[iy](e|s|t|ent)?|an[stz])|dis?mes|dis?tes|dir(o[iy])?(e|ent)?|di[cs]?t"%cd]
```

15 809 occurrences correspondent à cette requête dans le corpus *Epistemon* et 2 296 dans les six romans que nous avons sélectionnés pour l'analyse approfondie.

Dans de très rares cas, les occurrences correspondant à la requête n'ont pas de rapport avec le verbe *dire*, par exemple, le déterminant cardinal 'dix' graphié *dis*. Un peu plus souvent on trouve des infinitifs ou participes passés substantivés qui sont également à exclure. Néanmoins la grande majorité des occurrences sont bien des formes du verbe *dire*. Bien entendu, le verbe *dire* n'est pas toujours associé au discours direct ou indirect, et il peut se retrouver aussi bien dans l'annonce que dans l'incise, une annotation précise est donc nécessaire afin de caractériser la ponctuation des épisodes d'OR auxquels les occurrences du verbe *dire* sont associées. L'annotation exhaustive aurait été trop longue à réaliser et n'aurait sans doute pas été justifiée dans cette étude où nous cherchons à obtenir des indices quantitatifs simples. Nous avons donc choisi de nous limiter à l'annotation des 100 premières occurrences du verbe *dire* dans chacun des cinq romans sélectionnés. L'annotation a été réalisée sur concordances avec le logiciel *LibreOffice Calc*.

Pour chaque occurrence associée à un épisode d'OR nous avons noté s'il s'agit d'un discours indirect ou direct. Pour le discours direct, nous avons identifié la

localisation du verbe *dire* (annonce, incise ou rappel) et analysé la ponctuation des différentes « frontières ponctuables » (Lavrentiev 2009 : 97) : début et fin d'un épisode d'OR (monologue ou dialogue), changement de locuteur, début et fin de l'incise.

Un système de codes permettant de distinguer la forme des marques et leur « force » a été mis en place. La force de la ponctuation est déterminée, selon nous, par le caractère qui suit la ponctuation : la ponctuation est forte si elle est suivie d'une majuscule et faible si elle est suivie d'une minuscule. Dans les manuscrits médiévaux, ainsi que dans les imprimés du XVI^{e} siècle, il n'y avait pas de règle stricte d'association d'une marque à une force : un point pouvait être suivi d'une minuscule et un deux-points d'une majuscule. Une majuscule pouvait être utilisée seule et nous considérons ces emplois (hors noms propres) comme des ponctuations fortes.

En plus des marques de ponctuation, le retour à la ligne et les italiques (dans les imprimés) pouvaient participer au marquage graphique du discours direct. L'ensemble des codes utilisés dans l'annotation des ponctuations est présenté dans le tableau ci-dessous.

Code de ponctuation		Interprétation
forte	faible	
P	P	un point
V	V	une virgule
D	D	un deux-points
U	U	un point-virgule
B	B	une barre oblique
T	T	un tiret
I	I	un point d'interrogation
E	E	un point d'exclamation
G	G	une parenthèse gauche
R	R	une parenthèse droite
M		une majuscule seule
	0	aucune ponctuation
+it		début des italiques
–it		fin des italiques
+Q	+q	des guillemets
	+l	un retour à la ligne
Non		pas de frontière correspondante

Tableau 3. Codes d'annotation des marques graphiques du discours direct dans le corpus *Epistemon*.

Les italiques, les guillemets et parfois les parenthèses peuvent être utilisés en combinaison avec d'autres marques de ponctuation, un double code est alors attribué. Par exemple « P+l » signifie « un point suivi d'un passage à la ligne et d'une majuscule ». Le code « non » est utilisé lorsque l'épisode d'OR annoté ne contient pas la frontière

correspondante. Par exemple, il n'y a pas d'incise ou de changement de locuteur dans cet épisode.

Ces annotations nous ont permis de réaliser les décomptes que nous présenterons et analyserons plus bas.

4. Analyse des résultats

4.1. Fréquence relative du verbe *dire* et ses fonctions dans le marquage de l'oral représenté

Comme nous l'avons déjà indiqué, le verbe de parole le plus fréquent est incontestablement *dire*. Dans certains textes médiévaux, il est concurrencé par le verbe *faire* dans le rôle de l'incise, mais ce dernier n'est pas un verbe de parole et sa fréquence ne peut pas être utilisée comme indice de l'OR. Nous nous intéresserons donc au verbe *dire* pour évaluer le « poids » de l'OR dans les corpus analysés dans leur ensemble et dans certains genres de textes en particulier.

Puisqu'il s'agit d'une recherche qui porte sur un mot, nous pouvons utiliser l'ensemble de la BFM2019 et non seulement les quatre éditions de notre corpus restreint. Nous avons utilisé la même requête CQL que sur le corpus *Epistemon* (*cf.* plus haut) et obtenu 30 755 occurrences, soit près de 58 occurrences pour 10 000 mots et ponctuations. Nous utiliserons cette mesure (fréquence relative pour 10 000 occurrences) dans la suite de l'article.

Dans le corpus *Epistemon*, cette fréquence est considérablement plus basse, soit près de 40. Cela ne signifie pourtant pas la régression générale de la part d'OR (ou en tout cas du verbe *dire*) au XVI^e^ siècle. Si on limite le corpus aux genres littéraires narratifs (romans, nouvelles et récits brefs dans la BFM et romans dans *Epistemon*), on obtient pratiquement la même fréquence relative : 59 et 58 respectivement.

Toutefois, cette fréquence peut varier considérablement d'un texte à l'autre. Dans les cinq romans que nous avons sélectionnés pour annoter la ponctuation du discours direct, elle ne s'élève qu'à 51.

Pour aller plus loin, on peut se demander s'il est possible d'évaluer la part du discours direct et du discours indirect parmi les épisodes d'OR associés au verbe *dire* ? Même si une vérification attentive et l'annotation sont nécessaires pour disposer de données exactes, certains indices relativement fiables peuvent être obtenus grâce à des requêtes.

Ainsi, le discours indirect est le plus souvent introduit par la conjonction *que* ou par l'adverbe « subordonnant » *comment* qui suit directement le verbe *dire*. Dans les imprimés du XVI^e^ siècle la conjonction *que* est par ailleurs souvent précédée d'une virgule, tandis que dans la BFM, où la ponctuation est normalisée, il n'y a pas de ponctuation dans ce contexte. Parfois le discours indirect peut être séparé du verbe *dire* par un complément indiquant l'interlocuteur ou par pronom sujet, mais ces cas sont moins fréquents et plus difficiles à capter par une requête.

Nous avons donc cherché dans la BFM les occurrences de *dire* suivies des différentes formes graphiques de *que* et de *comment*. Dans le corpus *Epistemon*, la requête prévoit une virgule facultative entre ces deux éléments.

Le résultat est de 4 467 occurrences pour la BFM et 2 437 occurrences pour *Epistemon*, soit 14,5% et 15,4% des occurrences de *dire* respectivement. Même s'il s'agit de données brutes, on peut constater que la proportion de ces usages ressemblant au discours indirect reste stable. Dans l'échantillon annoté, la part du discours indirect s'élève à 12% en moyenne, mais des disparités assez importantes apparaissent entre les textes (de 8-9% dans *Alector* et les *Nouvelles recreations* à 19% dans *Pantagruel*).

En ce qui concerne le discours direct, son repérage est, comme nous l'avons déjà indiqué, difficile dans le corpus *Epistemon* en l'absence de ponctuation normalisée. On peut néanmoins estimer la quantité des incises, puisque dans la grande majorité des cas l'incise est composée de deux mots (le verbe *dire* ou *faire* à la 3^{e} personne du singulier ou du pluriel et le pronom, nom propre ou substantif sujet) entourées de marques de ponctuation. Il s'agit de parenthèses, de virgules ou de barres obliques dans le corpus *Epistemon* et de virgules dans la ponctuation normalisée de la BFM. Il convient de noter que dans les manuscrits médiévaux les incises n'étaient presque jamais ponctuées (Lavrentiev 2009 : 443), ce n'est donc que grâce à la ponctuation normalisée qu'il est possible de les repérer automatiquement.

Dans la BFM, le verbe *faire* est plus fréquent que *dire* dans les incises (2 512 occurrences contre 1 157), il est donc indispensable d'en tenir compte. Dans *Epistemon*, l'usage de *faire* dans les incises est beaucoup plus rare (40 occurrences contre 1 347 *dire*). Ce changement de verbe dominant dans les incises rend la comparaison des deux corpus délicate. Si on divise le nombre d'incises par la somme de toutes les occurrences de *dire* et des occurrences de *faire* dans les incises, on obtient près de 11% pour la BFM et un peu moins de 9% pour *Epistemon*. Les chiffres sont assez proches, compte tenu de l'approximation de la mesure. On peut également comparer la fréquence relative par rapport à toutes les occurrences dans la BFM et dans les romans d'*Epistemon*. Dans ce cas, les incises sont légèrement plus fréquentes dans *Epistemon* (9,4 occurrences pour 10 000 contre 6,9 dans la BFM), mais là aussi la différence ne semble pas significative.

On peut donc conclure que la fréquence des incises dans le discours direct n'évolue pas considérablement au sein des deux corpus dans leur ensemble.

4.2. Force de la ponctuation aux frontières des prises de parole : début et fin d'épisode, changement de locuteur

Cette mesure simple peut s'appliquer uniquement au corpus restreint annoté, soit les quatre éditions de la BFM et les extraits annotés du corpus *Epistemon*. Même si le nombre d'occurrences est très inégal dans les différents textes, on peut comparer le pourcentage des ponctuations faibles et fortes ou de l'absence de ponctuation dans

chaque position. Les figures 1, 2 et 3 présentent les données sous la forme d'histogrammes empilés 100%. Le nombre d'occurrences est également indiqué. Les textes sont classés dans l'ordre chronologique de leur composition. Pour l'*Image du monde*, texte où il n'y a que 7 prises de parole au total, les données ne sont pas significatives, mais elles semblent néanmoins révéler les mêmes tendances que les autres textes.

Le cas du changement de locuteur dans un dialogue, lorsqu'une nouvelle réplique suit directement la précédente, sans annonce ou incise interposée, est relativement rare. Nous n'en avons repéré aucune occurrence dans la *Vie de saint Alexis* et dans l'*Image du monde*. Par conséquent, ces textes n'apparaissent pas dans la Figure 3.

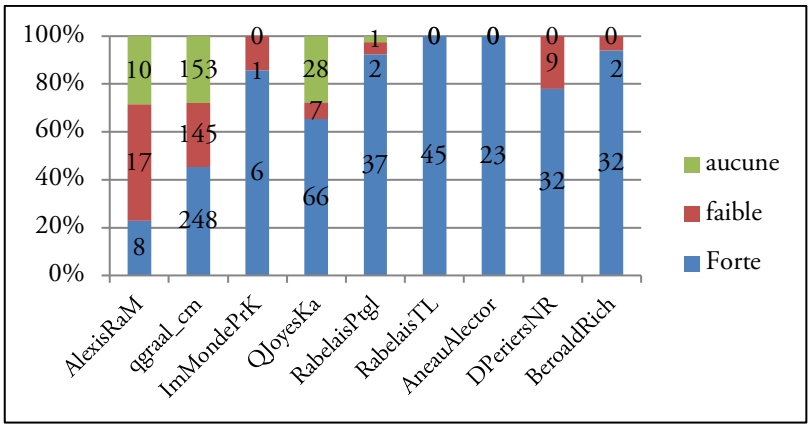

Figure 1. Ponctuation forte, faible ou absente au début des prises de parole.

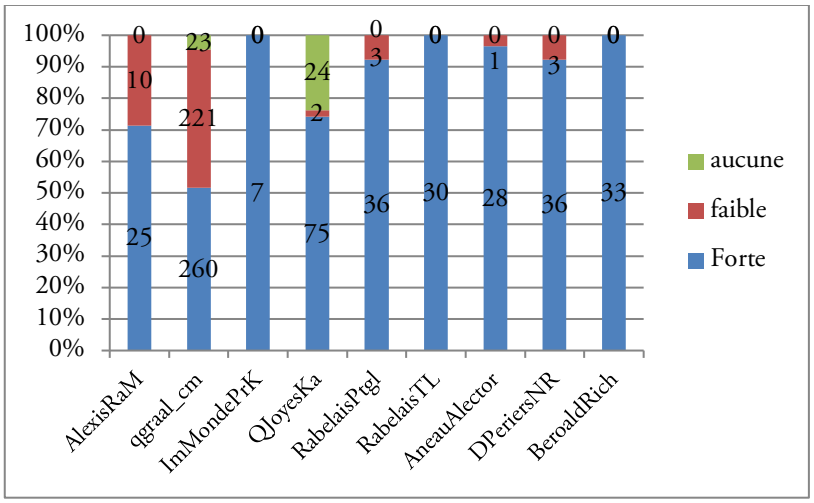

Figure 2. Ponctuation forte, faible ou absente à la fin des prises de parole.

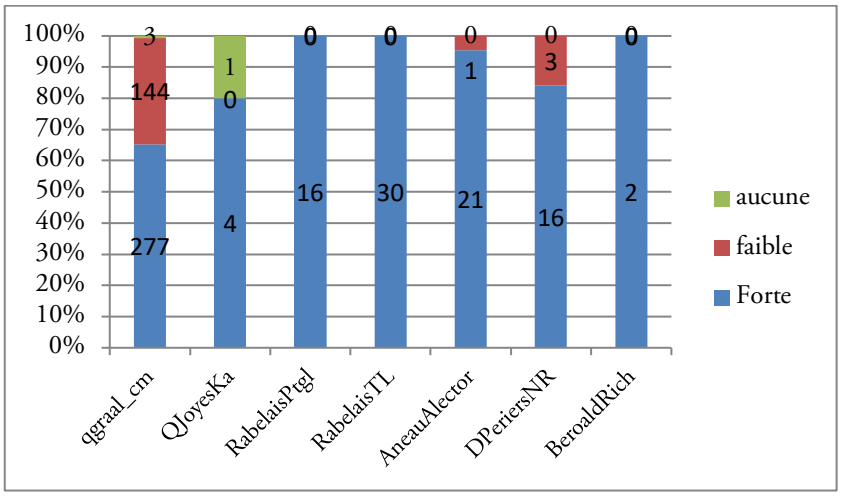

Figure 3. Ponctuation forte, faible ou absente en cas de changement de locuteur dans un dialogue.

Dans les trois positions, on voit une tendance à la dominance de la ponctuation forte qui se renforce au XVI^e^ siècle. L'absence de ponctuation est rare et devient exceptionnelle dans les textes du corpus *Epistemon*. La ponctuation faible et absente est constatée un peu plus souvent au début des prises de parole, en général entre l'annonce et le discours direct. Dans la *Vie de saint Alexis* cette tendance est particulièrement visible, puisque la part de la ponctuation faible passe de 47% au début à 28% à la fin des prises de parole. L'absence de ponctuation est constatée dans plus de 20% des cas au début des prises de parole dans la *Vie de saint Alexis* et dans la *Queste del saint Graal*. Elle est très rare en position finale : aucune occurrence dans le premier texte (mais le nombre total des occurrences n'est pas très élevé dans ce texte) et moins de 5% dans le dernier.

Le nombre de changements de locuteur est trop peu élevé pour pouvoir en tirer des conclusions fiables, mais l'usage de la ponctuation dans cette position ressemble plutôt à celui de fin de discours direct avec une dominance plus marquée de la ponctuation forte.

On peut noter une part relativement importante de l'absence de ponctuation dans les *Quinze joies de mariage* (près de 25% dans les deux positions). Ce manuscrit se caractérise en effet par un faible taux global de ponctuation et par un usage relativement fréquent de la majuscule seule que nous considérons comme une ponctuation forte. Le fait que les prises de paroles soient ponctuées d'une manière ou d'une autre dans ce texte montre que c'est l'un des types de frontières ponctuables qui « attire » le plus la ponctuation.

4.3. Les différentes marques

Si l'on s'intéresse plus en détail aux différents signes de ponctuation qui participent au marquage du discours direct, le contraste entre la période médiévale et le XVI^e^ siècle est évident.

Au Moyen Âge, la situation typique est l'existence d'une marque de ponctuation polyvalente (un point ou, dans certains manuscrits du XV^e^ siècle, une barre oblique) utilisée aussi bien avec une majuscule qu'une minuscule dans le mot qui suit. Dans certains manuscrits, on trouve une ou plusieurs marques complémentaires, dont l'usage est beaucoup plus rare et jamais obligatoire pour une fonction. L'étude de ces marques « périphériques » demanderait un corpus beaucoup plus important que le nôtre.

Dans la *Vie de saint Alexis* la seule marque de ponctuation utilisée est le point.

Dans la *Queste del saint Graal*, près de 97% des occurrences de marques de ponctuation sont des points. Cependant, on y trouve aussi des *punctus elevatus*, ou *comma* (204 occurrences, soit 2,5%), et *punctus interrogativus* (73 occurrences, soit 0,9%).

Ce dernier signe est clairement spécialisé dans le marquage des énoncés interrogatifs et se trouve souvent lorsque le locuteur change. Toutefois, plus de la moitié des énoncés interrogatifs (93 sur 169) se terminent par un simple point et dans 3 occurrences on trouve un *comma*.

Christiane Marchello-Nizia (2007) a d'ailleurs remarqué que près de 75% des occurrences du *comma* se situent dans le discours direct et a émis une hypothèse selon laquelle ce signe de ponctuation indique la présence d'une intonation montante. Rien ne contredit cette hypothèse, mais il est impossible d'affirmer avec certitude quel était le contour intonatif du discours au Moyen-Âge.

Il est certain que le *comma* est fortement associé à certaines interjections, en particulier *ha*, qui se trouvent au début de prise de parole et servent en quelque sorte de « ponctuation lexicale » du discours direct. Il est à noter qu'aucune occurrence du *comma* n'est située devant le premier mot du discours direct (en le séparant de l'annonce éventuelle), ni à la fin d'épisode d'OR. Trois occurrences se trouvent lors du changement de locuteur, après un énoncé interrogatif. Le *comma* joue donc ici le rôle de point d'interrogation. Dans certains manuscrits, d'ailleurs, les deux marques sont difficiles à distinguer l'une de l'autre.

Dans l'*Image du monde*, où les épisodes d'OR sont extrêmement rares, la fréquence relative du *comma* est pourtant plus élevée, soit près de 5,5% de toutes les marques de ponctuation. On trouve seulement deux occurrences du *punctus interrogativus* dans ce texte, et elles ne concernent pas des énoncés interrogatifs. Toutes les « prises de parole » sont précédées et suivies d'un point, à l'exception d'une occurrence où l'annonce est séparée du discours direct par un *comma* :

> (4) Lors **dist** || une m*ou*lt oscure parole **.'** || **ou** li diex **dist il** de na||ture sueffre grant tor||ment *et* gr*a*nt tort . Ou toz || li mondes se descorde . *et* se desjoint pour defail||lir . Comme cil q*u*i veult defenir . (Gossouin 2018 : 103d).

Dans les *Quinze joies de mariage* le point est également la marque de ponctuation dominante (1118 occurrences sur 1143, soit 97%), mais il est concurrencé par

l'usage de la majuscule seule (574 occurrences). Si on tient compte de ces « ponctuations sans marque », la part du point descend à 65%. La majuscule seule correspond à 33% des occurrences. Les 2% restants sont des occurrences de la barre oblique (25 au total) et une occurrence d'une marque qui ressemble à un deux-points que nous commenterons plus bas. La proportion des différentes marques est strictement la même aux frontières du discours direct (66% des points et 33% des majuscules seules), il ne semble donc y avoir aucun indice de spécialisation d'une marque de ponctuation dans la signalisation du discours direct.

Deux marques de ponctuation exceptionnelles pour le manuscrit des *Quinze joies de mariage* se trouvent à l'avant-dernière ligne du feuillet 105 recto. Il s'agit d'une sorte de deux-points et d'une barre oblique avec un point au-dessous (comme un point d'exclamation moderne incliné). Nous reproduisons l'image de cette ligne dans la figure 4 et nous donnons la transcription diplomatique du passage élargi dans l'exemple 5.

Figure 4. Marques de ponctuation exceptionnelles dans les *Quinze joies de mariage* (ms. Rouen, BM, 1052, f° 105r, l. 25).

(5) Lors le bon homme laccolle et trouue que elle est bien || chaude et il dit **:** **Voir !** Mes cest daultre mala||<105v>die quel ne dit et quil ne cuide quar el a par a||uenture songie que elle estoit auecques son amy et pour ce sue bien fort . (*Quinze joies de mariage* 2018 : 105r).
'Alors le bon homme l'accole et trouve qu'elle est bien chaude, et dit : « C'est vrai ! » Mais c'est [à cause] d'une autre maladie, dont elle ne parle pas et auquel il ne pense pas, car par aventure elle a songé qu'elle était avec son ami, et pour cette raison elle sue bien fortement'.

Ces deux marques sont placées autour d'une courte prise de parole : « Voir ! » ('C'est vrai !'). L'encre semble bien être celle du scribe, mais il est difficile de l'affirmer sur des signes aussi petits[4]. On ne peut pas non plus être sûr que le tracé de ces marques soit intentionnel. Dans le premier cas, il peut s'agir d'un point et dans le deuxième, d'une barre oblique. Quoi qu'il en soit, ces marques participent à la mise en évidence d'un passage au discours direct.

[4] Le manuscrit présente de nombreuses annotations marginales et corrections réalisées au XIX^e^ siècle. Selon C. Mira (2009, *Introduction* à l'édition des *Quinze joies de mariage*), la date « 22 déc. 1886 » est inscrite de la même main au verso de la page de garde. A. Coville (1935 : 132) date ces annotations de la première moitié du XIX^e^ siècle et les attribue, sans l'affirmer avec certitude, à André Pottier, bibliothécaire de la ville de Rouen. Il n'est pas exclu que certaines ponctuations aient été ajoutées par l'annotateur tardif.

Dans les cinq romans annotés du corpus *Epistemon*, la diversité des marques de ponctuation participant à la signalisation du discours direct est beaucoup plus importante, et la situation change selon le texte et la position analysée.

Afin d'illustrer cette diversité, nous présentons dans la figure 5 un histogramme empilé 100% des marques de ponctuation en début de discours direct.

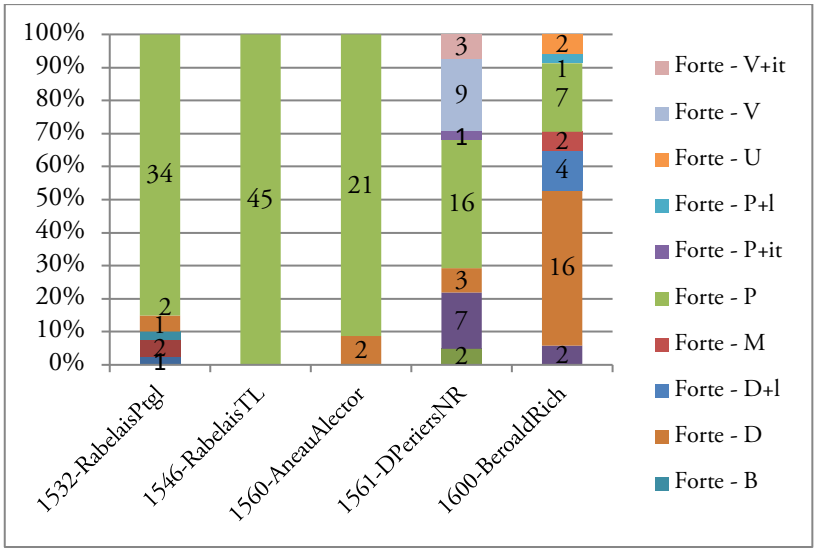

Figure 5. Répartition des marques de ponctuation forte et faible au début du discours direct de cinq romans du corpus *Epistemon* (voir le Tableau 3 pour la signification des codes).

Dans *Pantagruel* de Rabelais (1532) la marque dominante est le point (34 occurrences sur 40). Dans deux cas il s'agit d'un deux-points suivi d'une majuscule. La barre oblique dans cette position est suivie d'une majuscule une fois (mais il s'agit d'un nom propre) et deux fois d'une minuscule. Enfin, nous avons relevé une occurrence d'absence de ponctuation :

> (6) & apres commenca a dire || bon bon bon : car il ne scavoit encores pas bien parler / (Rabelais 1532 : [16]).

Notons qu'il s'agit d'une prise de parole très courte, située au début d'une ligne. Dans l'édition de 1542, il y a bien une virgule dans cette position.

Dans le *Tiers livre* (1546), la ponctuation du début du discours direct est très homogène : il s'agit d'un point suivi d'une majuscule dans 100% des cas (45 occurrences).

Dans *Alector* de Barthélemy Aneau (1560), la situation est très similaire : le point domine largement, mais on trouve tout de même deux occurrences de deux-points suivies d'une majuscule.

Dans les *Nouvelles recreations et joyeux devis* de Bonaventure Des Périers (1561) le changement de situation est radical : on trouve une grande diversité de marques : point, virgule et deux points, avec ou sans italiques, suivies ou non d'une majuscule. La ponctuation forte est dominante (près de 80% des cas), mais la part des

ponctuations faibles est la plus importante de notre corpus annoté. Le point est la marque la plus fréquente, mais il n'est plus majoritaire.

Dans *Le Tableau des riches inventions*, on constate la même diversité de marques, mais leur répartition est différente. Le deux-points suivi d'une majuscule devient la marque la plus fréquente ; on voit donc se dessiner l'usage qui est devenu normatif dans la typographie moderne (sans toutefois les guillemets). Le point reste présent, mais il n'est plus dominant dans cette position. Dans 5 cas, le discours direct commence sur une nouvelle ligne avec alinéa.

Pour la fin du discours direct la situation est assez similaire, mais la dominance du point est plus forte, y compris dans le *Tableau des riches inventions*. Toutefois, dans tous les textes sauf le *Tiers livre*, on trouve parfois d'autres marques : la barre oblique, le deux-points ou la virgule.

Le point est aussi la marque la plus fréquente en cas de changement de locuteur sauf dans le *Tableau des riches inventions*, où nous avons repéré seulement deux occurrences de ce type. Dans les deux cas, il s'agit du passage d'une question à la réponse marquée par un point d'interrogation. On retrouve d'ailleurs des points d'interrogation dans tous les autres textes. C'est toujours dans *Les nouvelles recréations* qu'on trouve la plus grande diversité de marques : le deux-points et la virgule s'ajoutent ici au point et au point d'interrogation, mais il y a trop peu d'occurrences de cette frontière ponctuable pour pouvoir tirer des conclusions à partir des fréquences observées.

La diversité de marquage est encore plus grande pour les incises, car pour chaque incise il faut analyser la marque qui la précède et celle qui la suit. La répartition des paires de marques de début et de fin d'incise est présentée dans la figure 6.

C'est dans *Pantagruel* que la situation est la plus confuse : il y a 11 combinaisons attestées et aucune ne présente plus de 5 occurrences dans le corpus annoté. Il convient toutefois de noter que la rupture avec le Moyen Âge est nette, car l'absence de ponctuation est rare. Dans le *Tableau des riches inventions*, le nombre total des incises est très faible, mais on peut toutefois constater l'absence de ponctuation dans 4 occurrences sur 5. C'est donc en quelque sorte un retour à la pratique médiévale que nous observons dans ce texte dont la date d'impression est la plus récente de notre corpus. Cela vient sans doute du fait que Béroalde réédite une traduction antérieure, celle de Jean Martin (1546), ce qui a pu influencer cet aspect de la ponctuation.

Dans les trois autres textes, on voit se dégager deux tendances : l'usage des parenthèses et d'une paire de virgules. Les parenthèses sont systématiques dans *Alector* et dominantes dans le *Tiers livre* (70% des occurrences). Dans les *Nouvelles recréations* presque toutes les occurrences présentent des virgules. Le plus souvent, il s'agit d'une paire de virgules, parfois, dans l'incise finale, d'une virgule et d'un point. Dans deux occurrences, la virgule manque au début ou à la fin de l'incise. On peut se demander s'il ne s'agit pas dans ce cas de coquilles typographiques. Dans deux occurrences, l'incise finale suit un énoncé interrogatif parqué par un point d'interrogation. Le marquage de

l'incise par une paire de virgules est également attesté dans le *Tiers livre*, mais cet usage reste minoritaire (9 occurrences, soit près de 19%).

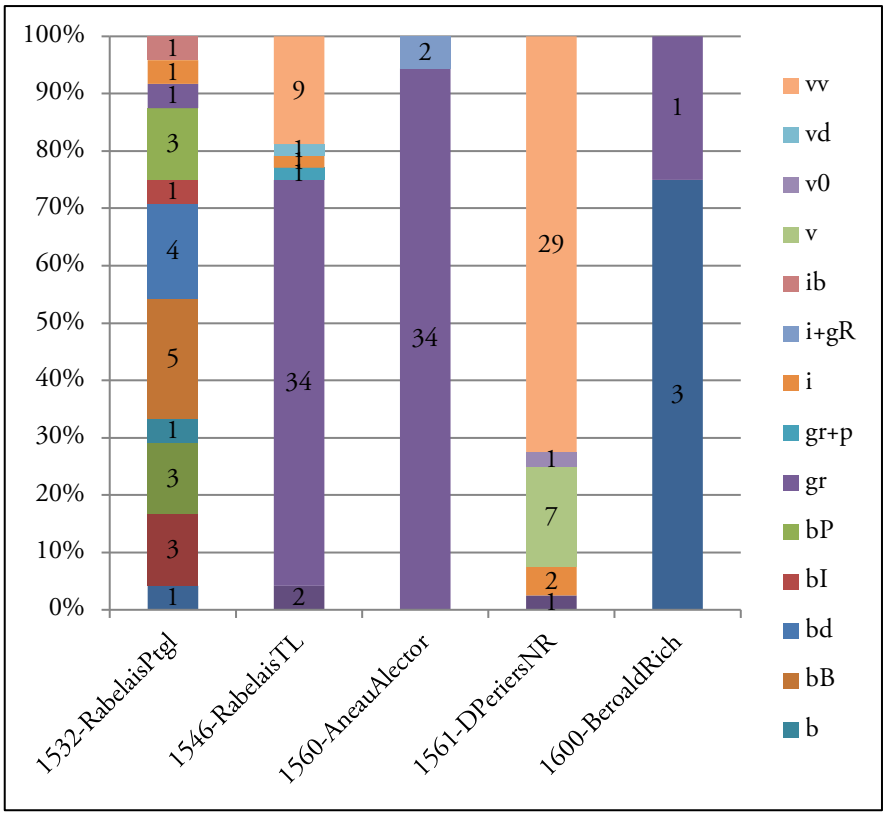

Figure 6. Marques de ponctuation de l'incise dans les cinq romans du corpus *Epistemon*.

Même si les observations que nous venons de présenter sont basées sur un corpus restreint, elles permettent de rendre compte d'une part de la diversité des pratiques et des grandes tendances qui se dessinent.

5. Marquage des citations dans le corpus *Epistemon*

L'analyse des pratiques de marquage de l'oral représenté serait incomplète sans la prise en compte de citations de sources écrites. Ces citations sont parfois présentées de la même façon que l'oral représenté avec des verbes de parole dans l'annonce ou dans l'incise. En même temps, des techniques spéciales sont développées notamment pour indiquer les sources citées. C'est d'ailleurs dans les citations que se développe, au XVI^e^ siècle, l'usage des marques graphiques spécialisées, à savoir les italiques et les guillemets.

L'origine des guillemets est liée aux *notæ* marginales utilisées pour attirer l'attention sur un passage important dans des textes religieux du Moyen-Âge (Parkes 1992 : 58). Inventées en 1500 par Alde Manuce à Venise, les italiques ont été rapidement adoptées par les imprimeurs français : dès 1502 à Lyon et dès 1512 à Paris

(Vervliet 2008 : 287). La concurrence entre les guillemets et les italiques peut s'expliquer en partie par des raisons matérielles : l'usage des guillemets (une paire de virgules imprimées en marge) pouvait servir à économiser les caractères italiques.

Les deux marques typographiques sont encodées dans le corpus *Epistemon* ; il est donc possible de les rechercher automatiquement dans l'ensemble du corpus et non seulement dans les extraits annotés des six romans (qui ne présentent d'ailleurs aucune occurrence de ce phénomène). Cette tâche se heurte toutefois à une double difficulté : la variabilité des pratiques propres à l'époque et les choix d'encodage parfois divergents pour un même phénomène.

Par exemple, les guillemets figurant en marge du document source sont selon les textes encodés par la balise TEI <note>, parfois par <label> et parfois, enfin, il s'agit d'une double virgule au début de ligne signalé par la balise <lb/>. Les italiques sont signalées par un attribut « rend » associé à la balise <cit> ou <hi>. La valeur de cet attribut qui indique les italiques est parfois « it » et parfois « italic ». Les italiques peuvent d'ailleurs se combiner avec l'alignement centré, ce qui augmente encore la variabilité de l'encodage possible.

L'ensemble de la citation et de l'indication éventuelle de sa source est en général balisé grâce à l'élément <quote>. Afin de faciliter l'analyse du corpus, nous avons tenté d'harmoniser l'encodage de ces citations par un procédé automatique. Il n'est pas impossible que certaines occurrences aient échappé à cette normalisation ou tout simplement qu'elles n'aient pas été balisées dans les transcriptions, mais nous considérons que les cas repérés sont suffisamment représentatifs pour dégager les grandes tendances des usages observés.

Nous utilisons la fonctionnalité « Progression » de TXM pour construire un graphique cumulatif (Figure 7). Chaque type de marquage est représenté par une courbe qui monte à chaque occurrence. Les parties plates correspondent aux segments du corpus qui ne présentent aucune occurrence du motif recherché. Ces absences peuvent s'expliquer par l'absence de citations dans le texte ou par l'absence de balisage. Nous avons choisi de travailler sur un sous-corpus composé uniquement de textes qui présentent au moins une occurrence de <quote>, ce qui permet de limiter le biais provoqué par l'absence de balisage. Ce sous corpus est composé de 27 unités textuelles et contient 1 724 427 occurrences (soit 44% de l'ensemble du corpus *Epistemon*).

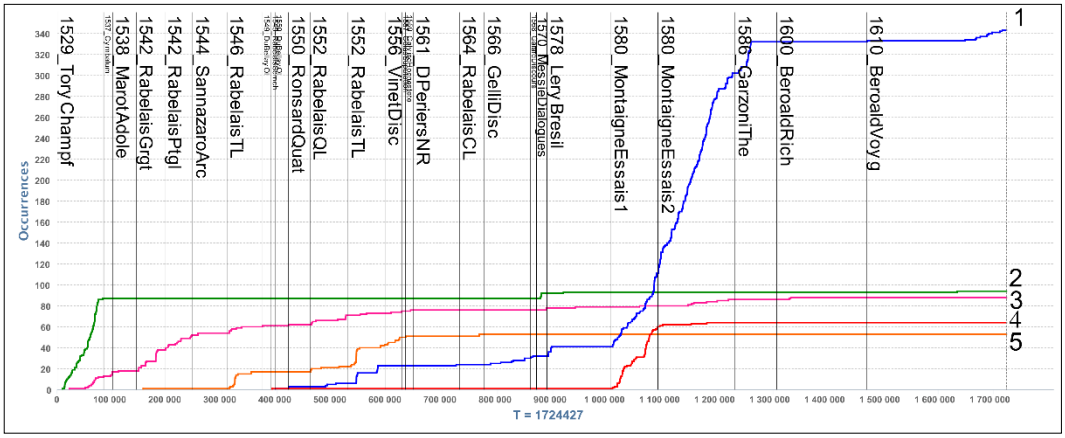

Figure 7. Graphique de progression des différentes techniques de marquage de citations dans les imprimés du corpus *Epistemon*.

Les lignes verticales dans le graphique correspondent aux frontières des unités textuelles. Nous avons également numéroté les courbes afin de faciliter la référence. La courbe 1 correspond à la requête <quote_rend="it.*"> [] (premier mot d'une citation en italiques). La progression des citations marquées par des guillemets en marge est représentée par la courbe 2 (requête <quote_rend="quote-margin"> []). La courbe 3 est générée par la requête <quote_rend="undef"> [] (premier mot de citation sans marquage particulier). La courbe 4 correspond à la requête <quote_rend="center_it.*" []> (citations marquées par des italiques centrés). Enfin, les citations en forme de vers dans un texte en prose (requête <quote_rend="verse"> []) sont représentées par la courbe 5.

Ce qui frappe dans le graphique c'est la forme abrupte des courbes. Des « plats » sans aucune occurrence alternent avec des zones de forte densité de citations.

La courbe 3 (citations non marquées) est à considérer avec beaucoup de précaution, car il s'agit souvent d'occurrences marquées par des guillemets ou des majuscules que la procédure de normalisation automatique n'a pas permis de détecter. On peut juste constater que la fréquence de ces cas non marqués ou non détectés baisse nettement à partir de l'*Arcadie* de Jacopo Sannazaro (1544).

Les italiques (courbe 1) et les italiques centrés (courbe 4) sont surtout utilisés dans les *Essais* de Montaigne (1580). Les deux types de marquage sont présents dans le premier livre, mais l'alignement centré semble disparaitre dans le livre 2. Un examen plus attentif des deux documents montre toutefois qu'il n'y a pas de vraie différence d'alignement de citation entre eux. Les citations sont normalement calées à gauche, avec parfois des retraits assez importants de la première ou de la dernière ligne. C'est ce retrait qui a parfois provoqué l'erreur d'interprétation et d'encodage réalisé par le prestataire des BVH.

Ce changement de mise en page est d'autant plus étonnant que les deux volumes ont été imprimés la même année chez le même imprimeur (Simon Millanges à Bordeaux).

Les guillemets en marge (courbe 2) sont utilisés régulièrement dans le *Champ fleury* de Geoffroy Tory (1532) et dans les *Trois dialogues* de Pero Mexia (1570). On trouve également quelques occurrences isolées dans l'*Histoire d'un voyage fait en la terre du Bresil* de Jean de Lery (1578) et dans l'*Histoire veritable* de François Béroalde de Verville (1610).

Enfin, des citations en vers au milieu de passages en prose sont reconnaissables par leur mise en page (saut de ligne à chaque fin de vers, retrait négatif en cas de rejet). C'est donc une forme de marquage typographique particulière. La quasi-totalité des occurrences de ce type de citation se trouve dans le *Tiers livre* et dans le *Quart livre* de Rabelais, ainsi que dans le *Discours non plus melancoliques* d'Élise Vinet (1556). Dans l'édition de 1552 du *Tiers livre* ce mode de marquage des citations alterne ou se combine avec les italiques (en particulier lorsque les citations sont en latin). L'édition de 1546 est en revanche entièrement composée en italiques, ce qui ne permet pas d'utiliser ces caractères pour mettre en relief les citations.

Les observations que nous venons de présenter témoignent à la fois de la difficulté d'extraction et d'analyse de données hétérogènes et du potentiel de l'outil numérique dans l'étude de l'évolution des pratiques typographiques.

6. Conclusion

Le corpus que nous avons pu étudier ne peut pas être considéré comme représentatif ni équilibré. C'est donc avec une extrême prudence qu'on peut tenter de tirer des conclusions générales à partir des résultats observés. Nous pouvons néanmoins préciser certaines hypothèses concernant les tendances de l'évolution générale et mettre en place un ensemble d'outils d'analyse qui pourront être appliqués au fur et à mesure de l'augmentation de la quantité et la qualité des données disponibles.

L'encodage « multi-facette » utilisé dans les éditions de la BFM est particulièrement intéressant de ce point de vue, car il permet d'utiliser la ponctuation moderne pour repérer les passages au discours direct et d'analyser la ponctuation originale des sources sans recourir à une annotation supplémentaire. Dans le corpus *Epistemon*, où le balisage est avant tout patrimonial et où les pratiques d'encodage varient d'un texte à l'autre, le repérage automatique est plus complexe, mais possible pour certains phénomènes précis (comme les occurrences des verbes de parole ou les citations marquées d'une manière ou d'une autre).

La fréquence relative du verbe *dire* reste assez stable tout au long de la diachronie observée, en tout cas au sein des textes littéraires du genre romanesque. Le marquage des frontières du discours direct par des ponctuations fortes est également une grande tendance observée dans l'ensemble du corpus. La spécialisation des marques (notamment le deux-points entre l'annonce et le discours direct et les parenthèses ou une paire de virgules autour des incises) se profile dans certains textes du corpus

Epistemon. Enfin, les marques spéciales, les italiques et les guillemets, se développent dans un premier temps dans les citations de sources écrites.

RÉFÉRENCES BIBLIOGRAPHIQUES

a) Bases de données et corpus

ANEAU, Barthélemy (1560) : *Alector*. Lyon, Pierre Fradin, *in Epistemon*.

Base textuelle du moyen français (2019). URL : http://zeus.atilf.fr/dmf/

BFM = *Base de français médiéval*, ENS de Lyon, URL : http://txm.bfm-corpus.org

BFM2019 = *Corpus BFM2019*, *in* BFM, ENS de Lyon, URL : http://bfm.ens-lyon.fr/spip.php?rubrique117

BVH = *Les Bibliothèques Virtuelles Humanistes,* Université de Tours, URL : http://www.bvh.univ-tours.fr)

Charrette (2007) = *The Princeton Charrette Project,* URL : http://www.princeton.edu/~lancelot/ss/

COLONNA, Francesco (1600) : *Le Tableau des riches inventions… représentées dans Le Songe de Poliphile*, éd. François Béroalde de Verville [d'après la traduction de Jean Martin]. Paris, Matthieu Guillemot, *in Epistemon*.

DES PERIERS, Bonaventure (1561) : *Nouvelles recreations et joyeux devis*. Lyon, Guillaume Rouillé, *in Epistemon*.

Epistemon = *Corpus de textes de la Renaissance*, *in* BVH, Université de Tours http://www.bvh.univ-tours.fr/Epistemon/index.asp

GOSSOUIN DE METZ (2018, en ligne) : *Image du monde*. Édition numérique interactive par Naomi Kanaoka. Lyon, ENS de Lyon, http://catalog.bfm-corpus.org/ImMondePrK.

MUFI = *Medieval Unicode Font Initiative*, https://folk.uib.no/hnooh/mufi.

Queste del saint Graal (2019, en ligne) : Édition numérique interactive par Christiane Marchello-Nizia et Alexei Lavrentiev. Lyon, ENS de Lyon, http://catalog.bfm-corpus.org/qgraal_cm.

Quinze joies de mariage (2018, en ligne) : Édition numérique interactive par Naomi Kanaoka sur la base de l'édition imprimée établie par Michèle Guéret-Laferté, Sylvain Louis et Camille Mora (Rouen, Presses universitaires de Rouen et du Havre, 2009). Lyon, ENS de Lyon, http://catalog.bfm-corpus.org/QJoyesKa.

RABELAIS, François (1532) : *Pantagruel*. Lyon, Claude Nourry, *in Epistemon*.

RABELAIS, François (1546) : *Le tiers livre*. Paris, Chrétien Wechel, *in Epistemon*.

Vie de saint Alexis (2018, en ligne) : Édition numérique interactive par Thomas Rainsford et Christiane Marchello-Nizia. Lyon, ENS de Lyon, http://catalog.bfm-corpus.org/AlexisRaM.

b) Études

COVILLE, Alfred (1935) : *Recherches sur quelques écrivains du XIV^e^ et du XV^e^ siècle*. Paris, Librairie E. Droz.

DUFOURNAUD, Nicole *et al.* (2017) : *Manuel d'encodage XML-TEI Renaissance et temps modernes (Imprimés – Manuscrits)*. Version 4. Tours, *Bibliothèques virtuelles humanistes*, http://www.bvh.univ-tours.fr/XML-TEI/index.asp

GUILLOT, Céline ; Sophie PREVOST & Alexei LAVRENTIEV (2014) : « Oral représenté et diachronie : étude des incises en français médiéval », *Actes du CMLF 2014 – 4e Congrès mondial de linguistique française*. Berlin, EDP Sciences, 259-276. URL : http://dx.doi.org/10.1051/shsconf/20140801284

GUILLOT-BARBANCE, Céline, Bénédicte PINCEMIN & Alexei LAVRENTIEV (2017), « Représentation de l'oral en français médiéval et genres textuels ». *Langages*, 208: 4, 53-68.

GUILLOT-BARBANCE, Céline, Alexei LAVRENTIEV, Serge HEIDEN, & Bénédicte PINCEMIN (2018) : « Diachronie de l'oral représenté: délimitation et segmentation interne du dialogue (IX^e^-XV^e^ siècle). », *in* Wendy Ayres-Benett, Anne Carlier; Julie Glikman, Thomas Rainsford, Gilles Siouffi & Carine Skupien Dekens (éds), *Nouvelles voies d'accès au changement linguistique. Actes du colloque de la SIDF*. Paris, Classiques Garnier, 279-296.

HEIDEN, Serge ; Jean-Philippe MAGUE & Bénédicte PINCEMIN (2010) : « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *in* Sergio Bolasco, Isabella Chiari & Luca Giuliano (éds), *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*, Rome, Italie, Edizioni Universitarie di Lettere Economia Diritto, 1021-1032.

HEIDEN, Serge ; Bénédicte PINCEMIN & Matthieu DECORDE (2015, en ligne) : *Manuel de TXM*. Version 0.7. Lyon, ENS de Lyon. URL : https://halshs.archives-ouvertes.fr/halshs-01223949.

LAVRENTIEV, Alexei (2007) : « Base de français médiéval et transcriptions de manuscrits : recherche de complémentarité », *in* David Trotter (éd.), *Actes du XXIV^e^ Congrès International de Linguistique et de Philologie Romanes*. Tübingen, Max Niemeyer, 405-410.

LAVRENTIEV, Alexei (2009) : *Tendances de la ponctuation dans les manuscrits et incunables français en prose, du XIII^e^ au XV^e^ siècle*. Thèse de doctorat en sciences du langage sous la direction de Christiane Marchello-Nizia. Lyon, ENS-LSH. URL : http://tel.archives-ouvertes.fr/tel-00494914

LLAMAS-POMBO, Elena (2010) : « Marques graphiques du discours rapporté (Manuscrits du Roman de la Rose, XV^e^ siècle) », *in* Céline Guillot, Bernard Combettes, Alexei Lavrentiev, Evelyne Oppermann-Marsaux & Sophie Prévost (éds), *Le changement en français : Études de linguistique diachronique. Actes du colloque international DIACHRO-IV (22-24 octobre 2008, Madrid)*. Bern, Peter Lang, 249-270.

MARCHELLO-NIZIA, Christiane (2007) : « Le comma dans un manuscrit en prose du 13^e^ siècle : grammaticalisation d'un marqueur de corrélation, ou marquage d'intonation ? », *in* Olivier Bertrand, Sophie Prévost & Michel Charolles (éds), *Discours,*

diachronie, stylistique du français. Études en hommage à Bernard Combettes, Berne, Peter Lang, 293-305.

MARCHELLO-NIZIA, Christiane (2012) : « L'oral représenté : un accès construit à une face cachée des langues 'mortes' », *in* Céline Guillot, Bernard Combettes, Alexei Lavrentiev, Évelyne Oppermann-Marsaux & Sophie Prévost (éds), *Le changement en français. Etudes de linguistique diachronique*. Bern, Berlin et Bruxelles, Peter Lang, 247-264.

MARCHELLO-NIZIA, Christiane ; Alexei LAVRENTIEV & Céline GUILLOT-BARBANCE (2015) : « Édition électronique de la *Queste del saint Graal* », *in* David Trotter (dir.), *Manuel de la philologie de l'édition*, Berlin et Boston, Mouton & de Gruyter, 155-176.

MARNETTE, Sophie (2006) : « La signalisation du discours rapporté en français médiéval ». *Langue française*, 149, 31-47.

PARKES, Malcolm B. (1992) : *Pause and effect: an introduction to the history of punctuation in the West*. Aldershot, Scolar Press.

VERVLIET, Hendrik D. L. (2008) : *The palaeotypography of the French Renaissance. Selected papers on sixteenth-century typefaces*. Leiden, Brill, (Library of the written word).