Studii de lingvistică 15, 2025, 99 - 130
DOI : 10.61215/SL.2025.15.1.06

Commencer et finir – une approche quantitative des binômes verbaux en français préclassique

To Begin and to End: a quantitative approach to verbal binomials in Preclassic French

Quentin Feltgen[1]

Abstract: Preclassic French may be regarded as a culmination in the use of synonymic binomials, reaching their peak of frequency, showing signs of frozenness, and starting to be denounced as a clumsy rhetorical device that dwindles accordingly. In this paper, we adopt a quantitative approach on verbal binomials in Preclassic French based on data from the Frantext corpus, evidencing key structural properties of the system: the two sets of verbs that are combined are different enough to ensure a greater diversity of binomials, while the associations between these sets are constrained so as to ensure a high level of informativity between the two verbs, and therefore a better degree of coherence in the use of the binomial construction. We also offer a simple yet novel statistical method to select the most relevant and significantly recurrent verbal binomials in the corpus.

Keywords: Zipf's law, statistics, mutual information, verbal binomials, preclassic French

1. Introduction

Le procédé rhétorique du binôme, en particulier sous sa version synonymique, est intéressant à plus d'un titre du point de vue stylistique, il présente une riche histoire diachronique (Buridant 1980, Doualan 2014) ; du point de vue grammatical, en particulier en grammaire des dépendances, il soulève d'intéressantes difficultés d'analyse (Mazziotta 2012) ; du point de vue de la production du discours, la multiplicité de ses réalisations présentent un caractère formulaïque remarquable (Norrick 1988), même si leur variation laisse aussi transparaître une dimension innovative et individuelle (Papahagi 2021 : 178) ; du point de vue structural, il se rapproche

[1] Université de Gand & Université Catholique de Louvain; quentin.feltgen@ugent.be.

des systèmes constructionnels, dont l'organisation donne à voir les principes à l'œuvre dans l'organisation de la langue, comme cela a été esquissé, en anglais, pour les tournures de redoublement d'un nom temporel articulé par une préposition (Sommerer & Baumann 2021), en néerlandais, pour le système des minimiseurs (Van den Heede & Lauwers 2023), en espagnol, pour le paradigme des auxiliaires inchoatifs (Van Hulle *et al.* 2025).

C'est dans cette dernière optique que nous nous intéresserons à cet ensemble d'unités linguistiques, en nous limitant toutefois aux binômes verbaux, précisément pour le rôle que joue le verbe comme tête de l'énoncé en grammaire des dépendances ; nous nous focaliserons en outre sur la période du français préclassique, particulièrement intéressante dans l'histoire du binôme synonymique (et, à titre d'hypothèse de travail, du binôme de manière plus générale) en ce qu'elle voit s'opérer un tournant dans la perception qu'en ont les acteurs linguistiques de l'époque (Leclercq 2008, Petrequin 2009), qui tendent tantôt à le proscrire, tantôt à en revendiquer le pouvoir expressif (Siouffi 2012). Cette critique s'associe à un déclin et à un figement des binômes, même si, comme en atteste l'exemple suivant, le procédé peut encore être exploité avec générosité :

(1) Et le lien qui unit deux natures si distantes, est si intime à la divinité, qu'il a identité et est une mesme chose auec l'essence divine ; et est rendu si propre à nostre humanité, qu'il **entre et penetre** ; qu'il **actuë et viuifie** ; qu'il **sanctifie et deïfie** toute la nature humaine, le corps, l'ame, et toutes les parties et puissances de ce petit monde, ou plustost de ce grand monde qui est Iesvs, et ce jusqu'au fond, au centre et en l'intime de son essence humaine (Pierre de Bérulle, *Discours de l'estat et des grandeurs de Jésus par l'union ineffable de la divinité avec l'humanité*, 1623)

Dans cet article, nous proposons une vision radicalement empirique et quantitative du binôme verbal (qu'il soit ou non de nature synonymique), pour comprendre d'une part les principes structurels sous-jacents à l'utilisation de cette tournure rhétorique, et, d'autre part, pour déterminer lesquels de ces binômes peuvent être considérés comme significativement récurrents d'un point de vue purement statistique. Il s'agit là d'une question méthodologique spécialement épineuse, notamment abordée par Siouffi (2012) dans le cas de *sans trouble ni nuage*. Après avoir présenté notre requête d'extraction d'occurrences à partir du corpus Frantext, les données obtenues, et la fréquence associée, nous procéderons à l'étude du système, que nous conceptualisons comme la mise en relation de deux paradigmes de variation distincts ou deux *jeux* de verbes, V1 et V2, dont la double réalisation détermine chaque occurrence individuelle

de binôme verbal. Nous caractériserons chacun de ces jeux de manière indépendante d'abord, pour caractériser leur distribution de fréquence associée (c'est-à-dire la manière dont les occurrences de chacun de ces jeux se répartissent entre leurs différents membres ou *types*) ainsi que leur productivité. Nous étudierons ensuite la manière dont ces deux jeux interagissent en pratique, en spécifiant notamment la différence de hiérarchie et de contenu entre les deux et leur degré d'information mutuelle. Nous présenterons ensuite notre méthode statistique pour sélectionner les binômes verbaux les plus pertinents en français préclassique, au-delà de la combinaison libre entre les deux jeux de verbes mis en relation par le binôme (c'est-à-dire les binômes dont la fréquence ne peut être expliquée par deux sélections successives indépendantes au niveau de chaque paradigme). Nous dresserons alors une typologie rapide des binômes rencontrés, afin de donner un aperçu de la diversité des emplois du binôme verbal. Enfin, nous évaluerons l'impact du genre textuel sur le recours au binôme verbal, qu'il soit formulaïque ou non.

2. Extraction d'occurrences du corpus Frantext

Pour mener à bien notre étude quantitative du paradigme des binômes verbaux en français préclassique, nous nous sommes appuyés sur les données du corpus de français préclassique de la base Frantext (ATILF 1998-2025), comprenant 353 textes entre 1550 et 1649, pour un total de 15 millions de mots. On notera que la délimitation proposée par Frantext ne fait pas nécessairement consensus (voir Ayres-Bennett & Caron 2016 pour une discussion critique de la question), mais comme nous le verrons, elle reste intéressante en ce qu'elle permet de visualiser une période de stabilité suivie de l'amorce d'un déclin dans l'emploi des binômes verbaux.

2.1. Requête et données

La requête que nous avons employée est la suivante : [pos="V"] [lemma="et"] [pos="CLO" | word="en" | word="ne" | word="n'" | word="y"]{0,2} [pos="V"], permettant d'extraire toute occurrence figurant la conjonction, via le lexème *et*, de deux verbes conjugués. Dans la perspective théorique proposée par Masini (2006) et dans laquelle nous nous insérons, nous considérons en effet le binôme (ici verbal) comme une possible construction associée à une palette de fonctions, parmi lesquelles l'usage synonymique bien identifié dans la littérature, et que nous ne distinguons pas d'entre les autres fonctions pour l'étude de la construction de façon plus globale. Notre requête opère néanmoins des choix implicites, notamment : ne pas considérer des binômes verbaux à l'infinitif (nous nous intéressons en effet au

cas où le binôme est la tête de la phrase au sens de la grammaire des dépendances), et ne pas considérer non plus une situation avec un binôme de participes partageant une même occurrence de l'auxiliaire :

(2) Ce que je feis, pour ce que ne le voulois desdire en rien comme **nous avions juré et promis** ensemble. (D. Zécaire, *Opuscule tres-eccelent de la vraye philosophie naturelle des metaulx*, 1550)

En outre, nous permettons qu'un ou deux clitiques s'interposent entre la conjonction marquée par *et* et le second verbe, comme en (3) :

(3) Tous les regnes mondains **se font & se defont**, // Et au gré de fortune ils viennent & s'en vont, // Et ne durent non plus qu'une flamme allumée // Qui soudain est esprise & soudain consumée. (Pierre de Ronsard, *Œuvres complètes: XI. Discours des misères de ce temps*, 1562)

Cette requête a renvoyé 7726 occurrences. Nous avons ensuite extrait les lemmes correspondant à chacun des deux verbes V1 et V2 (sans distinguer la forme réflexive et la forme standard), avec 1650 lemmes pour V1 et 1696 lemmes pour V2. Le lemmatiseur de Frantext n'étant pas entièrement fiable, surtout compte tenu de la diversité des formes graphiques à l'époque préclassique, nous avons ensuite corrigé manuellement les lemmes, avec 283 lemmes corrigés et 11 supprimés pour V1, et 298 lemmes corrigés (en plus des corrections déjà relevées pour V1) et 13 supprimés pour V2. Cela nous a permis de retenir 7682 occurrences et 1454 (resp. 1465) lemmes pour V1 (resp. V2).

Un point important concerne la question des traductions, qui sont nombreuses dans le corpus considéré (par exemple *Les douze livres de Lucius Junius Moderatus Columella des choses Rusticques* traduits par Claude Cotereau ou la volumineuse *Histoire des plantes* de Rembert Dodoens, traduite par Charles de l'Escluse). Nous avons décidé de les inclure pour deux raisons ; d'abord, même si la production du binôme peut être due à la langue source, et donc ne pas constituer un reflet fidèle de l'usage du français préclassique, la diffusion de ces livres influence et à ce titre participe néanmoins du paysage linguistique de l'époque. Ensuite, le binôme verbal constitue un procédé utile pour rendre un lexème complexe ou abstrait du texte (généralement latin, mais aussi, par exemple, italien, cf. Thorel 2010), le binôme constituant dès lors une innovation du traducteur par rapport au texte original (Willems 2003, Buridant 2003) ; dans l'extrait suivant (4), le binôme *sortira et germera* traduit le monôme *egerminent* (*ut quidam oculi trigeminis palmis egerminent*, dans le texte original). A ce titre, les binômes dans les œuvres traduites peuvent être des productions originales.

(4) Souvent il advient que d'un oeil **sortira et germera** troys boutons, dont il en fauldra oster deux, affin que l'aultre soit mieulx nourry. (Claude Cotereau, *Les douze livres de Lucius Junius Moderatus Columella des choses Rusticques*, 1551)

Deux autres points nous invitent à ne pas considérer les traductions comme une catégorie nécessairement à part. Tout d'abord, toutes les traductions ne relèvent pas d'une même langue-source (*L'Histoire des plantes* déjà mentionnée est traduite du « bas allemand », c'est-à-dire du néerlandais, *L'Histoire de la décadence de l'empire grec et establissement de celuy des Turcs* est traduite du grec). Parfois même, les traductions depuis le latin sont des traductions d'œuvres contemporaines écrites par un locuteur français (ainsi du *Traité du vin et du sidre*, écrit en latin par Julien Le Paulmier en 1588, ou quand Pierre de Boaistuau présente son *Théâtre du Monde* comme traduit depuis un original qu'il aurait lui-même composé en latin). On ne peut pas donc facilement imputer aux traductions l'influence directe et constante de traits spécifiques au latin. Ensuite, la plupart des traductions sont des traités ; or, en anticipant sur les résultats de l'article, les traités constituent un genre textuel qui se distingue par le recours abondant aux binômes verbaux. Il est donc difficile de démêler l'impact de la traduction de l'impact du genre textuel, représenté par une plus large diversité de textes. S'ajoutent à ces réserves théoriques une difficulté pratique : toutes les traductions ne sont pas distinguées comme telles dans Frantext, et l'auteur mentionné pour un texte est tantôt l'auteur de l'œuvre originale (p. ex. Rembert Dodoens pour *L'Histoire des plantes*, et non Charles de l'Escluse, son traducteur), tantôt le traducteur en français de cette œuvre (Antoine du Pinet pour *L'Histoire du monde* de Pline Second).

2.2 Fréquence

Cette requête permet de considérer la fréquence des binômes verbaux sur cette période et par là de s'assurer du caractère bien établi de ce procédé rhétorique. Au-delà du dénombrement d'occurrences, d'autres quantités peuvent servir à mesurer la fréquence (Loiseau 2015, Feltgen 2022), en particulier la prévalence, à savoir la proportion de textes (entre 0% et 100%) dans lesquels la forme se rencontre. Cette dernière est intéressante, en ce que le binôme, en tant que procédé rhétorique, peut constituer un marqueur stylistique de certains auteurs, et donc se trouver confiné dans certains textes. La Figure 1, qui montre à la fois la fréquence d'occurrence par million de mots (gauche) et la prévalence (droite), par décennie, sur toute la période considérée, révèle qu'il n'en est rien et que ce procédé est largement répandu dans les textes préclassiques.

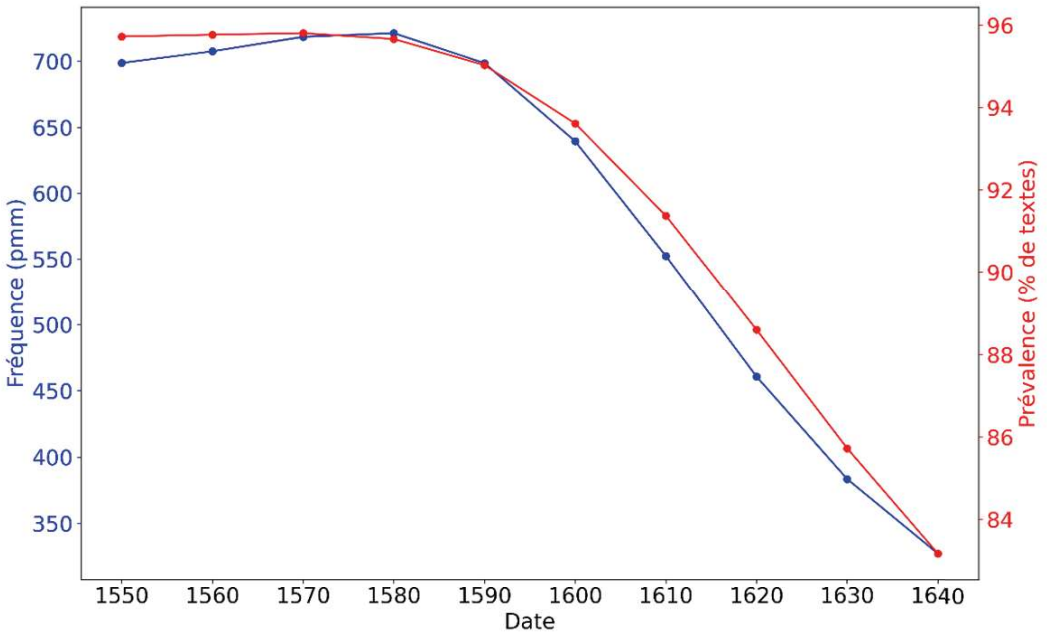

Figure 1: Fréquence par millions de mots (pmm) et prévalence (% de textes présentant au moins une occurrence de la forme) sur la période 1550-1649 d'après les données de la base Frantext. Les données ont été lissées au moyen d'un kernel gaussien d'écart-type 20.

On notera donc que le binôme verbal est une construction très fréquente, avec 515 occurrences par millions de mots sur l'ensemble de la période. Pour donner un ordre d'idée, sur la même période, *en* présente entre 17 000 et 15 000 occurrences par millions de mots, et *dans* entre 1 000 et 3 000 (Fagard & Combettes 2013 : 96), la construction *se faire* 17 occurrences par millions de mots (Lauwers & Duée 2010 : 103). Plus généralement, une préposition complexe est considérée comme ayant un score maximal de figement et d'enracinement à partir de 50 occurrences par millions de mots (Stosic & Fagard 2019, Table 1), score ramené à 40 par Vigier & Kahng (2022, 18). Rapportée à la taille du corpus frTenTen23 (23.8 milliards de mots), et située dans la liste des mots de ce corpus rangés par fréquence, cette fréquence correspond au 142ème mot le plus fréquent, *rien*, ce qui montre qu'il s'agit d'un procédé rhétorique particulièrement usité.

Mentionnons également que les binômes verbaux connaissent un net déclin à partir du XVIIe siècle, passant d'un peu plus de 700 occurrences par millions de mots à moins de la moitié (327) en l'espace de soixante ans. La prévalence connaît une décrue obéissant à la même dynamique, mais reste haute, passant de 96% à 83%. Cette décrue fait écho aux observations de Willems (2003), qui constate que, dans les traductions du latin, les traducteurs emploient moins de binômes

au XVI^e^ siècle qu'en moyen français, et coïncide notamment avec la critique du binôme synonymique déployée par François de Malherbe (Petrequin 2009 : 80), même si elle est amorcée avant cela. Nous ne nous pencherons cependant pas sur cette dynamique diachronique et adopterons dans toute la suite un point de vue synchronique sur la période préclassique prise dans son intégralité.

3. Etude quantitative du système binomial

Dans cette section, nous abordons l'étude des binômes verbaux en français préclassique sous un angle exclusivement quantitatif, en tentant de répondre aux questions suivantes : comment sont organisés les deux jeux de verbes (ou paradigmes variationnels) mis en relation dans le système binomial ? Quelle est leur productivité ? Ces deux jeux de verbes sont-ils comparables, c'est-à-dire reflètent-ils le même jeu de verbes sous-jacent, ou présentent-ils des différences statistiquement significatives ? Comment quantifier la relation entre ces deux jeux de verbes ? Nous terminerons enfin cette section par l'apport central de l'article, à savoir une méthode d'analyse statistique permettant de déterminer quels binômes verbaux sont les plus statistiquement significatifs, c'est-à-dire les plus enracinés dans l'usage linguistique en tant que tels, autrement dit, les moins prédictibles à partir de la libre combinaison des deux jeux de verbes qui composent le système binomial.

3.1. Loi de Zipf

Sans nécessairement suivre Masini (2006, 2016) dans sa lecture constructionnelle du binôme verbal (nous préférons laisser ouverte la question du statut grammatical de ce procédé rhétorique dans le cadre de la grammaire des constructions), nous nous inspirerons néanmoins de cette perspective théorique pour analyser l'organisation du système. En particulier, de nombreux travaux ont souligné la pertinence de la loi de Zipf pour comprendre l'organisation des termes (aussi appelés *types*) pouvant intervenir dans le schéma paradigmatique libre d'une construction (Goldberg *et al.* 2004, Ellis & Fereira-Junior 2009, Ellis 2012, Zeldes 2012, Ellis *et al.* 2014).

La loi de Zipf est une loi visant à caractériser la distribution de fréquence associée à un schéma paradigmatique : elle stipule la manière dont les occurrences individuelles de la construction/du schéma se répartissent entre les différents types. Dans notre cas, la distribution de fréquence représente, pour chacun des deux jeux V1 et V2 individuellement, le nombre d'occurrences de chacun des types. Par exemple, pour V1, on aura d'abord *aimer* avec 121 occurrences, puis *être*, *aller*, *voir* et *pouvoir* avec respectivement 117, 107, 98 et 90 occurrences. Pour V2, les cinq termes prépondérants sont *faire*,

être, *dire*, *avoir* et *mettre* avec respectivement 287, 197, 107, 102, et 88 occurrences. Le principe de la loi de Zipf consiste à dire que, si on range les types par ordre de fréquence décroissante (comme nous venons de le faire), alors cette fréquence décroît comme une loi de puissance du rang dans la hiérarchie. Une distribution zipfienne est donc typiquement associée avec quelques termes très fréquents et un très grand nombre de termes peu fréquents, notamment une large proportion de *hapax legomena*. Par ailleurs, des travaux ont montré qu'une correction de cette loi aux hautes fréquences, la loi de Zipf-Mandelbrot, était généralement plus pertinente (Evert 2004), c'est pourquoi nous l'adoptons ici. Cette loi se traduit par l'expression mathématique suivante, reliant la fréquence f d'un type à son rang r dans la hiérarchie :

$$f(r) = \frac{A}{(r+b)^{\alpha}}$$

Elle dépend de deux paramètres : α, qui décrit à quel point la décroissance de la fréquence en fonction du rang est rapide, et b, la correction de Mandelbrot, qui aplatit la courbe aux hautes fréquences. Le paramètre A est un facteur permettant de normaliser cette loi. S'agissant d'une loi de puissance, qui met en jeu des fréquences avec une forte disparité (allant de 1 à plusieurs centaines ici), on la représente plus usuellement en échelle logarithmique, ce que nous avons fait Figure 2 où les profils zipfiens des deux jeux de verbes V1 et V2 ont été représentés. Nous avons considéré seulement les 350 premiers termes, les termes de basse fréquence présentant une forte dégénérescence (c'est-à-dire que de nombreux types ont la même valeur de fréquence), ce qui rend caduque la notion même de rang entre les types.

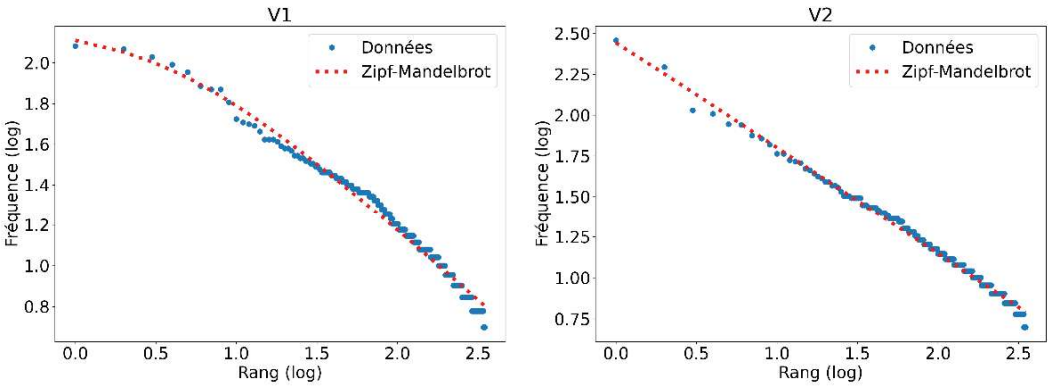

Figure 2: Profils zipfiens des 350 premiers termes des deux jeux de verbes V1 et V2

Obtenir les paramètres de la distribution zipfienne est notoirement compliqué, et nous avons suivi ici pour les obtenir la méthode de Koplenig (2018) basée sur la maximisation du *log-likelihood* associé à la distribution. La valeur des paramètres obtenue avec cette méthode est généralement peu fiable, mais elle présente l'avantage didactique d'offrir un bon accord avec les données empiriques sur le plan visuel. Les paramètres sont les suivants : α = 0.72 et b = 4.67 pour V1, α = 0.65 et b = 0.05 pour V2. La correction de Mandelbrot est donc inexistante pour V2, qui présente un profil plus purement zipfien, alors que V1 offre une plus grande variété de termes aux hautes fréquences, ce que caractérise la plus forte valeur du paramètre b. Nous verrons que cette différence, qualitativement visible sur la figure, a son importance vis-à-vis de la productivité de chacun des deux jeux de verbe.

3.2. Productivité

La productivité est une notion qui permet de mesurer la richesse lexicale associée à l'usage d'un motif morphologique ou syntaxique. Néanmoins, sa mesure fait l'objet de divergences : si le ratio nombre de hapax / nombre d'occurrences est l'un des plus populaires (Baayen 2009), le ratio nombre de types / nombre d'occurrences est également utilisé (Stefanowitsch & Flach 2017, Flach 2021), mais également le ratio nombre de hapax / nombre de types (Van Wettere 2022), ainsi que d'autres mesures, plus ou moins liées entre elles (Van den Heede & Lauwers 2023). Ces mesures sont par ailleurs sensibles de manière critique à la taille d'échantillon (Gaeta & Ricca 2006), mais cela ne constituera pas ici un écueil, puisque nous comparons deux jeux de verbes dont la taille d'échantillon est, par construction, identique (puisqu'elle est donnée par le nombre de binômes verbaux, chaque binôme ajoutant une occurrence à l'un et l'autre jeu de verbes).

Ces mesures ne prennent cependant pas en compte la distribution de fréquence dans son ensemble, dont on a vu pourtant qu'elle obéissait à une structure intéressante, bien rendue par la loi de Zipf-Mandelbrot. Pour y remédier, on peut considérer l'entropie associée à cette distribution, laquelle consiste à mesurer l'incertitude (en termes de résultat d'un tirage aléatoire) associée à la distribution de fréquence. Elle se définit comme suit :

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

C'est-à-dire que, pour une variable aléatoire X, on calcule la somme, sur toutes les réalisations possibles de X, du produit de la probabilité de cette réalisation et de son logarithme (traditionnellement

en base 2 pour mesurer l'entropie en bits d'information). Concrètement, pour l'un des deux jeux de verbes, par exemple V1, les réalisations possibles sont les différents types rencontrés dans V1, et leur probabilité, la proportion d'occurrences associée. L'entropie est comprise entre 0 (pour une variable dont l'issue est déterministe : il n'y a donc aucune incertitude), et $\log_2 K$, où K est le nombre de réalisations possibles de la variable aléatoire (dans la configuration où toutes les réalisations sont équiprobables et l'incertitude sur sa réalisation, maximale). Cependant, ce nombre de réalisations possibles, dans le cas d'un schéma linguistique, est ambigu : on le considère parfois égal au nombre de types (Römer & Gardner 2019), mais en réalité, chaque occurrence pourrait instancier un nouveau type, et sous cette perspective, le nombre de réalisations possibles est en réalité $\log_2 N$, où N est le nombre total d'occurrences. L'entropie étant difficile à interpréter pour ces raisons, nous préférons ici l'appréhender à travers la variable de perplexité (Bochkarev *et al.* 2023), qui se définit simplement par :

$$U = 2^{H(X)}$$

Cette variable de perplexité mesure le nombre de types équiprobables effectif, c'est-à-dire entre combien de types différents s'effectue réellement un tirage aléatoire d'une occurrence. La distribution étant zipfienne, les termes de haute fréquence réduisent en effet la diversité des occurrences rencontrées : c'est précisément ce biais que prend en compte la perplexité, *via* l'entropie qui en sous-tend la définition.

Le Tableau 1 récapitule ces différentes mesures. On notera d'abord que, suivant les trois mesures traditionnelles, V2 est très légèrement plus productif que V1 ; mais suivant la perplexité, c'est V1 qui est nettement plus productif. Pour évaluer l'importance de ces différences, nous avons procédé à un test statistique consistant à mélanger les occurrences des deux jeux de verbes entre eux, c'est-à-dire que nous regroupons ensemble toutes les occurrences de V1 et V2 réunis, et les répartissons aléatoirement en deux jeux de taille égale, identique à celle de V1 et V2. Pour cette nouvelle répartition aléatoire des occurrences en deux jeux distincts, nous pouvons calculer la différence de productivité (en valeur absolue : peu importe lequel des deux jeux est plus productif). En répétant cette procédure 1 000 fois, cela donne une distribution de la différence de productivité à laquelle on peut comparer la valeur observée. Il se trouve que seule la différence de perplexité est significative ; plus encore, aucune des répartitions aléatoires des occurrences ne donne une différence de perplexité plus grande.

	V1	V2	V1 – V2	IC	*p*	Binômes
Ratio Types/Occs	0,19	0,19	- 0,001	[0,001 – 0,008]	0,72	0,78
Ratio Hapax/Occs	0,07	0,08	- 0,003	[0,001 – 0,008]	0,37	0,67
Ratio Hapax/ Types	0,39	0,40	- 0,01	[0,005 – 0,034]	0,31	0,86
Perplexité	671	606	65	[6 - 34]	< 0,001	4829

Tableau 1: Différences de productivité entre V1 et V2 suivant quatre mesures de productivité. L'intervalle de confiance (IC, correspondant à 95% des valeurs de la distribution associée à la différence, en valeur absolue) et la valeur *p* associée permettent d'évaluer la significativité de cette différence.

Comme on le voit, la perplexité est une mesure de productivité plus fine en ce qu'elle est sensible à la distribution de fréquence dans son ensemble. Cette différence fait écho à la Figure 2 où une différence qualitative claire distinguait les deux jeux de verbes : alors que pour V2, le haut de la distribution est dominé par les verbes les plus génériques, dans une hiérarchie stricte, pour V1 la courbure liée au paramètre de Mandelbrot permet à plusieurs verbes de coexister aux plus hautes fréquences, et ces verbes sont plus atypiques. Cela traduit une plus grande diversité et une plus grande spécificité dans le choix du premier terme que dans le choix du second. En somme, dans le binôme, les deux jeux de verbes mis en relation ne présentent pas les mêmes propriétés structurales liées à l'usage.

Mentionnons par ailleurs que, si l'on considère la productivité de l'ensemble des combinaisons binomiales rencontrées (pour lesquelles nous avons encore une fois et par construction la même taille d'échantillon que précédemment, à savoir 7682 occurrences), on trouve ici une perplexité de 4829, ce qui est d'une part très élevé en comparaison de la productivité de chacun des jeux de verbes, mais également en comparaison du nombre de combinaisons rencontrées si les occurrences sont réparties de manière aléatoire entre les deux jeux de verbes comme dans l'analyse précédente (on a alors un intervalle de confiance compris entre 4557 et 4593).

3.3. Différences entre les deux hiérarchies verbales

Dans quelle mesure les deux jeux de verbes sont-ils spécifiques de chaque emplacement ? Dans quelle mesure correspondent-ils et s'informent-ils mutuellement ? S'il s'agit là de questions naturelles, y répondre avec rigueur est difficile compte tenu du caractère zipfien de l'organisation de ce paradigme, comme nous allons l'illustrer ici et dans la sous-section suivante.

Tout d'abord, les deux emplacements V1 et V2 ont une proportion élevée d'éléments en commun (c'est ce qu'on appelle le

coefficient de recouvrement) : sur 1 454 éléments de V1 (resp. 1465 éléments de V2), 1 008, soit 69%, se retrouvent également dans V2 (resp. 69 % se retrouvent dans V1), ce qui paraît élevé. On peut en outre considérer le rang, en termes de fréquence, des verbes de chacun de ces deux jeux V1 et V2, et considérer dans quelle mesure les verbes en commun ont le même rang de l'un à l'autre jeu. Par exemple, le verbe *aimer*, qui a le 1er rang dans V1, avec 120 occurrences, est au rang 33 dans V2, avec 28 occurrences ; le verbe *être* est au rang 2 dans V1 et dans V2, avec respectivement 115 et 178 occurrences, etc. Pour mesurer la similitude entre ces deux hiérarchies de fréquence, on peut calculer le coefficient de corrélation Spearman, qui mesure à quel degré deux jeux de données sont ordonnés de manière similaire. Ce coefficient est égal à 0,68 (avec une valeur p de 10^{-135}, ce qui est bien sûr hautement significatif). Cela laisse entendre que ces deux emplacements constituent deux instanciations indépendantes du même jeu de verbes sous-jacent (c'est-à-dire que les deux verbes seraient successivement choisis au sein d'un même paradigme variationnel), ce qui constituerait une « hypothèse nulle » dans la description du paradigme binomial.

Pour tester cette hypothèse, nous avons procédé à un test statistique, en répartissant toutes les occurrences aléatoirement mais équitablement dans l'un et l'autre jeu, comme dans la section précédente. Nous avons alors calculé le coefficient de Spearman et recommencé cela un grand nombre de fois (1000) afin d'en déduire une distribution des valeurs prises par ce coefficient de corrélation sous l'hypothèse que V1 et V2 sont aléatoirement dérivés du même jeu de données.

Il s'avère que, sous cette hypothèse, les valeurs prises par le coefficient de Spearman sont comprises dans l'intervalle de confiance [0,74 ; 0,77], et le minimum obtenu est 0,71 : la valeur trouvée précédemment (0,68) est en fait très significativement faible par rapport à l'hypothèse de deux jeux de données équivalents. De même, le coefficient de recouvrement entre les deux jeux de verbes (69%, on l'a vu), est associé à l'intervalle de confiance [72% ; 74%] avec un minimum à 70%. En d'autres termes, même si ces coefficients de corrélation et de recouvrement paraissent élevés, ils sont faibles au regard de ce que l'on pourrait attendre si les deux jeux de verbes étaient issus d'un même ensemble sous-jacent – autrement dit les deux jeux diffèrent significativement et par leur contenu, et par la hiérarchie qu'ils impriment sur ce qu'ils ont en commun.

3.4. Information mutuelle

Nous venons d'établir que les deux jeux de verbes, correspondant respectivement à la première et seconde positions du binôme, diffèrent effectivement, tant du point de vue du contenu que du point de vue

de leur organisation hiérarchique. Nous souhaitons nous tourner désormais vers la relation binomiale elle-même : en quoi ces deux jeux de verbes interagissent-ils ?

Là encore, il est possible d'essayer de quantifier cette interaction. Pour ce faire, on peut exploiter une mesure simple, proposée par la théorie de l'information : l'information mutuelle. L'information mutuelle entre deux variables probabilistes X et Y mesure la réduction d'incertitude de Y si la valeur de X est connue (ou vice-versa, la définition étant parfaitement symétrique). Si X et Y sont deux variables indépendantes, alors connaître la valeur de X « n'apprend rien » sur la valeur de Y : l'incertitude sur Y (mesurée par son entropie) reste inchangée et l'information mutuelle est nulle. Si au contraire Y dérive exactement de X, alors la valeur de X détermine exactement la valeur de Y : il n'y a plus d'incertitude sur celle-ci et l'information mutuelle est égale à l'entropie de Y.

On peut calculer cette information mutuelle comme suit :

$$I(X;Y) = H(Y) - H(Y|X)$$

Ici, l'entropie conditionnelle $H(Y|X)$ se calcule de la manière suivante :

$$H(Y|X) = \sum_x p(X = x)H(Y|X = x)$$

Concrètement, cela signifie que l'on considère chaque valeur possible de V1 et que l'on calcule ensuite l'entropie du sous-ensemble de V2 correspondant aux binômes commençant par ce verbe (par exemple pour la valeur *gâter* de V1, on obtient pour V2 11 occurrences de *corrompre* et 18 autres verbes de fréquence 1 : on calcule alors l'entropie associée à cette distribution de fréquence, soit 3,5 bits) ; on somme ensuite l'ensemble de ces entropies calculées sur chaque sous-ensemble de V2 correspondant à un choix de V1, pondérées par la probabilité de ce choix. L'exemple de *gâter* donne par ailleurs une bonne illustration de la manière dont V1 peut informer.

La valeur obtenue pour l'information mutuelle entre V1 et V2 est de 6,38 bits (soit 69% de sa valeur maximale, l'entropie de V2, égale à 9,24 bits). Comment interpréter cette valeur ? Là encore, nous procèderons à un test statistique, en comparant cette valeur à l'information mutuelle entre V1 et une version aléatoirement réarrangée de V2 – c'est-à-dire que l'on conserve les deux jeux de verbes à l'identique, mais que l'on choisit aléatoirement de redessiner les liens binomiaux entre ces deux jeux. Il faut noter ici que ceci est foncièrement différent de l'analyse statistique précédente, qui consistait à répartir les occurrences entre les deux jeux de verbes, changeant

donc leur composition, alors qu'ici, la composition de chacun des jeux et la distribution de fréquence associée restent identiques, seuls les liens faisant l'objet d'un réarrangement. À priori, ce réarrangement enlève toute l'information que la donnée de V1 pourrait apporter sur V2 du fait du système binomial.

Le résultat de ce test est surprenant. D'une part, l'information mutuelle entre V1 et n'importe quelle version réarrangée de V2 est de 5,81 bits en moyenne (63% de la valeur maximale), avec un intervalle de confiance de [5,81 ; 5,82], ce qui signifie que V1 est informatif de V2, même en absence de toute relation binomiale d'origine. Comment expliquer ce fait ? Il faut ici reprendre la définition et considérer par exemple le cas où la réalisation de V1 est un hapax (*hapax legomena* qui pèsent pour 7% de la distribution). Peu importe l'arrangement aléatoire, le binôme constitué par ce verbe détermine la réalisation de V2 de manière unique et la contribution correspondante à l'entropie conditionnelle est par conséquent nulle, ce qui rapproche l'information mutuelle de sa valeur maximale. Plus généralement, V1 étant constitué d'un grand nombre de verbes de basse fréquence, en conséquence associés à des distributions sur des sous-ensembles particulièrement limités de V2, leur contribution à l'entropie conditionnelle reste marginale : c'est donc bien la structure zipfienne et le large nombre de verbes faiblement fréquents qu'elle implique qui entraîne cette valeur de référence étonnamment élevée pour l'information mutuelle[2].

Mais, d'autre part, l'information mutuelle observée pour le vrai système binomial (sans réarrangement aléatoire de V2 donc) est très significativement plus élevée que cette valeur obtenue par réarrangement aléatoire (cette valeur se trouve à 137 déviations standards de la moyenne de l'information mutuelle entre V1 et un réarrangement aléatoire de V2 !). Même si la différence d'information mutuelle est faible (de l'ordre d'un demi-bit d'information), elle est cruciale, car elle distingue très nettement le système binomial véritable de n'importe quelle organisation aléatoire obtenue en appariant au hasard les occurrences des jeux V1 et V2. Cela signifie donc bien que le paradigme des binômes verbaux présente une organisation, une structure, et que les associations qu'il met en jeu ne sont pas arbitraires.

Par ailleurs, nous avons calculé la productivité de l'ensemble des binômes obtenus sous un réarrangement aléatoire des liens du système. Nous trouvons que la perplexité devrait être comprise entre 7247 et 7306 avec une probabilité de 95%, ce qui est très au-dessus des 4829 observés. Cela paraît contredire pleinement les résultats exposés *supra* en 3.2. et mérite explication. En 3.2., le jeu des binômes est plus

[2] On peut d'ailleurs calculer l'information mutuelle minimale du système étant donné le profil zipfien de V1, égale à 5,73 bits.

productif que n'importe quel jeu obtenu à partir d'un réarrangement des occurrences et donc du contenu des jeux V1 et V2. Ici, il est moins productif que n'importe quel jeu obtenu à partir d'un réarrangement des liens entre V1 et V2, tout en laissant leur contenu respectif inchangé. Si l'ensemble des binômes est plus diversifié en 3.2., c'est parce que les contenus des jeux V1 et V2, on l'a vu, diffèrent significativement : ces deux jeux ne sont pas interchangeables, les verbes qu'ils prennent en argument et la hiérarchie qu'ils imposent sur ceux-ci diffèrent de l'un à l'autre. Les relations binomiales entre ces deux jeux sont donc plus productives que s'ils étaient similaires. En revanche, les combinaisons entre ces deux jeux sont plus contraintes que ce que permettent des associations aléatoires : c'est pourquoi l'information mutuelle entre ces deux jeux est très significativement plus élevée que sous des associations aléatoires. Comme les combinaisons sont plus contraintes, la diversité des binômes est moindre qu'en l'absence de telles contraintes.

Ce raisonnement ne serait pas complet sans considérer également l'information mutuelle entre les termes du binôme, pour un système construit par répartition aléatoire de l'ensemble des occurrences entre deux jeux de même taille, comme en §3.2. On trouve alors que l'information mutuelle est comprise entre 5,92 et 5,94 bits, soit à peine supérieure à l'information mutuelle entre les deux termes du binôme sous réarrangement aléatoire des liens entre V1 et V2, et une fois encore très significativement inférieure à l'information mutuelle véritable du système.

Pour résumer, l'asymétrie entre les deux termes du binôme assure la diversité du système, et les contraintes d'association entre ces deux termes assurent l'organisation de ce système : ce sont ces deux aspects complémentaires qui assurent l'informativité du système par rapport à l'aléatoire. Ce point illustre toute la subtilité, la richesse et la complexité de l'organisation linguistique telle que la révèle le fonctionnement du système binomial.

3.5. Seuil de fréquence

Nous avons établi dans la sous-section précédente le caractère motivé et non-aléatoire des relations qui sous-tendent le paradigme des binômes verbaux. Cependant, nous souhaiterions désormais nous munir d'un critère statistique permettant de sélectionner les binômes verbaux pertinents et significatifs, au-delà de l'association libre entre deux jeux de verbes obéissant chacun à sa hiérarchie propre.

Un premier point est de prendre en considération les fréquences des verbes impliqués dans le binôme, afin de favoriser les associations récurrentes entre verbes rares, et pénaliser les associations entre

verbes génériques. Par exemple, en considérant le couple (*v1*, *v2*), de fréquences respectives *n1* et *n2*, on peut considérer que le nombre de liens moyens attendu entre ces deux verbes, en distribuant aléatoirement les liens entre les membres de V1 et V2, est de *n1* x *n2* / *N*, où N est le nombre total d'occurrences de binômes[3]. Ainsi, pour les deux verbes les plus fréquents de chaque jeu de verbe (respectivement *aimer* avec 121 occurrences pour V1 et *faire* avec 287 occurrences pour V2), on trouve que le nombre de liens est significatif s'il est d'au moins neuf. Or on ne trouve qu'une seule occurrence de ce binôme, qui n'est donc pas significatif :

(5) Ce qui me porte à la presomption qu'elle n'est point sans une affection, c'est qu'elle est jeune, elle est fine, elle est belle. Certes elle **ayme et fait** en sa ruelle ce que je pense et par discretion je ne dis pas. (Claude Malleville, *Œuvres poétiques*: 1, 1649)

En revanche, sous cette perspective, tous les binômes hapax-hapax, par exemple *crocheter* et *fureter* (6), même s'ils n'apparaissent par définition qu'une seule fois, sont hautement significatifs (on a alors $p = 1/N^2$ pour la distribution binomiale ; une fréquence de 1 est supérieure à la moyenne de 87 fois l'écart-type), ce qui ne convient pas.

(6) Horace ne se contente point d'une superficielle expression, elle le trahiroit. Il voit plus cler et plus outre dans la chose; son esprit **crochette et furette** tout le magasin des mots et des figures pour se représenter; et les luy faut outre l'ordinaire, comme sa conception est outre l'ordinaire. (Michel de Montaigne, *Essais*: t. 2, 1592)

C'est pourquoi nous proposons dans cet article d'affiner cette manière de procéder en considérant non pas la probabilité d'observer un certain nombre d'occurrences k d'un binôme donné, mais la probabilité d'observer ce nombre d'occurrences k *conditionné au fait que le binôme apparaît dans le corpus*, quantité que l'on peut obtenir à partir de la distribution générale *via* la relation de Bayes :

$$P(k\,|n_1, n_2, N, k \geq 1) = \frac{P(k|n_1, n_2, N)}{1 - P(0|n_1, n_2, N)}\ si\ k \geq 1\ ou\ 0\ sinon$$

En notant F la fonction cumulative associée à la loi $P(k|\,n_1,n_2,N)$, on peut alors calculer la valeur p associée à un binôme présentant k

[3] Il s'agit ici d'une application de la loi binomiale, chaque occurrence de binôme pouvant associer les deux verbes avec une probabilité $p = n1x\ n2\ /\ N^2$. Le nombre de liens peut être considéré significatif s'il est supérieur de deux écart-types au moins à la moyenne, la moyenne et l'écart-type étant respectivement données par Np et $\sqrt{Np(1-p)}$.

occurrences[4] dans le corpus *via* :

$$p = 1 - \frac{F(k-1|n_1, n_2, N) \; - \; P(0|n_1, n_2, N)}{1 \; - \; P(0|n_1, n_2, N)}$$

Cette dernière méthode résout les difficultés mentionnées précédemment : notamment, par construction, la valeur p est de 1 dans le cas d'un binôme n'apparaissant qu'une seule fois dans le corpus. Elle présente en outre l'avantage de ranger les binômes verbaux dans un ordre allant des binômes les plus fortement ancrés (c'est-à-dire dont le nombre d'occurrences observé est le moins probable sous l'hypothèse d'une association libre entre les jeux de verbe) au moins ancrés.

Reste à déterminer un critère pour retenir ou rejeter les binômes sur la base de cette valeur p, avec tout ce qu'une telle approche a d'arbitraire. Il est usuel de retenir comme significatives les observations associées à une valeur p inférieure à 0,05. Ici, sur les 5963 binômes, 666 (11%) seraient retenus selon ce critère. Il est cependant recommandé d'appliquer la correction de Bonferroni à ce seuil de 0,05, pour tenir compte du fait que l'on n'effectue pas une seule observation (nombre d'occurrences d'un binôme donné), mais 5963 (nombre d'occurrences de chacun des binômes). Cette correction consiste à ajuster la valeur seuil (0,05), en la divisant par le nombre d'observations (5963). En faisant cela, on retient 101 binômes verbaux, soit un peu moins de 2% du total. Ces 101 binômes verbaux totalisent 846 occurrences, soit 11% du total.

4. Typologie des binômes verbaux

Nous présenterons pour terminer les 101 binômes sélectionnés par l'analyse statistique comme bien ancrés dans l'usage en tant que binômes, au-delà de la combinaison libre permise par les deux jeux de verbes qui composent le système. Ces binômes présentent une large diversité fonctionnelle, que sous-tend la polysémie inhérente à la conjonction elle-même (Badiou-Monferran 2020). Nous les classerons en cinq catégories, en nous inspirant de la proposition de Masini (2006) : binômes antonymiques, binômes liés par une relation de séquentialité, binômes liés par une relation de complémentarité, binômes synonymiques[5]. Ces catégories manifestent entre elles une

[4] La valeur p est ici la probabilité d'observer (sous l'hypothèse de combinaisons libres entre chacun des deux jeux de verbes, tenant compte de la fréquence de leurs types respectifs) un nombre d'occurrences supérieur ou égal au nombre d'occurrences du binôme considéré. Sans notre correction cette valeur p est plus classiquement donnée par 1 - F($k|n_1,n_2,N$).

[5] On trouvera une typologie plus étoffée chez Legrand (2022), appliquée aux binômes (nominaux, verbaux, adjectivaux) rencontrés dans *l'Histoire d'un voyage fait en la terre*

certaine porosité (nombre de binômes pourraient facilement figurer dans plusieurs d'entre elles) et nous avons ici tenté de les agencer en un spectre sémantique allant du contraste à la répétition. Au sein de chacune d'elles, les binômes sont rangés par ordre décroissant de significativité. Par ailleurs, deux binômes présentent la répétition d'un même verbe, avec polyptote sur le temps verbal : il s'agit de *être et être* (7) et *avoir et avoir* (8). La formule *je suis et serai*, en particulier, est particulièrement formulaïque en clôture de lettre dans la correspondance :

(7) C'est pourquoy je finis en vous suppliant très humblement de croire que je **suis et serai** de tout mon coeur, et toute ma vie, monsieur, vostre très humble serviteur. (Guy Patin, *Lettres* : t. 1 : 1630-1649, 1643)

(8) Les cieux et les humains enflamez de courous, // **N'ont et n'eurent** jamais de justice pour nous (Alexandre Hardy, *La Force du sang*, 1626)

4.1. Binômes antonymiques

Certains binômes (12 types) fonctionnent comme un couple de termes antonymiques : *croître et décroître*, *commencer et finir*, *entrer et sortir*, *monter et descendre*, *hausser et baisser*, *ouvrir et fermer*, *vivre et mourir*, *rire et pleurer*, *fermer et ouvrir*, *englacer et enflammer*, *avancer et reculer*, *lier et délier*. On remarquera que *ouvrir et fermer* et *fermer et ouvrir* se rencontrent tous les deux. En dehors de l'effet fréquent de mérisme – expression de la totalité par conjonction des opposés (Martí Solano 2024 : 2) –, ces couples n'ont pas nécessairement une fonction bien établie et peuvent prendre leur sens en fonction du contexte. Pour *commencer et finir*, par exemple, le binôme peut exprimer la similarité voire l'identité entre les termes d'un procès ou d'une figure géométrique (9), mais aussi l'absence de durée du procès ou sa brièveté (10) :

(9) [les sieges] faisoient comme une parfaite couronne qui **commençoit et finissoit** où estoit Diane. (Honoré d'Urfé, *L'Astrée* : t. 3 : 3ème partie : livres 1 à 12, 1631)

(10) Jamais on ne m'apprit lequel des doits on moüille // quand pour gagner sa vie on file sa quenoüille : // une femme à filler ne fait pas grand butin ; // ma besongne **commence et finit** au matin ; // qu'on ne s'attende pas que je couse ou tapisse, // le plus aisé travail pour moy n'est qu'un supplice ; // puisque j'ay de quoy vivre et de quoy m'habiller, // qu'on me parle de rire, et non de travailler; (Jacques du Lorens, *Satires*, 1646)

---

du Brésil de Jean de Léry.

4.2. Binômes séquentiels

La seconde catégorie (10 types) concerne les binômes de verbes dont l'un prend la suite de l'autre : *passer et repasser*, *tourner et retourner*, *penser et repenser*, *hausser et abaisser*, *mûrir et romper*, *délibérer et résoudre*, *rompre et pousser*, *dire et redire*, *assiéger et prendre*, *mûrir et faire*. Dans une certaine mesure, cette dimension séquentielle est déjà présente dans les couples antonymiques. Ici, la séquentialité peut venir de la répétition, notamment soulignée par le redoublement d'un verbe au moyen du préfixe *re-* (11), ou bien du fait que l'un des termes apparaît comme la conséquence de l'autre (12). Les deux binômes en *mûrir* (lemmatisation par Frantext de *meurir*) peuvent surprendre : ils sont dus à la traduction de l'*Histoire des plantes* de Rembert Dodoens, où le terme *meurir* prend une signification technique et médicale qui semble se distinguer de celle du mûrissement rencontrée dans les autres traités de l'époque.

(11) La roue d'Ixion est le mouvement de ses desirs, qui **tournent et retournent** continuellement de haut en bas, et ne donnent aucun repos à son esprit. (Pierre Charron, *De la sagesse : trois livres*, 1601)

(12) Le pauvre escolier tremblant comme la feuille de l'arbre [..], **se delibera et resolut** aller vers la dame, quelque peril qu'il luy deust advenir. (Philippe d'Alcripe, *La Nouvelle fabrique des excellents traicts de verité*, 1579)

4.3. Binômes complémentaires

Il s'agit ici de la catégorie la plus importante (47 types), où nous avons regroupé tous les binômes où la conjonction des termes permet de raffiner et de préciser une idée, d'étoffer la description d'une action en décrivant la multiplicité de ses facettes : *aller et venir*, *boire et manger*, *guérir et consolider*, *pouvoir et devoir*, *pouvoir et vouloir*, *voir et connaître*, *donner et communiquer*, *manger et boire*, *consolider et guérir*, *promettre et jurer*, *aimer et estimer*, *jurer et promettre*, *tuer et pousser*, *aimer et honorer*, *devoir et pouvoir*, *dire et faire*, *vivre et régner*, *échauffer et dessécher*, *suivre et accompagner*, *voir et sentir*, *reconnaître et adorer*, *vouloir et pouvoir*, *nourrir et entretenir*, *faire et dire*, *voir et entendre*, *reprendre et accroître*, *fuir et abhorrer*, *écouler et évanouir*, *craindre et fuir*, *piller et brûler*, *empêcher et détourner*, *être et pouvoir*, *aimer et désirer*, *voir et ouïr*, *devoir et vouloir*, *ouïr et voir*, *vouloir et entendre*, *vouloir et oser*, *embrasser et caresser*, *empêcher et troubler*, *paître et délecter*, *gêner et déchirer*, *charmer et endormir*, *pleurer et soupirer*, *vouloir*

et ordonner, *avoir et adorer*, *estimer et honorer*, *maintenir et nourrir*.

L'exemple de *boire et manger* est caractéristique de cet usage : les deux verbes ne se recoupent pas, mais participent d'un même contexte, d'une même scène, sans relation temporelle spécifique entre eux, et permettent en conjonction une description plus complète que ce qu'autorise un seul de ces deux verbes, un procédé stylistique appelé *hendiadys* (Martí Solano 2024 : 3). On remarquera par ailleurs que certains de ces binômes sont réversibles. Dans cet emploi se rencontrent également de nombreux verbes de perception, ainsi que des verbes de modalité, ce qui permet là encore, par le jeu de la combinatoire, d'affiner la caractérisation du mode (13). Certains de ces binômes ont enfin une valeur formulaïque ; c'est le cas notamment de *vivre et régner* (14), utilisé par 5 auteurs différents, toujours dans le même contexte théologique.

(13) Mais je ne promets rien, par ce qu'un coeur auguste // **Ne veut et ne peut** rien promettre que de juste. (Pierre de Ruyer, *Thémistocle*, 1648)

(14) Car la doctrine dela foy […] nous enseignent iournellement que Dieu **vit et regne** en l'vnité du s. esprit, dans lequel il a sa vie et son repos […] et en qui se termine et accomplit heureusement l'vnité, la fecondité et la societé parfaitte des personnes diuines ensemble. (Pierre de Bérulle, *Discours de l'estat et des grandeurs de Jésus par l'union ineffable de la divinité avec l'humanité*, 1623)

4.4. Binômes synonymiques

La dernière catégorie (29 types) regroupe les binômes répétant la même idée au moyen d'une variation synonymique: *gâter et corrompre*, *supplier et conjurer*, *savoir et connaître*, *tourner et virer*, *embrasser et baiser*, *dédier et consacrer*, *prier et supplier*, *régir et gouverner*, *prier et conjurer*, *pleurer et lamenter*, *piller et saccager*, *priser et honorer*, *rompre et briser*, *baiser et embrasser*, *suivre et accompagner*, *adoucir et amollir*, *reconnaître et confesser*, *tromper et abuser*, *digérer et ressoudre*, *vouloir et désirer*, *reconnaître et avouer*, *louer et remercier*, *engendrer et produire*, *honorer et chérir*, *plaindre et soupirer*, *résoudre et digérer*, *renouveler et rafraîchir*, *dire et déclarer*, *consacrer et dédier*. La répétition synonymique est parfois mise au service d'un effet d'emphase (15), mais en d'autres cas il semble s'agir d'un emploi formulaïque, par exemple en (16) où le binôme intervient dans une parenthèse qui n'a *a priori* pas vocation à apporter rien d'autre qu'une précision factuelle incidente.

(15) Voila aussi pourquoi le Prophete dit, que telles gens escorchent la peau, mangent la chair, **rompent et brisent** les os du peuple de Dieu, comme s'ils les faisoyent bouïllir dans une chaudiere. (Jean de Léry, *Histoire d'un voyage faict en la terre du Brésil*, 1578)

(16) c'estoit de la croyance des Paiens d'estimer que la Fortune (à la quelle à céte cause ils **consacroient et dedioient** des temples) maistrisoit sur les plus sages comme une tres-puissante Déesse […]. (Scipion Dupleix, *La Logique ou l'Art de discourir et raisonner*, 1607)

5. Hétérogénéités dans la distribution des binômes

Il est possible d'aller un pas plus loin dans la description quantitative de l'usage des binômes verbaux en français préclassique : nous avons jusqu'ici laissé de côté la question du genre textuel, or celle-ci est absolument capitale. Dans la tradition de la linguistique de corpus, il a par exemple été montré que, dans un contexte d'apprentissage, les variations liées au registre (une notion légèrement différente de celle de genre textuel, mais soulevant des écueils méthodologiques très similaires) étaient largement prépondérantes par rapport à la langue mère des scripteurs (Larsson *et al.* 2021), ou encore, que les scripteurs étaient capables d'adapter largement les caractéristiques personnelles de leur production écrite à différents registres (Larsson *et al.* 2024). Dans la même veine, en diachronie, la variation diatopique est considérée comme de moindre importance comparée à la variation entre registres et situations sociolinguistiques (Glessgen & Schøsler 2018). L'écueil que peut représenter une composition peu représentative du corpus en termes de genres et les interprétations erronées qui peuvent découler d'une mauvaise composition du corpus en termes de genre textuel a par ailleurs été souligné (Prévost 2015 : 31), de même que l'impact de la donnée du genre sur la dynamique du changement (Ayres-Bennett 2025).

La sensibilité du recours aux binômes au genre textuel a d'ailleurs été explicitement remarquée. Ainsi, Legrand (2022 : 12 et 16) souligne que le genre du récit de voyage (et plus généralement de la chronique) motive le recours aux binômes (ici nominaux), et notamment à certaines formules stéréotypées. Siouffi (2012) évoquait déjà la possibilité que des binômes récurrents soient spécifiquement associés à des genres discursifs donnés. Aussi peut-on s'interroger plus généralement si l'emploi de nos binômes verbaux (ici pris dans leur ensemble plutôt qu'en référence à tel ou tel binôme particulier) est caractéristique de genres discursifs spécifiques.

Dans le corpus Frantext, cette question se heurte néanmoins à un problème méthodologique majeur, à savoir la variation inter-individuelle considérable entre les textes ; ainsi, le nombre

d'occurrences individuelles propre à chaque texte va de 517 occurrences à 0, ou rapporté à la taille des textes en termes de fréquence par millions de mots, de 2923 à 0 (le maximum étant dû à *De la sagesse* de Pierre Charron). Si l'on se restreint aux binômes retenus dans la section précédente, le maximum est de 63 occurrences pour l'*Histoire des plantes* de Rembert Dodoens et en termes de fréquence, 1629 occurrences par millions de mots pour l'épître dédicatoire de *Théodore, vierge et martyre* de Pierre Corneille. Ce dernier résultat est un artefact statistique : l'épître ne comporte qu'une seule occurrence d'un binôme verbal (*suis et serai*), mais ne compte que 614 mots, d'où une fréquence par millions de mots disproportionnée. Ce point illustre bien l'hétérogénéité du corpus et les difficultés qu'il pose pour un traitement quantitatif.

Pour passer outre ce problème et évaluer la sensibilité des binômes aux genres textuels, nous proposons une méthode en deux temps : d'abord, nous calculons une propension individuelle à l'usage des binômes, qui tient compte à la fois du nombre d'occurrences et de la taille du texte. Ensuite, nous menons une analyse par régression linéaire pour évaluer l'effet du genre sur cette propension individuelle. En prenant pour variable dépendante la propension individuelle, comprise entre 0 et 1, nous gommons les trop fortes disparités de fréquence mentionnées plus haut qui biaiseraient drastiquement la régression linéaire. Une autre possibilité pourrait consister à prendre le logarithme de la fréquence, qui en amortit les disparités, mais compte tenu de ce que le logarithme de 0 n'est pas défini et que 46 textes (soit 13% du corpus) ne présentent aucune occurrence de binôme verbal, cela reviendrait à exclure une partie des données, alors même que l'absence de tout binôme verbal est en soi informative d'une certaine spécificité stylistique.

5.1. Propension d'usage individualisée

Pour calculer la propension individuelle de l'usage des binômes verbaux, nous proposons de comparer le nombre d'occurrences observé dans chaque texte à la distribution aléatoire qui régirait ce nombre si les occurrences étaient réparties de manière aléatoire entre les textes. Cette distribution présente l'avantage d'obéir à une forme mathématique exacte, la distribution hypergéométrique (Lafon 1980). Considérant le nombre total d'occurrences de binômes verbaux (7682), la taille totale du corpus (15M de mots), et la taille d'un texte (par exemple les 614 mots de l'épître), on peut déduire de cette distribution la probabilité d'observer k occurrences de binômes verbaux, ou d'en observer plus ou moins que ce nombre. Nous définirons alors la propension d'usage du binôme pour un texte donné comme la probabilité d'observer moins d'occurrences que les k occurrences

observées, soit, avec F la fonction cumulative associée à la distribution hypergéométrique :

$$P(X < k) = \begin{cases} F(k-1)\ si\ k > 0 \\ 0\ si\ k = 0 \end{cases}$$

Pour l'épître dédicatoire de Corneille, cela revient à considérer la probabilité de n'observer aucune occurrence, soit 0,73. Pour donner un second exemple, considérons l'*Infidèle confidente* de Pichou (1631), avec 4 occurrences de binômes verbaux, pour un total de 17 955 mots. La propension à utiliser des binômes est alors proche de 0,02 – le nombre moyen d'occurrences attendu pour un texte de cette taille est de 9, soit plus du double de ce qui est observé.

En Figure 3, nous avons rapporté la distribution de ce score de propension (figure de gauche). Cette distribution est particulièrement frappante car particulièrement concentrée sur les scores extrêmes, proches de 1 et proches de 0. Cela signifie que la distribution des occurrences de binômes verbaux est très hétérogène, et que celles-ci sont essentiellement le fait d'un petit nombre de scripteurs : ainsi, les 50 textes (14% du corpus) qui présentent un score de propension supérieur à 0,95 contribuent à hauteur de 57% des occurrences de binômes verbaux. Si l'on applique par ailleurs la correction de Bonferroni pour un critère de significativité particulièrement restrictif, le seuil étant alors relevé de 0,95 à 0,9999, 28 textes (8% du corpus) passent ce critère, totalisant à eux seuls 46% des occurrences de binômes verbaux, soit presque la moitié. Inversement, un grand nombre de textes ne présentent aucune occurrence de binômes verbaux (46 textes, soit 13% du corpus).

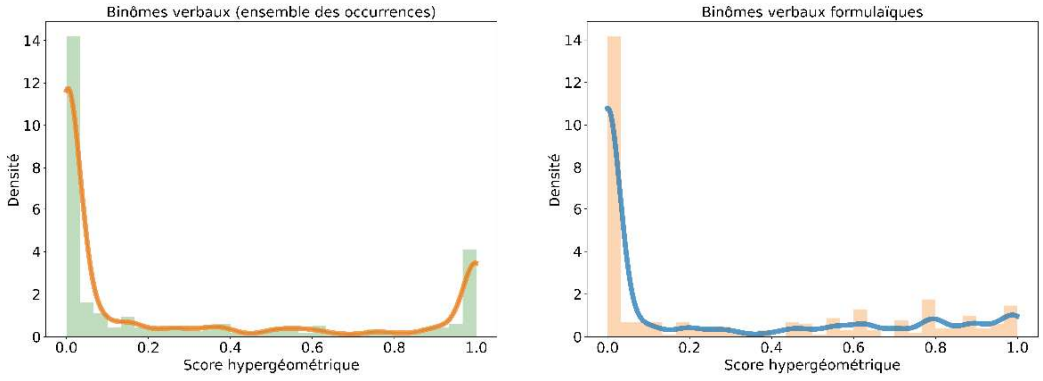

Figure 3 : Distribution à travers les textes du score de propension à utiliser les binômes verbaux compte tenu de la contribution de chaque texte au corpus (gauche) ; distribution à travers les textes du score de propension à utiliser les binômes verbaux formulaïques compte tenu de la contribution de chaque texte à l'ensemble des occurrences de binômes verbaux (droite)

Qu'en est-il à présent des binômes récurrents de nature formulaïque que nous avons identifié dans la section précédente ? Il serait possible de procéder à une analyse similaire, en ne tenant compte que des 846 occurrences de ces binômes en place des 7 682 occurrences totales de binômes verbaux. Cependant, la probabilité que ces occurrences se rencontre dans un texte donné dépend déjà de la probabilité de rencontrer une occurrence de binôme verbal, si bien que l'analyse ne ferait que recapturer les tendances générales de celle que nous venons de développer. Pour tenir compte donc de ce que le fait de trouver une occurrence de binôme formulaïque est directement conditionné au fait de trouver plus généralement une occurrence de binôme verbal, nous avons là encore procédé en calculant un score de propension d'usage à l'aide de la distribution hypergéométrique. Considérant une taille totale de 7 862 occurrences (plutôt que la taille du corpus), et 846 occurrences distribuées dans cet ensemble (plutôt que la manière dont les occurrences de binômes verbaux se distribuent dans l'ensemble du corpus, on considère ici la manière dont les occurrences de binômes récurrents se distribuent dans l'ensemble des occurrences de binômes verbaux), on peut calculer à l'aide de la distribution hypergéométrique la probabilité de rencontrer k occurrences de binômes récurrents parmi les N occurrences de binômes verbaux d'un texte donné.

L'asymétrie précédente de la distribution des scores de propension ne se retrouve pas ici, comme on peut s'en rendre compte sur la Figure 3 (figure de droite). Notons par ailleurs que, même si le graphique présente un fort pic de densité à 0, ne trouver aucune occurrence dans un texte n'est pas spécialement significatif : d'après la définition de la propension que nous avons donnée plus haut, un score de 0 peut être atteint alors même que ne rencontrer aucune occurrence de binôme récurrente est l'observation la plus probable. Pour évaluer la significativité de ce faible nombre d'occurrences, il faudrait considérer un autre score, donné par $P(X > k) = 1 - F(k)$. Dans ce cas, seuls deux textes présentent significativement peu d'occurrences formulaïques (sous la correction de Bonferroni) : il s'agit des deux tomes des *Essais* de Montaigne, lequel emploie donc la tournure d'une façon particulièrement créative.

Nous avons rapporté ci-dessous la liste des dix textes présentant la propension la plus marquée, tant pour les binômes verbaux de manière générale (Tableau 2) que pour les binômes formulaïques (Tableau 3). Un point particulièrement intéressant de ces listes est qu'elles ne présentent qu'un seul texte en commun (*L'histoire des plantes*), ce qui montre qu'une propension très élevée à utiliser les binômes verbaux n'implique pas un recours particulièrement prononcé aux binômes récurrents ou formulaïques (de fait, la corrélation entre les deux scores de propension est de -0,03, c'est-à-dire nulle). On notera

par ailleurs la présence de nombreuses traductions parmi les textes montrant une forte propension à utiliser ces binômes formulaïques ; si on peut imputer cette tendance au latin source pour *L'Histoire du monde*, de Pline le Second, les autres traités sont cependant traduits du bas allemand et du grec.

Titre	**Auteur**	**Date**	**# de mots**	**# occurrences**
Traité du vin et du sidre	Julien le Paulmier (traducteur)	1589	42 993	103
Essais t. 2	Michel de Montaigne	1592	135 499	205
Traicté de l'oeconomie politique	Antoine de Montchrestien	1615	153 522	164
Histoire des plantes	Rembert Dodoens (traduction)	1557	230 985	295
Discours de l'estat et des grandeurs de Jésus	Pierre de Bérulle	1623	146 202	309
De la sagesse	Pierre Charron	1601	176 895	517
L'Histoire prodigieuse du docteur Fauste	Pierre-Victor Palma-Cayet	1598	46 267	83
Essais t. 1	Michel de Montaigne	1592	303 139	284
Alector ou Le Coq : histoire fabuleuse : t. 1	Barthélémy Aneau	1560	70 118	96
Institution de la religion chrestienne t. 3	Jean Calvin	1560	217 507	207

Tableau 2 : Liste des 10 textes présentant le score le plus élevé de propension à utiliser la figure du binôme verbal compte tenu de leur contribution générale au corpus

Titre	Auteur	Date	# de binômes	# récurrents
L'Histoire du monde	Antoine du Pïnet (traducteur)	1562	55	23
Histoire des plantes	Rembert Dodoens (traduction)	1557	295	63
L'Astrée t. 4	Honoré d'Urfé	1627	107	30
Lettres t. 7	Nicolas de Pereisc	1637	40	15
Lettres t. 4	Nicolas de Pereisc	1637	18	9
L'Histoire de la décadence de l'Empire grec	Blaise de Vigenère (traducteur)	1577	97	22

Discours de l'estat et des grandeurs de Jésus	Pierre de Bérulle	1623	309	51
L'Astrée t. 2	Honoré d'Urfé	1610	64	15
Nouvelles récréations et joyeux devis	Bonaventure des Périers	1558	16	6
Lettres t. 1	Guy Patin	1649	49	11

Tableau 3 : Liste des 10 textes présentant le score le plus élevé de propension à utiliser des binômes formulaïques compte tenu de leur usage du binôme verbal

5.2. Effet du genre textuel

Nous nous sommes désormais munis d'une métrique permettant d'associer à chaque texte un score qui caractérise la propension de ce texte à recourir au procédé rhétorique du binôme verbal. Ce score, dont les valeurs restent maîtrisées entre 0 et 1, peut désormais servir de point de départ à une analyse de l'effet du genre textuel sur le recours au binôme verbal.

Le principe de l'analyse est simple en ce que celle-ci se résume à une régression multilinéaire classique. Le but du modèle est de prédire la variable dépendante (le score de propension) à partir de la seule donnée du genre. Pour ce faire, nous nous basons sur l'étiquetage de Frantext, où chaque texte est associé à un genre particulier. L'étiquetage de Frantext soulève cependant quelques difficultés : il présente en effet différents niveaux d'étiquetage, allant du plus général (écrits fictionnels *vs* non-fictionnels), au plus particulier (pour le théâtre, par exemple, l'étiquetage distingue la tragédie, la comédie, la tragi-comédie…). Les difficultés et les incohérences associées à cet étiquetage ont notamment été relevées par Ayres-Bennett (2025 : 14).

Idéalement, il conviendrait de choisir un niveau d'étiquetage, et de s'y tenir. Le deuxième niveau, qui distingue les genres principaux (théâtre, romans, essais), paraît convenir, mais les récits de voyage et la correspondance se trouvent alors regroupés sous l'étiquette 'Ecrits personnels'. Nous avons donc pris pour point de départ l'étiquette la plus fine et refléché certaines étiquettes trop spécifiques vers un genre plus général (par exemple le journal et le récit de voyage ont été regroupés avec les mémoires, et les articles scientifiques avec les traités), passant ainsi de 25 étiquettes à 11, incluant l'étiquette 'non renseigné' de Frantext, qui regroupe des textes de nature diverse. Cela permet de disposer d'au moins 5 textes pour chaque étiquette générique.

Le score de propension a été remplacé par sa version z-scorée (centrée par rapport à sa moyenne, et mise à l'échelle via son écart-

type). Chaque texte est codé en fonction de 11 facteurs correspondant à chaque genre, qui valent 1 si le texte appartient au genre correspondant et 0 sinon. Ainsi, le poids que reçoit chaque facteur dans la modélisation multilinéaire renseigne quant au fait que les textes appartenant à un genre donné présentent en moyenne un score de propension plus élevé (poids positif) ou plus faible (poids négatif) que la moyenne globale de ce score.

Le modèle obtenu explique 28% de la variance associée à la variable de propension (ce qui signifie que le genre joue un rôle certain, mais non prépondérant). Dans le Tableau 4, nous avons rapporté l'effet individuel de chaque genre (c'est-à-dire le poids du facteur associé), ainsi que la valeur *p* de significativité (entre parenthèses). Le seuil de significativité utilisé habituellement est $p < 0,05$, mais si la correction de Bonferroni est utilisée pour tenir compte de ce que les effets de multiples genres sont considérés simultanément, ce seuil s'abaisse à $p < 0,004$. Dans ces conditions, seuls trois genres présentent une tendance significative : les écrits scientifiques (en particulier les traités qui composent l'essentiel de cette catégorie) et les mémoires sont associés à une propension significativement plus élevée que la moyenne, tandis que le théâtre est associé à une propension significativement plus faible. Pour les mémoires, cette tendance est en particulier due aux récits de voyage (2 textes seulement).

Genre	# de textes	Propension binômes verbaux	Propension binômes formulaïques	Propension à l'évitement
Correspondance	22	- 0,32 (0,06)	0,29 (0,17)	- 0,41 (0,04)
Discours	12	0,56 (0,02)	- 0,03 (0,95)	- 0,25 (0,35)
Ecrits scientifiques	28	1,15 (< 0,001)	0,13 (0,48)	0,53 (0,003)
Essais	23	0,13 (0,47)	0,16 (0,44)	0,16 (0,42)
Mémoires	5	1,31 (0,001)	0,01 (0,97)	0,84 (0,04)
Nouvelles	5	0,00 (0,99)	0,70 (0,11)	- 0,33 (0,42)
Paratexte	9	- 0,40 (0,007)	0,16 (0,63)	- 0,76 (0,02)
Poésie	45	0,02 (0,83)	- 0,26 (0,08)	0,14 (0,31)
Romans	26	0,20 (0,24)	0,49 (0,01)	0,03 (0,87)
Théâtre	106	- 0,46 (< 0,001)	- 0,13 (0,21)	- 0,30 (0,001)

Tableau 4 : Effet de la donnée du genre textuel sur la propension des auteurs (pour chaque texte indépendamment) à faire usage du binôme verbal, à faire usage de binômes formulaïques, ou à éviter ces derniers

Au-delà du recours aux binômes verbaux, on peut s'interroger sur l'influence du genre quant à l'usage des binômes que nous avons identifiés comme formulaïques, en répétant la même analyse sur le sous-corpus de textes présentant au moins une occurrence de binôme verbal, et en adoptant comme variable dépendante à modéliser la

propension pour un texte d'utiliser un binôme formulaïque, étant donné la contribution du texte à l'ensemble des occurrences de binômes verbaux. Ici, la donnée du genre n'explique que 6% de la variance de la propension à utiliser les binômes formulaïques ; qui plus est, aucun genre ne présente une sensibilité significative pour cet usage. Globalement, le faible nombre de textes, et le faible nombre d'occurrences formulaïques (moins de 3 occurrences par texte du sous-corpus en moyenne) ne permet pas de tirer de conclusion marquée.

Cependant, si l'on considère cette fois le score de propension à éviter les binômes formulaïques, le modèle se révèle un peu meilleur (17% de variance expliquée), et deux résultats sont significatifs. Tout d'abord, les écrits scientifiques présentent une tendance marquée à éviter les binômes formulaïques. Ensuite, le théâtre est associé à un faible évitement des binômes formulaïques ; cela peut sembler surprenant, mais révèle un biais dans le calcul de la propension à l'évitement : le nombre d'occurrences de binômes verbaux dans les textes dramaturgiques étant significativement faible, on l'a vu, il reste probable de ne trouver aucune occurrence de binôme formulaïque dans un tel texte. Même si aucun texte dramaturgique ne présentait d'occurrence formulaïque, le score d'évitement moyen pour le genre théâtre serait ainsi de 0,40. A l'inverse, si aucun écrit scientifique ne présentait d'occurrence formulaïque, le score de propension à l'évitement moyen serait de 0,89. Seul le premier résultat est donc pertinent.

Pour conclure, les écrits de nature scientifique ou académique (en particulier les traités) et les mémoires (incluant les récits de voyage) sont associés à un recours plus marqué aux binômes verbaux, alors que le théâtre les emploie significativement peu, et cela tient au niveau des sous-genres également. Par ailleurs, aucun genre ne se démarque par sa sensibilité aux binômes formulaïques identifiés plus haut, même si les écrits de nature scientifique ou académique ont significativement tendance à ne pas les exploiter. Ces tendances statistiques ne doivent cependant pas masquer la forte variabilité interindividuelle, qui domine largement les effets de genre ; ainsi parmi les écrits scientifiques (hors traduction), trouve-t-on *Les Dialogues de Guy de Brués contre les nouveaux académiciens* de 1557, avec un score de propension pour les binômes formulaïques de 0,82, à côté de *La pyrotechnie, ou L'art du feu* de 1556, avec un score quasiment minimal de 0,03.

6. Conclusion

Notre article s'est attaché à proposer une analyse qualitative et statistique du paradigme des binômes verbaux en français préclassique, époque à laquelle ce procédé rhétorique est

caractérisé par une forte popularité, néanmoins déclinante. Nous avons d'abord établi certaines propriétés structurales du système, en montrant que les deux jeux de verbes mis en relation par la construction binomiale se distinguent par leurs contenus et par la hiérarchie organisationnelle qu'ils imposent à leurs membres. Cette distinction permet d'assurer une plus grande richesse combinatoire pour les binômes, ce que caractérise en particulier une plus grande productivité et une plus grande information mutuelle entre les deux jeux de verbes. En même temps, cette combinatoire est plus contrainte qu'une libre association entre les deux jeux de verbes, ce qui limite certes la productivité, mais là encore permet une plus grande informativité du système : en d'autres termes, le système est organisé de manière non-triviale, bien différent d'associations arbitraires, et de sorte à permettre à la fois une large variété de binômes, et une certaine codification de ce qui peut faire binôme.

Nous avons en outre mis en place une méthode permettant de sélectionner les binômes les plus pertinents en comparant leur fréquence observée à la probabilité de cette fréquence sous l'hypothèse d'une association libre, conditionnée à l'occurrence d'au moins un exemplaire de ce binôme. Cette méthode nous a permis de mettre en évidence une centaine de binômes bien établis dans l'usage, qui illustrent la palette d'emplois associés à ce procédé : synonymie à des fins d'insistance et d'emphase, complémentarité ou séquentialité permettant une plus grande finesse d'expression, antonymie ouvrant à une large variété d'effets de sens. Dans toutes ces catégories, on rencontre également des occurrences plus formulaïques, qu'elles relèvent de la prose religieuse, du code de la correspondance, ou d'unités quasi-lexicalisées, comme *aller et venir*. La richesse de ces unités formulaïques illustre bien que, en parallèle de la créativité offerte par ce procédé rhétorique, le langage reste intimement structuré par des unités plus enracinées qui permettent justement aux occurrences les plus innovantes, en résonnant avec ces prototypes, de fonctionner et de faire sens.

Références bibliographiques

ATILF (1998-2025), *Base textuelle Frantext*, ATILF-CNRS & Université de Lorraine ; https://www.frantext.fr/

Ayres-Bennett, W. (2025), "Tracing change through different text types and genres: successes and challenges", *Journal of French Language Studies*, 35, p. e5.

Ayres-Bennett, W., Caron, P. (2016), "Periodization, translation, prescription and the emergence of Classical French", *Transactions of the Philological Society*, 114/3, p. 339-390.

Baayen, R.H. (2009), "Corpus linguistics in morphology: morphological productivity", in Lüdeling, A., Kytö, M., *Corpus linguistics. An international handbook*, Mouton De Gruyter, Berlin/Boston, p. 900-919.

Badiou-Monferran, C. (2020), « Sémantique des coordonnants *et*, *ou*, *ni* », in Marchello-Nizia, C., Combettes, B., Prévost, S., Scheer, T. (éds), *Grande Grammaire Historique du français*, De Gruyter Mouton, Berlin/Boston, p. 1654-1678.

Bochkarev, V. V., Shevlyakova, A .V., Solovyev, V. D., Rakhilina, E. V., Paramei, G. V. (2023), "Linguistic mechanisms of colour term evolution: A diachronic investigation of 'Russian browns' buryj and koričnevyj", *Diachronica*, 40/4, p. 492-531.

Buridant, C. (1980), « Les binômes synonymiques. Esquisse d'une histoire des couples de synonymes du Moyen Age au XVIIe siècle », *Bulletin du Centre d'Analyse du discours*, 4, p. 5-76.

Buridant, C. (2003), « Le rôle des traductions médiévales dans l'évolution de la langue française et la constitution de sa grammaire », *Médiévales*, 45, p. 67-84.

Doualan, G. (2014), « Eléments pour une lecture de l'histoire de la synonymie », *SHS Web of Conferences*, 8, p. 409-424.

Ellis, N.C. (2012), "Formulaic language and second language acquisition: Zipf and the phrasal teddy bear", *Annual review of applied linguistics*, 32, p. 17-44.

Ellis, N. C., Ferreira-Junior, F. (2009), "Construction learning as a function of frequency, frequency distribution, and function", *The Modern language journal*, 93/3, p. 370-385.

Ellis, N. C., O'Donnell, M. B., Römer, U. (2014), "Does Language Zipf Right Along?", in Connor-Linton, J., Wander Amoroso, L. (éds), *Measured language : quantitative studies of acquisition, assessment, and variation*, Georgetown University Press, Washington D. C., p. 33-50.

Evert, S. (2004), "A simple LNRE model for random character sequences", in Purnell, G. *et al.* (éds), *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2004)*, Presses universitaires de Louvain, Louvain-la-Neuve, p. 411-422.

Fagard, B., Combettes, B. (2013), « De *en* à *dans*, un simple remplacement? Une étude diachronique », *Langue française*, 178, p. 93-115.

Feltgen, Q. (2022), « Ce que les variations de fréquence nous apprennent des changements linguistiques: le cas de la construction *en plein N* », *Langue française*, 215, p. 61-80.

Flach, S. (2021), "From *movement into action* to *manner of causation* : changes in argument mapping in the *into*-causative", *Linguistics*, 59/1, p. 247-283.

Gaeta, L., Ricca, D. (2006), "Productivity in Italian word formation: A variable-corpus approach", *Linguistics*, 44/1, p. 57-89.

Glessgen, M., Schøsler, L. (2018), « Repenser les axes diasystématiques : nature et statut ontologique », in Glessgen, M., Kabatek, J.,Völker, H. (éds), *Repenser la variation linguistique, Actes du Colloque DIA IV*, Société de Linguistique Romane / Editions de Linguistique et de Philologie, Strasbourg, p. 11-52.

Goldberg, A. E., Casenhiser, D. M., Sethuraman, N. (2004), "Learning argument structure generalizations", *Cognitive Linguistics*, 15/3, p. 289-316.

Koplenig, A. (2018), "Using the parameters of the Zipf-Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis", *Corpus Linguistics and Linguistic Theory*, 14/1, p. 1-34.

Lafon, P. (1980), « Sur la variabilité de la fréquence des formes dans un corpus », *Mots. Les langages du politique*, 1/1, p. 127-165.
Larsson, T., Biber, D., Hancock, G. R. (2024), "On the role of cumulative knowledge building and specific hypotheses: The case of grammatical complexity", *Corpora*, 19/3, p. 263-284.
Larsson, T., Paquot, M., Biber, D. (2021), "On the importance of register in learner writing: A multi-dimensional approach", in Seoane, E., Biber, D. (éds), *Corpus-based approaches to register variation*, John Benjamins Publishing Company, Amsterdam/Philadelphia, p. 235-258.
Lauwers, P., Duée, C. (2010), « *Se faire/hacerse*+ attribut: Une étude contrastive de deux semi-copules pronominales », *Romance Philology*, 64/1, p. 99-132.
Leclercq, O. (2008), « Le rôle des manuels d'enseignement du français langue étrangère dans la construction du lexique aux XVIe et XVIIe siècles », *Éla*, 150/2, p. 195-206.
Legrand, R. (2022), « "Des choses aussi bigerres et prodigieuses" : usages et fonctions des binômes dans l'*Histoire d'un voyage faict en la terre du Brésil* de Jean de Léry (1580) », *Loxias*, 79, en ligne.
Loiseau, S. (2015), « Les différentes formes de la fréquence textuelle: proposition d'inventaire », *Langages*, 197, p. 5-21.
Martí Solano, R. (2024), « Répétition, analogie et variation lexicale comme sources des binômes phraséologiques en anglais », *Lexis*, 24, en ligne ; DOI: https://doi.org/10.4000/12cvq
Masini, F. (2006), "Binomial constructions: inheritance, specification and subregularities", *Lingue e linguaggio*, 2, p. 207-232.
Masini, F. (2016), "Binominal constructions in Italian of the *N1-di-N2* type: towards a typology of Light Noun Constructions", *Language sciences*, 53, p. 99-113.
Mazziotta, N. (2012), « Approche dépendancielle de la coordination des compléments du verbe en ancien français », *SHS Web of Conferences*, 1, p. 187-199.
Norrick, N. R. (1988), "Binomial meaning in texts", *Journal of English Linguistics*, 21/1, p. 72-87.
Papahagi, C. (2021), « Binômes et polynômes dans la Chançun d'Willame », *Studia Universitatis Babeş-Bolyai. Philologia*, 66/1, p. 173-190.
Petrequin, G. (2009), « La synonymie au XVIIe siècle: une évolution conceptuelle et pragmatique », *Pratiques*, 141-142, p. 79-97.
Prévost, S. (2015), « Diachronie du français et linguistique de corpus: une approche quantitative renouvelée », *Langages*, 197, p. 23-45.
Römer, U., Garner, J. (2019), "The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency", *International Journal of Learner Corpus Research*, 5/2, p. 207-230.
Siouffi, G. (2012), « Les binômes synonymiques et la question de la figure au XVIIe siècle: Quelques investigations dans l'usage et dans les remarques », in Berlan, F., Berthomieu, G. (éds), *La synonymie*, Presses de la Sorbonne, Paris, p. 367-379.
Sommerer, L., Baumann, A. (2021), "Of absent mothers, strong sisters and peculiar daughters: The constructional network of English NPN constructions", *Cognitive Linguistics*, 32/1, p. 97-131.
Stefanowitsch, A., Flach, S. (2017), "The corpus-based perspective on entrenchment", in Schmid, H.-J. (ed.), *Entrenchment and the Psychology of*

Language Learning: How We Reorganize and Adapt Linguistic Knowledge, De Gruyter Mouton, Washington D. C., p. 101-128.

Stosic, D., Fagard, B. (2019), « Les prépositions complexes en français: Pour une méthode d'identification multicritère », *Revue romane*, 54/1, p. 8-38.

Thorel, M. (2010), « Synonymie lexicale et niveaux de style à la Renaissance: la traduction française du *Libro del Peregrino* », *Synergies Italie*, 6, p. 25-33.

Van den Heede, M., Lauwers, P. (2023), "Syntactic productivity under the microscope: the lexical and semantic openness of Dutch minimizing constructions", *Folia Linguistica*, 57/3, p. 723-761.

Van Hulle, S., Enghels, R., Lauwers, P. (2025), "The many guises of productivity: A case-study of Spanish inchoative constructions", *Linguistics*, 63/5, p. 1225-1263; https://doi.org/10.1515/ling-2023-0087

Van Wettere, N. (2022), "The hapax/type ratio: An indicator of minimally required sample size in productivity studies?", *International Journal of Corpus Linguistics*, 27/2, p. 166-190.

Vigier, D., Kahng, G. (2022), « Catégoriser les prépositions complexes en français », *SHS Web of Conferences*, 138, en ligne ; https://doi.org/10.1051/shsconf/202213804003

Willems, M. (2003), « Les binômes synonymiques au XIVe et au XVIe s. Étude comparée d'un procédé traductif et stylistique dans deux états d'un même texte », in Sánchez Miret, F. (éd.), *Actas del XXIII Congreso Internacional de Lingüística y Filogía Románica*, De Gruyter Mouton, Berlin/Boston, p. 415-429.

Zeldes, A. (2012), *Productivity in argument selection*, De Gruyter Mouton, Berlin/Boston.