Analyse stylométrique
de l'œuvre de Claude Yvon

La stylométrie, ou sémiométrie, est l'analyse quantitative de textes par des méthodes statistiques. L'objectif de ce type d'analyse est de caractériser quantitativement le style d'un auteur. Ces méthodes sont couramment utilisées pour identifier les auteurs d'articles de journaux ou de blogs anonymes sur internet. Dans le domaine académique, elles sont utilisées pour déterminer si une œuvre anonyme peut être attribuée à un seul auteur, ou à un auteur en particulier, ou encore pour déterminer si le style d'un auteur varie au cours de sa carrière. Nous avons parcouru la littérature spécialisée, ouvrages et revues, pour identifier les méthodes qui pourraient nous être utiles pour caractériser quantitativement le style de l'abbé Yvon ; certaines de ces méthodes nous ont inspiré à en créer d'autres.

Notre objectif est ici d'identifier les articles de l'abbé Yvon parmi les articles anonymes de l'*Encyclopédie* de Diderot, et parmi les articles non signés publiés dans le *Journal Encyclopédique* de Pierre Rousseau. Et il y a, bien sûr, l'*Apologie de Monsieur l'Abbé de Prades*, dont la deuxième partie aurait été écrite par Yvon, et la troisième par Diderot.

Toutefois, avant de tenter de déterminer si certains des textes anonymes attribuables à Yvon sont bien de lui, il nous faut examiner si les différents critères stylométriques que nous avons relevés ou créés dans ce but permettent effectivement de distinguer les œuvres d'Yvon de celles d'autres auteurs.

1. La base de données

Pour tester l'efficacité de nos critères stylométriques à distinguer par leurs style des auteurs différents, nous avons constitué une base de données de textes de l'Abbé Yvon et de trois autres auteurs de son époque, Diderot, l'abbé Bergier et l'abbé Raynal. Des textes de l'abbé de Prades nous auraient été fort utiles ; par malchance, il n'a pas publié d'ouvrages après sa fameuse thèse.

Nous avons trouvé sur internet la quasi-totalité des œuvres de Claude Yvon. Les seuls textes bruts absents d'internet sont les deux *Lettres à M. Rousseau* (1763), que nous avons dû saisir entièrement, et *l'Accord de la philosophie avec la religion (1776),* dont nous n'avons même pas trouvé l'ouvrage scanné. Rappelons que *l'Abrégé de l'histoire de l'Eglise (1761-67)* est identique, à peu de choses près, au *Discours* de 1768, et que l'*Accord de la philosophie avec la religion (1782)* est inclus dans l'ouvrage de 1785. Pour constituer la base de données, nous avons téléchargé les textes bruts obtenus par Google par reconnaissance optique de caractères des ouvrages scannés. Nous avons aussi téléchargé les textes bruts obtenus par Gallica, mais ne les avons pas utilisés. En effet, nous avons constaté que le logiciel de Google, qui utilise probablement le contexte fourni par son immense bibliothèque de textes numérisés pour reconnaître les mots, fournit des textes bruts avec moins d'erreurs ou d'omissions que celui de Gallica. En particulier, ce dernier a tendance à omettre les marques de ponctuation, du moins les virgules.

Ces textes bruts étant loin de reproduire correctement le texte imprimé, nous avons dû lire l'ensemble de ces textes pour en corriger les erreurs de reconnaissance, ce qui représente un travail considérable, mais nécessaire. Ensuite, plutôt que de faire des relectures fastidieuses de l'ensemble des textes, nous avons extrait de chaque texte tous les mots différents, que nous avons ensuite rangés par ordre alphabétique. Les erreurs ressortent ainsi aisément, le mot erroné n'étant pas à sa bonne place, ou le caractère aberrant apparaissant de façon évidente. Nous avons fait des recherches systématiques de certains mots comme *fait*, qui parfois devrait être *sait*, ou vice versa, car le

problème majeur de ces textes réside dans la ressemblance des caractères typographiques pour les lettres *f* et *s*. Nous avons ensuite supprimé tous les titres et sous-titres, les références bibliographiques, les tables des matières, les errata, pour ne conserver que les phrases du texte proprement dit. Nous n'avons généralement pas uniformisé l'orthographe, qui n'est pas stabilisé au dix-huitième siècle ; en particulier les mots ne sont pas toujours accentués de la même manière. Nous avons corrigé les fautes typographiques évidentes. Nous avons supprimé toutes les citations d'autres auteurs, souvent longues de plusieurs pages chez Yvon. Nous avons cependant laissé quelques citations courtes, lorsqu'elles sont insérées dans une phrase, et participent au rythme de lecture. La plupart de ces manipulations de textes ont été faites avec des procédures rédigées en langage Perl, qui est bien adapté au traitement automatique de textes.

Pour l'analyse stylométrique, il est recommandé de lemmatiser les textes. Cela consiste à ramener les formes élidées à la forme sans élision, les formes verbales à l'infinitif, les substantifs au singulier, les adjectifs au masculin singulier. Une telle procédure est importante pour déterminer la richesse de vocabulaire des textes, qui peut éventuellement caractériser un auteur. Nous n'avons pas lemmatisé nos textes, parce que c'est un travail considérable ; il aurait fallu pour cela uniformiser l'orthographe et distinguer tous les cas ambiguës, comme par exemple les noms *être* et *fait* des verbes correspondants. Nous aurions pu effectuer des études de richesse de vocabulaire sur les textes non lemmatisés, en supposant que les élisions, variations d'orthographes, etc, ont à peu près le même effet sur les textes des différents auteurs. Mais c'est ignorer les effets d'accentuation aléatoire produits par la reconnaissance de caractères sur des textes mal imprimés, ou la fiabilité variable des correcteurs de l'époque. Ainsi, nous n'avons pas utilisé la richesse du vocabulaire comme un critère stylométrique, nous l'avons tout au plus utilisé à titre indicatif.

Nous avons jugé que l'élision affecte tous les textes à peu près de la même manière, et qu'il n'est donc pas indispensable de ramener les mots à leur forme non élidée pour étudier la distribution du nombre de mots par phrase des différents auteurs. Suivant la même logique, nous n'avons pas non plus supprimé les traits d'union syntaxiques. Nous avons aussi conservés tels quels les mots composés, bien qu'il pourrait être judicieux de séparer ceux composés avec *très* ou *même* pour faire nos études de fréquences de mots.

Les textes de l'abbé Yvon ont été complétés par quelques textes de comparaison. Diderot n'ayant pas rédigé d'ouvrage religieux, et ses romans contenant beaucoup de dialogues, nous nous sommes rabattus sur ses articles dans l'*Encyclopédie*. Nous avons concaténé les trois plus longs : *Eclectisme*, *Encyclopédie* et *Philosophie des Juifs*. Dans ce texte, nous avons supprimé toutes les références bibliographiques (généralement données entre parenthèses) et les renvois à d'autres articles de l'Encyclopédie (*Voyez l'article* …) . Nous avons aussi supprimé toutes les phrases ne contenant qu'un numéro. Pour Bergier, nous avons choisi son *Traité de la vraie religion* (1780), les chapitres I à VI de la première partie. L'ensemble de l'ouvrage nous servira aussi pour identifier d'éventuels emprunts de l'abbé Yvon (il y en a). Pour l'abbé Raynal, nous avons d'abord traité son *Histoire philosophique et politique des établissemens & du commerce des européens dans les deux Indes*, livre sixième, mais nous avons en cours de route appris que nombreux auteurs ont participé à cet ouvrage. Ce texte doit donc être traité avec précaution. Comme cet abbé n'a pas écrit d'ouvrage religieux, nous avons ajouté l'*Histoire du divorce d'Henri VIII*.

Le tableau 1 ci-dessous décrit notre base de données. Les chiffres sont donnés à titre indicatif ; ils ne servent qu'à montrer les caractéristiques relatives de ces différents textes.

Tableau 1 : caractéristique de notre base de données

Titre	*Auteur*	*Nombre de mots et de ponctuations*	*Nombre de mots différents*	*Nombre de phrases*	*Nombre de pages A4 (texte brut)*
1754 La Liberté	Yvon	87180		2474	90
1763 Lettres Rousseau	Yvon	50110		1283	68
1768 discours	Yvon	352069		9978	533
1779 Histoire	Yvon	194067		5275	349
1785 Histoire	Yvon	194267		5094	294
3 articles de Encyclopédie	Diderot	102239	13862	2999	139
Traité historique	Abbé Bergier	211337		6680	319
Divorce Henri VIII	Abbé Raynal	25684		738	40
Histoire des deux Indes, livre VI	Abbé Raynal	34608		2395	48

2. La longueur des phrases

La distribution du nombre de mots par phrase a été utilisé avec plus ou moins de bonheur comme critère stylométrique, depuis les débuts de cette science, pour identifier un auteur. Nous ne ferons pas ici de bilan, estimant que, si la validité de ce critère n'est pas universellement reconnue, elle peut l'être dans des cas particuliers, en appui à d'autres critères.

Cette distribution montre toujours une asymétrie positive, avec une queue plus importante vers les longues phrases, et peut être modélisée, entre autres, par une loi lognormale. En portant sur un graphe la distribution du logarithme du nombre de mots par phrase en fonction de sa fréquence dans un texte, on obtient en général une distribution gaussienne des points, c'est-à-dire une belle courbe en cloche, au moyen de laquelle on peut attribuer une moyenne et un écart-type à la distribution. C'est la loi statistique que nous adopterons. Notons que d'autres lois, comme l'hyper-Pascal ou des variantes de la loi de Poisson sont également possibles.

Avant tout, il faut définir la phrase et sa longueur. Dans notre corpus, elle se termine par un point, un point d'interrogation ou un point d'exclamation. Comme le point est aussi utilisé pour abréger certains mots, comme Monsieur ou Saint, ou pour terminer un chiffre, par exemple ceux qui suivent

le mot chapitre, le nom d' un roi, celui d'un pape, il est nécessaire de neutraliser ces indicateurs avant de découper nos textes en phrases par une procédure automatique. Nous avons concaténé avec la phrase suivante les rares phrases ne comportant qu'un ou deux mots suivis d'un point d'exclamation, ou plus rarement d'interrogation, comme par exemple *Ho !* ou *Mais quoi !*. Ceci ne concerne que les textes d'Yvon.Nous définissons la longueur de la phrase comme le nombre de mots distincts des signes de ponctuation dans la phrase.

Il faut savoir que nous avons concaténé un certain nombre de noms propres composés, que nous avons comptés comme un seul mot, comme par exemple les noms précédés d'un prénom, les M. Diderot, les Saint Paul, les P. Buffier, etc. En pratique, ces modifications ne concernent que quelques centaines de mots, et sont sans effet notable sur la statistique.

Nous avons concaténé les cinq textes d'Yvon, et ajusté une loi lognormale à la distribution des longueurs de phrase sur ce texte global de 24105 phrases. La courbe ajustée, tracée en rouge sur les figures 1 et 2, a une moyenne de 30,88 mots par phrase et un écart-type de 20,33 mots.
La figure 1 montre la comparaison de la distribution de longueur de phrase dans les cinq ouvrages d'Yvon avec le modèle ; en portant en abscisses le logarithme népérien du nombre de mots par phrase, la courbe moyenne est une gaussienne. Les phrases des cinq ouvrages d'Yvon suivent relativement bien cette courbe moyenne, avec souvent un excès aux phrases de cinq mots ou moins.. On note toutefois de légères différences : un écart-type un peu plus faible dans les ouvrages de 1763 et 1779, un déficit de phrases de 11 et 12 mots dans celui de 1785, et un excès de phrases de 9 et 10 mots dans celui de 1754. Ces deux dernières légères asymétries peuvent se corriger en adoptant une moyenne et un écart-type un peu différents pour ces ouvrages. On peut aussi admettre que ces légères différences ne sont que l'effet de fluctuations statistiques. C'est ce que nous ferons, en notant que les distributions de longues phrases se superposent toutes bien, ce qui ne serait pas le cas si les écarts-types étaient différents.

En revanche, la comparaison de la distribution des phrases de Diderot, Bergier et Raynal avec la courbe modèle révèle des différences notables, comme le montre la figure 2. Chez Diderot, les phrases courtes (2, 3, 5, 6 et 7 mots) sont nettement plus nombreuses que chez Yvon. Chez Bergier, les phrases courtes sont également plus fréquentes ; en outre, la distribution est décalée vers les phrases plus longues. Les phrases de Raynal auteur unique ont une distribution voisine de celles d'Yvon, mais décalée vers les longues phrases.. Les phrases de Raynal et coauteurs ont une moyenne et un écart-type plus faibles.

Cette première analyse démontre que, dans notre corpus, la longueur des phrases est un critère valable pour distinguer des textes d'auteurs différents. C'est une condition suffisante, mais non nécessaire : deux textes ayant la même distribution de longueur de phrase peuvent être d'auteurs différents.

Nous avons remarqué que les textes plus courts ont tendance à avoir un écart-type plus faible, ce qui semble indiquer que la longueur du texte influe sur la distribution de longueur de phrase.

Nous appliquons maintenant ce critère au seul texte assez long pour une telle analyse, et dont l'auteur est douteux, l'Apologie de Monsieur l'Abbé de Prades. La première partie est la thèse elle-même, dont l'auteur est a priori l'abbé de Prades. Dans le traitement préliminaire de cette partie, nous avons supprimé tout le texte latin. En outre, nous avons effectué un lissage glissant sur 3 lignes de données pour atténuer le bruit. La distribution de longueur de phrase est comparée avec la courbe modèle en bas des figures 1 et 2. La distribution pour L'*Apologie 1* (la thèse) ressemble beaucoup à celle pour l'ouvrage de 1785 d'Yvon. La distribution pour L'*Apologie 2* a un excès notable de phrases courtes. Ces deux premières partie de l'*Apologie* ont toutes deux un déficit de phrases de 10 mots. Toutes ces propriétés suggèrent l'influence d'Yvon dans ces deux premières

parties de l'*Apologie*. La figure pour l'*Apologie 3* (Diderot) est trop bruitée pour permettre de conclusion fiable ; nous notons toutefois un excès important de phrases courtes, une caractéristique des textes de Diderot.

Avant de clore cette section, il convient de noter que, la distribution de longueur de phrase ayant une asymétrie positive prononcée, la moyenne ne représente pas la valeur la plus fréquente de la distribution. Si la moyenne de nos texte oscille autour de 30 mots par phrase, le mode de la distribution, c'est-à-dire la valeur la plus probable, a une valeur nettement plus faible, de l'ordre de 18 mots par phrase (sauf pour les oeuvres de Bergier et Raynal).

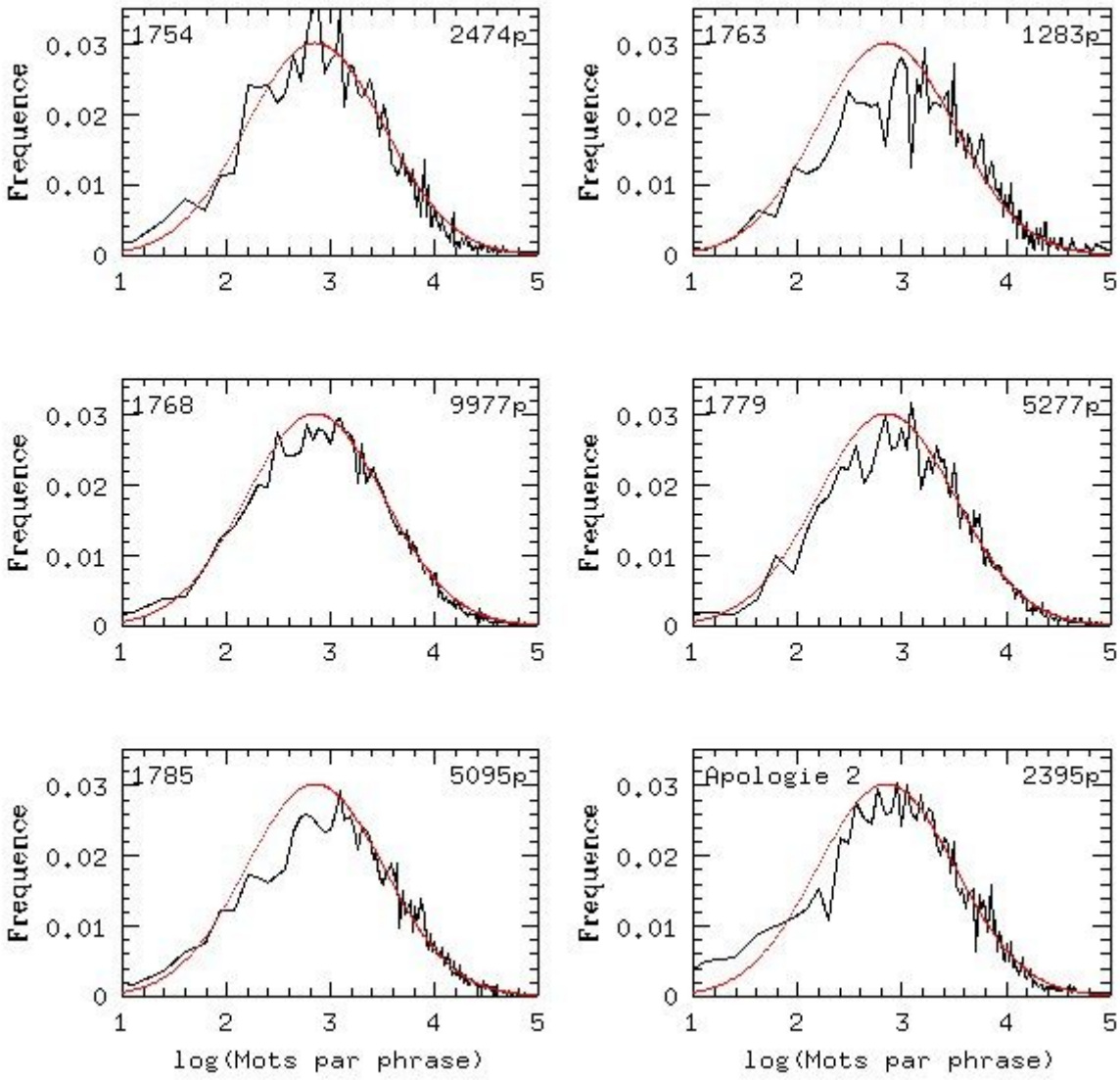

Figure 1. Distribution de la longueur de phrase dans les ouvrages de l'abbé Yvon, et dans la deuxième partie de l'*Apologie de Monsieur l'abbé de Prades*. Le logarithme népérien du nombre de mots par phrase est porté en abscisses, et sa fréquence en ordonnées, normalisée de telle sorte que l'intégrale de la distribution (la surface sous la courbe) soit l'unité. La courbe rouge est la gaussienne ajustée à l'ensemble des 24105 phrases des ouvrages d'Yvon. L'année de l'ouvrage est indiqué en haut à gauche de la figure, et le nombre de phrases dans l'ouvrage en haut à droite.

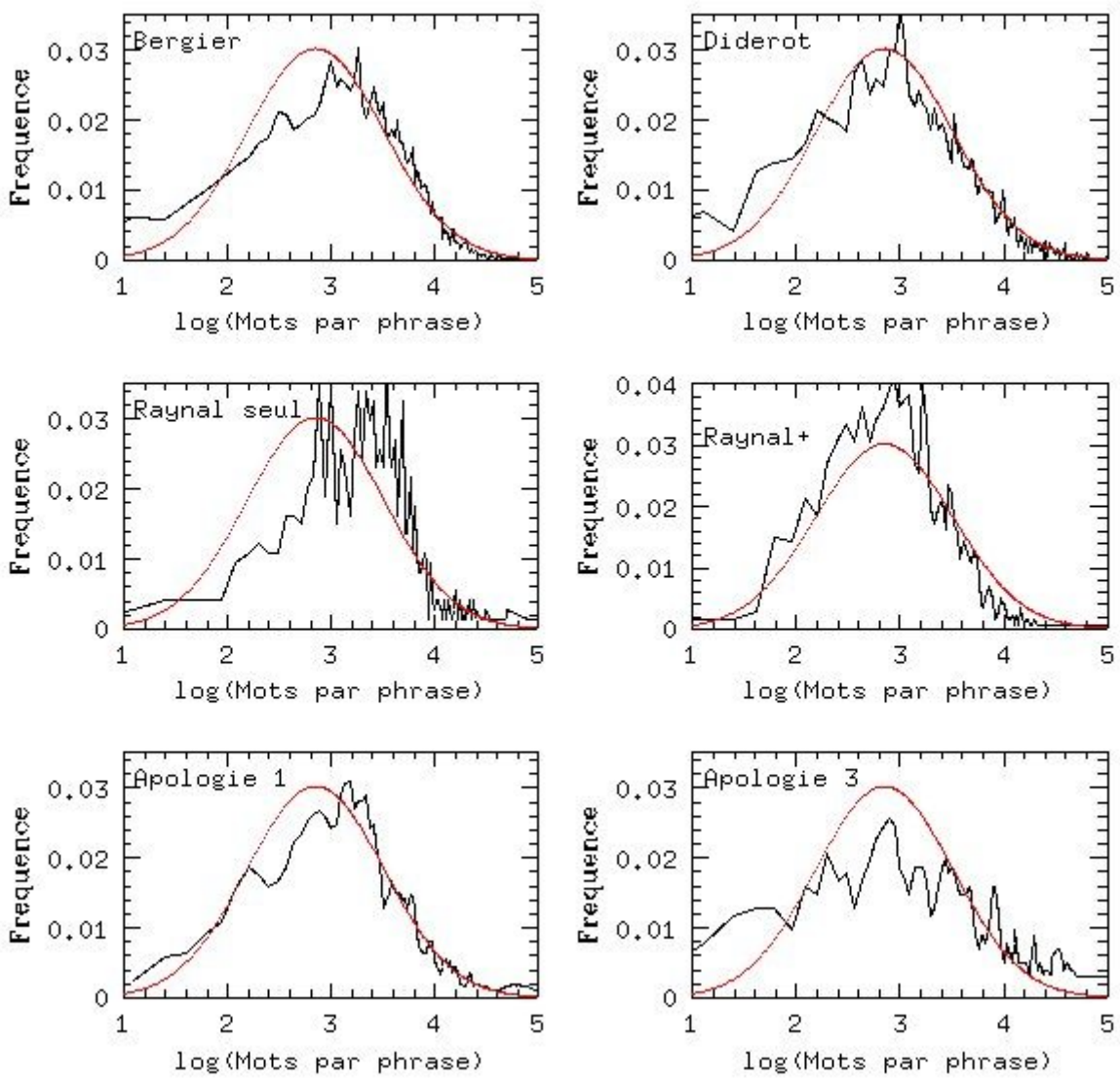

Figure 2. Distribution de la longueur de phrase dans les ouvrages de comparaison, et dans l'*Apologie de Monsieur l'abbé de Prades* (1ère et 3è parties). Le logarithme népérien du nombre de mots par phrase est porté en abscisses, et sa fréquence en ordonnées, normalisée de telle sorte que l'intégrale de la distribution (la surface sous la courbe) soit l'unité. La courbe rouge est la gaussienne ajustée à l'ensemble des 24105 phrases des ouvrages d'Yvon. L'auteur est indiqué en haut à gauche de la figure.

3. La ponctuation

Nous avons compté le nombre de virgules, de points-virgules et de signes de ponctuation dans les phrases, ainsi que la proportion de phrases interrogatives dans notre corpus. Les signes de ponctuation sont la virgule, le point-virgule, le point, le deux-points, les parenthèses et les points d'exclamation et d'interrogation. Nous ne comptons pas le signe de ponctuation final de chaque phrase, uniquement ceux qui sont à l'intérieur de la phrase. La proportion de phrases exclamatives est trop faible pour avoir une valeur statistiquement significative.

Plusieurs méthodes sont possible pour estimer la fréquence d'occurrence d'un signe de ponctuation donné. Une méthode consiste à compter le nombre total d'occurrences du signe, puis de diviser par le nombre total de mots. Nous avons préféré calculer le nombre moyen d'occurrence par phrase, en divisant le nombre de signes de ponctuation par le nombre de mots dans la phrase. Nous ramenons ainsi toutes les phrases à une longueur fixe de 1 mot, et en multipliant ensuite par 100, nous obtenons des pourcentages d'occurrence du signe dans la phrase. Notons que le résultat n'aurait pas été très différent si nous avions adopté la première méthode ci-dessus.

Le tableau 2 ci-dessous donne les résultats. Les chiffres des colonnes 2 à 5 donnent le pourcentage moyen de mots de la phrase qui sont un signe de ponctuation donné. La colonne 6 donne le pourcentage de phrases interrogatives dans le texte, et la dernière, qui est le nombre total de phrase, est indicatrice de la taille du texte.

Le mot *de* est le mot le plus fréquent de la plupart des textes en français. C'est pourquoi nous avons inclus la statistique de ce mot dans le tableau, bien que ce ne soit pas un signe de ponctuation. La deuxième colonne en donne le pourcentage moyen dans la phrase : environ 4 % des mots sont des *de*.

Tableau 2 : pourcentages moyens d'occurrence des signes de ponctuation et du mot *de*

Titre	*de*	*,*	*;*	*Ponctuation*	*?*	*N phrases*
1754	4,632	6,464	0,451	7,005	11,20	2474
1763	4,433	7,645	0,450	8,243	11,61	1283
1768	4,755	6,999	0,402	7,591	7,737	9978
1779	4,764	7,501	0,407	8,063	8,399	5272
1785	4,302	8,045	0,666	9,077	12,90	5094
Tous Yvon	4,632	7,309	0,467	7,983	9,533	24105
Diderot	4,305	6,801	1,322	8,858	7,069	2999
Bergier	4,185	6,724	1,364	8,535	9,835	6680
Raynal seul	4,951	5,964	0,429	6,726	0,5420	738
Raynal et al	3,929	6,146	0,570	7,093	16,58	2395
Apologie-1	4,378	7,030	0,592	7,746	10,04	837
Apologie-2	3,929	6,143	0,570	7,093	16,58	2395
Apologie-3	5,099	6,182	1,145	7,615	20,71	338

Dans ce tableau 2, deux pourcentages moyens insolites ressortent : celui des points-virgule et celui des signes de ponctuation chez Diderot et Bergier. Les deux auteurs utilisent nettement plus de points-virgules que les autres auteurs. Ils utilisent aussi davantage de signes de ponctuation ; toutefois cela s'explique simplement par l'excès de points-virgules chez ces auteurs. On retrouve dans la troisième partie de l'*Apologie* l'excès de points-virgules, un indice supplémentaire que ce texte est de Diderot. En revanche, certains des pourcentages relevés pour la deuxième partie de cet ouvrage s'écartent sensiblement de ceux d'Yvon.

Le pourcentage moyen de la virgule et celui du point d'interrogation sont assez fluctuants dans l'œuvre d'Yvon, et risquent de ne pas être très utiles comme indicateurs stylométriques.

Le pourcentage moyen est un indicateur statistique parmi d'autres de la fréquence de ces chiffres dans un texte. Nous inspirant du critère bien connu de la distribution de longueur de phrase, nous en avons créé un autre : la distribution statistique du nombre de signes de ponctuation dans les phrases. Tout comme la longueur de phrase, dont il n'est pas indépendant, le nombre de signes de ponctuation par phrase n'a pas une distribution symétrique avec une moyenne proche de la valeur la plus probable. La distribution est fortement asymétrique. Nous n'avons pas trouvé de différences significatives dans les distribution d'occurrence de la virgule et du point obtenues pour les différents auteurs du corpus. Ces distributions ne seront donc pas utilisées dans la suite de notre analyse stylométrique.

En revanche, nous avons trouvé des différences significative dans la distribution du mot *de* de nos différents auteurs. Comme pour la longueur de phrase, nous avons normalisé la fréquence du mot (portée en ordonnées) de telle sorte que l'intégrale de la distribution (la surface sous la courbe) soit l'unité.

La figure 3 montre une homogénéité remarquable dans la distribution du mot *de* dans l'œuvre de l'abbé Yvon, qui contraste avec celle des autres auteurs. Il est surprenant que la distribution pour l'*Apologie* 3 diffère significativement de celle pour Diderot. On note que la distribution de la première partie de l'*Apologie* ressemble fort à celle d'Yvon, alors que celle de la deuxième partie s'en écarte sensiblement, pour se rapprocher de celle de Bergier. Les différences sont les plus notables pour de faibles occurrences du nombre par phrase, c'est pourquoi nous avons limité la figure à des abscisses inférieures à 5.

Les indicateurs stylométriques à retenir à l'issue de cette analyse sont le pourcentage moyen de points-virgules dans la phrase et la distribution du mot *de* dans la phrase (en particulier la valeur des fréquences de 0, 1 et 2 occurrences du mot). Nous notons également une certaine inadéquation de la deuxième partie de l'*Apologie* avec le style d'Yvon, alors que la première partie s'en rapprocherait, du moins pour ces indicateurs. Rappelons que la ressemblance de style entre l'*Apologie 1* et l'œuvre d'Yvon porte aussi sur la distribution de longueur de phrase.

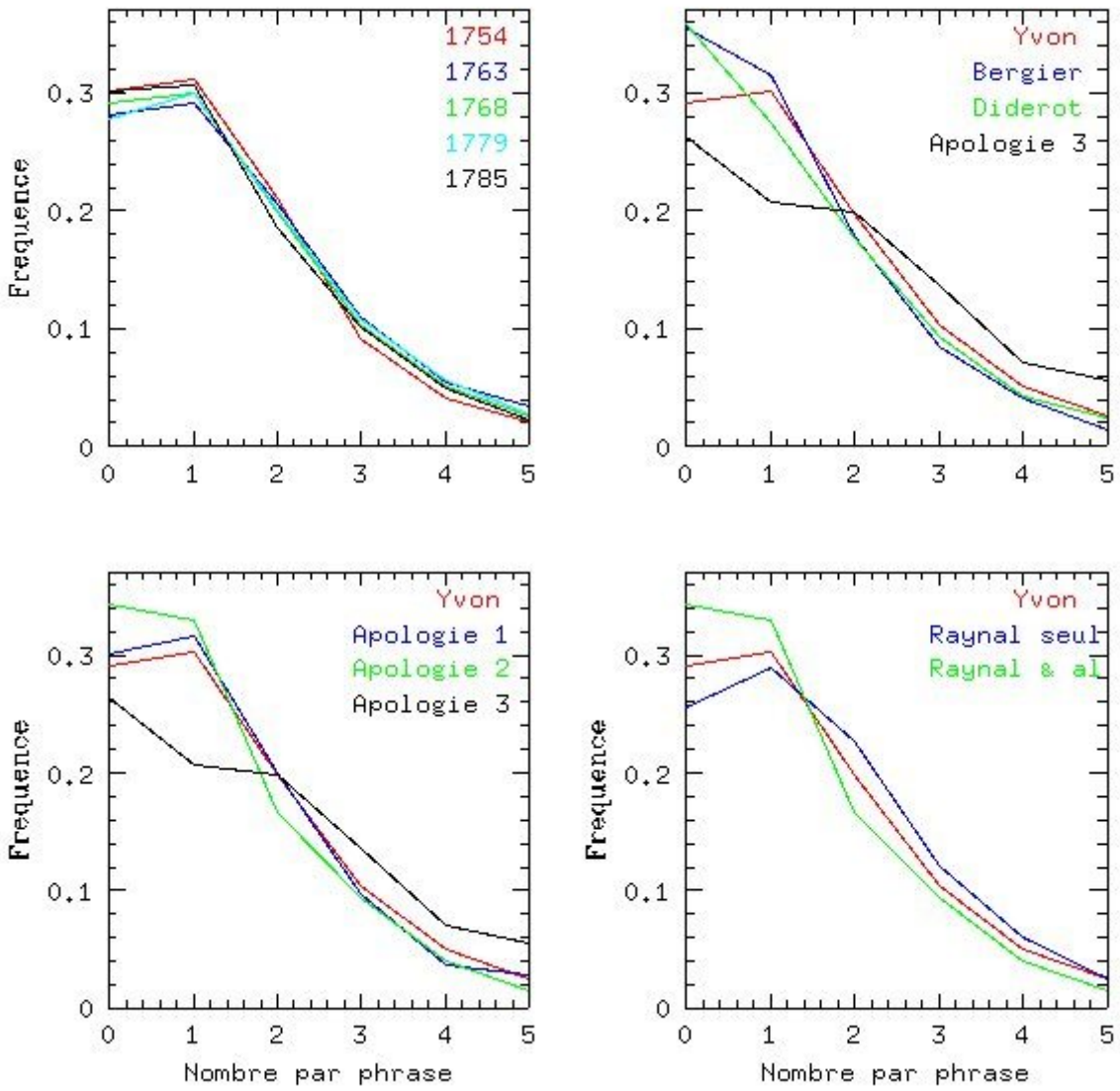

Figure 3. Distribution de la fréquence du mot *de* dans les phrases du corpus. Le nombre d'occurrence du mot *de* par phrase est en abscisses, et sa fréquence en ordonnées, normalisée par l'intégrale de la distribution. Les œuvres d'Yvon sont en haut à gauche, les œuvres de Diderot et Bergier en haut à droite, les trois parties de l'*Apologie* en bas à gauche, les œuvres de Raynal en bas à droite. La distribution moyenne des cinq œuvres d'Yvon est tracée en rouge pour comparaison, dans les trois derniers cas.

4. La longueur des mots

La longueur des mots est un autre critère utilisé en stylométrie. Nous avons à nouveau innové, et, continuant dans la voie tracée plus haut, nous avons déterminé la distribution de la longueur des mots dans les textes de notre corpus. Du point de vue technique, il s'agit de compter le nombre d'octets par mot. Auparavant, nous avons supprimé tous les accents, parce que les caractères accentués sont codés sur deux octets. Rappelons que nous avons concaténé un certain nombre de noms propres, ce qui rallonge un peu la queue de distribution.

La figure 4 montre les résultats. nous avons porté en abscisse la longueur de mot et en ordonnées sa fréquence, et, comme précédemment, nous avons normalisé les distributions de telle sorte que leur intégrale soit l'unité. La distribution de l'ensemble des œuvres d'Yvon est porté en rouge dans les comparaisons avec d'autres auteurs. Nous constatons à nouveau que les distributions pour les œuvres d'Yvon ont toutes à peu près la même distribution, et que celle de l'*Apologie* 1 coïncide avec la distribution moyenne d'Yvon. Il y a une différence marquée entre Yvon et Raynal, mais pas entre Yvon, Diderot et Bergier. Les différences se remarquent au niveau d'un plateau, ou du moins d'une portion de courbe de moindre pente, qui survient aux mots de 6, 7 et 8 caractères.

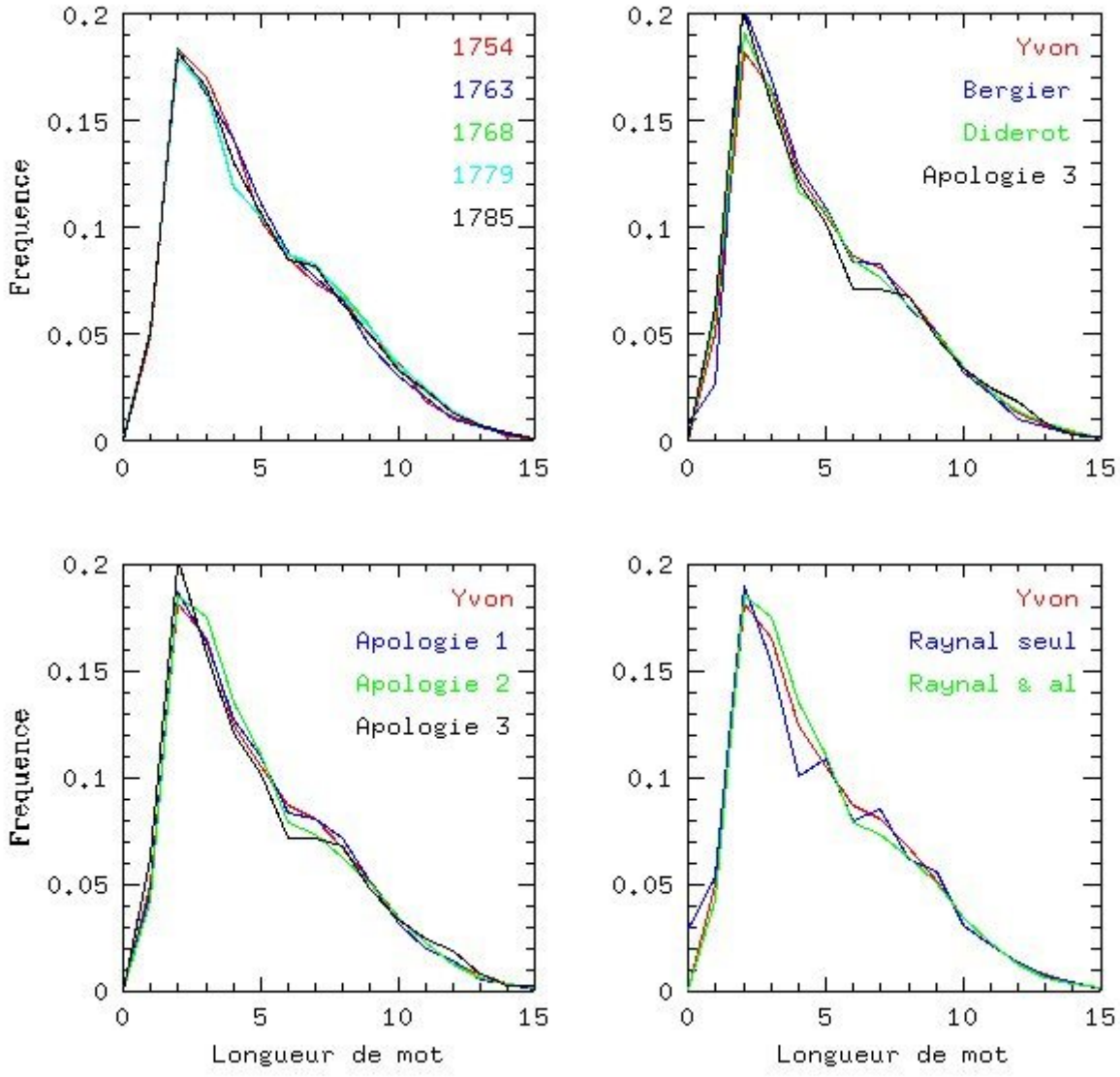

Figure 4. Distribution du nombre de caractères par mot dans les textes de notre corpus. Le nombre de caractères est porté en abscisses, et la fréquence de ce nombre en ordonnées. Les distributions

sont normalisées de telle sorte que leur intégrale soit l'unité.

5. Le premier mot de la phrase

La façon de commencer et de terminer une phrase peuvent caractériser le style d'un auteur. Nous avons rapidement constaté que le dernier mot d'une phrase est, la plupart du temps, contextuel. Dans les textes d'Yvon et de Bergier, ce sont des mots comme *Religion* ou *Dieu*, ce qui n'est pas surprenant, parce que ce sont aussi les mots contextuels les plus fréquents de ces textes. Nous avons donc dû renoncer à utiliser ce critère stylométrique.

En revanche, le premier mot d'une phrase est le plus souvent non contextuel, et nous avons étudié en détail sa distribution. Nous avons déterminé la fréquence des cent mots les plus fréquents dans l'ensemble de l'œuvre d'Yvon. Auparavant, nous avons supprimé de cette liste les premiers mots contextuels les plus fréquents (*Dieu, L'esprit, L'Eglise, L'homme*, *Platon* et *L'Empire)*. Les mêmes premiers mots, ainsi que leur fréquence, ont été extrait des autres textes et mis dans le même ordre. Cet ordre est donné dans le tableau 3 ci-dessous.

Tableau 3. Les cinquante mots non contextuels les plus fréquents dans l'œuvre de l'abbé Yvon

	1	2	3	4	5	6	7	8	9	10
1-10	Il	Les	Mais	Le	La	Ce	C'est	Si	On	Ils
11-20	En	Cette	Ainsi	Je	Or	Comme	Pour	Ces	Dans	Que
21-30	Elle	A	De	Car	Et	Cependant	Un	S'il	Par	Nous
31-40	Tout	Cet	Vous	Quand	Après	Une	Voilà	C'étoit	Au	Pourquoi
41-50	Tous	Comment	Ceux	Quoique	Quant	Son	Qui	Leur	Sous	Tel

Les résultats sont montrés dans la figure 5.

Nous remarquons tout d'abord une évolution stylistique dans l'œuvre d'Yvon. Le mot *Les* (rang 2) est moins souvent utilisé en début de phrase dans les deux premiers ouvrages (1754 et 1763) que dans les suivants, et les mots *C'est* (7), *Si* (8) et *Je* (14), y sont au contraire plus fréquents.

Les trois parties de l'*Apologie* montre pareillement un déficit de *Les* et un excès de *Je*. C'est pourquoi, dans la figure 5, nous avons comparé ces fréquences avec l'ouvrage de 1754 d'Yvon, plutôt qu'avec l'ensemble de son œuvre. L'excès de mots *Je* dans l'*Apologie* est certainement un effet de style polémique.

Les différences entre Yvon et Bergier et Diderot résident dans l'usage de *Il*, nettement plus fréquent chez Diderot, et nettement moins chez Bergier. Quant à Raynal, contrairement aux autres, il ne commence jamais une phrase avec *Mais, C'est* ou *En*.

Le premier mot de phrase s'avère donc un critère stylométrique utile, à condition de tenir compte de l'évolution stylistique d'Yvon dans ce domaine.

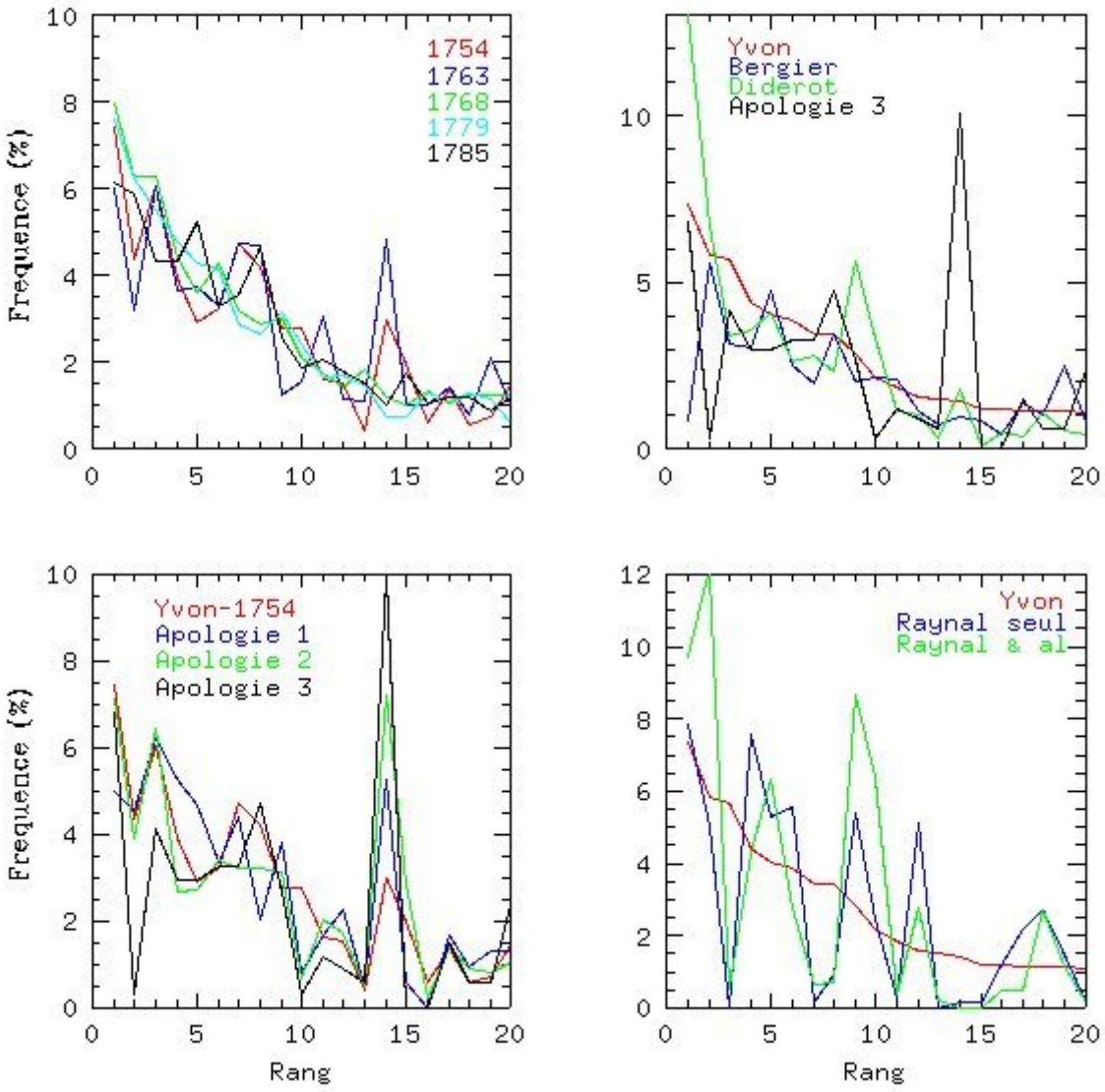

Figure 5. Fréquence des premiers mots de phrase (en pourcents) en fonction du rang. La correspondance entre rang et mot est donnée dans le tableau 3. Ainsi, l'excès au rang 14 dans les trois parties de l'Apologie est un excès de *Je*. La fréquence des premiers mots de l'ensemble de l'oeuvre d'Yvon est tracée en rouge dans les comparaisons avec d'autres auteurs, sauf pour l'*Apologie*, où la référence est son ouvrage de 1754 (à cause de l'évolution stylistique).

6. Les mots non-contextuels dans le texte

Le dernier paramètre stylométrique que nous utiliserons est la fréquence des mots non-contextuels dans les textes. Un certain nombre de méthodes innovantes ont été proposées dans la littérature spécialisée pour étudier la distribution statistique des mots non-contextuels, que nous n'avons pas jugé nécessaire de mettre en œuvre. Nous avons préféré adopter la méthode que nous avons créée pour l'analyse des débuts de phrase, à ceci près que nous ne tenons pas compte de la casse (les majuscules sont remplacées par des minuscules). En outre, nous avons supprimé les mots *être* (rang 60) et *bien* (rang 69), qui peuvent être un nom contextuel.

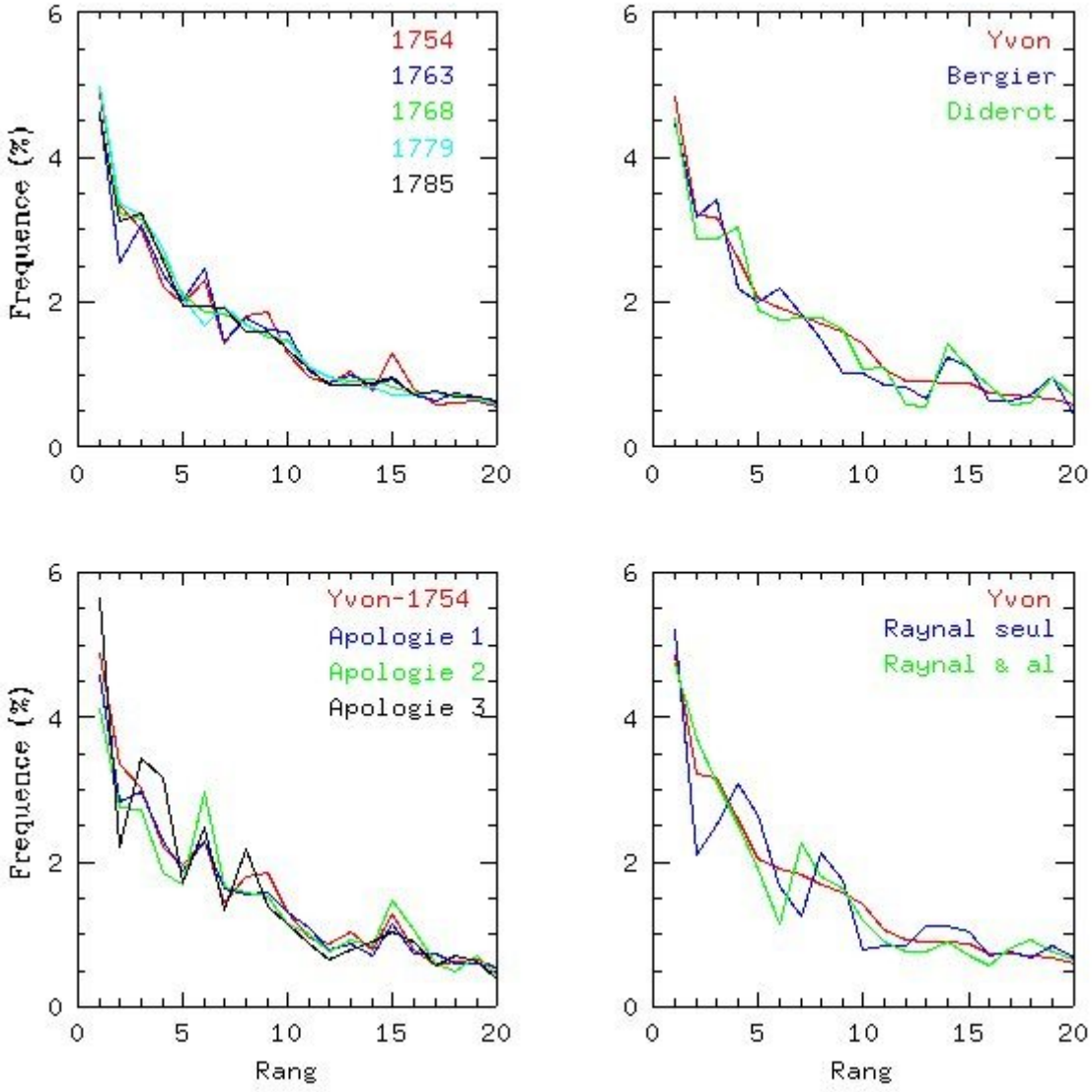

Figure 6. Fréquence des mots du texte (en pourcents) en fonction du rang. La correspondance entre rang et mot est donnée dans le tableau 4. La fréquence des mots de l'ensemble de l'œuvre d'Yvon est tracée en rouge dans les comparaisons avec d'autres auteurs, sauf pour l'*Apologie*, où la référence est son ouvrage de 1754 (à cause de l'évolution stylistique).

L'évolution stylistique que nous avons constatée pour les premiers mots des phrase est également présente parmi les autres mots de la phrase. Les mots *que* (rang 6) et *ne* (rang 15), et dans une moindre mesure *pour* (rang 13), sont plus fréquents dans les deux premiers ouvrages que par la suite. Ces excès se retrouvent dans les trois parties de l'*Apologie*.

L'examen de la figure 6 montre que les textes de Bergier et Diderot se distinguent modérément de ceux d'Yvon, avec en particulier un excès des mots *il* (rang 14) et *se* (rang 20). La différence est plus marquée pour les deux œuvres de Raynal, qui par ailleurs se distinguent très nettement l'une de l'autre, confirmant que *Les deux Indes* n'a certainement pas été rédigé par Raynal seul.

Tableau 4. Les 50 mots non-contextuels les plus fréquents dans l'œuvre de l'abbé Yvon

	1	2	3	4	5	6	7	8	9	10
1-10	de	les	la	&	le	que	des	à	qui	dans
11-20	en	par	pour	il	ne	ce	du	plus	un	se
21-30	une	leur	est	qu'il	a	pas	sur	si	son	comme
31-40	ses	mais	nous	ces	cette	au	avec	ils	aux	sa
41-50	on	lui	leurs	qu'ils	tous	même	sont	c'est	avoit	dont

Le bilan de cette première exploration stylométrique est positif : les méthodes présentées ci-dessus permettent, dans la plupart des cas, de distinguer les œuvres d'Yvon de celles des autres auteurs. Nous allons maintenant procéder à l'analyse stylométrique des textes dont l'abbé Yvon pourrait être l'auteur.