Mieux vaut *n-gram* qu'introspection. *Google Ngram Viewer* et parémiologie[1]

Mario MARCON
Università degli Studi di Udine & Università degli Studi di Padova (Italia)
mario.marcon16@gmail.com

Recibido: 29-05-2012 | Aceptado: 28-06-2012

Résumé

La numérisation du patrimoine libraire offre des masses de données textuelles. Cet article montre l'utilisation de ces données en parémiologie française à l'aide de l'interface en ligne *Google Ngram Viewer*. Nous expliquons brièvement ses caractéristiques pour mettre en relief les avantages et les inconvénients de son application aux études parémiologiques. Ensuite, nous présentons deux expériences qui portent sur l'essor du métaterme *parémie* et sa relation avec la dénomination plus connue de *proverbe*, ainsi que sur l'étude diachronique et sur l'utilisation actuelle de trois variantes proverbiales (*Il vaut mieux tard que jamais*, *Mieux vaut tard que jamais*, *Vaut mieux tard que jamais*) en diachronie. Pour chaque expérience, nous spécifions les techniques d'interrogation et suggérons des interprétations de graphiques. Nous montrons comment la consultation de *Google Ngram Viewer* et d'autres ressources linguistiques en ligne permettent aux parémiologues de vérifier rapidement et d'améliorer certaines intuitions. Cette opération de dépouillement d'archives documentaires virtuelles et les indications qu'elle suggère peuvent intégrer la méthodologie de recherche parémiologique et parémiographique.

Mots-clés
Parémiologie.
Parémie.
Diachronie.
Français.
TICE.

Título: «Más vale *n-gram* que instrospección. *Google Ngram Viewer* y paremiología».

Resumen

La digitalización del patrimonio bibliográfico ofrece un gran volumen de datos textuales. Este artículo muestra el uso de tales datos en la paremiología francesa con ayuda de la interfaz en línea *Google Ngram Viewer*. Explicamos brevemente sus características para destacar las ventajas e inconvenientes de su aplicación en los estudios paremiológicos. Posteriormente, presentamos dos experiencias que versan sobre la vida del archilexema paremia y su relación con la denominación más conocida, refrán, así como sobre el estudio diacrónico y la vigencia de tres variantes proverbiales (*Il vaut mieux tard que jamais*, *Mieux vaut tard que jamais*, *Vaut mieux tard que jamais*). De cada experiencia, detallamos las técnicas de interrogación y sugerimos interpretaciones de los gráficos. Mostramos cómo el uso de Gogle Ngram Viewer y otros recursos lingüísticos en línea permiten a los paremiólogos comprobar rápidamente algunas hipótesis y mejorarlas. Esta operación de vaciado de archivos documentales virtuales y las indicaciones que sugiere pueden incluirse en el método de investigación paremiológica y paremiográfica.

Palabras clave
Paremiología.
Paremia.
Diacronía.
Francés.
TIC

[1] Je tiens à remercier Cédrick Fairon pour m'avoir invité à explorer le potentiel des *n-grams* en parémiologie, ainsi que Mirella Conenna pour toutes ses remarques pointues et pour son encadrement. Je remercie également Margherita Romengo pour sa relecture attentive.

Title: «*N-gram* is better than instrospection. *Google Ngram Viewer* and paremiology»

Abstract

The digitization of books is providing masses of textual data. This paper shows how to use these data in French paremiology via the online interface *Google Ngram Viewer*. It briefly illustrates its features in order to underline its pros and cons for paremiological studies. We then present two experiences on the resurgent French metaterm *parémie* and its relationship with the well-known noun *proverbe*, as well as on the history and the vitality of three French proverbial variants (*Il vaut mieux tard que jamais*, *Mieux vaut tard que jamais*, *Vaut mieux tard que jamais*) in diachrony. For each experience, we specify our query techniques and suggest some interpretations of the graphs. We show how *Google Ngram Viewer* and other linguistic resources available online may enable paremiologists to quickly check and enhance some intuitions. The query on virtual textual archives and the indications that they suggest could be part of the paremiological and paremiographical research methodology.

Keywords
Paremiology.
Paremia.
Diachrony.
French.
ITC.

1. *GOOGLE NGRAM VIEWER* POUR LA RECHERCHE PARÉMIOLOGIQUE

Les projets de numérisation des patrimoines libraires mettent à la disposition de nombreuses données textuelles aux chercheurs en humanités numériques (*digital humanities*). L'œuvre entamée par *Google* en partenariat avec des bibliothèques universitaires dans le cadre du projet *Google Books*, a apporté des grands avantages à la recherche d'information de la part de tout usager, mais aussi à la recherche scientifique. Cet article veut illustrer des cas concrets d'exploitation du patrimoine libraire numérisé par *Google* en parémiologie française. Plus précisément, nous proposons d'utiliser l'application en ligne *Google Ngram Viewer* (Michel *et al*., 2011) pour extrapoler des indications dans le cadre d'études parémiologiques.

Tout d'abord, *Google Ngram Viewer* est une interface de recherche en ligne conçue pour le développement de la *culturomique* (*culturomics*), à savoir : « the application of high-throughput data collection and analysis to the study of human culture » (*ibid.*). Bien évidemment, les parémies sont des objets linguistiques et une expression du patrimoine culturel humain que ces collections de données permettent de mieux explorer.

Google Ngram Viewer utilise à peu près un tiers de la collection des livres numérisés par le projet *Google Books* (à savoir 5 millions de livres environ). Cet échantillon couvre 7 langues (allemand, anglais, chinois, espagnol, français, hébreu et russe) pour la période entre 1600 et 2008. La composante française compte environ 45 milliards de mots (ou *tokens*) dont on ignore, pourtant, la bibliographie qui est protégée par les droits de la propriété intellectuelle.

Les résultats des graphiques élaborés par *Google Ngram Viewer* découlent non pas des textes bruts numérisés, mais des tables de *n-grams*.[2] La véritable interrogation se fait sur des *n-grams* (de 1-*grams* à 5-*grams*), non pas sur l'intégralité des textes. Ces tables indiquent les fréquences exactes de chaque *n-gram* par année, ainsi que le nombre de pages et le nombre de livres concernés. Ces données sont visualisées sous forme de courbes où la fréquence des *n-grams* est normalisée, par défaut, sur la base du nombre de *tokens* par année. Les occurrences exactes restent ainsi inconnues aux usagers. De plus, tous les *n-grams* dont la fréquence (f) ≤ 40 ne sont pas représentés. Autrement dit, les hapax et les *n-grams* « rares » sont exclus lors de l'élaboration du graphique.[3]

[2] Un *n-gram* consiste en une séquence de chaînes de caractères (lettres, chiffres et signes de ponctuation), chaque chaîne étant séparée par des espaces en début et fin. Les tables de *n-grams* sont téléchargeables gratuitement (http://books.google.com/ngrams/datasets).

[3] Les tables de *n-grams*, par contre, incluent les *f* de tout *n-gram*. Des contre-vérifications peuvent être ainsi envisagées.

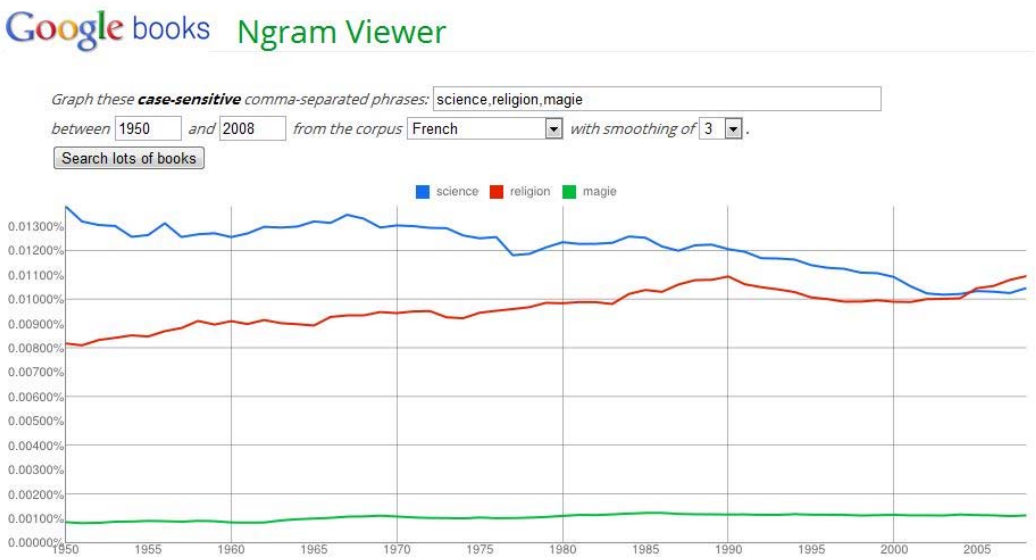


Figure 1. Exemple de graphique élaboré par *Google Ngram Viewer* pour 3 unités lexicales (*magie*, *religion* et *science*) pour la période 1950-2008 (lissage = 3).

La présentation graphique des résultats (Figure 1) se fait sur deux axes : l'axe *x* qui représente la ligne du temps et l'axe *y* qui indique les pourcentages normalisés de fréquence des *n-grams*. Il est possible d'observer une tendance à long terme (à savoir une courbe qui se rapproche d'une droite sur le graphique) grâce à un lissage par moyenne mobile.[4]

Le graphique est accompagné d'un découpage en périodes de temps. Chaque période renvoie aux contextes d'occurrence proposés par *Google Books*. Toutefois, ces contextes ne correspondent pas forcément au corpus exploité par *Google Ngram Viewer*.

D'après ces caractéristiques, les retombées en vue d'études parémiologiques et/ou parémiographiques ne peuvent être immédiates. Le dépouillement automatique d'une collection de textes (le véritable corpus) aurait facilité, par exemple, les études de fréquence (Arnaud & Moon, 1993), trop sous-estimées en parémiologie française. Pourtant, la méconnaissance des *f* brutes et de la véritable source textuelle des occurrences, ainsi que le silence imposé aux *n-grams* qui ont une *f* très basse, empêchent des estimations précises, sans pour autant exclure *a priori* des considérations génériques sur l'usage à l'écrit d'un proverbe ou, en général, d'une parémie.

Le fait que l'interrogation se fait sur des *n-grams*, et non pas sur des textes, permet de rechercher aisément tous les proverbes et les parémies dont la longueur (*l*) en *n-grams* ≤ 5, mais laisse de côté tous les autres, quitte à adopter des heuristiques de détection au cas par cas (§ 3).

Le repérage et l'extraction des contextes d'occurrence sont également biaisés par le détournement qui vire des tables des *n-grams* français vers la masse (encore moins connue) de *Google Books*, quoique des connaissances en littérature parémiologique et parémiographique ainsi que l'exploitation d'autres ressources bibliographiques (au format papier et/ou numérisées) pourraient ajuster le tir.

Il n'en reste pas moins vrai que, compte tenu de ses limites, *Google Ngram Viewer* peut suggérer des indications pour confirmer ou pour réfuter certaines hypothèses élaborées dans le cadre d'études parémiologiques et parémiographiques. Par l'appui d'autres ressources électroniques ainsi que des connaissances propres à l'acquis parémiologique, les graphiques peuvent promouvoir un meilleur déroulement de la réflexion. Par la suite, nous illustrons deux

[4] *Google Ngram Viewer* remplace un nombre *n* de *f* par année de chaque *n-gram* par leur moyenne arithmétique (http://books.google.com/ngrams/info).

manières d'intégrer l'interprétation des graphiques de *Google Ngram Viewer* parce que, comme le précisent Michel *et al*. (2011) : « Culturomic results are a new type of evidence in the humanities. As with fossils of ancient creatures, the challenge of culturomics lies in the interpretation of this evidence ».

Pour ces « fossiles dynamiques » que sont les proverbes, nos parcours d'interprétation conjuguent, pour autant que cela nous est possible, des connaissances en linguistique, notamment en syntaxe et en lexicographie, ainsi qu'en histoire de la parémiologie. En particulier, nous nous concentrons sur :

- l'essor du métaterme *parémie* et sa relation avec la dénomination *proverbe* (§ 2) ;
- l'*historique* (Conenna 2002) et la vitalité de trois variantes proverbiales en diachronie (§ 3).

2. EXPÉRIENCE N. 1 : MÉTALANGAGE PARÉMIOLOGIQUE

La classification des genres parémiques, ou le « problème insoluble » des parémiologues (Visetti & Cadiot, 2006: 317), a abouti à la reprise du métaterme *parémie*. Nous nous sommes demandé quelle relation *parémie* a entretenu et entretient avec *proverbe*.

Techniques d'interrogation

1. Nous avons saisi le singulier et le pluriel pour vérifier l'emploi des lemmes *parémie* et *proverbe* (*Google Ngram Viewer* n'est pas lemmatisé et permet seulement la recherche de formes exactes) ;
2. nous avons ignoré la casse parce qu'on estime qu'elle ne peut affecter de manière sensible les résultats de notre expérience ;
3. notre recherche couvre la période de 1600 à 2008 (lissage = 50).

Interprétation du graphique

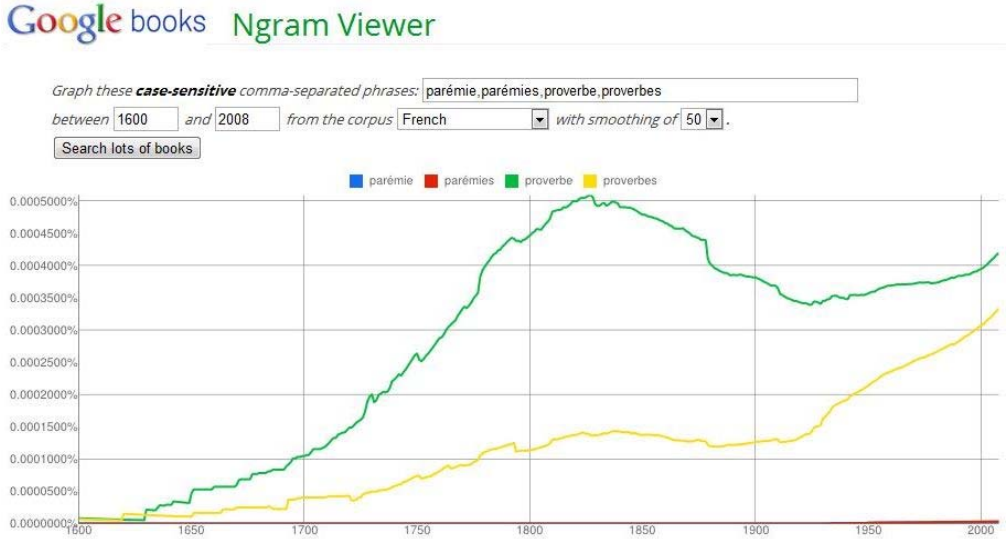


Figure 2. Graphique de *Google Ngram Viewer* pour *parémie* et *proverbe* (1600-2008).

On pourrait brièvement résumer que le proverbe est toujours (de plus en plus et quoiqu'on en pense) à la mode. Il est à la mode en tant que genre théâtral, et c'est d'ailleurs de cette manière qu'on peut aussi interpréter la trajectoire ascendante de la courbe verte tout au long du XVIII[e] siècle, et surtout le pic de la première moitié du XIX[e].[5] Il est à la mode en tant que genre

[5] « Genre mineur, le proverbe naît sous Louis XIII. C'est une petite comédie, jouée par des amateurs dans les salons mondains, dont l'intrigue est basée sur un proverbe que le spectateur doit deviner : c'est donc

gnomique codifié dans des œuvres parémiographiques. On peut également entrevoir la censure puriste des proverbes au cours du XVII[e] (Conenna, 1990), où les proverbes sont « dits par le peuple, dansés par le roi » (Conenna, 1991) et seulement recueillis pour l'apprentissage linguistique (Conenna, 1989 ; Julliani, 1990). Le tabou proverbial s'affaiblit au fil des siècles, jusqu'à l'affranchissement et à la commercialisation des recueils les plus disparates au XX[e].

Il est également évident que l'adaptation française de l'emprunt du grec παροιμία est une revitalisation linguistique qui s'affirme davantage au XX[e] siècle (épaisseur des lignes rousse et céleste). Il en ressort que cette création satisfait davantage les intérêts des parémiologues ainsi que l'autonomie disciplinaire de la parémiologie (comme son nom, d'ailleurs, le met en évidence). Autrement dit, les gens communs continuent à parler de « proverbes » pour faire référence à tout genre parémique (Marcon, à paraître), ce qui prouve à la fois le statut terminologique de *parémie* et un manque généralisé de compétence métaparémiologique. À l'heure actuelle, *parémie* est un technicisme propre à la parémiologie qui ne connaît pas de diffusion en dehors des discours scientifiques autour des genres textuels gnomiques. En ce sens, il est pertinent de comprendre le moment exact de son essor au XX[e] et quels ouvrages y contribuent.

Techniques d'interrogation

1. Nous saisissons le singulier et le pluriel (*parémie*, *parémies*), sans différences de casse ;
2. nous restreignons notre période de 1900 à 2008 ;
3. nous nous focalisons sur les fréquences exactes (lissage = 0) afin de mieux repérer les années d'attestation de *parémie* et ses pics d'usage.

Interprétation du graphique

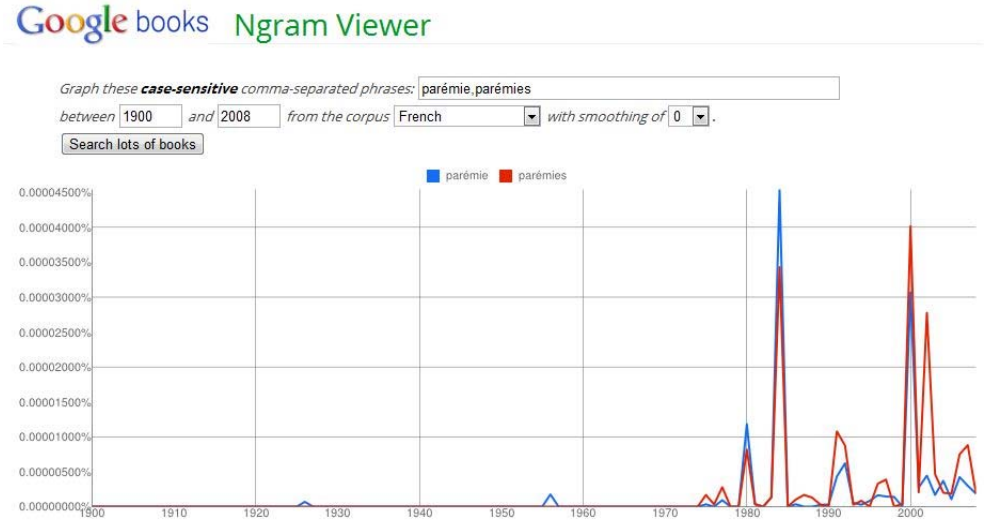


Figure 3. Graphique de *Google Ngram Viewer* pour *parémie*(*s*) (1900-2008).

D'après les livres numérisés, les années privilégiées d'usage correspondent à la première moitié des années 1980 et à la première moitié des années 2000. Plus précisément, le premier pic de fréquence est à la forme singulière en 1984. Comme le montre le renvoi aux pages de *Google Books*, le 1984 est l'année de parution de l'ouvrage *Richesse du proverbe* édité par Suard et Buridant en conclusion du 1[er] colloque international de parémiologie en 1981. Dans ce volume, Rodegem (1984 : 121) écrit : « Quel vocable convient le mieux pour désigner l'ensemble des formules consacrées par l'usage, qu'on appelle généralement « proverbes » ?

un jeu, au même titre que les devinettes et les charades. Il prend son essor au XVIII[e], et devient véritablement un sous-genre théâtral. » (Détrie *et al*., 1998:58).

[…] pour éviter les ambiguïtés, je réserverai le terme de parémie aux énoncés sentencieux pris globalement ».

Le deuxième pic est à attribuer à la forme plurielle en 2000 qui correspond, entre autres, à la parution :

- du numéro 139 de *Langages* consacré à la parole proverbiale, où Sevilla Muñoz illustre sa classification interlinguistique des parémies françaises et espagnoles ;
- et encore de la contribution de Sevilla Muñoz intitulée « Pour une saisie plurilingue des proverbes à partir de la langue espagnole » dans le volume *Micro- et macrolexèmes et leur figement discursif* édité par Gréciano après un colloque en 1998.

Loin d'être le fétiche métaterminologique de quelques parémiologues, *parémie* est déjà utilisé, d'après les résultats de *Google Books* entre 1900 et 1980, dans un article scientifique publié dans la *Revue de droit international et de législation comparée* en 1905. Malheuresement, l'intégralité de cette occurrence n'est pas disponible sur *Google Books*, ce qui nous fait poursuivre un chemin parallèle sur *Gallica*, expérience de numérisation du patrimoine libraire menée par la Bibliothèque nationale de France : « L'État repousse absolument toute action du dehors et règne positivement dans l'intérieur de ses limites. On a cherché à exprimer ces principes par la parémie : *Quidquid est in territorio est de territorio* » (Nys, 1905:238).

Parémie est utilisé dans un discours de droit où il est convenu d'appeler les formules latines telles *Quidquid est in territorio est de territorio* sous le nom de *brocards* ou d'*adages* (Conenna *et al*., 2011).

Un parcours en arrière dans le temps toujours via *Google Books* nous conduit au *Complément au Dictionnaire de l'Académie française* en 1842. Le dictionnaire définit le mot vedette *parémie* comme suit : « PARÉMIE. s.f. (littér.) Courte, allégorie. § Expression proverbiale »[6]. La circularité de la définition lexicographique nous ramène au *proverbe* ou, de toute façon, à la « proverbialité » d'une séquence formulaire. Cette définition dévoile l'opération d'étiquetage savant qui sert à encadrer les proverbes de manière systématique et scientifique, ainsi que toutes les formes gnomiques sentencieuses.

En gros, l'essor (relativement) récent du métaterme *parémie* est un héritage lexical revitalisé et relancé qui appartient, d'ailleurs, au passé de la parémiologie. Certes, une étude philologique minutieuse pourrait nous conduire aux premières attestations de *parémie* en langue française et mieux éclairer son chemin au long des siècles.

3. EXPÉRIENCE N. 2 : HISTORIQUE ET VITALITÉ PROVERBIALE

Prenons le cas des 3 variantes morphosyntaxiques suivantes :

(1) *Il/il vaut mieux tard que jamais*
(2) *Mieux/mieux vaut tard que jamais*
(3) *Vaut/vaut mieux tard que jamais*

et examinons leur présence en diachronie de 1600 à 2008 (lissage = 50) afin d'observer leur historique.

Techniques d'interrogation

1. Les variantes *Mieux vaut tard que jamais* et *Vaut mieux tard que jamais* sont des *5-grams* gérés par les tables de *Google Ngram Viewer*. Il suffit de saisir les deux variantes, faisant également attention à distinguer la casse minuscule/majuscule en début de variante.
2. Il n'en va pas de même pour *Il vaut mieux tard que jamais* (6-*gram*). Dans ce cas, nous émettons l'hypothèse que *l = 5* de *il vaut mieux tard que* peut nous aider à reconnaître cette variante et que le degré de superposition entre sa courbe et la courbe de la variante *vaut mieux tard que jamais* (*l = 5*) permet de détecter notre variante.

[6] Toute la famille des dérivés au préfixe *parémio-* est codifiée par les académiciens : *parémiographe*, *parémiographie*, *parémiographique*, *parémiologie*, *parémiologique*, *parémiologue*.

Interprétation du graphique

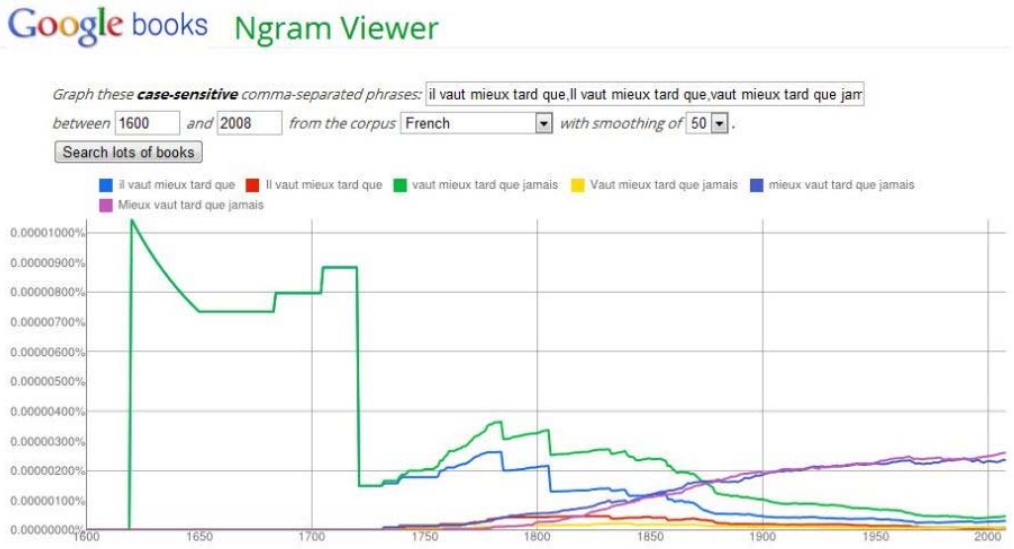


Figure 4. Graphique de *Google Ngram Viewer* pour les 3 variantes proverbiales : *Il vaut mieux tard que jamais* ; *Mieux vaut tard que jamais* ; *Vaut mieux tard que jamais* (1600-2008).

Le graphique suggère que *Il vaut mieux tard que jamais* est la variante préférée de 1600 jusqu'à la moitié du XVIII[e] siècle environ.[7] Les courbes céleste et verte sont superposées entre elles, ce qui équivaut à dire que le 5-*gram* : *il vaut mieux tard que* se superpose au 5-*gram* : *vaut mieux tard que jamais* et « rétablit » notre variante : *il vaut mieux tard que jamais*.

C'est à partir d'une dizaine d'années avant 1750 que se manifeste la concurrence entre les trois variantes.[8] En particulier, la variante *Mieux vaut tard que jamais* (courbes bleu et lavande) commence à s'installer dans le répertoire parémiologique et exacerbe sa compétition tout au long du XIX[e]. En effet, le graphique illustre qu'à la trajectoire descendante de la variante *Il vaut mieux tard que jamais* s'oppose la trajectoire ascendante de la variante *Mieux vaut tard que jamais*. On observe également la disjonction des courbes céleste et verte (respectivement *il vaut mieux tard que* et *vaut mieux tard que jamais*) au moment où la courbe rousse (*Il vaut mieux tard que*) monte. Autrement dit, l'écart qui se produit entre la courbe verte et la courbe céleste se traduit comme une indication du repérage de la variante graphique : *Il vaut mieux tard que jamais* par le 5-*gram vaut mieux tard que jamais*, mais aussi du repérage de la variante *Vaut/vaut mieux tard que jamais* qui s'affirme.

En ligne générale, pour la plus grande partie du XIX[e], la variante *Il vaut mieux tard que jamais* fait face à la concurrence de deux variantes qui prônent la concision formulaire : l'une qui propose l'érosion du pronom impersonnel ; l'autre qui veut un réagencement syntaxique initial, où l'adverbe préfère une position pré-verbale.

[7] Le pic correspond à un nombre relativement élevé de livres numérisés datés dans les deux premières décennies du XVII[e].

[8] Le *Dictionnaire du Moyen Français* prévoit un sous-paragraphe pour le proverbe *Mieux vaut tard que jamais* s.v. « tard » et repère une première occurrence en 1354 dans *Le Livre de Seyntz Medicines* par Henri de Lancaster : « *...et meultz vaut tarde qe jammes* ». Les autres attestations présentées, pourtant, reflètent la concurrence vivace entre les trois variantes et, par conséquent, la flexibilité de l'agencement syntaxique, de 1330 à 1500. Nous mentionnons ici les deux citations tirées des manuscrits de Christine de Pizan qui montrent non seulement cette concurrence, mais aussi le processus parémio-génétique par analogie et par juxtaposition de *Mieux vaut tard que jamais* au proverbe *Il vaut mieux tôt que tard* : « *car il vault trop mieulx tost que tard, et tart que jamais.* »(*Le livre des trois vertus* W.H., c.1405, 120) ; « *...comme mieux vaille tart que jamais* » (*Epistre de la prison de vie humaine*, 1416-1418, 17).

D'après les tables des *n-grams*, le tournant se réalise autour du 1875, quand l'usage à l'écrit de *Il vaut mieux tard que jamais* et la variante sans pronom impersonnel diminuent en faveur de *Mieux vaut tard que jamais*. Le renversement de la primauté de préférence se réalise et s'établit tout au long du XX^{e}. *Mieux vaut tard que jamais* semble s'imposer et remplacer de manière quasi radicale les deux autres variantes qui, toutefois, n'ont pas totalement disparu. Toutefois, la réduction de l'écart entre la courbe céleste et la courbe verte suggère une véritable concurrence entre *Mieux vaut tard que jamais* et *Il vaut mieux tard que jamais* au détriment de *Vaut mieux tard que jamais*.

On profite de cette considération pour remarquer que la concurrence ne cesse jamais et pour rappeler que les usages linguistique et parémique (à l'écrit) coexistent et se réaffirment au fil du temps. Les courbes plates ou descendantes ne témoignent pas de l'absence ou de la disparition *ipso facto* des variantes proverbiales : ni à l'écrit, ni à l'oral, ni dans la compétence parémique individuelle, voire communautaire. Plutôt les courbes renvoient l'image d'une codification, voire re-codification ou dé-codification, de ces variantes à l'écrit. Autrement dit, le graphique informe sur la transmission et sur l'œuvre de normalisation (parémiographique et non) à l'écrit des variantes proverbiales et, en large partie, il éclaire l'usage (ou le manque d'usage) dans le corpus pris en examen (important en termes quantitatifs, pour de vrai, mais à la fois peu transparent (§ 1)).

La préférence de la variante *Mieux vaut tard que jamais* par rapport à *Il vaut mieux tard que jamais* nous a amené à explorer l'usage en diachronie du verbe composé *valoir mieux* et des agencements proposés par nos variantes. Cette concurrence nous a fait supposer qu'il pourrait y avoir une corrélation entre la vitalité des variantes proverbiales et les préférences d'agencement syntaxique qui concernent de manière plus générale ce verbe.

Techniques d'interrogation

1. Nous avons saisi 3 couples de *n-grams* :
   a. *il vaut mieux* et sa variante graphique *Il vaut mieux* ;
   b. *mieux vaut* et sa variante graphique *Mieux vaut* ;
   c. *vaut mieux* et sa variante graphique *Vaut mieux* ;
2. nous avons sélectionné la même période de temps que pour la recherche sur les variantes proverbiales (1600-2008) et le même lissage (= 50).

Interprétation du graphique

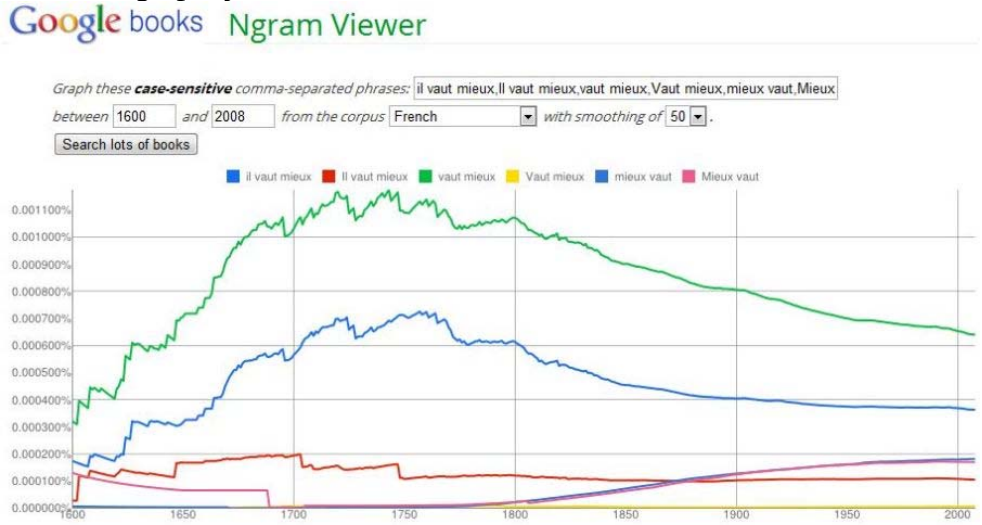


Figure 5. Graphique de *Google Ngram Viewer* pour 3 agencements syntaxiques du verbe composé *valoir mieux* (1600-2008).

Le graphique en Figure 5 montre que l'usage avec un pronom impersonnel domine toute la période analysée, quoique l'écart entre la courbe verte (*vaut mieux*) et la courbe céleste (*il vaut mieux*) laisse entrevoir des occurrences où le verbe est employé à la forme personnelle, même si en partie comblé par la variante graphique *Il vaut mieux* représentée par la courbe rousse. Comme on le disait auparavant, la courbe plate ou descendante est une indication de codification ou de normalisation, non pas de disparition. Les courbes bleue (*mieux vaut*) et rose (*Mieux vaut*) nous offrent un exemple en ce sens. Elles concurrencent l'agencement impersonnel dès le début du XVII^{e}, quitte à perdre de vigueur vers la fin du siècle et pour la plupart du XVIII^{e}, jusqu'au moment où le patron *adverbe + verbe* refait surface dès la fin du XVIII^{e} et continue à s'imposer, sans jamais dépasser l'usage impersonnel et personnel de *valoir mieux*.[9]

Si l'on compare les Figures 4 et 5, on comprend aisément que la trajectoire descendante de la courbe qui représente l'usage impersonnel du verbe est liée à la trajectoire descendante de la variante proverbiale impersonnelle. De même, on constate que la trajectoire ascendante de la courbe qui représente le patron *adverbe + verbe* est étroitement liée à la trajectoire ascendante de la variante proverbiale *Mieux vaut tard que jamais*. Après la perte de vigueur, la montée en codification de *mieux vaut* est évidemment influencée par l'essor de la variante proverbiale correspondante, comme le témoigne la fluctuation très similaire des courbes (courbes bleue et lavande en Figure 4, courbes rose et bleue en Figure 5) à partir du XVIII^{e}.

La transmission écrite des proverbes s'entrelace ainsi avec la syntaxe : les uns pèsent sur les sorts de l'autre. Mutuellement, les proverbes et les agencements syntaxiques influencent leur vitalité. D'une part, les proverbes (notamment leur manque d'usage à l'écrit) peuvent décourager certains agencements syntaxiques, comme la forme au pronom sujet impersonnel *il vaut mieux* dans notre expérience. D'autre part, les proverbes peuvent favoriser et revitaliser des agencements syntaxiques. Ainsi, ils n'héritent simplement de choix lexico-syntaxiques archaïsants, comme l'ont justement remarqué de nombreux parémiologues. Au fil du temps, les proverbes recyclent des éléments (en l'occurrence, des agencements syntaxiques) du répertoire linguistique qui continuent à coexister, malgré leur « disparition » apparente de la codification écrite, qu'elle soit soutenue (ou non) par les parémiographes et/ou par les lexicographes[10] ainsi que par les grammairiens.[11]

[9] Riegel *et al.* (2009:746) citent la *Grammaire nouvelle de l'ancien français* de Buridant (2000) où le linguiste constate que l'emploi du pronom impersonnel se manifeste au XIII^{e} et se répand tout au long du XVI^{e}. La construction *mieux vaut* serait, par contre, une parmi les « vestiges de l'époque où les verbes impersonnels […] étaient normalement dépourvus du pronom sujet » (*ibid.*) (§ note 7).

[10] Les relations entre proverbes et syntaxe peuvent être sous-représentées, voire altérées, par le prescriptivisme des recueils lexicographiques génériques de référence. Par exemple, s.v. « valoir » à l'acception 3d, le TLFi opère une distinction entre *valoir mieux* et *mieux valoir* au plan syntaxique. Le premier serait suivi par *que + nom* ou par *que de + verbe à l'infinitif*. En appui, le TLFi cite, entre autres, deux proverbes : *Bonne renommée vaut mieux que ceinture dorée* et *Un tiens vaut mieux que deux tu l'auras*. L'agencement *mieux valoir* se combinerait, par contre, seulement avec un verbe à l'infinitif, comme « prouvé » par le proverbe *Mieux vaut tenir que courir*. Pour renforcer et justifier ces agencements, il est intéressant de constater que le proverbe *Mieux vaut tard que jamais* est relégué s. v. « tard » et « jamais ». Sous la première voix, il est indiqué sous l'acception 1 en tant que proverbe, alors que sous l'acception 2b, il est mis en relation de variante de l'expression *Il n'est pas/jamais trop tard pour bien faire*, sans aucune référence aux propriétés syntaxiques. Sous la deuxième, *Mieux vaut tard que jamais* est mentionné comme exemple qui réalise le patron lexico-syntaxique : *mieux vaut* + mot temporel + *que jamais*. Ce prescriptivisme des recueils lexicographiques génériques peut également avoir une influence sur l'usage d'une variante proverbiale au détriment d'une autre. On remarque en effet qu'aucune des voix mentionnées ne cite la variante proverbiale *Il vaut mieux tard que jamais*.

[11] À ce propos, une analyse des grammaires françaises en diachronie pourra apporter quelques éclaircissements.

On remarque également que, de préférence, les variantes analysées privilégient un emploi en discours suivi plutôt qu'en position discursive initiale, et ce à l'exception de *Mieux vaut tard que jamais*, où la distinction de casse de l'initiale ne produit pas d'écart significatif. En ce sens, on pourrait supposer que la codification parémiographique a joué un rôle central pour la diffusion de cette variante ou que *Mieux vaut tard que jamais* se prêterait davantage à un emploi en position discursive initiale.

EN GUISE DE CONCLUSION

Google Ngram Viewer nous a servi de guide, tout comme d'autres ressources en ligne, pour l'élaboration et l'ajustement d'hypothèses, ainsi que pour formuler des considérations préalables à des études soignées concernant, en l'occurrence, le métalangage parémiologique et l'usage en diachronie de variantes proverbiales morphosyntaxiques. Nos expériences ne sont que deux parcours possibles d'application et d'interprétation des résultats proposés par *Google Ngram Viewer* qui présente des pour et des contre pour la recherche parémiologique (§ 1). De toute façon, nous espérons avoir montré comment la consultation d'un vaste patrimoine libraire numérisé (qu'il soit géré par un géant mondial du marché de l'information ou par une bibliothèque nationale) permet de tester rapidement des intuitions de recherche, qu'elles relèvent autant de la créativité personnelle du parémiologue que de la littérature parémiologique.

Nous souhaitons ainsi que cette opération de dépouillement virtuel de (grandes) archives documentaires intègre (de manière raisonnée et critique) la méthodologie de recherche parémiologique et parémiographique. *Google Ngram Viewer* ou des outils pareils peuvent apporter des indications précieuses lors de l'établissement d'une théorie ou d'une classification parémiologique, ainsi que pour la rédaction de recueils parémiographiques monolingues et multilingues. Pour ces derniers cas, son utilisation peut faciliter, entre autres choses, le choix de la forme canonique d'un proverbe et/ou des proverbes correspondants les plus appropriés entre deux (ou plusieurs) langues, non seulement aux plans lexical, sémantique et pragmatique, mais aussi en tenant compte de la concurrence des variantes proverbiales en macro- et en micro-diachronie.

RÉFÉRENCES BIBLIOGRAPHIQUES

ARNAUD, P.; MOON, R. (1993) : « Fréquence et emploi des proverbes français et anglais », in C. Plantin (réd) : *Lieux communs : topoï, stéréotypes, clichés*. Paris: Kimé : 323-341.

CONENNA, M. (1989) : « Les noms, les phrases et les proverbes de Julliani pour l'apprentissage des langues », *Quaderni del Seicento francese*, 9 : 41-62.

CONENNA, M. (1990) : « Proverbi fra censura e "bienséance" nel Seicento », in M. L. De Gaspari-Ronc, L. Pietromarchi, F. Piva (réd) : *Lo "spleen" nella letteratura francese. "Le mot déguisé": censura e interdizione linguistica nella storia del francese*. Atti del XVI Convegno della Società Universitaria per gli Studi di Lingua e Letteratura Francese, Trento 29 settembre/1 ottobre 1988, vol. I. Fasano : Schena, 189-202.

CONENNA, M. (1991) : « Dits par le peuple, dansés par le roi: les proverbes au XVIIe siècle », *Quaderni del Seicento francese*, 10 : 219-240.

CONENNA, M. (2002) : « Sur l'historique du proverbe », in G. Maiello & R. Stajano (réd). *Collage. Studi in memoria di Franca Calderari Bevilacqua*. Salerno/Milano : Oèdipus, 35-55.

CONENNA, M.; LAVERMICOCCA, F.; MARCON, M.; SERRONE, G.; VERGNE, M. (2011): « Un caso paremio-giuridico », Communication présentée à *I Jornadas Fraseología*

y Paremiología – Perspectivas y aplicaciones en Didáctica, Traducción y Lingüística de corpus, Bari, 5-6 mai 2011.
DÉTRIE, C.; MASSON, M.; VERINE, B. (réd) (1998) : *Pratiques textuelles*. Montpellier : Presses de l'Université Paul Valéry – Montepellier III.
JULLIANI (1990) : *Les proverbes divertissans*. Texte établi, annoté et présenté par Mirella Conenna. Fasano/Paris : Schena/Nizet.
MARCON, M. (à paraître) : « Another proverb in the wall. Perception parémique et intérêt parémiologique sur les fan pages de Facebook».
MICHEL, J.-B.; SHEN, Y. K.; AIDEN, A.P.; VERES, A.; GRAY, M. K.; BROCKMAN, W.; THE GOOGLE BOOKS TEAM; PICKETT, J. P.; HOIBERG, D.; CLANCY, D.; NORVIG, P.; ORWANT, J.; PINKER, S.; NOWAK, M. A.; LIEBERMAN AIDEN, E. (2011): « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science*, 331 (6014): 176-182 [paru en ligne avant impression: 16/12/2010].
RIEGEL, M.; PELLAT, J.-C.; RIOUL, R. (2009) : *Grammaire méthodique du français*. Paris: PUF (4e édition entièrement revue).
VISETTI, Y.-M.; CADIOT, P. (2006) : *Motifs et proverbes. Essai de sémantique proverbiale*. Paris: PUF.

SITOGRAPHIE

Dictionnaire du Moyen Français: http://www.atilf.fr/dmf/ [consulté: 28-05-2012].
Gallica: http://gallica.bnf.fr/ [28-05-2012].
Google Books: http://books.google.fr/intl/fr/googlebooks/about.html [28-05-2012].
Google Ngram Viewer: http://books.google.com/ngrams/ [28-05-2012].
Trésor de la Langue Française informatisé: http://atilf.atilf.fr/tlf.htm [28-05-2012].

RÉFÉRENCES CONTEXTES D'OCCURRENCE

HENRI DE LANCASTER (1354=1940) : *Le Livre de seyntz Medicines*. The unpublished devotional treatise of Henry of Lancaster. Text edited by E. J. Arnould.- Oxford: B. Blackwell.
NYS, E. (1905) : « La jurisdiction consulaire – Quelques notes », *Revue de droit international et de législation comparée*, 37(7): 237-252. http://gallica.bnf.fr/ [28-05-2012].
PIZAN, C. de (1405 c.) : *Le livre des trois vertus*. http://www.atilf.fr/dmf/ [28-05-2012].
PIZAN, C. de (1416-1418) : *Epistre de la prison de vie humaine*. http://www.atilf.fr/dmf/ [28-06-2012].
RODEGEM, F. (1984) : « La parole proverbiale », in F. Suard & C. Buridant (réd). *Richesse du proverbe. Vol. 2. Typologie et Fonctions*. Lille: Presses Universitaires de Lille: 121-135.
SEVILLA MUÑOZ, J. (2000a) : « Les proverbes et les phrases proverbiales français, et leurs équivalences en espagnol », *Langages*, 139: 98-109.
SEVILLA MUÑOZ, J. (2000b) : « Pour une saisie plurilingue des proverbes à partir de la langue espagnole», in G. Gréciano (réd) : *Micro- et macrolexemes et leur figement discursif Etudes de linguistique comparée francais/allemand*. Leuven: Peeters.