Nouvelle méthode d'analyse statistique d'apparition d'un mot particulier (études synchroniques et diachroniques)

Patricia Guilpin[1], Christian Guilpin[2]

[1] SYLED – CLA[2]T – ILPGA – Université de la Sorbonne Nouvelle Paris 3 – 19, rue des Bernardins – 75005 Paris – France

[2] Groupe de Physique des Solides – Universités Paris 6 et 7 – UMR 75 88 Campus Boucicaut 140, rue de Lourmel – 75015 Paris – France

guilpin@gps.jussieu.fr

Abstract

In this paper, a new statistical method is presented to study the frequency of a particular word at different times or in different styles (synchronic and diachronic studies). It is made use of equations one can easily compute. The example developed in order to validate our method deals with variations in the use of some determiners in classical Greek (by Aristophanes and Herodotus).

Résumé

Dans ce papier, nous développons une nouvelle méthode statistique afin d'étudier la fréquence d'un mot particulier à différentes époques ou dans différents styles (études synchroniques et diachroniques). Cette méthode originale utilise des équations que l'on peut aisément programmer. L'exemple choisi pour la validation de la méthode est l'étude des variations dans l'emploi des déterminants en grec classique chez Aristophane et Hérodote.

Mots-clés : loi de Poisson, loi binomiale, critère de Kolmogorov, populations parentes, variations linguistiques : synchronie et diachronie.

1. Introduction

Notre méthode a été élaborée dans le cadre d'une étude diachronique de l'article indéfini en grec au cours de laquelle nous avons constaté des variations très nettes dans l'utilisation des morphèmes d'une époque à l'autre (apparition ou disparition de formes, basculements vers un type d'emploi etc.). Pour corroborer nos observations, nous avons mis au point une méthode qui permet de comparer la fréquence d'apparition d'un morphème dans deux textes différents. Après avoir vérifié que la distribution des morphèmes obéit à la loi de Poisson dans chaque texte, nous appliquons un critère qui détermine si les populations sont parentes ou non. Ce test permet d'étudier des variations linguistiques aussi bien dans des corpus de petite taille que dans des corpus de grande taille (dès lors que ceux-ci sont jugés représentatifs pour la démonstration) et est adapté à l'étude de variations de lexèmes dans tous les types de textes. Cette méthode qui a pris naissance dans le cadre d'une recherche en grec peut s'appliquer à toutes les langues.

2. Méthode d'analyse statistique

2.1. *Position du problème*

On se propose d'effectuer une statistique sur la fréquence d'apparition d'un certain mot figurant dans un texte quelconque. Pour réaliser une telle étude, il est nécessaire d'introduire une **mesure** sur les textes, une mesure étant une application sur les nombres réels positifs. La mesure qui semble la plus naturelle et la plus employée repose sur l'ordre des mots comptés à partir du début du texte. Ainsi, la distance de deux mots est la valeur absolue de la différence de leurs rangs. La longueur d'un texte étant évidemment définie par le nombre de mots qu'il contient. À présent, on peut envisager une analyse statistique élémentaire sur la fréquence d'un certain mot X rencontré dans un texte A lequel comprend au total N_a mots. Le mot X est rencontré K_a fois et cela dans les positions notées $x_a(i)$ avec $i = 1, 2 \ldots K_a$.

Le problème que l'on désire résoudre est le suivant : on considère un autre texte appelé B qui contient au total N_b mots, dans lequel on recherche les occurrences du précédent mot X ; celles-ci apparaissent K_b fois et cela dans les positions notées $x_b(j)$ avec $j = 1, 2 \ldots K_b$. À partir de ces données, peut-on conclure à une différence significative ou non des deux populations de X rencontrées dans chacun des textes ? En d'autres termes, a-t-on affaire à la même **population parente** ou à deux populations parentes différentes. L'analyse statistique répond à cette question **sans préjuger des raisons qui peuvent expliquer le fait que les populations sont parentes ou non**.

Notations : dans la suite de ce propos les variables et quantités se rapportant au texte A seront indicées avec a et celles se rapportant au texte B avec l'indice b.

Tel que nous l'avons posé le problème relève de l'analyse statistique des séries d'événements dans laquelle la loi de Poisson joue un rôle fondamental. Soit ξ un nombre d'événements se produisant dans l'intervalle de longueur x. ξ suit une loi de Poisson de moyenne $(\lambda\xi)$, c'est-à-dire :

$$P(\xi = k) = \exp(-\lambda x)(\lambda x)^k / k! \text{ avec } k = 0,1,2...$$

C'est la probabilité que k événements apparaissent dans l'intervalle x. Une propriété importante de la loi de Poisson : les événements qui obéissent à la loi de Poisson sont distribués selon la **loi uniforme** c'est-à-dire que les événements sont équirépartis sur l'axe des abscisses x.

2.2. *hypothèse n° 1 - À l'intérieur d'un texte, les occurrences obéissent à la loi de Poisson*

L'expérience montre que, en règle générale, les occurrences d'un mot obéissent à la loi de Poisson et que cette hypothèse n'a jamais été rejetée lors de l'analyse de plus de *150* cas. Cependant, avant de poursuivre ce calcul, il convient de s'assurer de la validité de cette hypothèse dans chaque cas.

Pour ce faire, point n'est besoin d'estimer le paramètre λ, en effet, il suffit de vérifier que les occurrences sont distribuées uniformément dans le texte. Rappelons que l'on n'obtient jamais de réponse positive à une telle question, **mais on peut savoir si l'hypothèse n'est pas contredite par les données expérimentales**.

2.2.1. *Technique opératoire de la vérification de l'hypothèse N°1*

Il est aisé de construire un **histogramme** des occurrences du texte A. Pour cela il suffit de dénombrer les événements tombant dans chaque intervalle de regroupement. Le nombre I_a

d'intervalles de regroupement (cases) est donné par l'expression :
$I_a = \log_2 (K_a) + 1$, $\log_2$ désignant le logarithme en base *2,* et tous les intervalles de regroupement ont la même taille $h_a = 1/I_a$ (Aïvazian, 1970 ; Ch. Guilpin, 1999). La case $n° j$ contient alors n_j occurrences que l'on est en mesure de dénombrer. Évidemment, $\sum_{j=1}^{I_a} n_j = K_a$, ainsi la probabilité empirique de tomber dans la case $n° j$ s'écrit $p_j = n_j / K_a$. À partir des p_j, il est facile de construire la **fonction de répartition empirique**

$F_n = \sum_{j=1}^{n} p_j$ avec $n = 1, 2 \ldots I_a$, (on s'assure que $F_{I_a} = 1$).

Maintenant, il faut vérifier que l'hypothèse de la répartition uniforme n'est pas contredite par les données expérimentales. Pour cela, **nous allons faire usage du critère de Kolmogorov** pour lequel il convient de déterminer la quantité D_a donnée par l'expression :

$$D_a = \sup |F_n - F_n^*| \text{, où } F_n^* = \frac{n}{I_a} \text{ avec } n = 1, 2, \ldots I_a .$$

À partir de D_a, on calcule la quantité $v_{0a} = \sqrt{K_a} D_a$. v_0 est une valeur possible de la variable aléatoire v laquelle obéit à la loi de Kolmogorov (Ch. Guilpin, 1999). Ainsi, la probabilité pour que v puisse être supérieure à v_0 s'obtient par l'expression :

$$P(v > v_{0a}) = 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 v_{0a}^2) .$$

Si la probabilité calculée est inférieure au seuil de signification $\alpha = 0,05$, on rejette l'hypothèse d'une répartition uniforme avec *100* α chances sur *100* de la rejeter à tort. Signalons que le critère de Kolmogorov est sensible à la tendance que le processus étudié s'écarte d'un processus de Poisson (Cox *et al.*, 1969).

Une fois vérifiée l'hypothèse d'une distribution uniforme pour chacun des deux textes considérés appelés *A* et *B*, on cherche à répondre à la question de savoir si le mot *X* est extrait d'une même population parente ou non.

2.2.2. *Comparaison des taux d'occurrence de deux processus de Poisson*

On suppose donc que nous étudions **deux processus de Poisson indépendants** caractérisés par les paramètres λ_a et λ_b que nous cherchons à comparer.

Les deux processus de Poisson sont observés durant les intervalles fixes x_a et x_b avec $x_a = N_a$ et $x_b = N_b$ à l'intérieur desquels on trouve respectivement K_a et K_b occurrences. Ces deux dernières valeurs sont les **valeurs observées des deux variables aléatoires discrètes indépendantes ξ_a et ξ_b**, lesquelles obéissent chacune à une loi de Poisson de moyennes $\mu_a = \lambda_a x_a$ et $\mu_b = \lambda_b x_b$. Il s'ensuit que l'on peut écrire que la probabilité pour que $\xi_a = K_a$ et $\xi_b = K_b$ est donnée par l'expression suivante :

$$(1) \qquad P(\xi_a = K_a , \xi_b = K_b) = \frac{\exp(-\mu_a)\mu_a^{K_a}}{K_a!} \frac{\exp(-\mu_b)\mu_b^{K_b}}{K_b!} .$$

Pour comparer les deux processus, il est commode d'introduire le paramètre ρ ainsi défini :

$$\rho = \mu_a / \mu_b = \lambda_b x_b / \lambda_a x_a \ .$$

Rappelons que x_a et x_b sont des paramètres connus expérimentalement ainsi que K_a et K_b, par conséquent, une inférence sur ρ est équivalente à une inférence sur λ_b / λ_a . À présent, nous devons évaluer la probabilité conditionnelle que $\xi_b = K_b$, sachant que $\xi_a + \xi_b = K_a + K_b$. En définitive, on obtient :

$$P(\xi_b = K_b | \xi_a + \xi_b = K_a + K_b) = \frac{P((\xi_a = K_a).(\xi_b = K_b))}{P(\xi_a + \xi_b = K_a + K_b)} .$$

La probabilité $P(\xi_a + \xi_b = K_a + K_b)$ est donnée par le théorème de la somme de variables poissonniennes indépendantes, à savoir :

$$P(\xi_a + \xi_b = K_a + K_b) = \frac{(\mu_a + \mu_b)^{K_a + K_b}}{(K_a + K_b)!} \exp(-(\mu_a + \mu_b)) .$$

Il en résulte que : $$P(\xi_b = K_b | \xi_a + \xi_b = K_a + K_b) = \frac{\mu_a^{K_a} \mu_b^{K_b}}{K_a! K_b!} \frac{(K_a + K_b)!}{(\mu_a + \mu_b)^{K_a + K_b}} ,$$

puis, en posant $\theta = \frac{\rho}{1+\rho}$, on obtient :

(2) $$P(\xi_b = K_b | \xi_a + \xi_b = K_a + K_b) = C_{K_a + K_b}^{K_a} \theta^{K_b} (1 - \theta)^{K_a} \quad .$$

Cette probabilité est donnée, donc, par la loi binomiale.

2.3. *Hypothèse n° 2 : On se propose de vérifier que les deux lois de Poisson sont les mêmes, c'est-à-dire que* $\lambda_a = \lambda_b$.

Cela signifie que nous devons vérifier que les données expérimentales ne contredisent pas cette seconde hypothèse. Ainsi, si $\lambda_a = \lambda_b$, la loi donnée par l'expression (2) n'est rien d'autre que la loi binomiale de paramètre $\theta = N_b / (N_a + N_b)$.

À partir de (2), il reste à calculer la valeur particulière de ξ_b qui correspond au seuil de signification α que l'on choisit habituellement égal à $0,05$. On désigne par K_{inf} cette valeur limite dont l'usage est le suivant : si $K_b < K_{\inf}$ (pour $\alpha = 0,05$), on rejettera l'hypothèse selon laquelle les deux processus de Poisson relèvent d'une même population parente toujours avec 100α chances sur 100 de rejeter à tort l'hypothèse d'un même processus poissonnien. Dans le cas contraire ($K_b \geq K_{\inf}$), on conserve l'hypothèse qu'il s'agit du même processus poissonnien car elle n'est pas contredite par les données expérimentales ; elle est plausible. Pour déterminer K_{inf}, il suffit d'écrire les deux inéquations suivantes :

(3) $$\sum_{i=0}^{K_{\inf}} C_{K_a + K_b}^{i} \theta^i (1-\theta)^{K_a + K_b - i} \leq \alpha \text{ et } \sum_{i=0}^{K_{\inf}+1} C_{K_a + K_b}^{i} \theta^i (1-\theta)^{K_a + K_b - i} \geq \alpha .$$

Il n'y a aucune difficulté à calculer les valeurs K_{inf} pour la somme $K_a + K_b$ fixée à l'avance.

3. Validation de la méthode : Application à l'étude de deux textes grecs dont les résultats sont connus

3.1. *Introduction*

Cette méthode a été utilisée avec succès sur un très grand nombre de cas empruntés à la littérature grecque. À des fins d'illustrations simples, nous avons retenu des exemples en grec classique (synchronie) dont les résultats sont connus.

Nous étudions des variations dans l'emploi des déterminants du nom en grec ancien. Notre étude est synchronique et porte sur deux genres et deux styles différents au sein du groupe de dialecte ionien-attique. Nous avons choisi deux auteurs de la période classique : Aristophane (vers 445-386 avant J.-C.) qui utilise le dialecte attique du temps, enrichi d'une grande invention verbale, et d'autre part Hérodote (vers 485-425 avant J.-C.) dont le style évolue selon le contexte de la simplicité populaire au ton sentencieux ; son dialecte ionien est parcouru d'emprunts à Homère et à l'attique. Toute proportion gardée, ces deux auteurs sont représentatifs d'une langue qui peut être qualifiée de semi-savante. Pour nos calculs, nous avons utilisé les textes annotés (Habert *et al.*, 1997 et Guilpin, 2003 concernant les problèmes de méthode, codage et ressources concernant le grec) du projet américain Perseus (Tufts University).

3.1.1. *Corpus et données*

The Perseus Digital Library : http://www.perseus.tufts.edu

Aristophanes : *Les Grenouilles* [étude des 1003 premiers vers] - *Aristophanes Comœdiæ*, Hall F.W., Geldart W.M. (éd.), vol. 2, Oxford, Clarendon Press, 1907.

Hérodote : *Histoires* [étude du livre I, du début à LXXV] in Herodotus with an English translation by A.D. Godley, Cambridge, Harvard University Press, 1920.

Le texte *A* écrit par Aristophane comporte 4 873 mots, le texte *B* emprunté à Hérodote contient 11 859 mots. Nous avons $N_a = 4\,873$ et $N_b = 11\,859$.

3.1.2. *Objectifs*

Pour la validation, nous retiendrons la différence de genre entre les deux œuvres : la comédie d'Aristophane suit les conventions de la métrique, tandis que le travail d'historien d'Hérodote est composé en prose de façon à favoriser la mémorisation des textes selon la tradition orale. De ce point de vue, la littérature ionienne offre un support à l'essor intellectuel qui suit et à l'évolution des disciplines (Horrocks, 1997 : 21-3).

Nous nous proposons de vérifier deux phénomènes linguistiques connus en grec ancien a) les variations dans l'emploi du pronom-adjectif indéfini « tis, ti » dont l'apparition est d'autant plus marquée que le texte est savant ou sophistiqué, b) la stabilité des emplois de l'article défini « ho, hê, to » (Biraud, 1991 ; Guilpin, 2002). Nous limitons notre propos à l'étude des GN au nominatif et à l'accusatif. Cette contrainte nous permet par ailleurs de traiter de formes de déterminants communes à l'ionien et à l'attique (Chantraine, 1968).

3.2. *Variations dans l'emploi du pronom-adjectif indéfini « tis, ti »*

Nous partons de l'observation suivante : l'emploi du pronom-adjectif « tis, ti » est d'autant plus fréquent que la langue de l'auteur est savante et son style contraint par le rythme des vers. En revanche, les idiomes plus populaires tendent à effacer la présence de « tis ».

Nous souhaitons donc vérifier l'hypothèse selon laquelle l'emploi du pronom-adjectif « tis » diffère nettement entre l'œuvre versifiée d'Aristophane et l'œuvre en prose d'Hérodote.

On se propose d'analyser la forme neutre du pronom-adjectif indéfini X = « ti ». Précisons que « tis, ti » est traité globalement comme pronom-adjectif dans la mesure où les variations que l'on souhaite mettre en évidence concernent aussi bien les deux formes, notre visée n'étant ni historique, ni épistémologique. Au nominatif et accusatif singuliers. On dénombre alors $K_a = 58$ et $K_b = 25$ occurrences. Il convient de vérifier l'hypothèse de la distribution uniforme pour chacun des deux textes. Il est aisé de calculer que $I_a = 6$ $I_b = 5$ (aux arrondis près). Après avoir réalisé le cumul des fréquences pour obtenir la fonction de répartition, on trouve $D_a = 0,143$ et $D_b = 0,160$ ce qui permet de calculer $v_{0a} = 1,094$ et $v_{0b} = 0,358$. On peut alors obtenir les probabilités correspondantes de pouvoir dépasser ces valeurs : $P_a(v_a > 1,094) = 0,182$ et $P_b(v_b > 0,358) = 0,544$. **Ces résultats indiquent qu'il n'y a aucune raison d'abandonner l'hypothèse d'une répartition uniforme des x_a et x_b.**

Reste à envisager l'hypothèse d'une même loi de Poisson. On calcule alors la valeur. $\theta = N_b/(N_a + N_b) = 0,709$. Il suffit d'utiliser les relations (3) avec la valeur $K_a + K_b = 83$ pour trouver la valeur inférieure de K_b : $K_{\inf} = 52$. La valeur de K_b étant 25, **il nous faut renoncer à l'hypothèse d'une même population parente**.

Il est facile de recommencer la même application avec la forme commune du masculin et du féminin déclinée au nominatif singulier X = « tis » (cas particulier de déterminant « hermaphrodite » en grec). Alors on dénombre $K_a = 20$ et $K_b = 11$, puis on détermine les fonctions de répartition qui permettent d'obtenir $D_a = 0,250$ et $D_b = 0,250$; enfin, on calcule $v_{0a} = 1,118$ et $v_{0b} = 0,829$. À nouveau la probabilité de pouvoir dépasser ces valeurs est dans chaque cas supérieure à $0,05$ car $P_a(v_a > 1,118) = 0,164$ et $P_b(v_b > 0,829) = 0,498$. **On conserve donc l'hypothèse d'une distribution uniforme.**

Pour examiner l'hypothèse d'une seule population parente, il suffit de consulter les tables ou d'effectuer les calculs pour la valeur $K_a + K_b = 31$, on obtient alors : $K_{\inf} = 17$. Ici encore, **on devra renoncer à l'hypothèse d'une unique loi de Poisson** car $K_b = 11$.

Nous avons pu vérifier en terme statistique notre hypothèse : les emplois de « tis » ne sont pas parents. Au cours des siècles, ce phénomène s'accentue jusqu'à la disparition complète du morphème « tis, ti » (gr. byzantin « tinas ») dans la langue démotique. Elle s'amorce dans les dialectes au XV^{e} siècle et prend un tour définitif au XVII^{e} siècle au profit de « enas, mia, ena ». Ce morphème, que l'on trouve sous la forme « heis, mia, hen » jusqu'au X^{e} siècle, désigne en grec ancien le numéral « un » et s'emploie occasionnellement comme pronom indéfini. Dans la langue du Nouveau Testament, il a le statut supplémentaire d'adjectif pronominal à valeur indéfinie. Il ne s'agit pas encore véritablement d'un article. qui acquièrent pendant ce même intervalle de temps toutes les valeurs actuelles de l'article indéfini en grec (Guilpin, 2002). On pourrait vérifier ce phénomène au moyen de notre méthode.

3.3. *Stabilité des emplois de l'article défini*

Originellement, notamment chez Homère, « ho » est un démonstratif, ce qui entraîne dialectalement un emploi comme relatif, puis le mot devient un article, qualifié aujourd'hui de défini, dont les emplois sont bien établis en grec classique (Chantraine, 1968). C'est à partir

de l'examen de ce morphème que Denys le Thrace (170 – 90 avant J.-C.) fonde la catégorie grammaticale de l'article (gr. anc. « arthron », *articulation*, puis lat. « articulus »). Le terme d'*article* lorsqu'il est créé désigne strictement les emplois définis de « ho, hê, to », l'article indéfini n'apparaissant que bien ultérieurement (Guilpin, 2002).

Vérifions que les emplois de « ho, hê, to » sont stables quels que soient la nature du texte étudié et le style de l'auteur.

Intéressons-nous à la forme X = « ton », accusatif masculin singulier de « ho, hê, ton » (le choix de la flexion est arbitraire et le morphème « ton » traité de façon globale). On trouve alors $K_a = 48$ et $K_b = 144$. La détermination des fonctions de répartition fournit les résultats suivants $D_a = 0,055$ et $D_b = 0,0486$ à partir desquels on calcule $v_{0a} = 0,415$ et $v_{0b} = 0,583$. La probabilité de pouvoir dépasser ces valeurs est dans chaque cas supérieure à $0,05$ car $P_a(v_a > 0,415) = 0,99$ et $P_b(v_b > 0,183) = 0,886$. **On conserve donc l'hypothèse d'une distribution uniforme**.

Examinons à présent l'hypothèse d'une seule distribution parente. Il suffit de consulter les tables ou d'effectuer les calculs pour la valeur $K_a + K_b = 192$, on obtient alors pour $K_a : K_{\inf} = 45$. Or K_a vaut 48, **donc il n'y a pas de raison de rejeter l'hypothèse d'une même population parente** (on échange les indices a et b de la relation (2), donc $\theta = N_a / (N_a + N_b) = 0,291$).

Notre hypothèse est donc vérifiée : les emplois de l'article défini sont stables. Cette question va de soi et nous permet d'achever la validation de la méthode.

4. Conclusion

Au cours de la mise au point de cette méthode, nous n'avons jamais pas rencontré de cas où la loi de Poisson devait être rejetée. Toutefois, il est possible qu'un tel cas échoit, ce qui n'interdit pas la poursuite du calcul, mais les résultats devront être traités avec circonspection. On pourra consulter Cox pour l'étude de la comparaison des taux de processus non poissonniens.

Ici, nous avons choisi volontairement des exemples simples et mis en contraste des phénomènes caractéristiques dans la détermination du nom en grec ancien.

Cette méthode pourra être utilisée à des fins plus spécifiques quelle que soit la taille du corpus (dès lors qu'il est jugé représentatif pour la démonstration) et quelle que soit la langue étudiée. De plus, elle permet de traiter aussi bien des données en synchronie qu'en diachronie.

Références

Aïvazian S. (1970). *Étude statistique des dépendances*. MIR.

Biraud M. (1991). *La détermination du nom en grec classique*. Publications de la Faculté de Lettres et Sciences Humaines de Nice.

Chantraine P. (1968-1980). *Dictionnaire étymologique de la langue grecque*. 4 volumes. Klincksieck (reprint 1999).

Cox D.R. et Lewis P.A.W. (1969). *L'analyse statistique des séries d'événements*. Dunod.

Guilpin Ch. (1999). *Manuel de calcul numérique appliqué*. EDP Sciences.

Guilpin P. (2002). Το αόριστο άρθρο στα ελληνικά : Διαχρονική μελέτη (trad. L'article indéfini en grec : étude diachronique). In Clairis Ch. (Ed.), *Recherches en linguistique grecque*, vol. (1), L'Harmattan, *Actes du 5e Colloque International de Linguistique Grecque*.

Guilpin P. (2003). Les textes grecs des origines à nos jours (Ve siècle av. J.-C. – XXIe siècle) – Codage, outils et méthodes de travail. *Lexicometrica* (4).

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Armand Colin/Masson.

Horrocks G. (1997). *Greek : A History of the Language and its Speakers*. Longman.

Jannaris A.N. (1897). *An Historical Greek Grammar, chiefly of the Attic dialect*. Macmillan, 1897 (reprint 1987, Georg Olms : Hidelsheim, Zurich and New York).