

Quelques utilisations possibles de la modélisation du langage par des chaînes de Markov

Jean-Pierre Anfosso



Édition électronique

URL : <http://journals.openedition.org/corpus/46>

DOI : [10.4000/corpus.46](https://doi.org/10.4000/corpus.46)

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 15 décembre 2003

ISSN : 1638-9808

Référence électronique

Jean-Pierre Anfosso, « Quelques utilisations possibles de la modélisation du langage par des chaînes de Markov », *Corpus* [En ligne], 2 | 2003, mis en ligne le 15 décembre 2004, consulté le 07 septembre 2020. URL : <http://journals.openedition.org/corpus/46> ; DOI : <https://doi.org/10.4000/corpus.46>

Ce document a été généré automatiquement le 7 septembre 2020.

© Tous droits réservés

Quelques utilisations possibles de la modélisation du langage par des chaînes de Markov¹

Jean-Pierre Anfosso

1. Linguistique et critique des textes

- 1 Parmi les questions posées par l'étude des manuscrits et l'établissement des textes, il y a des problèmes d'attribution de fragments anonymes, des problèmes de reconnaissance d'interpolations dues à certains copistes, et plus généralement, de reconnaissance de scripteurs différents dans un corpus donné. Il s'agit donc, tantôt de rapprocher d'un corpus ou d'un scripteur connus un fragment inconnu, tantôt de déceler dans un corpus donné, un ou plusieurs passages qu'il faut attribuer à un autre scripteur.
- 2 Pour rapprocher des textes ou les différencier, on peut songer d'abord à une critique externe, fondée sur tous les renseignements qu'on peut avoir sur les supports (paléographie, datation et origine géographique, par des procédés qui vont de l'étude des poussières, des pollens, jusqu'à la recherche du code génétique de l'animal qui a fourni la peau du parchemin).
- 3 On peut songer aussi à une critique historique des textes, fondée entre autres sur l'étude des allusions à des événements, sur l'étude de la toponymie, etc. Mais tout cela peut être trompeur, et même volontairement : il suffit qu'un copiste ait interpolé un court passage faisant allusion par exemple à un règne connu par ailleurs, pour fausser la datation ...
- 4 On est donc conduit à essayer de faire aussi de la critique interne, et à essayer de faire dire au texte ce qu'il n'était pas destiné à dire.
- 5 Or les problèmes posés plus haut (attribution, etc.), replacés dans le cadre d'une critique qui ne s'intéresse qu'au texte, peuvent se ramener à caractériser un scripteur, voire une personnalité, à travers son langage. Par exemple en étudiant des termes ou des expressions qui apparaissent de manière récurrente et quasi automatique, quel que soit le sujet traité.

- 6 Cette critique interne peut se concevoir à divers niveaux.
- 7 Au niveau de la sémantique on pourrait utiliser, de manière systématique, des méthodes analogues à un test de Jung, en interrogeant le texte par la recherche des mots associés le plus fréquemment à certains mots pris dans une liste (par exemple l'ensemble des adjectifs associés à un nom). Ces méthodes rejoindraient d'ailleurs celles de la psychocritique (exposées entre autres par Charles Mauron). Une partie du texte étant alors considérée comme dictée par l'inconscient. Nous sommes là au niveau le plus élevé, celui de la signification.
- 8 Le scripteur a donc des habitudes, voire des tics, qui se traduisent par des mots et des formules récurrentes, qu'on peut aussi retrouver, sans tenir compte de leur signification, au niveau le plus bas, dans la répétition de certaines suites de lettres, ou au niveau syntaxique, dans l'enchaînement de certaines catégories de mots. Ce sont ces entités que nous allons utiliser dans la suite.

2. Choix d'une modélisation statistique

- 9 Si nous utilisons par exemple le dénombrement des *n-grammes* pour caractériser des textes (que ce soit du point de vue du scripteur, du genre, du sujet, etc.), pour les rapprocher ou pour les distinguer, les modèles statistiques peuvent se classer en deux catégories :
- soit on ne considère que le nombre de chaque *n-gramme* (gamme des fréquences), sans se préoccuper ni de leur position, ni de leur enchaînement, les *n-grammes* étant comme des boules mélangées dans une urne, ou comme les molécules d'un gaz (ce qui correspond à un modèle thermodynamique),
 - soit on considère le texte de manière séquentielle, comme un message produit par des tirages successifs de symboles (ce qui correspond à un processus discret).
- 10 Dans le deuxième cas, on peut alors essayer d'ajuster un modèle markovien d'ordre $n-1$ sur les *n-grammes*. Si c'est possible, et comme, sous certaines hypothèses, une chaîne de Markov est un processus ergodique², on aura des informations sur l'enchaînement d'un état à l'autre (probabilité de transition), et pour passer de tel à tel état en n transitions, on aura les mêmes propriétés qu'une fonction d'autocorrélation (dans le cas ergodique) : la probabilité ne dépendra pas de la position, mais seulement de la distance entre les deux symboles.
- 11 Quand on considère l'ensemble des textes ayant une même « gamme des fréquences » pour les *n-grammes*, ceux qui sont markoviens (homogènes) d'ordre $n-1$, constituent un sous-ensemble. Ces derniers ont des propriétés supplémentaires.
- 12 Si les deux types de modèle permettent de rapprocher des textes, avec par exemple une analyse de données à partir de la statistique des *n-grammes* dans le cas général, les chaînes de Markov permettent en outre de calculer la vraisemblance d'une séquence dans un modèle, c'est-à-dire la probabilité que cette séquence appartienne à ce modèle (ajusté sur un corpus, un scripteur, etc.).
- 13 Mais pour utiliser les propriétés des chaînes de Markov, on doit faire un test d'adéquation, que les textes (considérés comme suite de lettres, de mots, etc.) ne vérifient pas en général.
- 14 Nous sommes donc amenés à regrouper soit des lettres, soit des mots en catégories, ce qui diminue le nombre d'états et peut rendre la probabilité critique (du test) acceptable. On traite donc, à la place du texte réel, une séquence transcrite systématiquement, mais réductrice (par exemple, pour les catégories de lettres, en

utilisant un alphabet à quatre symboles : voyelle, consonne, semi-voyelle, espace). La même suite d'états pouvant renvoyer « en clair » à plusieurs séquences différentes, elle est indéchiffrable, et ce qu'on pourra mettre en évidence était invisible pour le scripteur lui-même.

- 15 Cette transcription entraînant une perte d'information, le problème est de savoir s'il restera une information suffisante pour faire la différence entre deux styles, et pour caractériser quelque chose. Seules les expériences peuvent répondre à cette question.

3. Les apports de la biochimie génétique

- 16 Nous allons voir que dans l'étude des génomes on retrouve des problèmes analogues à ceux que nous nous sommes posés au début.

- 17 Après « séquençage », une molécule d'ADN se présente comme une suite de bases (notées par les lettres *ACGT*). Elle se divise en régions « codantes » (les gènes), séparées par des régions « non codantes », chaque gène étant chargé de coder la synthèse d'une protéine. Les gènes seront transcrits dans l'ARN messenger, puis traduits en une suite d'acides aminés, qui constituent la protéine. Chez les eucaryotes (êtres possédant un noyau dans leurs cellules), le gène lui-même comprend des « exons », fragments qui seront traduits en acides aminés, et des « introns », fragments qui seront transcrits, mais non traduits.

- 18 Ces différentes zones (non codantes, codantes, introns, exons, etc.) peuvent en général être représentées par des modèles markoviens de divers ordres, et de divers paramètres. Chaque type de zone, dans un génome donné (ou dans des génomes de la même famille), a des paramètres suffisamment stables pour qu'on puisse ajuster un modèle sur des parties déjà connues et analysées. On aura donc un modèle pour les régions non codantes, un pour les gènes, etc., et un fragment d'ADN est censé pouvoir se découper en « plages » homogènes de deux à quatre paramètres différents.

- 19 Cela pose deux sortes de problèmes :

- en supposant qu'on ait délimité les différentes zones, et qu'on ait estimé leurs paramètres, on se donne un fragment inconnu assez court (supposé homogène) : comment savoir de quel type de zone il fait probablement partie ?
- comment, dans une séquence assez longue, délimiter les différentes zones qui s'y trouvent ?

- 20 Les méthodes utilisées par la bio-informatique pour répondre à ces questions, vont nous servir pour traiter des problèmes analogues sur des textes naturels.

- 21 Si on a par exemple, ajusté des modèles sur différents corpus (ou auteurs), et qu'on se donne un fragment inconnu, on pourra essayer, avec les procédés de la génomique pour résoudre la première question, de dire à quel corpus (ou auteur), il appartient vraisemblablement (problème d'attribution).

- 22 Si on suppose qu'un texte contient des passages dus à un scripteur différent, on pourra, avec les procédés utilisés par la génomique pour résoudre la deuxième question, chercher à délimiter les zones homogènes (par exemple pour trouver l'emplacement d'interpolations).

4. Méthodes et algorithmes

- 23 Pour la première question (attribution d'un fragment), la bio-informatique ajuste des modèles sur les différents types de zone avec des parties de génomes déjà analysées et découpées, calcule la vraisemblance du fragment inconnu dans les différents modèles, et attribue le fragment au modèle pour lequel la vraisemblance est maximale.

Les modèles étant markoviens, si on note $X = x_1 \dots x_h$ une séquence inconnue de longueur h , $\pi(x_i, x_j)$ les probabilités de transition, μ le vecteur invariant, sa vraisemblance dans un modèle q sera :

$$L(X, \text{modèle } q) = \mu_q(x_1) \prod_{i=1}^{h-1} \pi_q(x_i, x_{i+1})$$

- 24 L'ensemble des états et l'ordre k doivent être les mêmes, les modèles ne différant que par le paramètre.
- 25 La même méthode peut s'appliquer à des textes naturels, après le choix d'une transcription (par exemple catégories de lettres) et d'un ordre k , qui donnent une probabilité critique acceptable pour le test d'adéquation, les modèles étant ajustés sur différents corpus (corpus qui peuvent aller jusqu'à recouvrir, par des échantillons, tout le domaine connu d'un scripteur).
- 26 Pour la seconde question (découpage d'une séquence en zones homogènes différentes), il faut distinguer deux cas :
 - soit les modèles des différentes zones sont connus (il ne reste que le découpage à faire),
 - soit on n'a aucune connaissance *a priori*, ni sur les emplacements des zones, ni sur les modèles.
- 27 Dans le premier cas, on utilise la même méthode que pour la première question. Supposons qu'on ait n modèles indexés par q , et que la séquence totale Y puisse se découper en plages correspondant à ces modèles ; on prend au début, un échantillon Y' d'une longueur suffisante pour pouvoir l'attribuer à un type de zone, et on calcule pour tout q , $L(Y', \text{modèle } q)$. On déplace alors l'échantillon d'une position, et on reporte chaque fois les n vraisemblances sur un graphique. Quand Y' balaie toute la séquence Y , on trace ainsi n courbes, chacune correspondant à un modèle : dans chaque intervalle où une courbe dépasse toutes les autres, on peut considérer qu'on a une zone homogène pour ce modèle. Mais cette méthode ne peut s'appliquer à des textes naturels, pour y chercher des interpolations par exemple, que si on a une hypothèse sur les différents scripteurs d'un corpus donné.
- 28 Dans le deuxième cas (où les différents paramètres sont inconnus), on utilise les *modèles de Markov cachés (HMM)* pour représenter l'hétérogénéité d'une séquence. La suite des bases (en génomique) ou la suite des états pour les textes naturels, est « la séquence observée ». La séquence parallèle (de même longueur), qui, à chaque état, associe le numéro de la zone dans laquelle il se trouve, est la « chaîne cachée » (notée S).
- 29 En donnant les différentes matrices B^q correspondant à chaque zone q , et la matrice A ajustée sur la chaîne cachée à l'ordre 1, un HMM modélise une séquence hétérogène Y .
- 30 Si le HMM est connu, à partir de Y , on peut trouver la probabilité qu'à telle position on soit dans telle zone, et reconstituer la chaîne cachée la plus probable, c'est-à-dire la délimitation des différentes zones (par exemple avec les algorithmes *forward-backward*, ou celui de *Viterbi*).

Il est facile d'ajuster un HMM quand on connaît S , mais la chaîne cachée est précisément ce qu'on cherche. Il faut donc des méthodes pour l'ajuster uniquement à partir de la séquence observée Y (sans connaître ni l'emplacement des zones, ni les modèles correspondants).

Un HMM étant donné (même arbitrairement), les algorithmes *forward* ou *backward* permettent de calculer la vraisemblance de Y dans ce modèle : $L(Y, \text{modèle } \theta)$.

On est donc amené à chercher le modèle qui maximise cette vraisemblance :

$$\theta^* = \underset{\theta}{\operatorname{argmax}} L(Y, \theta)$$

Malheureusement dans ce cas, la méthode des multiplicateurs de Lagrange ne conduit pas à des équations qu'on peut résoudre directement. On a donc recours à des méthodes itératives qui, à partir d'un modèle donné θ^0 , permettent de construire un modèle θ^1 pour lequel la vraisemblance de Y est plus grande, c'est-à-dire :

$$L(Y, \theta^1) > L(Y, \theta^0)$$

méthode qu'on peut répéter jusqu'à atteindre un maximum.

Malheureusement dans ce cas, la méthode des multiplicateurs de Lagrange ne conduit pas à des équations qu'on peut résoudre directement. On a donc recours à des méthodes itératives qui, à partir d'un modèle donné θ^0 , permettent de construire un modèle θ^1 pour lequel la vraisemblance de Y est plus grande, c'est-à-dire :

$$L(Y, \theta^1) > L(Y, \theta^0)$$

méthode qu'on peut répéter jusqu'à atteindre un maximum.

31 La bio-informatique utilise entre autres l'algorithme de Baum-Welch. La difficulté est qu'il faut se donner un modèle initial θ^0 obtenu par un autre procédé. Certains spécialistes de génomique préconisent un modèle initial carrément arbitraire. Mais dans le cas qui nous occupe, celui des textes naturels, cela conduit souvent à des maxima locaux.

32 L'initialisation est donc un problème crucial. Toutefois on peut s'aider du fait que, si on se donne plusieurs modèles initiaux, l'algorithme *forward*, en calculant la vraisemblance de Y dans chacun d'eux, peut dire quel est le meilleur avant tout recours à une méthode itérative.

5. Essais et résultats

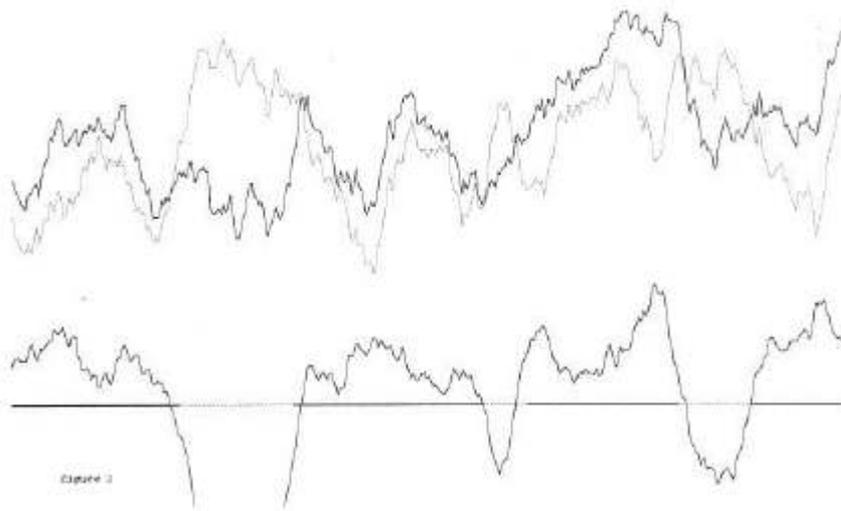
33 La première série d'expériences concerne le test d'adéquation. Il s'agit de savoir dans quelle mesure un modèle markovien peut être ajusté sur une séquence donnée.

34 On s'intéresse aux suites de deux transitions (*k+2-grammes*) ; la somme z_n des carrés des écarts réduits, mesure une distance entre la distribution théorique et la distribution observée. Cette distance suit une loi du χ^2 à k degrés de liberté. La valeur de k croissant avec la longueur de la séquence, cette séquence ne doit pas être trop courte, ce qui peut conduire à des valeurs de $z_n < 1$ et rendre le test impossible.

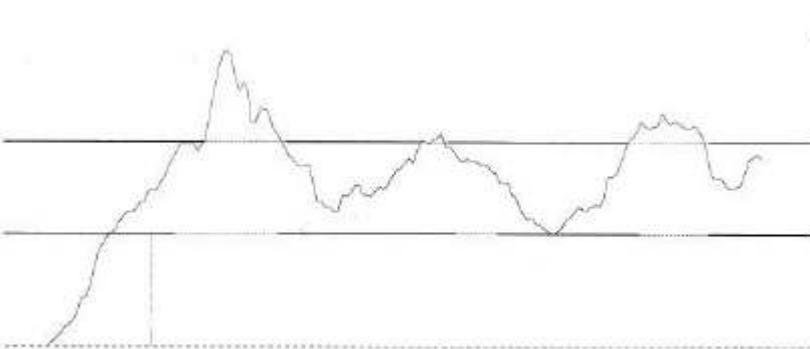
35 La probabilité critique est donnée par l'intégrale du χ^2 de z_n à $+\infty$; le programme la calcule si elle est supérieure à 0,05 (5%). Bien qu'elle puisse, dans certains exemples, dépasser 50%, on peut considérer qu'à cette valeur (qui correspond à $z_n = 1$, moyenne du

\mathbb{F}_2^2), l'adéquation est parfaite (puisque avec de vraies chaînes de Markov, en raison des fluctuations aléatoires, $z_n \xrightarrow[n \rightarrow \infty]{} 0$ quand $n \rightarrow \infty$).

- 36 Le premier essai porte sur un texte latin, *De oratore* de Cicéron, transcrit en catégories de lettres : voyelle, consonne, semi-voyelle, espace. On a des valeurs acceptables avec $k=1$ ou 2 , mais pour de petites séquences ; avec $k=3$, pour n allant de 100 à 1400, la probabilité varie entre 17% et 65%. Le test est aussi acceptable avec $k=4$, mais sur une plage plus réduite. Si on supprime les espaces, on a des résultats encore meilleurs, mais le texte transcrit contient moins d'information.
- 37 Le deuxième corpus est un autre texte latin, le *Satiricon* de Pétrone, transcrit en catégories de mots. Les catégories grammaticales sont fournies par le fichier du LASLA³, mais regroupées en 5 classes (noms, adjectifs-pronoms, verbes, prépositions, adverbes-conjonctions-interjections). Avec $k=1$ le test est acceptable pour de petites séquences (11% à 52% pour n allant de 100 à 250), avec $k=3$ pour n allant de 4000 à 6000, la probabilité varie entre 18% et 27%.
- 38 Ces exemples suffisent à montrer que, sous réserve de vérification, pour certaines transcriptions des textes, pour certaines valeurs de k et pour certaines longueurs de séquence, les modèles markoviens constituent une approximation utilisable.
- 39 La deuxième série d'expériences concerne l'attribution d'un fragment. Elles sont faites avec des textes latins transcrits en catégories de lettres.
- 40 Les premiers essais portent sur trois ouvrages de trois auteurs différents (*De oratore* de Cicéron, *Adelphi* de Térence, *Metamorphoseon* d'Apulée). Trois modèles sont ajustés sur des échantillons de 10000 caractères, et on prend dans chaque corpus des fragments en dehors des échantillons. L'attribution est bonne avec $k=2$ pour des fragments d'environ 1000 caractères. Avec $k=3$ on peut attribuer correctement jusqu'à 250 caractères, et avec $k=4$, jusqu'à 125 caractères (environ 25 mots). Mais cela revient à distinguer des œuvres particulières, qui diffèrent par l'auteur, mais aussi par l'époque, le genre, le sujet ...
- 41 Pour essayer de caractériser un scripteur, nous avons pris six œuvres de même genre appartenant à deux auteurs différents, ajusté un modèle avec des échantillons de deux œuvres morales de Cicéron (*De amicitia* et *De senectute*), un autre modèle avec des échantillons de deux œuvres morales de Sénèque (*Ad Marciam de consolatione* et *De brevitate vitae*), pris deux fragments d'environ 500 caractères dans d'autres ouvrages moraux de Sénèque (*Ad Helviam de consolatione*) et de Cicéron (*De officiis*). Chaque fragment est attribué correctement à son auteur, alors que les œuvres d'où ces séquences sont tirées n'ont pas servi à ajuster les modèles.
- 42 La troisième expérience concerne la recherche de zones homogènes différentes dans une séquence assez longue, les modèles étant supposés connus par ailleurs.
- 43 A titre d'exemple, nous avons pris une séquence de Cicéron (environ 23000 caractères), dans laquelle nous avons inséré, à trois endroits différents, des fragments de Térence (3000, 1200 et 2000 caractères). Nous utilisons la méthode présentée à la section 4, dont on peut voir les résultats sur la figure 1.

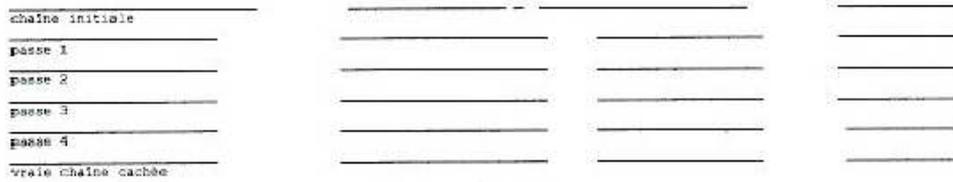


- 44 Pour l'ordre 4, et une séquence mobile de 1000 caractères, les deux courbes supérieures représentent la vraisemblance dans le modèle de Cicéron (noir) et dans celui de Térence (gris). La courbe du dessous représente leur différence, les parties positives étant censées déterminer ce qui est dû à Cicéron, les parties négatives, ce qui est dû à Térence. L'axe représente, en gras et en pointillé, le vrai découpage du texte (gras pour celui de Cicéron, pointillé pour les fragments insérés de Térence). La courbe de la différence repère visiblement les trois interpolations, avec des erreurs sur le début ou la fin, dont les plus grandes vont de 150 à 300 caractères environ (30 à 60 mots).
- 45 La quatrième expérience concerne la recherche de zones différentes, mais sans aucune connaissance des modèles.
- 46 Nous essayons de situer les interpolations de la séquence étudiée précédemment (sans utiliser la connaissance des textes qui la composent).



- 47 La courbe de la figure 2 représente une distance (du \mathbb{R}^2) entre un modèle ajusté au début et un modèle mobile, pour $k=3$ et une séquence mobile de 3000 caractères. Cela permet de faire un découpage approximatif (droite du haut), qu'on peut comparer au vrai découpage (droite du bas).
- 48 Ce découpage approximatif permet de construire une chaîne cachée initiale S_0 avec laquelle on ajuste un HMM à (Y, S_0) . Avec ce modèle \mathbb{R}^0 , l'algorithme *forward-backward* fournit, au bout de quatre itérations de l'algorithme de Baum-Welch, une chaîne cachée

améliorée qui rejoint pratiquement le vrai découpage, l'erreur la plus grande, qui porte sur la fin de la troisième interpolation, étant de 13 mots, sur les 400 environ de cette interpolation (figure 3).



6. Bilan et perspectives

- 49 Bien sûr certains exemples que nous avons donnés peuvent passer pour des exercices d'école. Ils sont toutefois suffisants pour montrer qu'on peut envisager d'utiliser les procédés que nous avons présentés, comme moyens d'« attaque » de certains problèmes posés par les textes, pour diverses transcriptions et dans diverses langues. D'ailleurs, des possibilités sont offertes pour approfondir ces méthodes et perfectionner ces outils.
- 50 Nous avons regroupé lettres ou mots en certaines catégories, mais la question d'autres transcriptions en corpus « sous-jacents » qui aient la meilleure adéquation possible à des modèles markoviens, reste ouverte (en particulier par agglomération d'états fondée sur les probabilités de transition).
- 51 De plus ces corpus sous-jacents pourraient être conçus de manière à filtrer les « harmoniques » d'une écriture et distinguer ce qui est dû à la langue, à un scripteur, à un genre, au sujet. Pour commencer, beaucoup de procédés et d'algorithmes de la bio-informatique restent à exploiter (Viterbi training, méthode du gradient, etc.). Mais tous les détours par l'analyse du génome, n'auront pour but, en définitive, que de répondre à cette question : un scripteur peut-il avoir une « empreinte génétique » ?

BIBLIOGRAPHIE

- Anfosso J-P., (2002). Contribution à une modélisation statistique du langage et à sa mise en œuvre informatique. Thèse, Université de Nice-Sophia Antipolis.
- Baldi P. & Brunak S. (1998). Bioinformatics. Cambridge, Massachusetts : MIT Press.
- Cellier D. (polycop). DESS Etude de génomes. « Analyse statistique de séquences ». Université de Rouen, janvier 2000.
- Durbin E. & Krogh M. (1998). Biological Sequence Analysis. Cambridge : Cambridge University Press.
- Manning C.D. & Schütze H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts : MIT Press.

NOTES

- 1.. Thèse dirigée par le professeur É. Brunet et soutenue à l'Université de Nice-Sophia Antipolis devant MM. L. Lebart et A. Salem (4 novembre 2002).
 - 2.. C'est-à-dire dont on peut relever les caractéristiques statistiques (qui restent les mêmes) sur n'importe quel extrait (d'une longueur suffisante).
 - 3.. *Laboratoire d'Analyse Statistique des Langues Anciennes* (Université de Liège) qui a lemmatisé et étiqueté une grande partie des textes de la latinité classique.
-

AUTEUR

JEAN-PIERRE ANFOSSO

« Bases, Corpus et Langage », UMR 6039