Histoire & mesure

XVIII - 3/4 | 2003
Mesurer le texte

Des chiffres et des lettres

Quelques pistes pour l'historien

Jean-Philippe Genet et Pierre Lafon

revues.org

Édition électronique
URL : http://histoiremesure.revues.org/819
DOI : 10.4000/histoiremesure.819
ISSN : 1957-7745

Éditeur
Éditions de l'EHESS

Édition imprimée
Date de publication : 2 décembre 2003
Pagination : 215-223
ISBN : 2-222-96740-6
ISSN : 0982-1783

Référence électronique

Des chiffres et des lettres

Quelques pistes pour l'historien

Jean-Philippe Genet et Pierre Lafon

1 Les historiens ont de tout temps utilisé les textes. Mais leur approche a longtemps été purement empirique. Au fond, l'objectif primordial de l'historien était de disposer d'un texte sûr, d'où l'accent sur la critique des sources d'une part, et sur l'établissement du contenu de l'autre : une fois le texte solidement établi, on pouvait l'utiliser avec le statut de source autorisée et sous la forme de citations. Le découpage sélectif produit par le choix des citations tenait lieu d'interprétation et permettait de se livrer ensuite, la conscience tranquille, aux joies de la « synthèse subjective », pour paraphraser Michel Foucault. Si l'on excepte le recours aux méthodes philologiques pour les périodes anciennes et médiévales (reconnu de bonne grâce dans la mesure où prendre appui sur une discipline aussi prestigieuse ne pouvait que conforter le statut de l'historien), le chercheur régnait ainsi sur une série de « sciences auxiliaires » vouées à l'aider au déchiffrement et à l'interprétation littérale des sources (épigraphie, paléographie, héraldique, sigillographie, diplomatique). Fort de son attelage pédagogique avec la géographie, il pouvait s'estimer dispensé de tout contact approfondi avec les autres sciences humaines et sociales. On le sait, Marc Bloch et Lucien Febvre, fascinés par la sociologie de Durkheim comme par l'anthropologie de Frazer et de Mauss, convaincus de la nécessité de travailler avec les économistes, ont changé tout cela. Mais la « révolution des *Annales* » n'a guère ébranlé la posture historienne face au texte.

2 De fait, celle-ci a longtemps résisté et les historiens des *Annales* ne se sont pas passionnés pour ce problème. C'est, en réalité, l'évolution de la discipline linguistique elle-même qui a conduit les historiens à réfléchir sur l'insuffisance de leur approche traditionnelle : l'irrésistible ascension de la linguistique, science phare du structuralisme triomphant dans les années soixante, est venue troubler la bonne conscience des historiens. Et si le texte était discours ? Et s'il devait donc être envisagé comme pratique, en tenant compte de ses conditions de production aussi bien que de toutes ses relations à ce qui est extralinguistique et qu'il faudrait alors tenter de prendre en compte par des concepts socio-culturels, tel celui de « champ » (Pierre Bourdieu)[1] ou celui de « formation

discursive » (Michel Foucault)[2] ? Et si ce discours ne pouvait se comprendre qu'en fonction d'une langue, dont Saussure nous assurait qu'elle avait pour seule fonction d'être un instrument de communication, une institution sociale par excellence fonctionnant comme un système, voire comme un code, ce qui impliquerait qu'un texte quelconque ne puisse faire sens que synchroniquement, par rapport à ce système ?

3 De telles interrogations pouvaient avoir des conséquences redoutables pour l'historien : puisque chaque élément linguistique, chaque signe, avait une valeur qui correspondait précisément à sa fonction et à sa place dans le système de la langue à un moment donné, il était donc impossible d'établir la valeur d'un signe quelconque sans comprendre l'ensemble du système. Du coup, la construction du discours historique, à partir d'une structuration le plus souvent diachronique des citations, le rendait hautement suspect. De ce recentrement de la linguistique sur la synchronie provient la notion de norme linguistique et, partant, celle de variation ou d'écart par rapport à cette norme. Tous ces problèmes ont été très bien exposés par Régine Robin – une historienne confrontée aux cahiers de doléances, dans un livre qui reste, aujourd'hui encore, la meilleure introduction à la problématique des rapports de l'histoire et de la linguistique[3].

4 Depuis les années soixante, cependant, les historiens ont beaucoup travaillé. D'une part, ils ont porté toute l'attention qu'elles méritent à la matérialité du texte et aux façons dont il s'inscrit et se transmet dans le système de communication. Comme les linguistes qui travaillent aujourd'hui sur la notion de brouillon ou sur l'écriture informatique, ils ont intégré cette exigence, et, par exemple, les travaux des médiévistes sur l'écrit et sur la codicologie témoignent largement de ces préoccupations : certains d'entre eux ont réalisé qu'il était absurde de reconstituer un texte hypothétique qui n'avait jamais existé en collectant les meilleures leçons de divers manuscrits. Si l'on voulait donner à un texte toute sa valeur en tant que source, il devenait préférable d'éditer le texte, même imparfait, d'un manuscrit donné, mais en prenant en compte sa matérialité (autographe, copie de luxe ou non, type d'écriture, disposition dans la page, densité des abréviations ou des ratures, commentaires marginaux, etc.), et son environnement précis (textes contenus dans le même volume, documents insérés etc.), tous éléments riches d'informations complémentaires. Quant au recours à la linguistique et à ses méthodes, il a surtout concerné la lexicologie (lexique) et la sémantique (étude des contextes), en faisant plus particulièrement appel à la lexicologie quantitative.

5 Les analyses de discours ont d'abord porté sur des textes courts (les ordinateurs étaient peu puissants et la difficulté de la saisie sur carte perforée militait pour la brièveté !) et à contenu dense, comme les pamphlets, les tracts, ou les articles de journaux. Des recherches de lexicologie quantitative ont ainsi été menées sur la Révolution américaine, la Révolution française, la Commune, le Congrès de Tours et, un peu plus tard, les évènements de mai 1968, quelques historiens venant alors travailler aux côtés des linguistes et s'initier à la linguistique. Dans les années soixante-dix, il a même paru naturel de créer une unité de valeur d'initiation à la linguistique, assurée par Frédéric François, à l'UFR d'Histoire de Paris I !

6 Hélas, depuis quelques années, il faut bien constater un reflux. Pourtant, les obstacles techniques ont disparu : rien de plus facile aujourd'hui que de scanner un texte – si du moins il a été imprimé dans la seconde moitié du XIX^e^ siècle ou après – et de le traiter avec un logiciel performant et convivial ! Ce reflux paraît avoir trois raisons majeures.

7 La première est l'évolution de la linguistique elle-même : les linguistes travaillent désormais sur des corpus énormes, et qui plus est sur des corpus étiquetés, c'est-à-dire comportant pour chaque mot les indications morpho-syntaxiques nécessaires : BNC (*British National Corpus*), par exemple, comporte cent millions de mots. Les historiens n'ont pas suivi, ne serait-ce que parce que l'étiquetage est un travail énorme qu'ils jugent, à tort ou à raison, hors de leur portée et parce que ces corpus sont exclusivement contemporains.

8 Deuxième raison, l'investissement méthodologique dans les recherches historiques a, contrairement à ce à quoi l'on aurait pu s'attendre, baissé. Si la recherche universitaire reste toujours dominée par le travail individuel dans le cadre de la thèse, celle-ci se fait désormais en temps limité, ce qui dissuade les doctorants de passer trop de temps à apprendre et à maîtriser des méthodologies complexes aux marges de leur discipline. S'ils sont docteurs, dans le meilleur des cas, il leur faut ensuite songer à une habilitation, un nouvel exercice individuel. Et, dans la plupart des cas, ils ne s'insèreront pas dans un véritable laboratoire où l'effort méthodologique peut être développé dans la continuité et grâce à une synergie pluridisciplinaire, simplement parce que de tels laboratoires n'existent pratiquement qu'au CNRS, ou en liaison avec lui, et qu'ils sont trop peu nombreux dans le champ de l'histoire en France ; d'ailleurs, leur nombre va diminuant.

9 Enfin, les résultats obtenus par les recherches utilisant les méthodes lexicométriques n'ont sans doute pas été assimilés et estimés à leur juste valeur par la communauté scientifique historienne. Sans doute auraient-ils pu être parfois exposés avec plus de clarté (le vocabulaire technique du linguiste déroute facilement le lecteur non initié), mais on peut aussi se demander si la vieille attitude subjectiviste des historiens n'est pas restée ou en tous cas redevenue dominante...

10 C'est pourquoi nous avons conçu ce numéro destiné à présenter une sorte d'état de la question. Dans ce reflux généralisé, marqué par la disparition quasi totale des grands projets collectifs, où en sont les historiens qui continuent à avoir recours à la lexicologie quantitative ? Les méthodes ont-elles évolué ? Des normes spécifiques aux historiens sont-elles apparues ? D'une façon générale, l'historien qui travaille sur des langues anciennes aura nécessairement besoin d'explorer longuement les aspects lexicologiques avant de passer à l'étude sémantique, alors que cette phase sera plus rapide pour celui qui peut disposer d'excellents dictionnaires. Pour celui qui n'est pas dans ce cas, il est nécessaire de construire d'abord un corpus, à partir duquel il pourra développer ses observations lexicologiques (présence ou absence d'un terme, fréquence d'usage, apparition-disparition, hapax) et sémantiques (contextes d'usage).

11 Il ne saurait être question d'aborder ici le problème des corpus en linguistique dans toute sa complexité, et certains linguistes (à commencer par Noam Chomsky) sont même hostiles par principe à la notion de corpus. Leur usage est néanmoins indispensable lorsque l'analyste d'une langue se trouve dans l'impossibilité de porter des jugements d'acceptabilité, ce qui est le cas pour la plupart des historiens travaillant sur des textes rédigés en langues anciennes, médiévales, et même modernes jusqu'au XVII^e^ siècle au moins, ou sur des textes utilisant ce que les linguistes appellent des « langages spécialisés ». Dans tous ces cas, s'il veut connaître ne serait-ce que le stock lexical de la langue particulière sur laquelle il travaille et pour laquelle les lexiques existants ne lui offrent pas de base solide[4], l'historien doit commencer par construire un corpus, avant de songer à l'exploiter.

12 En réalité, et c'est ce qui les différencie des linguistes, les historiens ne construisent pas leur corpus pour disposer de l'équivalent d'une norme de langue, mais pour répondre à des questions précises que leur pose leur documentation. Les corpus des historiens, comme les bases de données historiques d'une autre nature, sont à finalité essentiellement heuristique. La règle pour l'historien est de constituer un corpus comme un terrain d'expérience, en fonction de la problématique qu'il entend explorer. Bien sûr, il est de son intérêt de multiplier les angles d'attaque, et de contrôler la stabilité de l'ensemble construit – puisque le corpus est une construction –, en le soumettant à des expériences différentes. Les corpus des historiens ne répondront donc pas forcément exactement aux exigences strictes de la linguistique. Les proclamations des présidents mexicains sont certes formellement comparables mais elles s'ég rènent dans une diachronie qui interdit de penser que l'état de la langue n'a pas changé depuis Juarez ! L'autobiographie de l'empereur Charles IV, texte unique, n'est pas à proprement parler un corpus, et celui qu'étudie Aude Mairey est écrasé par le poids de l'une des poésies, *Piers Plowman*. Mais ces textes peuvent être réinsérés dans des corpus beaucoup plus vastes, des ensembles de textes historiographiques pour Charles IV ou de textes politiques anglais pour *Piers* : les normes ont alors changé, relativisant les écarts à la norme repérés dans la première configuration.

13 C'est la multiplication des mesures, relativement aisée à partir du moment où l'enregistrement des textes est cohérent, qui permet d'assurer l'interprétation. Naturellement, les exemples de corpus disparates sont d'autant plus nombreux que les textes sont rares : la richesse relative des archives des mouvements politiques et syndicaux, et surtout l'existence de la presse permettent de construire des corpus de plus en plus rigoureux au fur et à mesure que l'on progresse dans le temps, à partir de la fin du XVIII^{e} siècle. Mais, une fois le corpus constitué, il faut refuser de se lancer dans une lecture impressionniste des textes qu'il contient : il faut d'abord les mesurer, car c'est en refusant de les traiter banalement comme des sources qu'on pourra extraire tout leur apport en tant que source !

14 Ce numéro comporte plusieurs études fondées sur des dépouillements textuels. Elles sont destinées soit à éclairer une situation historique particulière, soit à tenter d'établir un diagnostic sur la proximité ou au contraire l'éloignement de textes ou de discours réunis dans un même corpus à des fins de conclusions historiques. Il s'agit dans tous les cas de mettre en œuvre des outils statistiques permettant de porter un jugement sur la ressemblance ou la dissemblance de textes, d'analyser l'évolution de séries textuelles, en mettant au jour les continuités ou les moments de fracture, bref d'établir des classifications en faisant apparaître des apparentements et/ou des oppositions entre les textes.

15 Mesurer un texte implique de remplacer la perception globale et singulière que l'on peut en avoir par un examen attentif du matériel lexical qui le compose. On se place, comme c'est toujours le cas aujourd'hui, dans la perspective d'un traitement automatique. Celui-ci suppose que les unités qui entrent en jeu dans le comptage soient déterminées avec précision. Quelles unités doit-on choisir ? On sait que ce point donne lieu à toutes sortes de pratiques divergentes. Certains s'en tiennent à l'apparence extérieure des signifiants et comptent des unités graphiques, d'autres adoptent un point de vue proche de celui des lexicographes et regroupent en lemmes. Mais on observe, d'une part que les pratiques lemmatisantes ne sont pas unifiées, d'autre part que, malgré la lemmatisation, toutes sortes de problèmes subsistent qui sont diversement réglés et donnent lieu à des

résolutions arbitraires. Parmi les problèmes les plus courants et les plus épineux pour les textes en français on trouve :

– le traitement des unités amalgamées telles que « au » « du » « des » « desquels » « lesquels » etc. qui sont tantôt éclatées en leurs unités sous-jacentes, tantôt systématiquement traitées comme une seule unité ;

– le traitement des formes verbales composées : « il a été mangé », deux, trois ou quatre unités ?

– le traitement des séquences figées, qu'il s'agisse de noms propres ou de lexies complexes, qu'elles soient lexicales, « pouvoir d'achat », « sécurité sociale », « affaires étrangères », « mettre en cause », « prendre en compte », ou grammaticales, « de telle sorte que », « d'ores et déjà », « au fur et à mesure ».

16 Les décisions prises concernant la norme du dépouillement sont conditionnées tant par les objectifs de la recherche que par le corpus étudié. Au sein d'un traitement brut en formes graphiques, la désambiguïsation de quelques homographies locales pour distinguer des emplois peut être jugée nécessaire mais suffisante (voir, par exemple, la façon dont Elsa Carrillo-Blouin distingue *estados* État et *estados* être). Enfin, certains travaux n'imposent pas de procéder à un dépouillement exhaustif du vocabulaire ; les mots grammaticaux, par exemple, ne sont pas pris en compte, voire seuls quelques « domaines lexicaux » sont retenus.

17 Quelle que soit la norme adoptée pour le comptage des unités, l'essentiel est évidemment de rester uniforme sur l'ensemble du corpus étudié de manière à éviter un biais dans les comparaisons. Plusieurs logiciels permettent d'obtenir un inventaire exhaustif ou au contraire sélectif du lexique d'un corpus. Ce numéro d'*Histoire & Mesure* comporte une brève présentation de trois d'entre eux[5] qui, en raison des fonctions multiples qu'ils offrent, sont parmi les plus couramment utilisés par les chercheurs de toute discipline (en particulier les historiens) lorsqu'ils souhaitent appréhender un corpus textuel par ce moyen. La première étape de la recherche consiste à constituer un inventaire fréquentiel du lexique des textes étudiés. Il est d'ailleurs possible de ne pas entrer véritablement dans la description statistique de cet inventaire et d'utiliser le logiciel comme un simple instrument documentaire d'exploration et d'aide à la lecture du texte : repérage immédiat des formes recherchées, obtention de concordances et de contextes étendus autour de certaines familles de mots, qu'ils soient envisagés sous l'angle des signifiants, des signifiés ou des référents. Mais si l'on accorde valeur de témoignage aux fréquences observées dans un texte, alors le recours à une méthodologie statistique s'impose pour les analyser.

18 Le cadre de la méthode est contrastif. En effet, avant d'avoir comparé à d'autres textes émis dans des situations de discours analogues, rien ne permet d'émettre un jugement sur une fréquence, car il n'existe nulle part une norme fréquentielle universelle à laquelle nous pourrions nous référer. Nous pensons que la fréquence n'est pas une caractéristique de langue, et qu'en conséquence un modèle fréquentiel de la langue reste une chimère. C'est la raison pour laquelle les corpus étudiés donnent lieu à des partitions en fragments adaptées aux objectifs de la recherche. On trouve des situations canoniques de ce type dans les discours présidentiels étudiés ci-après respectivement par Elsa Carrillo-Blouin et par Damon Mayaffre et Xuan Luong.

19 Dans le cas où les constats portent sur des phénomènes isolés, il s'agit de mesurer les écarts par rapport à l'hypothèse d'équi-répartition du phénomène au sein du corpus. C'est la totalité du corpus qui tient lieu de norme, celle-ci est donc endogène. Ainsi tous

les écarts sont mesurés par rapport à l'ensemble du corpus. La mesure de ces écarts donne, en général, lieu à des formulations probabilistes. On remarquera cependant que les probabilités ne sont jamais utilisées dans leur contexte habituel, soit pour permettre de faire des conjectures sur des événements incertains. Bien au contraire, les événements que nous examinons sont complètement déterminés. Les probabilités ne sont ici que des indicateurs permettant de porter un jugement sur les fréquences observées, et de sélectionner celles qui présentent les écarts positifs (sur-emploi) ou négatifs (sous-emploi) les plus importants. Ainsi la méthode dite des « spécificités » a une fonction heuristique ; elle permet de faire apparaître les événements fréquentiels les plus saillants dans les fragments du corpus.

20 Prendre en compte plusieurs phénomènes à la fois conduit à constituer un tableau croisé qui comporte autant de lignes que de phénomènes et autant de colonnes que de fragments dans le corpus. Dans les descriptions textuelles, le plus souvent une ligne représente la distribution d'un mot dans les fragments du corpus, tandis qu'une colonne représente un fragment du corpus. Cette disposition des données en tableau rectangulaire est extrêmement répandue dans toutes les enquêtes quantitatives en sciences humaines et sociales. Les tableaux issus des applications textuelles ont deux particularités notables. D'une part, ils ont en général un très grand nombre de lignes (autant que de mots différents dans le corpus étudié), mais des tableaux partiels peuvent être extraits du tableau global. D'autre part, la distribution des fréquences dans les colonnes du tableau est contrainte par la loi de Zipf[6]. À la suite des travaux de Jean-Paul Benzécri[7], la méthode la plus répandue à pour traiter des tableaux de ce type est l'analyse factorielle des correspondances. Nous en trouvons plusieurs exemples dans les articles de ce numéro. L'intérêt de cette technique est de produire des représentations graphiques sur les plans factoriels dans lesquels la proximité d'un mot avec un fragment témoigne de l'usage préférentiel de ce mot dans le fragment. Mais d'autres représentations d'une grande lisibilité peuvent être construites à partir des mêmes tableaux. Tel est le cas de l'analyse arborée présentée ici dans la contribution de Damon Mayaffre et Xuan Luong.

21 Nous venons de voir que les modèles statistiques qui sont mis en œuvre n'introduisent pas d'hypothèses supplémentaires relatives à la nature langagière de leurs objets. Ce point confère de la puissance à la méthode qui se trouve être applicable à tous les textes et fournit toujours des mesures contrastives donc des résultats. Mais ces dernières ne peuvent donner lieu à une interprétation et conduire à des conclusions que si le corpus dont ils sont issus a été construit autour d'une problématique explicite. En général, on estime la qualité d'un corpus à sa représentativité. En histoire, la représentativité fait référence parfois à une documentation plus vaste dont le corpus est un échantillon, mais en même temps à une problématique historique à étudier. Nous avons tenté de montrer l'incidence cruciale que la forme du corpus et son équilibre peuvent avoir sur la qualité des observations et des constats numériques. Placées sous la responsabilité de l'historien, l'élaboration du corpus et sa justification, pourtant situées en amont des traitements statistiques, conditionnent en fait leur bon fonctionnement.

22 Certains pourront objecter que la validité du résultat est étroitement dépendante de l'organisation du corpus, ce qui interdit tout cumul de résultats qui, parce qu'ils sont obtenus à partir de normes locales (puisque liées au corpus) sont incommensurables. L'objection est admissible, mais sa portée est réduite, à partir du moment où l'on admet que l'historien travaille dans une perspective heuristique, dans le cadre d'une problématique historique. En revanche, les historiens doivent observer avec attention

l'évolution des méthodes en linguistique, dans la mesure où les linguistes continuent à en développer de nouvelles. La plupart de celles qui sont présentées ici (analyse factorielle, spécificités lexicales) sont expérimentées depuis une trentaine d'années, mais nous avons surtout voulu suggérer la variété des applications possibles à des textes dont on n'imaginait pas *a priori* qu'ils puissent faire l'objet de telles approches (les cartulaires, les poèmes allitératifs, une autobiographie médiévale, les discours politiques étant par contre plus habituels).

23 Mais l'analyse arborée est un apport plus récent et les possibilités de travail sur de très vastes corpus (comme celui utilisé par Mayaffre et Xuang) apportent de nouvelles perspectives, même si les historiens ne s'y sont pas encore beaucoup aventurés. L'existence de FRANTEXT et de logiciels puissants et d'un emploi relativement aisé est à cet égard une incitation à s'engager dans cette voie. Ce numéro est à la fois un point sur l'état de l'art et une incitation à aller de l'avant.

BIBLIOGRAPHIE

BENZÉCRI, Jean-Paul & *alii*, *Pratique de l'analyse des données*, III, Linguistique et lexicologie, Paris, Dunod, 1981.

BLANCHARD, J. & QUEREUIL, M., *Lexique de Christine de Pisan*, Paris, Klincksieck, 1999.

BOURDIEU, Pierre, *Ce que parler veut dire. L'économie des échanges linguistiques*, Paris, Fayard, 1982 ; réédition augmentée *Langage et pouvoir symbolique*, Paris, Seuil, 2001.

FOUCAULT, Michel, *L'archéologie du savoir*, Paris, Gallimard, 1969.

JACQUART, D. & THOMASSET, Claude, *Lexique de la langue scientifique*, Paris, 1997.

LALANDE, D., *Lexique des Chroniqueurs français (XIV^e siècle, début du XV^e siècle)*, Paris, Klincksieck, 1995.

ROBIN, Régine, *Histoire et linguistique*, Paris, Larousse, 1973.

NOTES

1. BOURDIEU, P., 1982.

2. FOUCAULT, M., 1969, pp. 44-54.

3. ROBIN, R., 1973. Le texte de Michel Foucault auquel je viens de faire allusion est cité par Régine Robin, p. 5 et se trouve dans M. FOUCAULT, 1969, pp. 23-24.

4. D'où l'intérêt pour les historiens, et notamment pour les médiévistes, des « lexiques » spécifiques, par exemple ceux d'un auteur, d'une classe de textes, ou d'un domaine de pensée, comme ceux qui sont publiés dans le cadre de l'élaboration du *Dictionnaire du Moyen Âge Français*, sous la direction de Robert Martin. *Cf.*, par exemple, J. BLANCHARD & M. QUEREUIL, 1999 ; D. LALANDE, 1995 ; D. JACQUART & C. THOMASSET, 1997.

5. *Cf.* Emmanuel BONIN & Alain DALLO, *infra* ; pour Weblex, voir le site http://weblex.ens-lsh.fr/doc/weblex.pdf

6. La loi de Zipf modélise la distribution des fréquences pour tous les textes dans toutes les langues : en tête de la distribution on trouve un petit nombre de formes extrêmement fréquentes, puis l'effectif des classes de fréquence augmente à mesure que les fréquences diminuent, jusqu'aux hapax qui sont toujours les plus nombreux.
7. BENZÉCRI, J.-P. & *alii*, 1981.

INDEX

Mots-clés : analyse textuelle

AUTEURS

JEAN-PHILIPPE GENET

Lamop/Cnrs-Paris I

PIERRE LAFON

Umr 5191, Icar, Cnrs-Lsh