paru dans : Mellet S. et Vuillaume M.(textes rassemblés par), *Mots chiffrés et déchiffrés. Mélanges offerts à Etienne Brunet*,
Paris, Champion, Genève, Slatkine, 1998, p.41-65.

L'ANALYSE RHETORIQUE DE DONNEES TEXTUELLES : UNE COMPARAISON ENTRE TEXTES SCIENTIFIQUES DE DEUX DISCIPLINES

Philippe Cibois
Université de Picardie Jules Verne

Dans cet article on traitera de données textuelles mais on ne fera pas d'*analyse des données textuelles*, c'est à dire qu'on n'utilisera pas cet ensemble de techniques à base de travail statistique sur le *mot*. On utilisera au contraire une technique d'analyse de contenu se situant à un niveau très éloigné de la surface de la phrase puisqu'il s'agit de mettre à jour la rhétorique d'articles scientifiques. Avant d'opérer la comparaison entre des textes issus de deux disciplines, l'un de biologie moléculaire appliquée à l'évolution des oiseaux, l'autre d'analyse des données textuelles avec une étude d'Etienne Brunet consacré à l'histoire du mot "*latin*" dans la littérature française, nous allons tenter de justifier notre refus en montrant que l'aspect polémique de l'analyse des données textuelles ne peut être qu'une étape dans une recherche.

La situation polémique de l'analyse des données textuelles

On peut, face à un article scientifique avoir une attitude normative ou une attitude descriptive. Dans le premier cas, on peut vouloir vérifier que l'auteur ne va pas plus loin que ses données ne l'y autorisent et pour cela on peut formaliser son raisonnement : c'est la perspective de l'analyse logiciste (Gardin 1991). L'attitude est là normative ("non déguisée" dit d'ailleurs Gardin p.67) : il s'agit d'aider les archéologues à sortir leur raisonnement de l'à-peu-près où une tradition humaniste voulait les enfermer et à pratiquer une science, de l'objet historique certes, mais qui utilise le même type de raisonnement que d'autres sciences historiques reconnues comme certaines sciences de la nature, sans qu'il soit besoin de marquer la spécificité humaine de la discipline.

Traditionnellement, dans le cadre de l'analyse des données textuelles, la visée sera descriptive : on ne cherchera pas à traiter de la validité du raisonnement mais on tentera, en prenant un point de vue pertinent, de faire une description utile du texte envisagé. Mais que faut-il entendre par point de vue pertinent et quelles est la nature de l'utilité de la description ? Pour répondre à ces questions nous utiliserons l'expérience de l'analyse des données appliquée aux données sociologiques où un certain nombre de cas étudiés nous permettent d'ébaucher des réponses.

Un point de vue pertinent

La perspective de l'*Analyse des données*, qui dans sa version française se voulait polémique par rapport aux disciplines des sciences humaines (Cibois 1981), rejoint une perspective plus générale (Rouanet 1976) qui consiste à oublier pour un temps les problématiques qui ont permis le recueil des données et à les examiner pour elles-mêmes, indépendamment des hypothèses de recueil, pour "les laisser s'exprimer, dire ce qu'elles ont à nous dire". Par ces expressions issues du rapport interpersonnel, on veut manifester non pas une hypothétique intentionnalité des données, mais un dédoublement du chercheur qui d'un côté sait bien que ses données ne sont que l'opérationalisation d'hypothèses de recherche mais qui, d'un autre point de vue, se met dans la disposition d'esprit nécessaire pour réorganiser le réel (vu paradoxalement à travers ses données). Le chercheur peut vouloir découper différemment le

réel s'il en était besoin, s'il en éprouvait lui-même le besoin, s'il estimait que ses hypothèses de constitution des données en rendent mal compte. Le paradoxe de l'analyse des données où une construction (des données) remet en cause les principes qui ont été à l'origine de cette construction n'en est pas un. C'est la situation normale de toute recherche empirique où une observation conduit souvent à une réévaluation : la situation est cependant différente du cas de l'observation banale, où ce sont les catégories du sens commun qui servent de théorie constitutive des données ; ou du cas de l'expérimentation où l'on met plutôt le chercheur en position de vérifier la cohérence de ses hypothèses de constitution des données. Conclure d'une analyse des données que l'observation conduit à de nouvelles hypothèses, c'est dans un raccourci de recherche dire que l'observation des données a réfuté l'hypothèse qui avait servi à les construire et propose une nouvelle hypothèse plus cohérente avec le monde observé.

La pertinence se définit dans cette logique et il n'y a de pertinence que relative. On juge pertinent ce qui correspond à ses hypothèses de recherche ou, dans la lignée polémique de l'analyse des données, on juge pertinent ce que l'on tient comme plus tenable que ce à quoi tiennent les adversaires. Par exemple en analyse des données textuelles, on veut lutter contre la dérive herméneutique qui peut facilement s'emparer de n'importe quel texte pour lui faire dire n'importe quoi. On le fait en déclarant pertinent ce qui ne l'est pas au niveau d'analyse de la phrase, c'est à dire le mot. La statistique du mot, son comptage, l'examen de sa spécificité dans une portion de texte permet au chercheur de casser le sens obvie et de faire le cas échéant une démarche utile.

L'utilité de la description

Là encore l'utilité reconnue à la description est souvent polémique : quand on estime que ceux qui interprètent les textes le font d'une manière débridée, tout résultat même pauvre parait estimable car assuré. Pour l'enrichir, il faut souvent user d'une perfusion de théorie, c'est à dire injecter dans l'interprétation du résultat des hypothèses suffisamment riches pour que leur validation par le résultat forme un tout intéressant. On dira que l'on a fait que repousser la liberté interprétative et ses dangers : nous dirons plutôt qu'il était naïf de croire que l'on pouvait se passer d'interprétation et que l'on doit lutter contre la dérive herméneutique sur son terrain propre, en critiquant la dérive, non l'interprétation. Cependant il est souvent plus facile d'abandonner le terrain à l'adversaire et de créer un nouveau champ de recherche de toutes pièces en se forgeant un objet nouveau.

L'utilité obtenue par une analyse des données textuelles peut se voir aussi dans ses aspects pratiques : pour le sociologue qui est confronté à l'analyse d'une enquête où se trouvent de nombreuses questions ouvertes, il peut être intéressant de trouver un outil qui lui permette de regrouper d'une manière automatique des réponses qu'il validera ensuite comme semblables. Pour des chercheurs en littérature, des index de fréquences, des descriptions de la spécificité d'un mot à travers les époques sont des instruments de référence fabriqués sans a priori d'utilisation, de la même manière que l'on produit des tables astronomiques qui seront tout aussi bien utilisées pour la conquête de l'espace que pour rédiger des horoscopes.

Comment enrichir l'analyse

Il est possible à mon avis d'enrichir le résultat de l'analyse si l'on accepte de charger de théorie le trait pertinent : si l'on abandonne l'aspect polémique de l'analyse des données pour ne considérer que l'ensemble des techniques qui s'y trouvent, on jugera normal que tout enrichissement des données se traduise par une pertinence plus forte des résultats. Par

exemple le sociologue sait très bien fermer une question ouverte, c'est à dire déterminer par essai/erreur la typologie qui rendra le mieux compte des réponses textuelles rencontrées. Prétendre qu'un algorithme aura un résultat plus pertinent relève de la polémique, qui peut parfois se justifier mais qui ne saurait s'imposer en tout temps et en tout lieu.

En *analyse de contenu*, autre démarche tournée souvent vers les mêmes objets textuels mais sans a priori théorique, on pratique l'analyse thématique, c'est à dire que l'on s'autorise à reconnaitre dans des énoncés différents un thème identique. Les difficultés sont nombreuses mais la plus rude, pour un pratiquant de l'analyse des données textuelles est de franchir le pas car les bornes franchies, il n'y a plus de limites. C'est en tout cas l'impression du puriste qui se refuse souvent à franchir la borne du mot, et qui admet tout juste le lemme, ou le segment répété (Salem 1995).

Franchir la limite du mot en analyse textuelle, ce n'est pas se lancer dans l'univers sans loi de la surinterprétation, c'est entrer dans la cohérence d'un champ disciplinaire particulier que l'on a la possibilité d'assumer en le croyant perfectible. C'est refuser ainsi de faire table rase et assumer une certaine cumulativité du savoir. L'iconoclastie est certainement à certaines époques une attitude indispensable : il faut savoir dire non, refuser de prendre en compte des faux savoirs qui ne se tiennent que parce qu'ils ont colloques revues et chaires. Ce peut être une étape indispensable, ce ne peut être une attitude permanente, il faut accepter non pas le culte des images, mais l'attitude prudentielle qui le sous-tend, c'est à dire l'acceptation d'une incarnation. On ne peut jouer sans fin de la théologie négative : il faut un jour au l'autre accepter de parler de son sujet, reprendre chez les anciens ce qui s'est dit d'intéressant, assumer un discours théorique sur son objet.

Comment analyser un texte scientifique

Dans l'analyse textuelle des textes scientifiques qui va suivre on ne va pas chercher à examiner la validité de la démonstration mais d'une manière plus large sa rhétorique, c'est à dire tout ce qui peut concourir à la persuasion du lecteur. Cette perspective est davantage celle de la sociologie de la science dans la mesure où dans l'optique du paradigme kuhnien, le consensus des agents est vu comme fondateur de la validité scientifique. Nous ne discuterons pas ici de la controverse (Boudon 1994) portant sur l'aspect fondateur du consensus scientifique : nous noterons simplement que dans la perspective d'une discipline, la sociologie, dont l'objet propre est l'influence des groupes sociaux, il soit normal de mettre en avant cette étude quand on étudie la science. Ce qui est certainement contestable est de réduire la démarche scientifique au seul consensus : comme le dit Latour (1991) humains et non-humains doivent coopérer.

Les traits pertinents de l'analyse seront donc les effets de persuasion que l'on trouvera dans deux types d'endroits : les lieux explicites de persuasion, ceux qui explicitement permettent à l'auteur de tirer les conclusions de prémisses, et les lieux implicites, ceux qui renforcent l'autorité de la chose dite. Pour la recherche de premiers, il suffit de lire l'auteur qui nous les indique expressément ; pour les seconds, on étudiera tout ce qui a rapport à l'environnement social de l'auteur, sa position, l'institution où il travaille, celle qui a financé ses recherches, les auteurs avec lesquels il a collaboré explicitement ou non, ceux qu'il cite dans son texte ou sa bibliographie. Enfin on ne s'interdira pas, le cas échéant, de repérer des renforcements persuasifs dans des phrases qui ne rentrent dans aucune des catégories précédentes.

Nous étudierons deux textes : le premier nous est donné par la circonstance de parution du présent texte puisque nous voulons rendre hommage à un auteur. Nous prendrons donc un de

ses textes récents (Brunet 1996), assez court et sur un sujet qui nous intéresse puisqu'il fait partie de nos préoccupations de recherche (Cibois 1996). Il s'agit d'une étude qui porte sur l'évolution dans les siècles récents de la présence du latin dans la littérature française en prenant comme référence la base de données Frantext qui possède plusieurs milliers de textes et qui regroupe 160 millions de mots.

Face à ce texte qui cherche à rendre compte d'une évolution historique avec des outils informatiques et statistiques, nous avons voulu, en suivant une remarque de J.-Cl. Gardin (1993 : 154), prendre un texte scientifique voulant lui aussi rendre compte d'une évolution historique mais dans un domaine tout différent et appartenant à une discipline dont la scientificité n'est contestée par personne, la biologie. A cette fin nous avons utilisé un article utilisant la biologie moléculaire pour rendre compte d'un phénomène d'évolution chez les passereaux[1] (Edwards 1991).

Biologie moléculaire et évolution

La naissance de la biologie moléculaire, marquée par la découverte de l'ADN par Watson et Crick en 1953 est encore considérée comme un choc scientifique qui aura "ébranlé l'économie méthodologique et ontologique de la biologie, et l'onde s'en fait encore sentir" (Duchesneau 1997 : XII). L'article que nous étudions est souvent cité car c'est l'un des premiers qui a permis de qualifier une méthode se basant explicitement sur la biologie moléculaire pour aider à voir plus clair sur la systématique d'un ordre (entendu au sens d'unité de classification), et en le classant, à mieux voir l'histoire de l'évolution des diverses familles qui le constituent. Auparavant, les classifications de la systématique étaient basées sur des traits morphologiques et le but des nouvelles méthodes est de comparer la classification menée à partir de données de la biologie moléculaire à une classification menée en utilisant des traits morphologiques. Comme les deux classifications sont équivalentes, on pourra désormais utiliser la biologie moléculaire, plus vaste d'emploi que les comparaisons morphologiques qui ne sont pas toujours possibles. Il faut ajouter qu'aux yeux des chercheurs, l'utilisation de la biologie moléculaire, du fait de sa technicité, semble plus indépendante du chercheur qui doit, dans le recherche des comparaisons morphologiques, utiliser plus d'intuition et de flair, et dont les conclusions sembleront souvent moins objectives.

Ce texte s'intitule "Elucidation mitochondriale d'une branche profonde dans l'arbre généalogique des oiseaux percheurs"[2], il se divise, très classiquement dans ce genre d'articles en 4 parties : introduction, matériaux et méthodes, résultats, discussion.

1) introduction : le but de l'article est d'affiner les méthodes de reconstruction de l'arbre généalogique à différents niveaux de spécificité (ordre, famille, sous-famille, espèce)[3] en utilisant le séquençage de l'ADN. En particulier, les auteurs veulent critiquer une méthode différente utilisant l'hybridation de l'ADN selon la température, moins précise à leurs yeux que le séquençage. Ils veulent aussi améliorer la technique en précisant les morceaux d'ADN qu'il faut choisir pour amplifier le fragment qui servira à l'analyse[4], tous n'étant pas propice à être le support de variations dues à des mutations.

---

[1] Je remercie Alice Cibois pour ses explications et ses conseils dans ce champ.

[2] "Mitochondrial resolution of a deep branch in the genalogical tree for perching birds"

[3] Ce qu'on appelle une *phylogénie*

[4] Ce qu'on appelle les *amorces*

2) matériaux et méthodes : 14 espèces d'oiseaux percheurs sont choisies. Certaines appartiennent à la famille des oiseaux disposant d'un organe vocal complexe, d'autres non. Enfin à titre de comparaison, un pic doré des Andes qui n'appartient pas au même ordre est ajouté. On utilise des échantillons sanguins de ces 14 espèces dont on séquence une portion d'ADN. On dispose ainsi pour chaque espèce d'une séquence d'ADN de 924 positions. Dans environ la moitié des positions, il y a identité entre les 14 espèces, par contre il y a des variations sur les autres. L'analyse statistique consiste à faire une classification sous forme d'arbre entre les 14 espèces au vu des identités et des variations selon les positions. On teste statistiquement la validité des résultats.

3) résultats : ils concernent à la fois la séquence d'ADN et les classifications obtenues. Sur le premier point, plus méthodologique, il est montré à la fois la difficulté et cependant l'intérêt de la dernière des trois positions d'un bloc de séquençage[5]. En effet, cette dernière position, qui évolue plus vite que les autres, peut apporter du bruit dans les premiers niveaux de classements qui correspondent aux séparations de familles les plus anciennes au cours de l'évolution, mais peut apporter des informations pertinentes quand il s'agit de faire des différenciations sur des subdivisions plus récentes.

En ce qui concerne les comparaisons entre la classification obtenue et les classifications antérieures, on arrive à la conclusion que les techniques antérieures d'hybridation d'ADN sont compatibles avec les résultats antérieurement trouvés mais que la méthode employée offre plus de possibilités.

4) Discussion : elle consiste à tirer la conclusion que les méthodes utilisant l'ADN sont cohérentes quant à leurs résultats avec les classifications obtenues sur des bases morphologiques ou comportementales. D'autre part elles permettent de se faire une idée du plus ou moins long temps qui sépare les divisions à l'intérieur d'une famille.

Rhétorique implicite

Auteurs

L'article comporte 3 auteurs : Scott V. Edwards, Peter Arctander, Allan C. Wilson. Si l'on prend pour base les 44 références des articles cités qui regroupent 107 auteurs, l'article se trouve près de la moyenne de 2,4 auteurs par article. L'ordre des auteurs n'est pas alphabétique comme dans les 2/3 des articles à plusieurs auteurs.

Le premier auteur, Edwards est indiqué comme ayant pour rattachement institutionnel l'université de Berkeley à travers 2 laboratoires : un service de biologie moléculaire et un laboratoire de zoologie des vertébrés. Il est coauteur des deux articles les plus cités dans le corps de l'article, 7 fois chacun. Il l'est une fois en tête avec Wilson (donc presque dans la même configuration que le présent article) avec un article sur la philogénie des oiseaux chanteurs d'Australie en utilisant l'ADN, l'autre fois il est encore cité avec Wilson mais dans un groupe de 7 auteurs (non cités en ordre alphabétique) mais où il n'est qu'en 4e position.

Arctander n'ayant qu'un seul prénom et étant rattaché pour une part à l'Institut de biologie des populations de l'université de Copenhague semble danois mais est rattaché également au service de biologie moléculaire de Berkeley. Dans la bibliographie, il est cité 3 fois pour des articles utilisant l'ADN dont 2 fois seul pour des articles sur la phylogénie des oiseaux en général et une fois, collectivement sur celle d'une Pie-grièche de Somalie. Il est comme

[5] unité de traduction pour la séquence d'acides aminés appelé un *codon*

Wilson en même temps spécialiste d'un domaine de l'ornithologie et en même temps de la méthode mais par rapport à Berkeley, il est plus périphérique.

Wilson n'appartient qu'au service de biologie moléculaire de Berkeley : il est cité 9 fois dans la bibliographie, toujours dans des articles à plusieurs auteurs et jamais en première position. Il travaille toujours sur du séquençage d'ADN mais sur des sujets variés : sur une Drosophile d'Hawaii, sur des oiseaux Australiens (avec Edwards), sur des mammifères, sur des poissons, sur les techniques de séquençage, sur les statistiques associées. Wilson joue dans ces articles un rôle de technicien de la méthode par rapport aux spécialistes du domaine.

Ce qui ressort de cette présentation est la prééminence de l'Université de Berkeley qui joue un rôle central dans la composition de cette article ; ainsi que le couplage d'une compétence sur un domaine (ici l'ornithologie) et de la maitrise d'une technique (le séquençage de l'ADN), les deux étant reliés par une technique statistique. Edwards, étant relié à Berkeley même aux deux domaines se trouve de ce fait dans une position dominante par rapport à ses coauteurs : on comprend qu'il soit l'auteur de tête. Cette prééminence de Berkeley, université prestigieuse, qui réussit le couplage du domaine et de la méthode peut apporter à l'article une plus-value de réputation qui bien qu'implicite n'en est pas moins efficace.

Références bibliographiques

Les références bibliographiques données en fin d'article et qui sont au nombre de 44 ne sont que les références utilisées : toute référence donnée dans le texte se trouve en fin d'article et toute référence de fin d'article est utilisée dans le corps du texte (ou dans les figures). Il n'y a aucune place pour une bibliographie du sujet qui servirait de cadre de référence : ne sont présents que des articles réellement utilisés.

Ces articles 3 fois sur 4 ne sont cités qu'une fois mais pour le quart restant les citations peuvent aller jusqu'à 7 fois pour les articles d'Edwards déjà cités ou 5 fois pour l'article qui utilise l'hybridation de l'ADN, méthode que l'on veut améliorer. On trouve dans le texte 71 références ce qui pour un texte de 102 phrases fait 0,7 référence par phrase. Sur 10 phrases de texte, 7 d'entre elles possèdent une référence ce qui donne au lecteur le sentiment que pratiquement à chaque instant, les auteurs s'appuient sur des références.

Cette fréquence n'est pas constante : elle est plus forte dans l'introduction qui cadre la problématique par rapport aux autres auteurs (1,3 ref. par phrase), reste encore forte dans la partie matériaux et méthodes (1,0) ; elle est la plus faible dans la partie résultats (0,3) où l'accent est mis sur l'apport propre des auteurs et elle remonte dans la discussion (0,9) où à nouveau on confronte les résultats avec d'autres auteurs.

Ce qui ressort du point de vue de la rhétorique implicite est l'impression d'*encadrement* qui est donnée : les auteurs ne sont pas isolés mais travaillent en s'appuyant ou en contredisant une communauté scientifique Cette référence à d'autres membres de la communauté scientifique, référence faite plus souvent que dans une phrase sur deux, si elle hache la lecture, manifeste la liaison forte des auteurs avec leurs pairs.

Revues

L'article est paru dans les comptes-rendus de la Royal Society of London qui (pour la série B des *Biological sciences*) en est à son volume 243 ce qui indique une origine dans le temps assez ancienne. En effet (Shapin 1993) la Royal Society of London a été fondée en 1660 à Gresham College autour du programme expérimental dont les expériences de Boyle sur la pompe à air constituent le prototype. La revue est aujourd'hui spécialisée en biologie : son prestige dans le milieu est largement inférieur à celui des trois grandes revues généralistes

Cell, *Nature* et *Science*. Enfin, l'article a été reçu le 4 septembre 1990, accepté le 15 octobre et publié le 22 février 1991. Ces cours délais font partie de la stratégie de la revue qui déclare publier dans les trois mois tout article accepté, ce qui la conduit à doubler aujourd'hui son rythme de parution (24 numéros par an à partir de 1998)[6].

Si l'on examine les articles cités en référence, aucun n'a été publié par la Royal Society. La revue la plus présente est le *Journal of molec. Evol.* (9 articles) : viennent ensuite les comptes-rendus de l'Académie nationale des Sciences USA (4 articles) puis *Mol. Biol. Evol.* (2 articles). Les autres articles viennent de revues citées une fois ou de chapitres de livres. En regroupant les diverses sources on arrive aux chiffres suivants :

Biologie moléculaire	15
Revues généralistes	10
Evolution et systématique	9
Ornithologie	7
Logiciels	3
Total	44

Le domaine est clairement ciblé par importance décroissante : l'article porte 1) sur la méthode 2) en vue de résoudre des problèmes d'évolution 3) dans le cas des oiseaux. L'information implicite de cette distribution est bien la nature du champ où s'effectue la recherche. La méthode est récente, les articles aussi, les deux tiers ont moins de 5 ans. Il nous reste maintenant à voir comment cette recherche s'articule dans sa rhétorique explicite.

Rhétorique explicite

Cependant, si le sociologue se sent à l'aise pour se faufiler dans le détail des personnes, des institutions et des revues, pour y rendre explicite la rhétorique implicite (mais qui est parfaitement connue des protagonistes du milieu[7]), il peut légitimement se demander quel doit être son rôle quand il examine la démarche explicite de justification. Doit-il juger de la validité des conclusions à l'aune des prémisses ? C'est se situer en juge, ou au moins en arbitre alors que nous ne voulons être que descriptif. Nous allons essayer plutôt d'observer la chaine argumentative dans son coeur, c'est à dire au moment où l'on tire les conclusions de l'observation.

Ici le "réel" c'est le séquençage de l'ADN de 14 espèces pour 924 positions. On a donc un tableau de deux pages avec 14 lignes et 924 positions découpées en groupes de 3 où l'on trouve l'une des 4 base du système A, T, C, G. Ce séquençage est donné en détail, ce qui prend de la place : dans les articles plus récents, il est simplement indiqué la base de données où l'on peut le trouver.

La méthode employée est décrite en citant les programmes utilisés : son principe est celui de la parcimonie de l'école dite cladiste[8] : celle école construit des classifications des espèces en faisant l'hypothèse qu'il est plus économique (principe de parcimonie) d'attribuer une

[6] La Royal Society, en plus d'une revue d'histoire des sciences publie en biologie et en physique deux types de revues : les *Proceedings* qui accepte des articles et les *Philosophical Transactions* où les articles sont sollicités afin de faire des numéros thématiques.

[7] et qu'ils codifient dans des palmarès de revues scientifiques confectionnés annuellement à partir des *Science citations index*.

[8] du grec *klados*, petite branche rameau.

ressemblance à une origine commune qu'à des évolutions parallèle ayant conduit au même résultat. On choisira donc l'arbre de classification qui maximise l'origine généalogique des ressemblances. Quand une classification est établie, elle est interprétée d'une manière historique : plus les noeuds sont situés vers le sommet de la hiérarchie plus ils sont supposés anciens. Les méthodes de classification ne sont pas considérées comme purement descriptives mais comme fondées sur l'histoire de l'évolution des espèces (Duchesneau 1997). La validité des résultats est testée par une méthode bootstrap[9], par des tests statistiques jugés significatifs ou par des valeurs du log de la vraisemblance plus ou moins forts[10].

Nous ne discuterons pas des apports techniques de la méthode qui concernent les points 2 à 4 des résultats et qui sont consacrés à la chaine de séquençage et à ses aspects : nous nous centrerons sur le point numéro un qui est à propos de l'arbre phylogénique obtenu et de sa comparaison avec l'arbre obtenue avec la méthode précédente (hybridation de l'ADN). Pour cette comparaison on trouve dans l'article à la partie *résultats* 16 phrases de texte, une figure comportant 3 arbres et 14 phrases d'explication. Dans la partie *discussion*, ces résultats sont repris par 5 phrases.

L'articulation du discours est la suivante : plusieurs arbres ont été fournis en faisant varier les méthodes : soit en prenant une méthode de parcimonie, soit une méthode plus descriptive comme le maximum de vraisemblance et à l'intérieur de ces méthodes plusieurs variations viennent du fait de la prise en compte de diverses positions dans le séquençage. Ce qui est commun à tous ces arbres c'est qu'ils divisent rapidement les espèces des oiseaux percheurs en deux groupes : les oiseaux chanteurs et ceux qui ne le sont pas. De plus cette classification rejoint celle de la méthode précédente par hybridation de l'ADN. Cette convergence de méthodes renforce la thèse que l'apparition d'un appareil vocal chez l'oiseau est un évènement unique de l'évolution ayant conduit à la diversification des oiseaux.

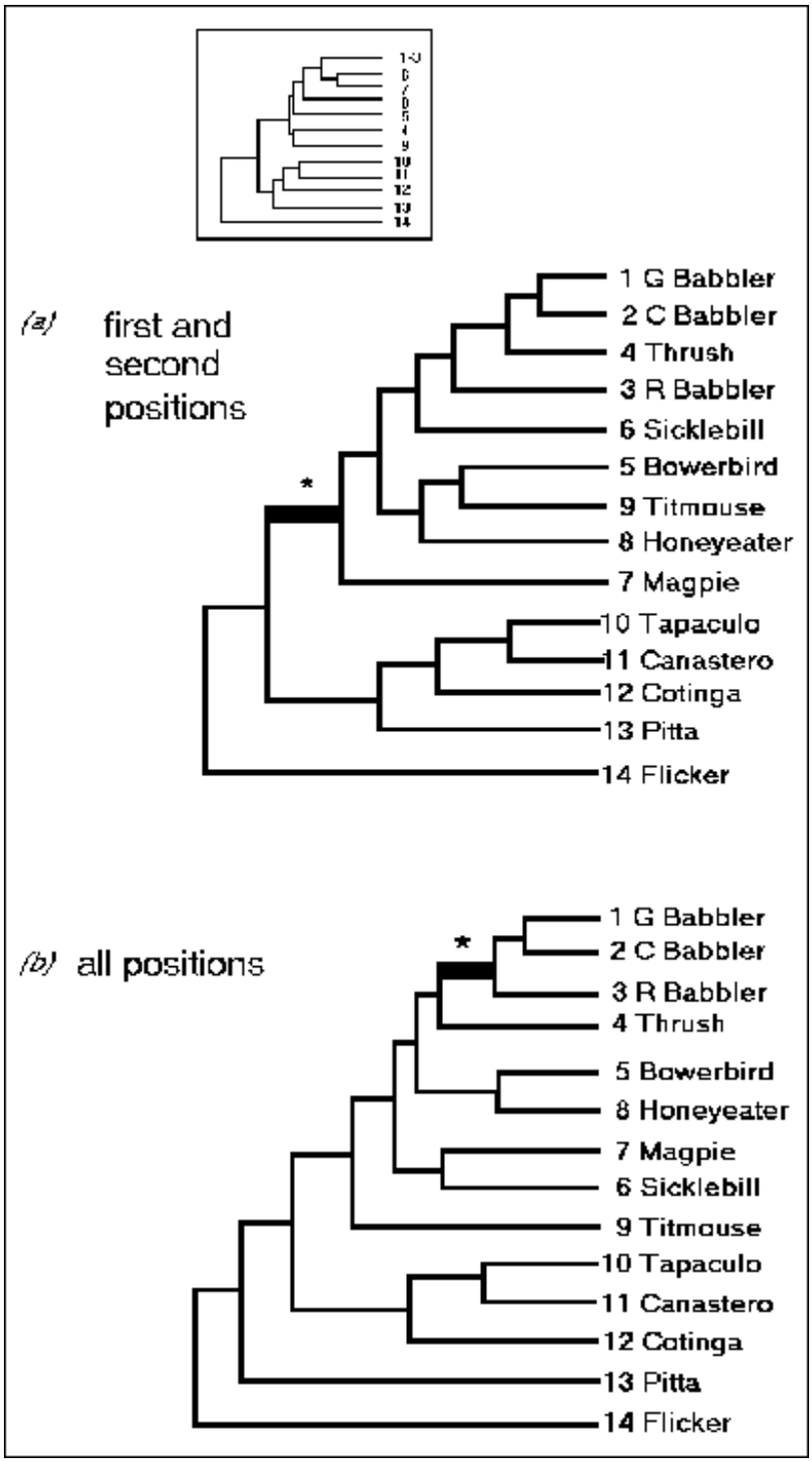
Cependant cet aspect de confirmation de résultat est surtout l'occasion pour les auteurs pour insister sur l'aspect méthodologique : ayant prouvé sur cet exemple la validité de la méthode, ayant amélioré ses conditions d'utilisations par des discussions sur les zones de séquençage à prendre en compte, les auteurs notent que "nos résultats renforcent l'idée que le séquençage de l'ADN sera d'une grande efficacité pour des études futures pour l'histoire de la généalogie des oiseaux". C'est bien là l'aspect de l'article qui a été retenu ultérieurement : il y est fait référence quand on utilise la méthode et afin de justifier cet usage.

Si l'on regarde maintenant avec plus de précision le détail de l'argumentation de similitude entre les arbres, cela revient à comparer visuellement les arbres de classification (cf Edwards figure 4) : les deux arbres *a* et *b* de l'auteur ont un certain nombre de variations mais regroupent également les espèces 1 à 9 et 10 à 13 qui correspondent respectivement aux

[9] Le terme *bootstrap* fait référence aux boucles ou aux lacets de chaussures par lequels Cyrano et quelques autres héros mythiques ont pu s'élever en l'air en tirant très fort dessus. L'idée de la méthode est de n'utiliser que les données dont on dispose mais en les tirant au sort des milliers de fois de façon à constituer des populations suffisantes pour un test statistique. On ne dispose que de ses propres données mais en les utilisant bien par duplication, on peut en tirer des informations pertinentes. On retrouve la même idée d'auto-lévitation dans les séquences boot-strap de démarrage d'un ordinateur où il faut, pour lancer un systême, un programme : qui lancera le programme qui lancera le programme ? C'est la procédure boot-strap.

[10] Dans la suite les classifications basées sur la parcimonie sont comparées aux classifications faite avec le principe du maximum de vraisemblance, méthode itérative qui permet de trouver la distribution la plus compatible possible avec les données

oiseaux chanteurs et à ceux qui ne le sont pas[11]. On constate des ressemblances entre les trois arbres aux niveaux les plus anciens (1-7 contre 10-13) mais aussi au niveau des ressemblances d'espèces (1-3). Entre les deux niveaux, il y un une certaine instabilité. La figure sert ici de preuve pour faire comprendre au lecteur les positions respectives de la stabilité (niveau profond) et de l'instabilité. Comme c'est le niveau profond, le plus ancien qui est utilisé pour faire des généalogies, c'est évidemment cet aspect qui est mis en avant.



[11] l'oiseau 14 est le hors-groupe : il est a priori différent et sert à étalonner les critères de séparation les plus anciens dans l'évolution.

figure 4 de l'article d'Edwards

Qu'est-ce qui est convaincant pour le lecteur ? On ne peut pas dire qu'un test de signification ou un chiffre soit réellement mis en avant : ce qui a valeur de preuve c'est la ressemblance au niveau profond des arbres. La phrase qui la décrit : "comme dans l'arbre associé à l'hybridation de l'ADN (encadré de la figure 4), l'arbre utilisant les séquences mitochondriales indique l'unité des séquences des oiseaux chanteurs par rapport à celles des autres oiseaux percheurs", cette phrase sans la figure serait difficilement compréhensible. La preuve est renvoyée à la visualisation de la structure de l'arbre ou la similitude topologique apparait aisément.

Un deuxième résultat mis en avant repose sur les différences entre nouvelle méthode et ancienne : dans les deux arbres de la nouvelle méthode l'espèce 4 (Thrush, la grive) se trouve toujours associée aux espèces 1 à 3 (Babblers australiens ou timaliidés), ce qui n'est pas le cas dans l'arbre ancien. Cette permanence n'est pas seulement défendue du point de vue statistique mais aussi pour des raisons morphologiques qui semblent emporter la décision.

Le latin dans la littérature française

L'article d'Etienne Brunet[12] dont nous allons comparer la rhétorique[13] avec celle de l'article précédent a pour horizon de questionnement la place du latin dans la culture française. Il s'agit là d'un débat social que nous pouvons repérer aujourd'hui à travers la question de l'enseignement du latin qui en est un bon reflet.

La question du latin[14]

A la Renaissance, le renouveau dans l'étude du latin (et du grec) a signifié trois choses (Garin 1968 : 238-239) :

1) le remplacement d'un latin devenu un jargon par le beau latin de Cicéron et de l'époque classique que L.Valla par ses *Elegantiae linguae latinae* avait fait redécouvrir.

2) la prise en compte du trésor de la science et de l'art de l'antiquité

3) une redécouverte de la valeur morale exemplaire des héros de l'antiquité.

A la fin du 18e siècle, les deux premières raisons ont disparues : les langues nationales européennes ont également leurs chefs-d'œuvre littéraires et leurs propres classiques. Il en est de même pour les arts : si les chefs-d'œuvre de l'antiquité restent une référence, ils s'inscrivent désormais dans un patrimoine qui accueille petit à petit toutes les productions artistiques de toutes les époques. Quant aux sciences, elles se sont développées à un point tel que la science grecque n'a plus d'intérêt en tant que tel, sinon historique. Par contre la valeur morale exemplaire des héros de l'antiquité est toujours reconnue comme devant faire partie de l'enseignement.

---

[12] Etienne Brunet, « Le latin dans la littérature française », in Actes du Colloque *La réception du latin*, Angers, 1996, p.125-141.

[13] Le terme rhétorique est ici entendu dans son acception aristotélitienne d'origine qui, comme l'a montré Paul Ricoeur (1975 : 13), comprend d'abord une théorie de l'argumentation avant une théorie de l'élocution et de la composition. Aristote adapte le syllogisme au cas où la déduction ne se fait plus sur des énoncés vrais ou faux : quand les prémisses sont tirées de vraisemblances ou d'indices, le syllogisme devient alors un *enthymême* (Rhétorique 1357). La rhétorique étudiée sera l'ensemble des lieux de persuasions, que leur statut soit celui de la preuve explicite (rhétorique explicite) ou qu'il soit celui du renforcement par la confiance donnée à celui qui parle (rhétorique implicite).

[14] Le débat est à ce point récurrent que ce titre a déjà été utilisé, à des fins humoristiques, par Guy de Maupassant qui fait allusion au débat de l'époque : la nouvelle est de 1886.

L'enseignement du latin, mis à la portée des enfants des élites par les Collèges (Compère 1985) a trouvé sa forme durable dans la *ration studiorum* des collèges jésuites. Il est frappant d'y trouver un ordre pédagogique d'enseignement des auteurs qui sera encore celui du *Traité des études* de Rollin qui est un bon reflet de la situation au début du 18e siècle : César est le premier auteur à étudier en 4e nous dit-il[15], ce qui est toujours le cas dans l'enseignement français des années 1960.

Au début 19e siècle on n'enseigne encore que le latin et les mathématiques, même si le professeur de latin donne des leçons d'histoire et de géographie "quand le latin voulait bien donner un instant de répit" dit Jules Simon[16], et celui de mathématiques des notions de physique, de chimie et d'histoire naturelle. Dans le courant du 19e les enseignements faits par des professeurs différents se sont multipliés : grec, philosophie, histoire, géographie, physique, chimie, histoire naturelle, langues vivantes. Dans sa circulaire aux proviseurs, Jules Simon montre que le latin doit céder une place aux nouvelles venues :

> Le latin, Monsieur le Proviseur, n'est complètement une langue morte que depuis notre âge. Il a d'abord été la langue d'un peuple, et ensuite celle de toute une classe d'homme savants et lettrés, qui l'employaient pour leurs écrits, pour leur correspondance et pour l'enseignement. (...) Mais le latin est maintenant une langue morte dans toute l'étendue du terme, et les progrès de l'enseignement des langues vivantes achèvent et complètent cette transformation. On étudiera désormais le latin pour le comprendre, et non pour le parler[17].

Pour pouvoir faire une place aux nouveaux enseignants, il faut faire moins de latin : on abandonnera donc le thème latin associé à la rhétorique antique qui va disparaitre avec lui. Désormais on ne parle plus le latin mais on le comprend : la version latine prend son essor en tant que discipline reine de latin : si la part de l'enseignement du latin a diminué, la version reste l'exercice roi. Pour justifier ce choix, on invente le slogan du latin comme *gymnastique de l'esprit*, thème qui fonctionne encore parfaitement aujourd'hui.

Les finalités sociales du latin changent : alors qu'il était la formation de base du notable, il devient moins indispensable, il est remis en cause : l'enseignement *moderne* peut se substituer en partie à lui. En faire faire à ses enfants ne s'impose plus : pourquoi alors de maintient-il ? Pour Edmond Goblot en 1925 (1984 : 84-85) la réponse est que la latin est une barrière sociale qui permet à la bourgeoisie de se distinguer du peuple. Ce thème de la distinction sera repris et élargi par Pierre Bourdieu (1979).

Une hypothèse plus économique que l'hypothèse par la fonction de distinction, et elle ne lui pas contradictoire, est de dire que l'institution scolaire perpétue ses enseignements disciplinaires pour ses propres raisons internes (Chervel 1988 : 70-71). L'extérieur peut agir directement, comme dans le cas de la réforme de l'enseignement secondaire des premières années de la IIIe république, soit d'une manière indirecte par la plus ou moins bonne réponse de l'enseignement à l'attente du monde extérieur.

Pour le latin, où l'attente d'apprentissage est nulle[18], on peut faire l'hypothèse économique que la poursuite du latin à d'abord un origine interne à l'école : c'est sous la pression des enseignants que sont peuplées les classes de langues anciennes. Le reste pourrait bien n'être qu'habillage variant avec les époques et selon les besoins : gymnastique de l'esprit, aide pour

---

[15] Charles Rollin, *De la manière d'enseigner et d'étudier les Belles-Lettres par rapport à l'esprit et au coeur*, dit *Traité des études*, Edition Savy à Lyon, 1808, T.1, p.164-165.

[16] Jules Simon, *La réforme de l'enseignement secondaire*, 1874, p.411.

[17] *ibid*., p. 412-413

[18] Tous les détracteurs du latin soulignent qu'on peut sans crainte oublier son latin quand on a fini ses études, cf par ex. Goblot à l'endroit cité.

l'apprentissage du français, découverte des chefs-d'œuvre de l'antiquité, éléments de culture religieuse pour ce qui concerne les arguments positifs ; perte de temps, distinction sociale, passéisme, élitisme, pour les arguments négatifs. L'hypothèse reste à vérifier.

Ce qui ne fait de doute pour personne, soit pour s'en réjouir, soit plus souvent pour s'en désoler[19], c'est que le latin perd de son importance sociale dans la formation scolaire, dans l'éducation, dans la culture et dans les lettres : c'est à propos de ce dernier point qu'Etienne Brunet envisage d'appliquer les techniques de l'informatique appliquées à un corpus de texte. Son article commence ainsi :

> On peut avoir le sentiment que le latin perd du terrain dans la culture française et qu'au fil des siècles la langue vernaculaire ayant progressivement rejeté la langue des savants et des lettrés, le latin a fini par ne plus trouver un refuge assuré dans les écoles, dans les prétoires, et dans les églises. Mais se maintient-il dans les textes et la littérature ? Ou bien observe-t-on là aussi un déclin inéluctable ?

Etienne Brunet pour éclairer la question va utiliser le corpus de la base Frantext et va de différentes manières interroger ce corpus. Examinons maintenant la rhétorique de son article.

Rhétorique implicite

Elle semble se réduire à peu de choses : un seul auteur (EB) avec une double référence : à l'Institut national de la langue française et à la ville de Nice. L'auteur aurait pu préciser le rattachement au CNRS et à l'Université : il ne l'a pas fait et a préféré une posture minimaliste. En ce qui concerne les références, celles dans le texte qui font référence à des œuvres littéraires où l'on cherche des traces latines sont nombreuses, celles qui font référence à d'autres entreprises analogues sont absentes. L'auteur travaille seul et avec ses propres forces propose seul des résultats.

Seul ? Non, un autre acteur est présent, valorisé, exalté : l'ordinateur. C'est lui qui a droit à la seule référence bibliographique par le biais du rappel de son nom de baptême en langue française. C'est lui qui d'une manière humoristique certes, mais appuyée, est pris comme arbitre permettant de répondre à la question de base. Il ne vient pas seul mais associé à la base de données Frantext, "qui est grosse de plusieurs milliers de titres et rend compte de 160 millions de mots" (p.125) qui est citée 8 fois dans le texte lui-même et qui fait l'objet de la chute de la conclusion du texte. On plaisante beaucoup de l'objet informatique dans ce texte car un bête machine ne peut jouir d'un grand prestige mais on en vante les mérites car la base de donnée utilisée, issue des travaux du Trésor de la Langue Française (TLF) a permis à l'auteur la plupart de ses travaux.

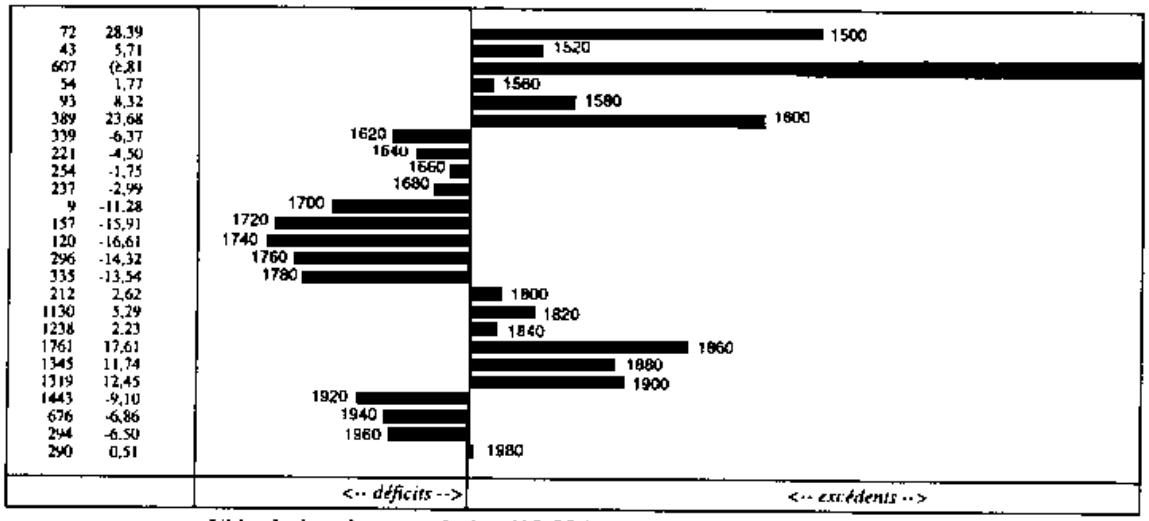
Rhétorique explicite

La méthode employée tout au long du texte est la suivante : on lance une interrogation de la base (ou d'une de ses parties), c'est à dire que l'on fait une requête sur un mot et/ou un type de mots. On compte les occurrences obtenues par tranche chronologique et on visualise le résultat une fois pondéré par la taille relative de la période donnée[20]. On commente le résultat et on l'affine en modifiant la requête. On précise les résultats par des listes d'importance décroissante de termes ou d'auteurs.

[19] Jacqueline de Romilly, *Lettre aux parents sur les choix scolaires*, Paris, Ed. de Fallois, 1994

[20] technique de l'écart réduit qui centre et réduit la valeur observée.

Graphique 19. L'évolution des mots latins



L'évolution des mots latins (12 934 occurr.). Courbe de l'évolution.
(25 périodes prises en compte. Taille du corpus : 117 326 493 occurrences)
Coefficient de corrélation chronologique : – 0,3577
(seuil à 5 % : 0,3961 pour 25 paires d'observation).

A titre d'exemple soit le dernier graphique portant sur l'évolution des mots latins : l'auteur dit y distinguer 4 paliers : élevés au 16 et au 19e siècles, bas au 18e et au 20e. Ce qui conforte l'interprétation est évidemment l'évidence visuelle : elle est corroborée par un test de corrélation chronologique qui est proche du seuil de signification à 5%. Mais ce test est simplement indiqué dans le graphique et non repris dans le texte. Ce qui est fondamental c'est la topologie du graphique, explicitée dans la conclusion : "L'avenir du latin dans les lettres françaises n'apparait plus voué à une progression linéaire, non plus qu'à une chute inéluctable. Ballotté par les vagues de l'histoire, le latin semble soumis à un mouvement cyclique dont la phase périodique s'étend sur un siècle entier" (p.140). Comme dans l'article précédent, le graphique par son évidence fait office de preuve. Comparons donc maintenant les démarches.

Comparaison

Dans les deux articles, le rapport à l'observé, au terrain, au "réel" est médiatisé par un graphique dont la bonne forme entraine la conviction du lecteur, les tests statistiques ne servant que de garde-fou pour les auteurs, manifestant par leur utilisation qu'ils savent prendre leurs précautions. Ce n'est donc pas au plan de la rhétorique explicite que se joue la différence entre l'article de science de la nature et celui de sciences humaines. Par contre au niveau de la rhétorique implicite, les différences sont importantes : dans le premier cas l'auteur est un élément d'une communauté de chercheurs dont les apports sont sans arrêt invoqués d'une manière implicite[21] par le biais de citations. Pour l'autre, il n'en est pas de même.

Dans le cas du latin, il ne fait aucun doute que si l'auteur est seul, c'est qu'il est en position d'opposition avec le domaine littéraire où il travaille : les techniques quantitatives qu'il emploie permettent des conclusions chronologiques analogues à celles des sciences de la nature mais elles sont mal acceptées, comme le note Charles Muller, qui a été l'initiateur des

[21] les références sont explicites mais leur contenu quant à lui est implicite ; la référence est un *pointeur*, non une attestation explicite.

techniques statistiques appliquées aux textes littéraires, dans une allocation d'ouverture lors d'un colloque de l'*Association for Linguistic and Literary Computing*[22].

> Je m'interroge sur nos échecs.(...) Ce qui m'inquiète, c'est la situation générale des travaux qui mettent l'informatique et la statistique (ou les deux) au service de la linguistique et de l'étude des textes.
>
> Car il faut bien constater que parmi les linguistes et surtout les "littéraires", c'est une minorité qui pratique ces méthodes, et que cette minorité reste très isolée, presque "marginale"; un peu comme une secte hérétique, tolérée mais suspecte. (...) pour ne citer qu'un exemple, je ne vois guère les index et les concordances figurer dans les bibliographie agrégatives, ni leur usage être exposé et recommandé à nos étudiants.

En termes kuhniens, si dans le cas de la biologie nous sommes dans une science *normale* où un consensus existe sur les méthodes (ou dans le cas présent, sur la manière de changer de méthode), dans le cas de l'analyse des textes, la science normale littéraire rejette encore les techniques formelles d'analyse des textes et de ce fait, ceux qui les utilisent ne peuvent pratiquer l'accumulation. Ils en sont réduits à pratiquer non une culture intensive mais ce que J.-Cl. Passeron appelait une "agriculture sur brulis successifs" (1991 : 365). Pour montrer la validité des méthodes qu'ils emploient ils multiplient les exemples mettant à la disposition des littéraires des outils que ceux-ci n'acceptent qu'avec réticence.

Ce n'est pas parce qu'en analyse formelle des textes les outils utilisés sont de nature analogue à ceux d'une science normale que cette utilisation la fait accéder au même statut : la conclusion que l'on peut tirer de cette comparaison est que la différence se joue plus dans l'implicite du consensus sur l'objet que dans l'explicite des méthodes pour le traiter. Dans le cadre de l'ornithologie qui servait de base au premier article, ce qui est à prendre en compte c'est le monde des oiseaux : il est sûr que c'est par référence à l'homme qui a de tout temps identifié ces objets volants naturels que la discipline s'est crée. Le découpage du champ à étudier se fait par affinement des catégories spontanées même si certaines remises en causes sont faites par rapport au sens commun. Le "corbeau" sauf s'il est "grand" est une catégorie du sens commun et des fables de La Fontaine puisque les corbeaux en question sont plutôt dénommés "corneilles". Les divisions fines et leurs regroupements peuvent être objets de discussions, le débat qui existe ne se situe qu'au plan académique.

Il n'en est pas de même dans le découpage du réel qui peut être opéré dans le cas des sciences humaines : prenons pour l'illustrer le cas du latin étudié précédemment. On peut bien sûr comme le fait Etienne Brunet apporter des éléments de réflexion qui, de la manière dont il les envisage, ne souffrent guère discussion, mais s'il a choisi cet objet, c'est qu'il était digne d'intérêt parce qu'objet de débat social. Parler du latin ce n'est pas comme parler des ressemblances de l'hirondelle et du martinet, c'est s'inscrire dans un débat chargé de passion où la neutralité n'est qu'apparente. Par exemple conclure comme Etienne Brunet que l'utilisation du latin est soumise aux vicissitudes de l'histoire peut être interprété par les uns comme un abandon, une trahison même de la part d'un littéraire. A l'inverse se réjouir de l'objectivité de sa conclusion, cela peut être considéré comme se faire le protagoniste d'une modernité militante qui s'oppose au patrimoine de la littérature.

Dans le premier cas, le découpage du réel est sans aspect polémique : l'objet (oiseau) vient du sens commun et il n'est pas objet de débat social. Dans le deuxième cas, il n'en est pas de même, la pratique scientifique elle-même voit ses résultats recyclés dans un débat social qui est sous-jacent au découpage du réel. On peut bien, comme J.- Cl. Gardin le suggère (1993 : 165) refuser la surinterprétation littéraire, on n'échappe pas, en sociologie en particulier, au fait que le découpage du réel, qui peut être fait de différentes manières est déjà un débat social

[22] Etienne Brunet (*ed.*), *Méthodes quantitatives et informatique dans l'étude des textes. En hommage à Charles Muller*, Genève Paris Slatkine Champion 1986, p.10.

et qu'en parler de manière neutre, c'est déjà une manière de prendre partie. C'est peut-être souhaitable politiquement si l'on considère que science et modernité ont à parcourir un chemin ensemble mais c'est certainement moins simple à envisager que lorsqu'on se livre à l'observation des oiseaux.

Philippe Cibois

Université de Picardie Jules Verne

Références bibliographiques

Boudon, Raymond, 1994. *Le relativisme est-il résistible ?*. Paris : Presses Universitaires de France.

Bourdieu, Pierre, 1979. *La distinction*. Paris : Minuit.

Brunet, Etienne, 1996. "Le latin dans la littérature française", *in* Actes du colloque *La réception du latin.* Angers, pp. 125-141.

Chervel, André, 1988. "L'histoire des disciplines scolaires : réflexions sur un domaine de recherches", *Histoire de l'éducation* n°38, pp. 59-119.

Cibois, Philippe, 1981. "Analyse des données et sociologie", *L'Année sociologique* 31, pp. 333-348.

Cibois, Philippe, 1996. "Le choix de l'option latin au collège", *Education & Formations* 48, pp. 39-51.

Compère, Marie-Madeleine, 1985. *Du collège au lycée (1500-1850).* Paris : Gallimard/Julliard (coll. Archives).

Duchesneau, François, 1997. *Philosophie de la biologie*. Paris, Presses Universitaires de France.

Edwards, Scott V., Arctander, Peter and Wilson, Allan, C., 1991, "Mitochondrial resolution of a deep branch in the genealogical tree for perching birds", *Proc. R. Soc. Lond. B* 243, pp. 99-107.

Garin, Eugenio, 1968. *L'éducation de l'homme moderne 1400-1600,* Paris : Fayard (édition originale italienne, Bari, 1957)

Gardin, Jean-Claude, 1991. *Le calcul et la raison.* Paris : Editions de l'EHESS.

Gardin, Jean-Claude, 1993. "Les embarras du naturel", *Archives Européennes de Sociologie* 34, pp. 152-165.

Goblot, Edmond, 1984. *La barrière et le niveau*. Paris : Montfort (édition originale 1925).

Latour, Bruno, 1991. *Nous n'avons jamais été modernes.* Paris : La Découverte.

Passeron, Jean-Claude, 1991. *Le raisonnement sociologique,* Paris : Nathan.

Ricoeur, Paul, 1975. *La métaphore vive*. Paris, Seuil.

Rouanet, Henry et Lépine, Daniel, 1976. "A propos de l'analyse des données selon Benzécri", *Année psychologique* 76, pp. 137-138.

Salem, André, 1995. "Les unités lexicométriques", *in* S.Bolasco, L.Lebart et A.Salem : *Analisi statistica dei dati testuali.* Rome, Cisu (actes des JADT 1995), pp. 19-26.

Shapin, Steven et Schaffer Simon, 1993. *Leviathan et la pompe à air*. Paris, La Découverte (édition originale 1985, Princeton University Press).