cahiers
FUNDAMENTA SCIENTIAE

SEMINAIRE SUR LES FONDEMENTS DES SCIENCES

N° 92

PROBLEMES LINGUISTIQUE ET SEMIOLOGIQUES

L'ANALYSE SEMIOLOGIQUE

Jeanne MARTINET

MATHEMATIQUES ET LINGUISTIQUE

M.P. SCHUTZENBERGER

*　　*

*

1979

UNIVERSITE LOUIS PASTEUR

STRASBOURG

MATHEMATIQUES ET LINGUISTIQUE

par

M.P. Schutzenberger *

* Monsieur M.P. SCHUTZENBERGER

Département de mathématiques
Université Paris VII
2 place Jussieu
75230 PARIS CEDEX 05

Quel titre imposant ai-je accepté dans un mouvement d'enthousiasme, pour les quelques remarques précaires que je suis seulement capable de soumettre à votre discussion ! Et d'ailleurs, que pourrais-je faire d'autre puisque pour traiter le sujet il faudrait être un mathématicien linguiste, si une telle créature peut exister, ou du moins un mathématicien et un linguiste ce qui n'est doublement pas mon cas.

Permettez que nous procédions d'abord à une série de négations visant à exclure ce dont on ne pourrait parler avec profit dans le temps d'une conférence, d'autant qu'il faudra bien raconter la Belle Histoire des langages context-free de Chomsky pour qu'une partie de l'audience ne soit pas trop frustrée dans ses pieuses attentes.

Tout d'abord, statistique et linguistique : je ne crois pas être le seul ici à croire que la statistique est l'une des plus belles applications du Calcul des Probabilités et que ce dernier lui-même est un admirable chapitre de la Physique traitant d'un certain phénomène physique (je répète), à savoir le hasard. Par conséquent, la statistique a une place prééminente partout où l'observation est rendue difficile par des facteurs parasites aléatoires, impossibles à contrôler. Elle a dominé la génétique et joue un rôle grandissant en astronomie. Il ne me semble n'y avoir aucun doute sur sa valeur dans toutes les branches de la linguistique, de la phonétique à la philologie, et sur le fait que nous sommes encore très très loin d'avoir essayé de l'employer partout où cela semble raisonnablement prometteur, comme par exemple en grammaire.

Rappelons pour mémoire quelle dette a contracté le Calcul des Probabilités vis-à-vis des sciences du langage et de la communication. Nous leur devons aussi bien les chaînes de Markov que le théorème du codage de Shannon, donc pour les plus élégants, la classification des shifts bernouilliens. Ici encore, nous proposons que le sujet mérite un séminaire et nous en tirons la conclusion qui s'impose.

A l'autre extrême s'étendent la logique élémentaire et les notations de la théorie des ensembles qui ont donné prétexte à de nombreux ouvrages dont les titres alléchants ont fourvoyé plus d'un étudiant. Précédées ou non d'un cours de probabilités du niveau du brevet de gymnastique, on y voit de longues chaînes de théorèmes aboutissant au résultat majeur que l'ensemble des mots de douze lettres qui commencent par un k est l'intersection de celui des mots dont douze est la cardinalité du nombre des lettres par le complément de celui des mots qui commencent par une autre lettre que k. J'ai, bien sûr, modifié l'exemple auquel je pense pour n'offenser personne, mais quiconque a fréquenté les mathématiques des sciences humaines aura la tentation d'en soumettre de plus drôles. Passons vite, non sans regretter pour les forêts scandinaves que Hjemslev n'aît pas connu Boole. Et souscrivons à l'opinion du prochain conférencier que les "math. modernes" présentent quelques mérites pour ce qui est de l'enrichissement du vocabulaire des marmots. Autre sujet de séminaire !

Nous avons du même coup annoncé que nous ne parlerions que des modèles de langage : mise à part la phonétique pure qui emploie avec succès l'acoustique, refermé les brèves parenthèses de la glottochronologie de Swadesh et du calcul de Lambek et annoncé avec déférence la Théorie sémantique et grammaticale de Thom, l'élaboration de ces modèles est, à ma connaissance, la seule branche de la linguistique où interviennent des mathématiques autres que la théorie statistique ou le vocabulaire des ensembles, ou enfin, mais ceci ne nous concerne pas ici, la logique mathématique.

Excusez-moi d'avoir été aussi expéditif mais je préfère consacrer le temps à plus de réflexion préliminaire sur ce que signifie le fait même d'appliquer les mathématiques à une discipline, comme on disait autrefois ou des mathématiser celle-ci comme on le dit plus volontiers aujourd'hui. Pour aborder ce point il est commode de s'interroger en toute généralité sur les principes qui peuvent permettre de juger que la mathématisation (totale ou partielle)d'une discipline a été ou non réussie. C'est une question qui ne rencontre guère d'échos chez les mathématiciens bien que les plus sages semblent avoir pour critère le nombre des postes d'enseignants et de chercheurs qu'elle permet de faire créer.

De fait le succès prodigieux (succès "au delà de toute raison" dit Wigner) de la mathématisation de la physique ne laisse place chez la plupart des mathématiciens à aucun doute sur la validité ni même la possibilité de l'extension de cette réussite à toutes les autres disciplines. Il est cependant nécessaire de serrer la question de plus près quand ce ne serait que pour marquer les progrès accomplis et dans ce but je propose (sans originalité) que les deux objectifs majeurs sont l'<u>explication</u> la <u>prévision</u>. J'entends par là que la théorie mathématique d'un phénomène peut réaliser (et souvent simultanément) deux buts : donner à ceux qui la connaissent le sentiment d'avoir mieux compris ce phénomène, c'est l'explication - et de sa valeur, seuls peuvent être juges ceux qui éprouvent ce sentiment d'intime illumination-; ou bien, prévoir des conséquences observables, non triviales, qualitatives ou quantitatives.

La physique newtonienne atteint indiscutablement ces deux buts (et aussi un autre bien plus fondamental concernant l'activité de <u>définition</u> qui exigerait à elle seule une longue discussion). Il en est de même,de façon bien plus modeste,de la mécanique de l'hérédité mendélienne. Par contre la génétique des populations peut-être considérée comme purement explicative (avec sans doute quelque exagération dans un sens qu'il est inutile de préciser). Le moindre modèle doit recéler tant de paramètres et chaque observation est tellement perturbée par des facteurs transients ou séculaires que l'accord entre les chiffres (ou les faits) calculés et observés ne peut qu'être la conclusion d'une étude particulière sans permettre de prédiction pour un autre cas. Par contraste, la physique possède en plus de ses théories générales une foule de lois semi-empiriques qui n'expliquent, ni ne prétendent expliquer, rien mais qui conduisent à des prédictions précises et efficaces.

Peut-être pourrait-on soumettre toute mathématisation à l'épreuve suivante inspirée par la méthode de falsification de Popper. Utilise-t-elle un théorème dont les observations permettraient de conjecturer sérieusement la vérité si il se trouvait que la preuve mathématique en fût encore inconnue ?

Un très bel exemple est celui du théorème de Poisson sur l'annulation du potentiel à l'intérieur d'une sphère chargée. Un autre me semble être celui du théorème de Nash (sur la convergence vers le point-selle) qui pourrait être vérifié expérimentalement en faisant disputer des milliers de parties à des joueurs professionnels. Naturellement, il s'agit ici d'une expérience rêvée mais dont le caractère utopique n'a rien à voir avec l'essence du processus étudié.

Un contre-exemple intéressant est l'équation différentielle de Volterra (sur les oscillations des systèmes de populations) qui a fait la joie de tant d'enseignants du PCB à court d'exemples biologiques. Elle ne paraît pas pouvoir passer avec succès notre épreuve.

La question mathématique étant de savoir si les solutions de tel type d'équation différentielle possèdent ou non tel caractère oscillatoire, aucune observation de population ne peut y apporter de réponse. En effet il faut toute la fraîcheur d'âme d'un mathématicien pour croire que le modèle particulier proposé a un rapport plus singulier avec les processus effectifs que mille autres possibles qui en seraient radicalement différents du point de vue mathématique, et, par conséquent, l'observation dans la nature d'oscillations plus ou moins analogues ne dit <u>rien</u> sur l'équation en cause (sinon qu'elle n'est pas trop grossièrement contredite par les faits).

Autrement dit la dynamique des populations offre une illustration du comportement d'un être mathématique et il semble bien vraisemblable que c'est ainsi que l'avait vue l'illustre géomètre. Dans le même ordre d'idée je rappelle à votre attention l'immense littérature médiévale sur les nombres parfaits et le caractère si contemporain de ce mode de discours.

Je suis sûr que nos collègues qui pratiquent l'économie mathématique pourraient singulièrement enrichir de beaux exemples toute cette discussion et après ces préliminaires (trop longs, trop sommaires et trop partiaux) nous en venons à la question impliquée par le titre de la conférence : Existe-t-il des théorèmes conduisant à prévoir des phénomènes linguistiques de façon non triviale ?

J'avance que la réponse est positive et voici pourquoi. Comme on sait, la mode était il y a vingt ans de considérer que pour l'essentiel les phénomènes langagiers étaient de nature markoviens, les états de la chaîne étant les dernières syllabes ou mots prononcés et la matrice de transition étant par définition la langue considérée. Modèle excellent pour les fins qui l'avaient suggéré (la transmission du signal) mais peut être un peu sommaire pour le reste. Comme la théorie des chaînes de Markov à une infinité d'états était aussi fort à la mode, le remède du passage à la limite (par augmentation du nombre des états) était sous la main, et dans les Collèges les plus illustres on entendit déjà quelques années plus tard promettre des explications savantes du langage fondées sur cette approche.

Au tournant de cette époque était apparue une autre théorie. Abandonnant toute tentative de production séquentielle de la chaîne du discours, N. Chomsky proposait d'engendrer l'ensemble des phrases grammaticalement acceptables par un algorithme différent reposant sur des substitutions itérées. La chose est trop connue pour que nous nous attardions à plus de détails.

Pour un raisonnement mathématique il en déduisait plusieurs conséquences. La première est que si son modèle était correct toute tentative d'approximer les phrases grammaticalement acceptables par un système markovien devait buter irrémédiablement sur l'obstacle d'une croissance trop rapide (de type exponentiel) du nombre des états requis avec la longueur des phrases à analyser. Une seconde (à vrai dire plutôt indiquée qu'imposée par le modèle) était l'abondance d'ambiguïtés inhérentes dans l'analyse grammaticale.

L'une et l'autre ont été amplement confirmées par les expériences qui de Leningrad à Washington ont consacré des dizaines de millions à vérifier l'impossibilité de la traduction automatique. Pour les besoins de cet exposé je préfère les considérer comme une possibilité de "vérifier empiriquement" un théorème trop technique pour être énoncé ici et dont on doit la substance à Kleene et à Chomsky.

Une troisième conséquence observable et mathématiquement dérivable du modèle de Chomsky concerne la géométrie des relations de dépendance grammaticale : à un certain niveau d'approximation elles sont représentables par des graphes planaires. Phénomène remarquable que Tesnières avait vu il y a une quarantaine d'années dans des travaux anticipant ce qui se fait aujourd'hui. De nouveau, je vous invite à voir dans le fait que quasiment toutes les langues naturelles examinées présentent ce caractère de planarité, une confirmation empirique de la preuve d'un énoncé mathématique non trivial.

Tout ceci date d'il y a vingt ans. Depuis, la théorie mathématique de ce que l'on appelle maintenant les "langages" (=parties d'un monoïde libre) "algébriques" (= solutions formelles d'équations polynomiales en variables non commutatives) s'est épanouie et la brillante école que M. Nivat a réussi à former autour de lui est connue pour tenir le premier rang parmi les nombreux chercheurs qui cultivent ce sujet dans le monde.

Un facteur essentiel de ce développement est que la structure générative proposée par Chomsky comme modèle de grammaticalité pour les langues naturelles est à peu près parfaitement réalisée dans les langages de programmation tels que le Fortran ou l'Algol. Les questions mathématiques que posent les "langages" algébriques sont donc liées de façon très intime aux problèmes dits "de compilation". De plus, une technique informatique importante en analyse non numérique (la "gestion de pile") se trouve être en correspondance avec l'un des algorithmes de base de la théorie algébrique. Le livre de J. Berstel (sous presse chez Teubner) fait à merveille le point des résultats acquis.

Sur un autre versant les rapports entretenus avec les mathématiques les plus classiques sont assez profonds. Par exemple c'est avec joie que nous avons vu R. Lyndon (qui ne s'est jamais, je crois, soucié des langages algébriques) faire intervenir avec succès des considérations de planarité (cf. plus haut) dans la théorie dite combinatoire des groupes libres (par exemple dans la preuve du lemme de Greendlinger). Ici encore je dois me borner à être abusivement allusif en jalonnant des noms de Lentin, Tutte et Cori les très larges routes qui relient ces résultats.

Nous ne sommes pas tout à fait sorti du sujet puisqu'il s'agissait de langages pour communiquer avec les ordinateurs mais il vaut mieux revenir à la linguistique stricto sensu, et la deuxième question qui me paraît se poser maintenant est celle de la continuation de contacts non triviaux entre celle-ci et ses modèles mathématiques.

Autrement dit, mises à part les applications du calcul des probabilités et de la statistique, existe-t-il une discipline que l'on puisse sérieusement appeler linguistique mathématique ?

Bien sûr, je ne parle pas d'un enseignement portant ce nom puisque tout ce que j'ai déjà exposé (et quelques autres sujets non négligeables que j'ai laissés dans l'ombre) justifient amplement l'existence d'une propédeutique à contenu mathématique (et informatique) contribuant à la formation de linguistes.

Bien sûr aussi, je ne parle pas de ce qui peut ou pourrait être demain. Très concrètement je parle de recherches actuelles actives et sérieuses dans lesquelles la solution d'un problème mathématique donnerait la réponse à une question linguistique, ou en poserait une nouvelle, comme cela se passe quotidiennement en physique.

La réponse me paraît (aujourd'hui) être non et ceci pour la base de deux faits. Le premier est le développement de l'Ecole de Chomsky. Ainsi que j'ai essayé de le montrer plus haut les premiers résultats mathématiques de cet auteur avaient une signification linguistique certaine et immédiate.

Depuis, malgré une audience mondiale, des imitateurs nombreux et l'atmosphère vigoureusement pro-mathématique et interdisciplinaire de l'époque, et spécialement de M.I.T., rien de nouveau ne semble devoir être ajouté à ce qui se trouve déjà dans les chapitres qu'il avait écrit en 1963 avec G. Miller dans le <u>Manuel de Psychologie Mathématique</u>. Les travaux de l'école chomskienne ont porté sur l'usage du modèle génératif dans la discussion subtile de phénomène langagiers.

Pour ce qui est des techniques mathématiques employées, les arbres de Tesnière y suffiraient largement. D'ailleurs, développant une autre méthode proposée par Zellig Harris, ce modèle trop étroit a vite été remplacé par un modèle transformationnel dont les propriétés mathématiques sont nulles. Ceci ne signifie évidemment pas qu'il soit vide ou inintéressant mais simplement qu'il ne possède pas une structure abstraite assez précise et assez riche pour bénéficier de raisonnements mathématiques excédant ceux des "Math. Modernes". D'autres recherches de la même école se sont constamment appuyées sur la logique mathématique, mais ceci n'est en rien le propre de la linguistique puisqu'une interaction de cette nature est un trait général de la philosophie anglo-saxonne. A titre de confirmation, je vous renvoie aux ouvrages les plus récents des exposants autorisés français de la linguistique générative et transformationnelle, et je précise qu'il en est de même en ce qui concerne les écoles dissidentes (schismatiques ou hérétiques).

Tel me paraît être le constat relatif au courant le plus porté à la mathématisation et un deuxième ordre de faits me conduit à croire que ceci a des causes internes propres au langage lui-même.

Je veux parler ici des travaux de mon ami M. Gross. Parti il y a quinze ans du désir d'appliquer les méthodes dont nous venons de parler il a inventorié <u>exhaustivement</u> les phénomènes qu'il voulait étudier. Principe habituel dans les sciences de la nature mais idée extrêmement neuve et féconde dans une discipline où (comme dans bien d'autres!) le raisonnement coutumier procède par voie d'exemples et contre-exemples sans le moindre souci de leur extensionnalité. Or, à sa surprise, les faits rassemblés sont apparus très différents de ceux qu'aurait prescrit la théorie initiale : utilisant les mêmes critères ou des critères analogues à ceux employés dans la casuistique des exemples, les huit mille verbes de français se révèlent à peu près tous être chacun <u>unique</u> quand aux propriétés grammaticales. Sautant à la conclusion, nous retrouvons à un autre niveau la même contradiction qui avait conduit à écarter les modèles markoviens. Même à un niveau d'approximation sommaire, la réalité linguistique se révèle trop complexe pour être utilement modélisée par les grammaires génératives ou transformationnelles. L'étude fine des propriétés mathématiques ou logiques de ces dernières est donc aussi inadéquate au but poursuivi que le serait la théorie de l'approximation des nombres algébriques de Baker à la théorie des gaz nonobstant la loi de Mariotte et la définition des logarithmes.

Depuis, l'école de M. Gross a retrouvé dans d'autres langues que le français, les mêmes phénomènes et l'on sait que cette richesse de complexité s'est traduite par un renouvellement de la problématique et la solution de nombreuses questions linguistiques selon des méthodes originales qui (malheureusement dira-t-on) n'ont rien de mathématique.

Résultat inattendu ? peut-être moins, si l'on avait mieux compris les indications de Zellig Harris (dont et Chomsky et Gross sont les élèves). En tous cas je tire de ces deux ordres de faits la conclusion que pour ce qui est du passé et du présent, la linguistique et les mathématiques ont eu un point de tangence - peut être même d'osculation. Je ne sais ce que réserve le futur.

Veuillez ne voir dans ces propos, ni pessimisme ni regret. L'osculation (comme toute la géométrie différentielle) est une très belle chose et de voir les recherches linguistiques lancées dans une direction nouvelle ne peut que susciter l'enthousiasme.

Les mathématiciens non plus ne doivent pas pleurer. Si leur but est,comme il se doit,de faire progresser la connaissance des objets mathématiques (et non d'enrichir ou de raffiner une nouvelle rhétorique) ils ont acquis toute une série d'intuitions nouvelles sur des objets classiques et quelques objets nouveaux qui ne semblent pas déraisonnables. Vous pourrez juger sur pièces quand sera paru le volume C du traité de Eilenberg, si les travaux publiés jusqu'ici vous semblent n'avoir pas encore atteint ce niveau de perfection abstraite qu'exigent les canons de l'époque.

*

* *