MATHÉMATIQUES ET SCIENCES HUMAINES

SOLOMON MARCUS

Mathématique et linguistique

Mathématiques et sciences humaines, tome 103 (1988), p. 7-21

<http://www.numdam.org/item?id=MSH_1988__103__7_0>

Math. Inf. et Sci. hum. (26e année, n°103, *1988*, pp. 7-21).

MATHEMATIQUE ET LINGUISTIQUE

Solomon MARCUS[1]

DEUX PRECURSEURS : DESCARTES ET LEIBNIZ.

L'histoire des relations entre mathématiques et linguistique commence bien avant l'apparition de ce qu'on appelle, depuis quelques décennies, la linguistique mathématique et continue même lorsqu'on cesse de parler de ce domaine interdisciplinaire. Il y a d'abord la longue période qui précède l'apparition de la linguistique comme discipline scientifique. On sait que cet événement s'est produit dans le siècle passé et s'identifie avec le développement de la linguistique historique, notamment des méthodes de la grammaire comparée indo-européenne. Mais des idées logiques et mathématiques dans l'étude et la construction des langues apparaissent quelques siècles auparavant. L'ancien rêve, dont l'origine est l'épisode de la Tour de Babel, se prolonge par le projet de construction de langages artificiels pour la communication internationale. Le premier projet scientifique de ce type appartient à René Descartes qui, en 1629, propose la création d'une langue reposant sur une grammaire formée de règles logiques et numériques et ne connaissant aucune irrégularité. Des tentatives similaires sont reprises au XVIIème siècle par Dalgano, Urquhart, Wilkins et Leibniz. Il s'agit de systèmes artificiels *a priori*, dont la finalité principale est de permettre et de stimuler la réflexion philosophique, (pour plus de détails, voir Mario A. Pei, "Artificial languages ; International (auxiliary)", in *Current Trends in Linguistics* (ed. Th. A. Sebeok), vol. 12, XX, La Haye - Paris, Mouton, 1974, p. 999-1017).

Il est important de remarquer que les tentatives de Descartes et de Leibniz se produisent justement au cours de la période où prend naissance le langage mathématique comme une entité bien différente de la langue naturelle. Certains philosophes contemporains voient dans ce moment le commencement du divorce entre le langage scientifique et le langage humain. C'est ce qu'affirme, par exemple, Georges Steiner (dans son livre *Language and Silence*, New York, Atheneum, 1967 ; version française, Paris, Editions du Seuil, 1969), pour qui l'histoire du langage scientifique est celle d'une déshumanisation et d'une incommunicabilité progressive. En opposition avec cette attitude, l'histoire récente du langage scientifique montre que beaucoup de langages artificiels sont, dans un certains sens, plus humains que les langues naturelles. En effet, comment détermine-t-on le caractère plus ou moins humain d'un langage ? Selon sa genèse, son contenu, sa structure, sa finalité ou bien selon son degré d'accessibilité ?

1 Université de Bucarest, Faculté de Mathématique, 14 Academici str., 70109 BUCAREST, Roumanie.

L'accessibilité est le résultat d'un processus d'apprentissage et d'une accommodation progressive. Ce qui détermine le caractère humain d'un langage c'est son contenu et sa finalité. On ne peut pas opposer les langages artificiels aux langages naturels de la même façon qu'une prothèse ne peut être opposée à l'organe qu'elle prolonge ou remplace.

L'ESSOR DES LANGAGES ARTIFICIELS.

La géométrie analytique de Descartes et le calcul différentiel et intégral de Newton et Leibniz marquent une étape où le raisonnement scientifique a besoin d'une précision et d'une rigueur que la langue naturelle, formée dans la communication spontanée des gens, ne peut plus satisfaire. Remarquons que Descartes et Leibniz sont, comme on l'a rappelé ci-dessus, justement les premiers auteurs de projets de langages logiques ; cette coïncidence donne beaucoup à réfléchir.

Mais si la langue naturelle n'est plus suffisante, il ne faut pas oublier qu'elle reste toujours nécessaire dans toutes les étapes de l'activité scientifique. Il est prouvé que l'élaboration des idées scientifiques repose sur la langue naturelle. La communication scientifique, même en mathématique, doit se prévaloir de la langue naturelle pour au moins 50%, si elle veut se rendre efficace (voir Paul R. Halmos, *How to write Mathematics*).

Mais si les langues naturelles ont besoin d'être prolongées et renforcées par la prothèse des langages artificiels, notamment du langage mathématique, alors le même mariage est nécessaire lorsqu'on édifie le métalangage utilisé dans l'étude des langues naturelles. Si l'on accepte que des sciences comme la physique, la chimie ou la biologie aient besoin d'un métalangage mixte, dont une composante est naturelle et l'autre est formelle, alors on ne peut dénier cette même exigence à la linguistique. C'est là une motivation fondamentale de la linguistique mathématique.

L'intérêt de plus en plus grand accordé aux langages artificiels formels renouent maintenant avec les traditions du XVIIème siècle. Le terrain était préparé par G. Peano, qui vers la fin du XIXème siècle tentait, pour la première fois, de formaliser la composante naturelle du langage mathématique. De telles tentatives sont restées jusqu'à aujourd'hui sans résultat, peut-être parce que la langue naturelle est organiquement impliquée dans notre comportement et ne peut pas être remplacée, dans son intégralité, par quelque chose de plus simple. Avec B. Russel, A.N. Whitehead et D. Hilbert, on formalise l'idée de démonstration et on prouve qu'il s'agit ici de ce qu'on appelle, dans la terminologie actuelle, une structure de langage formel. En effet, un système formel est une collection de quatre objets : un vocabulaire V et quatre langages L_1, L_2, T et A sur le vocabulaire V, tels que L_1 et L_2 sont disjoints et $A \subset T \subset L_2$, L_1 est l'ensemble des termes, L_2 l'ensemble des relations, T est l'ensemble des théorèmes et A est l'ensemble des axiomes. On indique aussi une procédure constructive pour obtenir les théorèmes, à l'aide d'un système K de règles. On appelle *texte démonstratif* toute suite finie de relations où chaque relation est ou bien un axiome, ou bien se déduit des relations précédentes dans la suite, à l'aide du système K.

Chaque relation qui est un terme d'un texte démonstratif constitue un théorème. Mais les textes démonstratifs forment, à leur tour, un langage et on constate ici comment les idées de Hilbert anticipent celles de Chomsky. Les textes démonstratifs sont remplacés par les dérivations, les règles de déduction sont remplacées par les règles de dérivation, les théorèmes sont remplacés par des suites correctement formées. Si, grâce à G. Boole, la logique a reçu, au siècle passé, un modèle mathématique, grâce à Hilbert, elle a reçu, au cours de notre siècle, une structure de langage formel. Comme une confirmation, la biologie et la psychologie moderne placent les activités logiques et linguistiques sous le contrôle de l'hémisphère cérébral gauche, comme variantes des activités de nature séquentielle. Le logique et le linguistique vont ensemble pour des raisons bien plus profondes que celles qu'on pouvait soupçonner.

MATHEMATIQUE DE LA PHONOLOGIE.

Mais l'intérêt pour les langages artificiels formels se développe maintenant dans des directions multiples. Il faut d'abord signaler la diffusion de la méthode axiomatique-déductive, non seulement dans les sciences naturelles, mais aussi dans les sciences humaines. Pour la linguistique, il faut mentionner L. Bloomfield, "A set of post slates for the science of language" (*Language*, vol. 2, 1926, p. 26-31), qui précède de onze ans la tentative de H. Woodger (*The Axiomatic Method in Biology*, Cambridge Univ. Press, 1937) d'axiomatiser la biologie. Mais la tentative de Bloomfield, assez naïve, rappelant la manière dont Baruch Spinoza essayait, au XVIIème siècle, d'axiomatiser l'éthique, est suivie de quelques tentatives plus approfondies pour axiomatiser la phonologie : B. Bloch ("A set of postulates for phonemic analysis", *Language*, vol. 24, 1948, n° 1), J.H. Greenberg ("An axiomatization of the phonologic aspect of language", *Symposium on Sociological Theory*, ed. L. Gross, New York, Evanston, 1959), pour culminer dans les travaux des logiciens comme Tadeusz Batog ("Logiczna rekonstrukcja pojecia fonemu", *Studia Logica*, vol. 11, 1961, p. 139-183) et Stig Kanger ("The notion of a phoneme", *Statistical Methods in Linguistics*, 1964, n° 3, p. 43-48), des linguistes comme G. Ungeheur ("Das logistiche Fundament binärer Phonemklasifikationen", *Studia Linguistica*, 1959, n° 2, p. 69-97), R.G. Piotrovskiĭ ("Eščo raz o differencialnyh priznakah fonemy", *Voprosy Jazykoznanÿa*, 1960, n° 6), S.K. Šaumjan ("Problemy teoretičeskoĭ fonologi", *Izd. Akad. Nauk SSSR*, Moscou, 1962), V.N. Belcozerov ("Formulnoe opredelenie fonemy", *Voprosy Jazykoznanÿa*, 1964, n° 6, p. 54-60), I.I. Revzin ("K logičeskomu obosnovaniju teorii fonologičeskih priznakov", *Voprosy Jazykoznanÿa*, 1964, n° 5, p. 59-65). V.A. Vinogradov ("Nekotorye voprosy teorii fonologičeskih oppozic i neutralizacci", *Problemy lingvističeskogo analiza*, Izd. Nauka, Moskva 1966, p. 3-25), M.I. Lekomceva ("K opisaniu fonologičeskoi sistemy staroslavianskogo jazyka na osnove ternarnogo principy", *Lingvističeskie issledovanija po obščei i slavianskoi tipologii*, Izd. Nauka, 1966, p. 117-123) et F.H.H. Kortlandt (*Modelling the Phoneme*, La Haye, Mouton, 1972), qui donne une synthèse critique de toutes les contributions de l'Europe de l'Est dans le domaine de la phonologie, et des mathématiciens comme Solomon Marcus ("Un model matematic al fonemului", *Studii şi Cercetări Matematice*, vol. 14, 1963, n° 3, p. 405-421 ; "Sur un ouvrage de Stig Kanger concernant le phonème", *Statistical Methods in Linguistics*, 1966, n° 4), V.A. Uspenskiĭ ("Odna model dlja ponjatija fonemy", *Voprosy Jazykoznanijy*, 1964, n°6, p. 39-53), Ladislav Nebeský ("K odnoi matematičeskoi modeli fonemy", *Revue Roumaine de Mathématiques Pures et Appliquées*, vol. 11, n°4 ; "On the notion of relevant features", *Prague Bulletin of Mathematical Linguistics*, vol. 6, 1966, p. 35-43 et Barron Brainerd ("On Marcus definition of phoneme", *Cah. de Ling. thé. et appl.*, vol. 6, 1969, p. 15-41).

Une mention spéciale pour le modèle proposé par Brainerd, où l'on trouve un traitement général de la distinction entre *unité étique* et *unité émique*, entre *système étique* et *système émique*, formalisant et généralisant la distinction bien connue, mais jamais complètement clarifiée, entre *phonétique et phonémique*. Il est intéressant de voir, dans la succession de ces modèles, comment la présence de la sémantique devient de plus en plus forte. La sémantique est absente dans le modèle de Revzin, faible dans le modèle de Marcus, forte dans le modèle de Nebeský.

A ces travaux, on doit ajouter les collaborations interdisciplinaires, d'abord le fameux article de E.C. Cherry, M. Hall et R. Jakobson ("Toward the logical description of languages in their phonemic aspects", *Language*, vol. 29, 1953, n°1), puis F. Harary, H.H. Paper ("Toward a general calculus of phonemic distribution", *Language*, vol. 33, 1957, p. 143-169), S. Marcus - E. Vasiliu ("Théorie des graphes et consonantisme de la langue roumaine", *Revue de Mathématiques Pures et Appliquées*, vol. 5, 1960, n° 2 et n° 3-4), G.E. Peterson - F. Harary (*Foundations of phonemic theory*, Proceedings of the Symposia of Applied Mathematics, vol. 2, 1961, p. 139-165).

Nous insistons sur l'exemple de la phonologie, car il est l'une des meilleures réussites de collaboration de la linguistique avec la logique et la mathématique. Toutes les théories phonologiques importantes ont bénéficié de modèles mathématiques. Des questions cruciales telles que : comment choisir, parmi les traits acoustiques ou articulatoires d'un son du langage, ceux qui sont pertinents du point de vue linguistique ? (point de vue de Jakobson) ou : dans quelles conditions deux séquences sonores correspondent-elles à un même phonème (point de vue de la linguistique descriptive) ?, ont reçu des réponses profondes et révélatrices (bien que pas encore élaborées dans tous les détails). On a montré, par exemple que, contrairement à ce qu'on attendait, la relation binaire qui s'établit entre deux séquences sonores définissant un même phonème n'est pas une relation d'équivalence, car elle n'est pas transitive. En fait, le statut même du phonème, comme entité conceptuelle, s'est clarifié seulement après l'intervention des mathématiques, bien que certaines zones restent encore obscures ou controversées. Il y a, par exemple, en phonologie, une dialectique très fine, entre le discret et le continu, qui a engendré des stratégies d'une grande diversité. Certains chercheurs japonais (M. Iri, "The logarithmic vowel triangle based on the phonetic-geometrical structure of the Japanese and English languages", *RAAG Memoirs of the Unifying Study of Basic Problems in Engineering and Physical Sciences by means of Geometry*, vol. 3, Division H, 1962, p. 551-566 ; K. Kondo, "A geometrical analytical approach to the mathematical foundation of Phonetics", *ibidem*, p. 567-606 ; K. Kondo, "Quaternionion Phonology", *ibidem,* p. 606-646 ; K. Kondo, M. Takata, Schun-ichi Amari, "Analytical structures of speechsounds mechanical composition and decomposition of phonemes and the design principle of primary phonetical automata", *ibidem*), ont employé la géométrie analytique et différentielle pour explorer systématiquement la réalité phonétique et son évolution vers une réalité phonologique. Une autre représentation continue de la réalité phonétique se trouve chez Tadeusz Batog (op. cit.), où l'on emploie dans ce but la méréologie de Lesniewski, qui concerne les relations du type *partie-tout*, très différentes des relations du type *élément-ensemble*. Il est intéressant de remarquer que la même méréologie de Lesniewski est à la base de l'axiomatique proposée par J.H. Woodger (*op. cit.*) pour la biologie et de l'axiomatique proposée par V.F. Rickey (*A survey of Lesniewski's logic*, Dept of Math., Bowling Green State Univ. Ohio, 1976) pour la syntaxe. Il y a ici une réplique à Ferdinand de Saussure. L'objet de la linguistique est-il seulement le discret de la langue ou bien aussi le continu de la parole *et* le passage de celui-ci au discret de la langue ?

L'évolution des dernières décennies a écarté le phonème du premier plan de la phonologie, car, comme on le sait, Roman Jakobson et Noam Chomsky ont choisi, pour des raisons différentes les traits distinctifs binaires comme l'information phonologique pertinente et opérationnelle. Mais le phonème reste néanmoins un des concepts-clé de la linguistique.

BENZECRI ET THOM

Parmi les partisans d'une vision continue il faut mentionner Jean-Paul Benzécri, qui utilise la théorie de la mesure, l'analyse factorielle et la topologie algébrique ("Physique et langue", *La Traduction Automatique*, vol. 4, 1963, n° 2, p. 31-50 ; *Linguistique Mathématique*, Université de Rennes, 1964) et surtout René Thom, qui, il y a vingt ans, commence à publier une série de travaux consacrés surtout à la sémantique et qui se réclame d'une philosophie où le continu est en premier plan. Thom a réussi à prouver que le discret perceptible dans la sémantique du verbe français est l'effet d'un processus dont la dynamique ne peut être saisie que dans un contexte continu. En général, on peut dire que le continu linguistique est chez Thom ce que depuis longtemps l'analyse mathématique est pour la théorie des nombres. Les travaux de linguistique et de sémiotique de René Thom sont réunis, avec certains autres articles du même auteur et consacrés, tous, à la théorie des catastrophes, dans son livre *Modèles mathématiques de la morphogénèse* (Paris, Christian Bourgois, 1980). L'attitude de René Thom doit être analysée dans un contexte plus général, où se place, par exemple, aussi son point de vue en ce qui concerne l'enseignement de la géométrie. Il faut remarquer que la plaidoirie de René Thom pour

l'utilisation, en biologie et en linguistique, des modèles de la topologie différentielle est concomitante avec une plaidoirie similaire de S. Smale en ce concerne les sciences économiques. Peut-être doit-on placer dans le même ordre d'idées l'utilisation de l'analyse non-standard de Abraham Robinson dans l'étude des économies d'échange. Toute cette offensive du continu dans les sciences de l'homme se développe après une période où la plupart des chercheurs considéraient que les sciences humaines et sociales ont besoin surtout des mathématiques discrètes, car elles ont à résoudre le problème de bien choisir leurs unités fondamentales et étudier la combinatoire de ces unités. La provocation que René Thom a lancée au monde scientifique est d'autant plus audacieuse qu'elle se passe à un moment où, stimulées par l'informatique, les mathématiques discrètes connaissent un développement sans précédent et où beaucoup d'auteurs les placent, en importance, avant les mathématiques continues (voir, par exemple, Anthony Ralston, "The Decline of Calculus - The Rise of Discrete Mathematics", in *Mathematic Tomorrow*, editor Lynn Arthur Steen, New York, Springer, 1981, p. 213-229). Il faut aussi prendre en considération les résultats récents de la biologie et de la psychologie, qui attribuent à l'hémisphère gauche du cerveau humain une priorité en ce qui concerne le contrôle des activités de nature séquentielle, dont le langage et la logique sont les plus importantes. Mais justement ce dernier fait peut nous orienter dans le débat si délicat sur le discret et le continu. On accepte aujourd'hui qu'une des conditions de santé psycho-somatique de l'homme soit celle de l'équilibre entre les activités des deux hémisphères cérébraux, donc entre les activités séquentielles et les activités non-séquentielles, de concomitance (où sont incluses les émotions, les intuitions et l'affectivité). Il y a donc, au niveau biologique, une solidarité entre le continu et le discret, un besoin d'équilibre entre eux. Chaque mathématicien sait que l'analyse mathématique utilise des approximations discrètes et mêmes finies, tandis que l'algèbre et la théorie des nombres ont besoin de l'analyse. Que serait la théorie des nombres premiers sans l'hypothèse de Riemann ? On l'a vu déjà ci-dessus, mais nous allons le voir encore : l'étude du discret linguistique a bénéficié de l'apport de l'analyse, de la topologie et de la théorie de la mesure. L'interaction du discret et du continu est inévitable. Voir maintenant, à ce sujet, Jean Petitot-Cocorda, *Les catastrophes de la parole, de Roman Jakobson à René Thom*, Paris, Maloine, 1985.

DEUX SOURCES DE L'AXIOMATISATION EN LINGUISTIQUE

Le développement de la pensée axiomatique en phonologie a deux sources, ce phénomène étant valable pour l'ensemble de l'abord axiomatique - déductif en linguistique. Il s'agit, d'une part, du fait que l'axiomatisation et la formalisation sont à la mode dans notre siècle, à la suite du développement de la logique et de la mathématique. Il s'agit, d'autre part, du fait que le développement de la linguistique structurale, avec Ferdinand de Saussure et l'Ecole de Prague, en Europe, avec E. Sapir, L. Bloomfield, Z. Harris et d'autres, en Amérique, a préparé, à l'intérieur de la linguistique, la voie vers la mathématisation. En ce qui concerne les modèles mathématiques en phonologie, on peut dire que les travaux de Batog et Kanger ont une origine du premier type tandis que ceux de Bloch, Revzin et Šaumjan sont du deuxième type. Mais dans les plupart des recherches les deux types se combinent, ce qui explique aussi la fréquence des travaux qui résultent de collaborations interdisciplinaires. Nous allons suivre ces développements, mais il faut remarquer une troisième ligne de pensée, peut-être la plus importante.

Avec le développement moderne de la linguistique structurale, la plupart des applications des mathématiques en linguistique concernaient l'aspect statistique au niveau du vocabulaire. Certains dictionnaires de fréquences datent du commencement de notre siècle et même du siècle passé. Les mathématiques sont aristocrates ; elles aident mieux les riches. Plus une discipline est mûre, plus elle peut bénéficier des idées, des résultats et méthodes mathématiques. Plus une science est capable d'expliciter ces aspects structuraux, plus elle se prête à être explorée à l'aide des mathématiques. Il se trouve que le développement de la pensée structurale est presque parallèle en linguistique et en mathématique. La linguistique a son Bourbaki, ou peut-être

plusieurs Bourbaki, car le structuralisme linguistique européen a développé son propre jargon, bien différent de l'américain.

LES ANNES CINQUANTE

Dans les années cinquante de notre siècle, le développement de la cybernétique, de la théorie de l'information et des ordinateurs concrétise une commande sociale de plus en plus pressante, concernant le besoin de traitement automatique de l'information, des textes et des langues. Des travaux caractéristiques de cette période : Vitold Belevitch, *Langages des machines et langage humain* (1949) ; J. Apostel, B. Mandelbrot, A. Morf, *Logique, langue, théorie de l'information* (Paris, Presses Universitaires de France, 1957). En Occident, on met l'accent sur les relations entre langage et information (rappelons que Claude Shannon publie en 1951 "Prediction and entropy of printed English", *Bell System Technical Journal*, 30, 1951, p.50–60, tandis que, en 1957, paraît la première édition du livre de Colin Cherry, *On Human Communication*, Cambridge, M.I.T. Press, Mass., où la relation du langage avec les aspects techniques et mathématiques de l'information et de la communication est présentée d'une façon systématique et détaillée). En même temps, les premières expériences de traduction automatique se déroulent avec enthousiasme à l'Ouest et à l'Est, mais la réaction théorique est différente. Une photographie fidèle de ce moment, en Union Soviétique, est l'article de Robert Abernathy ("Mathematics Linguistics", dans Current Trends in Linguistics, Th. A. Sebeok ed., vol. 1, La Haye, Mouton, 1968, p. 113-131). Plusieurs mathématiciens soviétiques se préocuppent, à ce moment, de l'élaboration de modèles mathématiques permettant une représentation rigoureuse de l'information linguistique : A.N. Kolmogorov (la catégorie du cas), R.L. Dobrušin (la catégorie grammaticale élémentaire), V.A. Uspenskiĭ (la partie du discours) et surtout O.S. Kulagina, qui dans un article programme ("Ob odnom sposobe opredelenija grammatičeskih ponjatiĭ na baze teorri množestv", Problemy Kibernetiki, vol. 1, 1958, p. 203-214) trace les lignes de base de ce qu'on va appeler le modèle ensembliste de la langue. Les bases linguistiques de ce modèle se trouvent dans la linguistique descriptive américaine (la notion de famille introduite par Kulagina est une formalisation des classes de distribution de Z.S. Harris) et dans le courant d'idées du Cercle de Prague, en ce qui concerne les aspects paradigmatiques. Mais le modèle de Kulagina doit être aussi rapporté à la structure de la langue russe et, en général, des langues slaves, caractérisées par une morphologie riche, en contraste avec l'anglais (et même le français), où l'intérêt est orienté vers la syntaxe. Même si les motivations immédiates de ces modèles tenaient de "la préparation des langues naturelles pour leur traitement automatique", on peut maintenant juger leur apport en général.

IMPORTANCE DU MODELE DE KULAGINA

Il y a deux idées profondes dans le modèle de Kulagina :

a) le concept de dérivation d'une partition du vocabulaire, qui se propose de formuler une hypothèse en ce qui concerne le passage de la partition en classes flexionnelles à la partition en parties du discours (il s'agit, naturellement, d'une approximation, car généralement ni les classes flexionnelles ni les parties du discours ne forment des partitions) ;

b) le concept de configurations syntaxiques de divers ordres, qui se propose de décrire le mécanisme logique permettant de distinguer des dépendances de divers ordres (par exemple, dans *très beau livre*, la dépendance de *beau* par rapport à *livre* devient manifeste seulement après avoir tenu compte de la dépendance de *très* par rapport à *beau*, en remplaçant *très beau* par la résultante *beau* ; on obtient ainsi le syntagme *beau livre*, dont la résultante est *livre*). Chacune de ces idées a engendré des recherches riches et intéressantes. Les premiers pas ont été enregistrés dans notre monographie parue en 1967 (*Algebraic Linguistics ; Analytic Models,*

New York, Academic Press), mais les recherches ultérieures n'ont pas encore été synthétisées. Les résultats pourraient être résumés de la façon suivante : création du concept de catégorie morphologique (qui, comme dénominateur commun des catégories morphologiques spéciales, telles le genre, le nombre, le cas, etc. n'existe pas dans le linguistique structurale) ; découverte de mécanismes formels expliquant le phénomène d'homonymie contextuelle et permettant de comparer deux formes du point de vue de leur degré d'homonymie ; élaboration de modèles algébriques de la partie du discours, du cas et du genre grammatical, appliqués aux langues slaves de l'Est par O. Karpikaja ("Metody tipologičeskogo opisanija slavianskih rodovyh sistem", *Lingvističeskie issledovanija po obščei i slavianskoi tipologii*, Izd. Nauka, 1966, p. 75-116) et au Tchèque et au Slovaque par J. Horecky (*Uvod di matematickej jazykoved*, Univ. Komenského, v. Bratislave, 1969) ; élaboration d'un typologie linguistique raffinant la typologie traditionnelle ; un ouvrage significatif de cette orientation est celui de I.I. Revzin ("Metod modelirovania i tipologya slavianskih jazykov", *Izd. Nauka*, Moskva, 1967). La théorie des configurations syntactiques a été développée par A.V. Gladkiĭ, Miroslav Novotny et Maria Semeniuk ; mais le problème de connaître l'ordre maximal que peuvent avoir les configurations syntactiques dans les langues naturelles n'est pas encore résolu.

LE MOMENT DOBRUŠIN - SESTIER

La relation de domination contextuelle, introduite par R.L. Dobrušin en 1957, s'est avérée riche en conséquences. On dit que la phrase x domine la phrase y par rapport au langage L si chaque contexte qui accepte x dans L accepte aussi y dans L. Si x domine y, alors l'homonymie contextuelle de x n'est pas supérieure à celle de y. L'algèbre de cette relation de domination s'est avérée un chapitre intéressant de la théorie des monoïdes libres, en convergence avec certaines notions et résultats étudiés préalablement, sans aucun rapport avec les langues naturelles. Indépendamment de Dobrušin, un auteur mystérieux, A. Sestier ("Contributions à une théorie ensembliste des classifications linguistiques", *Premier Congrès de l'Association Française de Calcul*, Grenoble, 1960, Paris, 1961, p. 293-305) a proposé un abord purement contextuel à l'aide de ce qu'on appelle aujourd'hui une fermeture de Sestier. En partant d'une collection A de phrases, si $c(A)$ est ensemble de contextes qui acceptent chaque phrase de A, alors la fermeture de Sestier $f(A)$ est l'ensemble des phrases acceptées par chaque contexte de $c(A)$. Une étude approfondie des fermetures de Sestier comme modèle algébrique des catégories grammaticales a été développée par Jurgen Kunze, dans une série de cinq articles sous le titre "Versuch eines objectivierten Grammatik Modells" (*Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, vol. 20, 1976 et les suivants). La catégorie grammaticale de Dobrušin est la réunion d'un ensemble A de mots et de l'ensemble des mots dominés par tous les mots de A. Les fermetures de Sestier s'avèrent un modèle plus convenable des catégories grammaticales que celui de Dobrušin. Kunze montre qu'il y a en allemand des catégories grammaticales qui admettent le modèle de Sestier, mais pas celui de Dobrušin.

L'ASPECT ALGEBRIQUE

Les idées de Kulagina, Dobrušin et Sestier ont stimulé des recherches bien au-delà des intentions de ces auteurs. La fermeture de Sestier est un opérateur de fermeture, ce qui a provoqué l'intérêt pour le rôle de tels opérateurs dans l'étude du langage. On sait maintenant quels sont les types de domination contextuelle engendrés par des opérateurs de fermeture (voir, par exemple, Dora Pană, "Opérateurs de fermeture et relations de domination dans la linguistique algébrique" (en roumain), *Studii şi Cercetări Matematice*, vol. 27, 1975, n°4, p. 461-470). La fermeture de Sestier a attiré aussi l'attention sur le rôle des connexions de Galois dans l'étude des langages. La connexion de Galois qui intervient entre phrases et contextes est devenue le point de départ d'une théorie générale où relations de domination, opérateurs de fermeture et connexions de Galois sont toujours en interaction. Le théoricien le plus profond de cette orientation est

Miroslaw Novotný (voir, par exemple, son étude "Algebraic Structures of Mathematical Linguistics", *Bull. Math. de la Société Math. de la R.S.R.*, vol. 12, 1968, n°3). Le point de vue des relations binaires comme cadre général de la linguistique algébrique analytique a été développé par J.A Šreider ("Algebra binarnyh otnošenii", annexe au volume de S. Marcus, *Teoretiko-množestvennye modeli jazykov*, Moska, Izd. Nauka, 1970, p. 300-330). Ces trois orientations, connexions de Galois, fermetures et relations binaires, ont bénéficié de l'article déjà classique de Jacques Riguet ("Relations binaires, fermetures, correspondances de Galois", *Bull. de la Soc. math. de France*).

Brian H. Mayoh ("Simple structures defined on a transitive and reflexive graph", *Revue Roum. Math. Pures appl.*, vol. 11, 1966, n° 1, p. 43-51) voit dans la théorie des graphes le meilleur cadre des modèles de catégories grammaticales, mais peut-être le chapitre mathématique le mieux stimulé par les modèles analytiques de langage est la théorie des demi-groupes libres. Certains aspects de ces modèles apparaissent, indépendamment et sans leur motivation linguistique, dans des recherches purement algébriques ; voir, par exemple, B.M. Schein ("Homomorphisms and subdirect decompositions of semi-groups", Pacific J. of Mathematics, vol. 17, 1966, n° 3, p. 529-547) et Joseph Zapletal ("Distinguishing subsets of semi-groups and groups", *Spisy Prirodov, Fak. Univ. J.E. Purkajne v Bone*, vol. 4, 1968, p. 241-252). Ladislav Nebeský ("K ponijatiu Konteksta", *Prague, Studies in Math. Linguistics*, vol. 2, 1967, p. 177-185 ; "A theorem on partitions and two linguistic applications", *Prague, Bull. of Math. Ling.*, 7, 1967, p. 52-57 ; "Language Systems", *Revue Roum. Math. pures appl.*, vol. 12, 1967) prend comme point de départ un véritable calcul des contextes. Un autre point de vue permettant d'englober la presque totalité des phénomènes analytiques de langages a été développé par M. S. Burgin et E.S. Burgina ("Poisk informaticii i mnogoznačnye razbienija v jazykah", *Kibernetika*, Kiev, 1982, n°1, p. 30-43) et repose sur le concept de "décomposition multivaluée" d'un alphabet A (ou langage L), définie comme une relation P sur A (ou sur L) et un certain ensemble F, tels que P soit contenu dans le produit cartésien $A \times F$ (ou dans $L \times F$) ; F est l'ensemble des traits de P. L'abord de Burgin et Burgina couvre aussi les phénomènes concernant le traitement automatique de l'information. Un autre exemple intéressant est la tentative de Jerzy Pogonowski ("Tolerance Spaces with applications to Linguistics", *Universytet Im. Adama Mikiewicza w Poznaniu*, seria Jezykoznawstwo Stosowane, nr. 3, Poznan, 1981, 103 pp.) d'unifier une grande variété de faits de langage à l'aide des relations de tolérance (= relations binaires réflexives et symétriques) introduites par Zeeman, dans son fameux article sur la topologie du cerveau. L'algèbre des relations de tolérance a été approfondie, dans plusieurs articles, par Bohdan Zelinka.

L'intérêt de ces recherches va d'habitude au-delà des problèmes linguistiques et engendre parfois des questions d'intérêt mathématique incontestable. Les racines de beaucoup de ces recherches se trouvent, comme on l'a déjà remarqué, dans les idées et les méthodes de la linguistique distributionnelle américaine. Le paradoxe qui apparaît ici mérite d'être signalé. En dépit des traditions de la linguistique distributionnelle et du fait que Chomsky a débuté comme étudiant et collaborateur de Harris (en l'aidant à achever la forme définitive de son livre Structural Linguistics, Chicago Univ. Press, 1961), le développement des modèles formels dans la linguistique américaine des années 1956-1986 a presque ignoré l'héritage descriptif-distributionnel, se consacrant surtout à la linguistique générative et transformationnelle. Il est vrai que les grammaires de structure de la phrase dont Chomsky parlait au commencement de sa théorie avaient comme une de leurs motivations, la formalisation de l'analyse en constituants immédiats, si importante chez des distributionnalistes comme Bloch, Harris et Wells ; comme on sait, cet épisode a été rapidement dépassé, la théorie générative transformationnelle devenant très polémique à l'adresse de la linguistique descriptive. En revanche, certains pays d'Europe dont les relations avec le structuralisme linguistique américain étaient assez faibles se sont montrés, à la fin des années cinquante et au commencement des années soixante, prêts à continuer, à un niveau supérieur de formalisation, les théories américaines des années quarante et cinquante.

LA PROJECTIVITE SYNTACTIQUE

Il s'agit d'une propriété mise en évidence par Yves Lecerf et P. Ihm (Eléments pour une grammaire générale des langues projectives, Rapport CETIS n° 1, Euratom, 1959, p. 1-19), K.E. Harper, D.G. Hays ("The use of machines in the construction of a grammar and computer program for structural analysis", *Proc. Intern. Congr. Information Processing*, UNESCO, Paris 1959), L. Hirschberg (*Le relâchement conditionnel de l'hypothèse de projectivité.* Rapport CETIS, n° 35, Euratom, 1961) et I. Lynch (*Suggestions for modification of Lecerf theory of projectivity and of his stemmas, for the purposes of their application to "non-projective" Russian sentences*, Rapport CETIS, n°35, Euratom, 1961), au cours de leurs recherches sur la traduction automatique. Il s'agit d'une contrainte rencontrée dans la structure de dépendances de la plupart des propositions d'une langue naturelle. La variante la plus connue est celle proposée par Lecerf et Ihm et affirme que la subordination (qui est la clôture transitive et réflexive de la relation de dépendance syntactique) d'un terme *a* par rapport à un autre terme *b* implique la subordination de *b* à chaque intermédiaire. Intéressante en soi, cette propriété a aussi d'autres points d'attraction. Du point de vue mathématique, elle est à l'origine d'un chapitre nouveau de la théorie des graphes, notamment de la théorie des arborescences ; Ladislav Nebeský lui a consacré un livre (*Algebraic properties of trees*, Praga, Univ. Karlova, 1969) et un grand nombre d'articles, tandis que d'autres mathématiciens comme Ju. A. Šreider ("The property of projectivity of language", *Naučna Techničeskaja Informacija*, 1964, n°8, p. 38-41), M. Mihaly ("Sur la notion de projetivité syntactique", *Cah. Ling. théor. et appl.*, 1968, p. 159-171 ; "Asupra unor tipuri de proiectivitate sintactică", *Studii si Cercetări Matematice*, vol. 26, 1974, n°2, p. 179-196), Jurgen Kunze ("The treatment of non-projective structures in English and German", *Computational Linguistics*, vol. 7, 1968, p. 67-77) ont étudié comparativement les divers types de projectivité, et Dušan Pospišl ("On a linearization of projective w-trees", *Prague Bull. of Math. Ling.*, vol. 6, 1966, p. 35-43) a donné une méthode de linéarisation des graphes projectifs. Du point de vue linguistique, on constate que la projectivité syntactique présente un intérêt qui dépasse de loin sa motivation initiale. Des linguistes comme E.V. Paducčeva, L.N. Iordanskaja dans les années soixante, John Anderson-Charles Jones ("Three theses concerning phonological representations", *J. of Linguistics*, vol. 10, 1974, n°1, p. 1-26) et John Anderson ("Serialisation dependency and the syntax of possessives in Moru", *Studia Linguistica*, vol. 33, 1979, N°1, p.1-25) ont mis en évidence l'apport de la projectivité syntactique en phonologie et dans la syntaxe des langues naturelles. Plusieurs travaux ont utilisé la non-projectivité comme critère de distorsion syntactique et de poéticité (Lucreţia Vasilescu-Şerban Gavrilă, "Modèle d'analyse automatique des types de projectivité", analyse de la poésie "Crayon" de Tudor Arghezi, *Cah. de Ling. théor. et appl.*, vol. 13, 1979, n°2, p. 613-625 ; Nicos Cosmas, "Confrontation between various types of syntactic projectivity", Rev. Roum. de Linguistique, *Cah. de Ling. théor. et appl.*, vol. 22, 1985, n°2, p. 119-120 ; Syntactic projectivity in Romanian and Greek poetry (Bacovia and Karyotakis), Rev. Roum. de Linguistique, *Cah. de Ling. theor. et appl.*, vol. 23, 1986, n°2, p. 89-94).

La projectivité syntactique est un des chapitres les plus intéressants des grammaires de dépendance ayant leur origine chez Lucien Tesnière (*Eléments de syntaxe structurale*, Paris, Klincksieck, 1959), développées ensuite par D.G. Hays ("Dependancy theory : A formalism and some observations", *Language*, vol. 40, 1964, n° 4, p. 511-525) et confrontées à d'autres types de grammaires par Maurice Gross ("On the equivalence of models of language used in the fields of mechanical translation and information retrieval", *Inform. Storage Retrieval*, vol. 2, 1964, p. 43-57). Celui-ci a montré que tous les modèles de langages utilisés dans les recherches de traduction automatique et le traitement de l'information sont équivalents à ce qu'on appelle aujourd'hui les grammaires context-free. Il s'agit d'une restriction dont la nature reste encore à éclaircir et qui évoque d'autres restrictions du type de l'hypothèse de Yngve-Miller sur le "magical number 7" (voir V.H. Yngve, "A model and an hypothesis for language structure", *Proc. Am. Phil. Soc.* 104, 1960, p. 444-466 ; "The depth hypothesis", *in Structure of Language and its Mathematical Aspects, Proc. Symp. Applied Math.*, 1961, p. 130-138 ; George A. Miller, "Human memory and the storage of information", *IRE Trans. Inform.*

Theory, vol. 2, 1965, n° 3, p. 129-137) ou de l'hypothèse de l'unicité du régent due à Tesnière.

A la fin d'un paragraphe où l'on a discuté des idées nées surtout des recherches sur la traduction automatique, il faut dire quelques mots des grammaires formelles de Y. Bar-Hillel (voir leur version formalisée chez Y. Bar-Hillel, C. Gaifman et E. Shamir, "On categorial and phrase structure grammars", *Bull. Res. Council Israel*, Sec. F.vol. 9, 1960, p. 1-16) et du calcul syntactique de Joachim Lambek ("On the calculus of syntactic types", in *Structure of Language and its Math. Aspects, Proc. 12th Symp. Appl. Math.*, American Math. Soc., 1961, p. 166-178). Stimulées par les préoccupations sur la traduction automatique de l'époque, ces recherches développent une idée du logicien polonais Kasimir Ajduliewicz ("Die Syntaktische Konnexität", *Studia philosophica* I, 1935, p. 1-27). Les mots et les groupes de mots sont associés, chez Bar-Hillel, à des catégories qui peuvent être de deux types, catégories de base et opérateurs. Chez Lambek, on travaille avec des types syntactiques. Le calcul de Lambek a été formalisé à l'aide de la théorie des catégories (Ana Burghelea, "Syntactic types and Theory of Categories", *Rev. Roumaine Math. pures appl.*, 1968). D'ailleurs, la théorie des catégories devenait, dès le commencement des années soixante, un langage fréquent dans les recherches d'informatique et de linguistique mathématique (voir, par exemple, les articles publiés dans les années soixante par Jacques Riguet et Jacques Roubaud).

Toutes ces recherches sur les grammaires de dépendance, sur les grammaires catégorielles et sur le calcul des types syntactiques attirent l'attention sur la nature générative des mécanismes de description syntactique, nature qui devait devenir manifeste avec Noam Chomsky.

LA LINGUISTIQUE GENERATIVE ET TRANSFORMATIONNELLE

La linguistique structurale, cette antichambre de la linguistique mathématique, avait apporté à la linguistique des succès incontestables. Le *Cours de Linguistique Générale* de Ferdinand de Saussure (3ème ed., Paris, 1931), les *Prolegomena to a theory of language* (Baltimore, 1953) de L. Hjelmslev, les *Principes de Phonologie* (traduits de l'allemand par J. Cantineau, Paris, 1957) de N.S. Troubetzkoy, les *Eléments de linguistique générale* (Paris, 1960) de André Martinet, les travaux de Roman Jakobson avaient conduit à une situation où, comme ce dernier auteur l'avait remarqué, "la linguistique est devenue la mathématique des sciences humanistes et sociales". Méthodologiquement, la linguistique se situait, en effet, au point le plus élevé. Toutes les autres disciplines sociales apprenaient de la linguistique comment choisir leurs unités et relations de base. L'anthropologie de Claude Lévi-Strauss illustre bien cette orientation. Roland Barthes avait raison de proclamer que méthodologiquement la sémiologie était subordonnée à la linguistique. Un article de l'époque, "Linguistic as a Pilot Science", de J.H. Greenberg, en est aussi une illustration.

Mais ce succès social de la linguistique allait être dépassé par la linguistique générative et transformationnelle lancée par Noam Chomsky dans la période 1956-64 et connaissant plusieurs phases successives. Un syntagme comme *révolution chomskienne* devient fréquent chez des auteurs de prestige comme John Lyons. Maintenant, on a la perspective nécessaire à une évaluation raisonnable. La publicité immense qu'on a fait à cette orientation nous dispense de rappeler ses notions de base. Mais il est important de souligner que les origines et les racines de la linguistique générative sont multiples. Il y a d'abord une source linguistique, expliquée d'une façon détaillée par Noam Chomsky (*Syntactic Structures*, Gravenhage, Mouton, 1957). Il y a ensuite une source logique qui, partant de la théorie des systèmes formels de Hilbert, va jusqu'aux systèmes combinatoires présentés comme préambule à la théorie des langages formels par Maurice Gross et André Lentin (*Notions sur les grammaires formelles*, Paris, Gauthier-Villars, 1970). A ces deux motivations initiales, dont seulement la première est explicite chez Chomsky, s'ajoutent, lors de l'évolution ultérieure, quelques motivations supplémentaires, mais

de la même importance. La linguistique générative et transformationnelle se réclame d'une philosophie qui la transforme en un chapitre de la psychologie cognitive, se constituant comme une réplique non seulement à Skinner (Review of "Verbal Behaviour", *Language*, vol. 35, 1959, p. 26-58), comme Chomsky le montrait dans son fameux article, mais aussi, comme on allait voir plus tard, à Piaget (voir *Théorie du Langage, théories de l'apprentissage*, éditeur Massimo Piattelli-Palmarini, Paris, Editions du Seuil, 1979). Le théorique et l'empirique, l'inné et l'acquis se trouvaient en confrontation. Dans une monographie de synthèse, M.L. Moreau et M. Richelle (*L'acquisition du langage*, Bruxelles, Mardaga, 1981) considèrent que Chomsky a échoué dans sa tentative d'expliquer les mécanismes que l'enfant utilise pour apprendre le langage. Cette position extrémiste a été suivie par la réplique de Marie Labelle ("Pertinence du modèle transformationnel en linguistique appliquée", *Cahiers Linguistiques d'Ottawa*, n°13, 1985, p.1-36), qui, en dépit du désaccord avec le modèle transformationnel accentue le rôle fondamental de la théorie chomskienne dans l'élaboration de la distinction compétence-performance, dans la compréhension du rôle des règles dans le fonctionnement de la langue, dans la formulation nouvelle du problème de la nature du langage humain et des universaux linguistiques. Une synthèse antérieure de Judith Grenne (Psycholinguistics, Chomsky and Psychology, Harmondsworth, Penguin Books, 1972) montraient que seulement une partie des expériences a confirmé l'utilisation, par les enfants, de règles du type chomskien. La discussion reste, sans doute, ouverte, mais il faut rappeler qu'un modèle scientifique n'est pas obligatoirement un modèle de simulation. Le fameux exemple de la boîte noire des cybernéticiens est valable ici aussi. Même si la grammaire chomskienne ne reproduit pas le processus biopsychique de l'apprentissage, elle reste intéressante si elle est capable de nous dire quelque chose sur le résultat de ce processus.

MODELES MATHEMATIQUES DE L'APPRENTISSAGE

A la lumière de ces clarifications, examinons quelles sont les conséquences de la théorie chomskienne en ce qui concerne les modèles mathématiques de l'apprentissage. Les modèles stochastiques habituels, bien que très développés du point de vue de la technique mathématique utilisée (processus de Markov, processus à liaisons complètes de Onicescu, Mihoc et Iosifescu), sont tributaires d'une représentation empirique du processus de l'apprentissage, vu exclusivement comme une interaction entre stimulus et réponses (les synthèses les plus récentes de ce point de vue sont effectuées par M. Iosifescu - R. Theodorescu, Random processes of learning, Berlin, Springer, 1969, et par M.F. Normann, Markov processes and learning models, New York, Academic Press, 1972). Mais ces modèles ne sont pas capables d'expliquer l'apprentissage des concepts. Un tel apprentissage exige un processus génératif infini, que seulement l'orientation chomskienne est capable d'éclaircir. C'est ce qui se produit dans les recherches de Y. Uesaka, T. Aizawa, T. Ebara, K. Ozeki ("A theory of learnability", *Kybernetik*, vol. 3, 1973, p. 123-131), T. Aizawa, T. Ebara, K. Ozeki, Y. Uesaka ("Sur l'espace topologique lié à une nouvelle théorie de l'apprentissage", Kybernetik, vol. 14, 1974, p. 144-149).

Un objet à apprendre est représenté ici par une fonction $f : N \to N$, donc par une suite infinie de paires ordonnées de nombres entiers positifs, où le couple $(i, f(i))$ signifie que $f(i)$ est la réponse donnée au stimulus i. Par exemple, apprendre le concept de nombre impair c'est apprendre la suite infinie des couples (1,1), (2,3),(3,5),....,$(n+1, 2n+1)$,.... .

L'activité empirique seule ne peut aller au-delà d'un nombre fini de tels couples, seule l'articulation de l'empirique et du réflexif peut fournir l'apprentissage de l'idée générale de nombre impair. Une section fini d de la suite f est la virtualité d'une infinité d'objets à apprendre, donc de suites qui commencent avec d. Désignons par $\pi(d)$ l'ensemble de ces suites compatibles avec d. Si F est la totalité des objets à apprendre et si S et T sont des parties de F, alors on dit que l'objet f (situé dans F) peut être appris avec la connaissance "f

dans T" et sous l'information innée "f dans S" s'il existe une section finie d de f, telle que $f \in S \cap \pi(d) \subset T$.

On définit sur F une topologie τ dont la base des ouverts est $\{\pi(d)$; d parcourant toutes les suites finies de paires ordonnées de nombres entiers positifs$\}$. On peut apprendre la connaissance "$f \in T$" sous la présupposition "$f \in S$" si et seulement si $T \cap S$ est un voisinage de f dans la topologie de (F, τ) relativisée à S. D'autre part, la distance entre deux objets à apprendre f et g peut être mesurée par :

$$\delta_\alpha (f,g) = \sum_{x=1}^{\infty} \frac{1}{\alpha X} \sigma (f(X), g(X)), \text{ où } \sigma (X,Y) = \begin{cases} 0, X = Y \\ 1, X \neq Y \end{cases}$$

où $\alpha > 1$. L'espace (F, δ_α) est complet et homéomorphe à (F, τ). La relation avec la théorie des langages formels s'obtient à l'aide d'une construction due à V.G. Bodnarčuk ("Metričeskoe prostranstvo sobytii", *Kibernitika*, Kiev, vol. 1, 1965, n° 1, p. 24-27). Si A est un alphabet fini et A^* est le monoïde libre généré par A, soit l'ensemble $\mathfrak{P}(A^*)$ des parties de A^* organisé comme espace vectoriel normé sur le corps $K_2 = \{0,1\}$, où pour deux langages E,G sur A on met $E+G = E \Delta G$ (la différence symétrique de E et de G), $O.E = \varnothing$, $1.E = E$. En désignant par $\lambda(E)$ le minimum des longueurs des mots de E, on définit la norme $\|E\|$ de E par

$\|E\| = 2^{-\lambda(E)}$, $\|\varnothing\| = 0$.

Cette norme induit une distance $d(E,G) = \|E - G\|$. L'espace métrique $(\mathfrak{P}(A), d)$ ainsi obtenu est un espace complet, séparable et compact ; il est appelé l'espace de Bodnarchuk. On démontre que cet espace est homéomorphe à une partie de l'espace (F, τ) de l'apprentissage, tandis que l'espace (F, τ) est homéomorphe à une partie de l'espace $(\mathfrak{P}(A^*), d)$.

LES LANGAGES DE PROGRAMMATION ENTRENT EN SCENE

L'exemple ci-dessus ne fait qu'anticiper, par la structure de langage formel envisagée, une troisième motivation des grammaires chomskiennes : leur emploi dans l'étude de la syntaxe des langages de programmation. Le langage ALGOL apparaît en 1960 et il est le premier dont la définition utilise un système formel, la notation de Backus. L'année suivante, S. Ginsburg et H.G. Rice (*Two families of languages related to ALGOL*, SDC Technical Report, January 1961) démontrent que la notation de Backus est équivalente aux grammaires context-free. C'est un moment décisif, qui tourne l'attention des informaticiens vers les langages et les grammaires de Chomsky. Bien qu'on ait remarqué dès le début que l'ALGOL, dans son ensemble, n'est pas context-free (les conditions sémantiques assurant la correction des programmes ALGOL ne pouvant pas être modelées à l'aide d'une grammaire context-free), les grammaires context-free sont restées jusqu'à aujourd'hui d'une utilisation permanente, en ce qui concerne la définition et la manipulation des langages de programmation. Comme le dit S. Ginsburg : "Nous visons ou nous mourons en fonction des langages contex-free". La dénomination de langages algébriques donnée par certains auteurs français à ces langages est extrêmement heureuse, mais cette métaphore (avec la métaphore "langages rationnels" pour les langages réguliers) est aussi un programme de recherche concernant toutes les motivations possibles de l'analogie langage-nombre réel.

Pour les informaticiens, la théorie de Chomsky n'est plus la théorie générative-transformationnelle, mais la théorie des langages formels. Les linguistes ne se sont pas attardés

sur les langages algébriques, dès qu'a été constatée leur inadéquation aux langues naturelles. Mais récemment, on est revenu sur cette question et on s'est demandé de nouveau si les langues naturelles étaient context-free (voir les articles sur ce sujet publiés dans les dernières années dans Computational Linguistics). Les problèmes d'adéquation semblent plus compliqués que ceux concernant l'exactité formelle. En ce qui concerne les grammaires context-sensitives, elles ont aussi été abandonnées, l'intérêt principal se dirigeant vers les transformations. On a eu besoin de plusieurs années pour remarquer que la théorie des langages formels n'était plus orientée vers la linguistique, mais vers l'informatique. Cette déroute est visible même pour un regard superficiel. Le livre de John P. Kimball (*The formal theory of grammar*, New Jersey, Prentice Hall, 1973) est une présentation formelle attractive des idées de Chomsky, mais il ne s'agit pas du côté formel ; on peut mettre en contraste ce livre avec un autre, paru la même année et ayant un titre similaire : Arto Salomaa, *Formal languages*, New York, Academic Press, 1973. Ce dernier est vraiment formel, mais sans relation avec les langues naturelles. Le livre de Robert Wall, (*Introduction to Mathematical Linguistics*, New Jersey, Prentice Hall, 1972) présente les éléments de la théorie des ensembles, le calcul propositionnel, le calcul des prédicats, les relations binaires, quelques éléments de la théorie des graphes, des structures algébriques, les définitions récursives, l'axiomatique, quelques éléments de la théorie des langages formels et de la théorie des graphes, des structures algébriques, les définitions récursives, l'axiomatique, quelques éléments de la théorie des langages formels et de la théorie des automates et une présentation sommaire des grammaires transformationnelles. Evidemment, il ne s'agit pas d'un livre de linguistique mathématique, mais d'un livre de mathématique pour les linguistes.

La déroute est présente aussi dans les journaux internationaux de références tels que *Mathematical Reviews* et *Zentralblatt für Mathematik*, où la sous-section *Linguistics* de la section *Computer Science* comprend 90% des références sur des travaux concernant la théorie des langages formels, donc une théorie dont les relations avec la linguistique sont presque rompues, en tout cas très faibles. Un exemple de tout autre facture est le livre de M.C. Barbault et J.P. Desclés (*Transformations formelles et théories linguistiques*, Doc. de Linguistique quantitative, 11, Association Jean Favard pour le développement de la linguistique quantitative, Univ. de Paris VI, 1972), où la rigueur et l'adéquation se combinent pour réaliser la meilleure présentation des grammaires transformationnelles, avec les motivations linguistiques nécessaires.

LES LANGAGES ARTIFICIELS, OBJETS DE LA LINGUISTIQUE ?

Mais la déroute dont on a parlé ci-dessus peut être le symptôme d'une distribution nouvelle des problèmes entre divers domaines. Traditionnellement, la linguistique est la science dont l'objet est constitué par les langues naturelles. Mais l'évolution récente a obligé de plus en plus la linguistique à inclure aussi dans ses intérêts les langages artificiels. La linguistique s'oriente vers le langage humain dans toutes ses manifestations. Le langage cosmique de Hans Freudenthal présente un intérêt linguistique immense, devenant un laboratoire où sont testées les possibilités d'un langage qui veut devenir son propre métalangage. Le théorème d'incomplétitude de Gödel offre l'expérience d'un métalangage qui, voulant prendre la distance par rapport au langage-objet de l'arithmétique finit justement dans la situation opposée : la coïncidence entre le métalangage et le langage objet. N'oublions pas non plus que la numération de Gödel évoque le paradoxe de Richard, dont la nature linguistique est évidente. Richard Montague traite l'anglais comme un langage formel et ne voit pas une différence essentielle entre le naturel et le formel, tous les deux admettant les mêmes mathématiques (R. Montague, "Universal grammar", *Theoria*, vol. 36, 1970, p. 373-398). Noam Chomsky soutient une idée similaire, refusant même l'utilisation de l'expression "langue naturelle" ; il parle des langages humains, comme systèmes sémiotiques englobant le naturel et l'artificiel à la fois ("Human language and other semiotic systems", *Semiotica*, 1979, n°1-2). Les langages de programmation sont-ils des langages artificiels formels ? Une réponse du type oui ou non à cette question n'est plus possible. On croyait parfois que les langues naturelles se caractérisent par le statut flou (non-rigoureux des séquences

bien formées de mots (ou morphèmes). On ne connaît pas de condition nécessaire et suffisante pour qu'une suite de mots soit correcte, en contraste avec ce qui se passe, par exemple, dans le langage (context-free) du calcul propositionnel à un nombre fini de variables. Mais dans les langages de programmation assez évolués les suites bien formées ont un caractère flou ; on ne connaît pas de condition nécessaire et suffisante pour qu'une suite d'instructions forment un programme d'ordinateur. Une autre propriété qui semblait caractériser les langues naturelles est la présence d'une sémantique intégrative, opposée à la sémantique additive des langages artificiels (où la signification d'une chaîne s'obtient par concaténation des significations des termes de la chaîne). En fait, on constate que les "expressions idiomatiques" abondent aussi dans les langages de programmation (voir C. Calude, "Sur quelques arguments pour le caractère non-formel des langages de programmation", *Cah. de Ling. theor. et appl.*, vol. 13, 1976, p. 257-264) et dans le langage mathématique (il est aisé de voir que la notation de l'intégrale d'une fonction sur un certain intervalle est une "expression idiomatique").

Des auteurs comme Hans Freudenthal ("Cosmic languages", dans *Current Trends in Linguistics*, Th. A. Sebeok ed., vol. 12, La Haye, Mouton, 1974, p. 1019-1042) attribuent aux langages artificiels, la caractéristique d'avoir une destination très spéciale, en contraste avec l'utilisation universelle des langues naturelles. Mais les langages de programmation, en contraste avec leur destination initiale, se sont montrés très utiles dans certaines études de logique et dans l'investigation de la sémantique des langues naturelles. D'autre part, les langues naturelles ne peuvent satisfaire les exigences de rigueur de la science actuelle.

Selon André Martinet, la double articulation est spécifique aux langues naturelles (voir *Le langage*, sous la direction d'André Martinet, Paris, Gallimard, 1968, p. 31-33). Mais le langage génétique (ce langage n'est pas artificiel, mais son modèle mathématique, envisagé ici, est une construction artificielle ; voir S. Marcus, "Linguistic structures and generative devices in molecular genetics", *Cah. de Ling. théor. et appl.*, vol. 11, 1974, n° 2, p. 77-104) semble avoir une double articulation, le rôle des morphèmes étant rempli ici par les codons, tandis que les bases nucléotides sont les phonèmes génétiques.

La présence de la double articulation dans certains langages artificiels reste à être discutée en fonction de ce qu'on considère comme essentiel. Si l'on voit dans la double articulation surtout l'existence d'un petit nombre d'éléments dépourvus de signification, mais capables de conduire, par combinaison syntagmatique, à un très grand nombre d'éléments doués de signification, alors on ne peut refuser, à certains langages artificiels, le privilège de la double articulation. S'agissant d'un phénomène surtout combinatoire, il est plausible d'avoir ici les effets de certaines restrictions générales concernant les capacités humaines de traitement de l'information. En ce qui concerne la possibilité de l'autoréférence, attribuée par Paul Garvin exclusivement aux langues naturelles, il faut rappeler les recherches de R. Smullyan ("Languages in which self-reference is possible", J. of Symbolic Logic, vol. 22, 1957, p. 55-67) concernant justement la possibilité des langages logiques autoréférentiels.

Par utilisation, les langages artificiels deviennent de plus en plus humains et leur structure s'approche de la structure des langues naturelles.

Mais la question qui forme le titre de ce paragraphe admet une question symétrique : les langues naturelles, objet de l'informatique. On sait bien que le traitement automatique des langues naturelles forme un chapitre important de l'Intelligence artificielle. Maintenant, il y a des projets concernant la programmation en langue naturelle. Edsger W. Dijkstra (*How do we tell truths that might hurt ?* , ACM Sigplan Notices, vol. 17, 1982, n°5, p. 13-15) affirme que ces projets sont destinés à l'échec. Cela peut étonner, à un moment où les préoccupations concernant le dialogue en langue naturelle avec l'ordinateur connaissent des succès. Mais le diagnostic de Dijkstra est également dur en ce qui concerne les langages de programmation : FORTRAN, le "désordre enfantin" ; PL1, "une maladie fatale - surtout un problème plus qu'une solution" ; BASIC, "mutile la pensée", etc. .

UNIVERSALITE DES STRUCTURES DE LANGAGE

L'un des plus importants aspects de la formalisation en linguistique est le rôle de la linguistique formelle comme catalyseur de contacts les plus imprévisibles et, à la fois, profonds. A vrai dire, c'est la mathématique qui est le médiateur de collaborations entre les structures de langages et les disciplines naturelles ou sociales les plus diverses. Les métaphores linguistiques utilisées en génétique moléculaire se sont transformées en modèles mathématiques ; la syntaxe génétique est approximée par une grammaire context-free dont le langage des dérivations (qui est une approximation meilleure) est seulement context-sensitive. La praxiologie se prévaut couramment des grammaires formelles (voir Maria Nowakowska, *Language of action and language of motivation*, La Haye, Mouton, 1973) et on peut écrire, comme exercices plus ou moins simples, la grammaire d'un coup de téléphone ou la grammaire du coiffeur. La chimie (surtout la chimie organique) emploie souvent les grammaires formelles, comme on peut voir dans des journaux tels que *MATCH* (*Mathematical Chemistry*) et *Chemical Documentation.*

Même en mathématique les applications sont assez vieilles. On se rappelle qu'il y a vingt ans Jean Friant montrait que le langage $\{a^n\}$ (n nombre premier) est context-sensitive, mais pas context-free et que le grand théorème de Fermat peut être formulé comme un problème de décision en théorie des langages formels. Le livre de M. Gross et André Lentin *Notions sur les grammaires formelles* (Paris, Gauthier-Villars,) contient une section consacrée aux applications des grammaires formelles dans la géométrie combinatoire. Une foule d'autres applications doivent être mentionnées. Gh. Paun (*Generative grammars of economic processes*, Foundations of control Engineering, vol. 1, 1975) est l'auteur d'un livre *Mécanismes génératifs des processus économiques* (en roumain, Bucarest, Editura Tehnică, 1980). Il y a beaucoup d'applications en ethnologie, folklore, narrativité, jeux énigmatiques, relations internationales, théâtre ; même le jeu de tennis et la cuisine ont été analysés sous l'aspect génératif. Dans toutes ces recherches, il s'agit d'un nouveau type de lecture des phénomènes, que l'on pourrait appeler *lecture générative*, une manière d'approximation du fini par l'infini, un approfondissement du combinatoire par le génératif. Les répétitions ne sont plus envisagées sous l'aspect quantitatif de la fréquence, mais sous l'aspect qualitatif du type de répétition et de la nature des éléments itératifs. Le paradoxe de l'induction, mis en évidence par C. Hempel et Nelson Goodman, a son équivalent ici. Il y a, comme on le voit, une tendance du langage à imposer ses structures à une foule d'autres phénomènes et processus. Le langage fonctionne ici par rapport à d'autres systèmes sémiotiques, comme l'espace euclidien par rapport à tout autre espace que la science propose ; il est un terme universel de référence.