L'ANALYSE FORMELLE DES LANGUES NATURELLES

ÉCOLE PRATIQUE DES HAUTES ÉTUDES - SORBONNE
SIXIÈME SECTION: SCIENCES ÉCONOMIQUES ET SOCIALES

MATHÉMATIQUES ET SCIENCES DE L'HOMME VIII

MOUTON/GAUTHIER-VILLARS

NOAM CHOMSKY / GEORGE A. MILLER

L'ANALYSE FORMELLE DES LANGUES NATURELLES

(Introduction to the Formal Analysis of Natural Languages)

Traduction de

PH. RICHARD / N. RUWET

MOUTON/GAUTHIER-VILLARS

L'analyse formelle des langues naturelles est une traduction des chapitres 11 et 12 du Volume II du *Handbook of Mathematical Psychology*

Traduction française publiée en 1968

Gauthier-Villars
Mouton
École Pratique des Hautes Études

Diffusion en France: Dunod, 92, rue Bonaparte - Paris 6

2e tirage 1971

Printed in the Netherlands

TABLE DES MATIÈRES

I INTRODUCTION À L'ANALYSE FORMELLE DES LANGUES NATURELLES 1

1. Les limites de la discussion 3
2. Quelques aspects algébriques du codage 9
3. Quelques concepts linguistiques de base 15
4. Une classe simple de grammaires génératives 26
5. Grammaires transformationnelles 31
   5.1 Quelques insuffisances des grammaires de constituants . 32
   5.2 Caractérisation des transformations grammaticales . . . 36
   5.3 La structure de constituants des suites transformées . . 40
6. Structure phonique 43
   6.1 Le rôle de la composante phonologique 44
   6.2 Phones et phonèmes 45
   6.3 Les conditions d'invariance et de linéarité 48
   6.4 Quelques règles phonologiques 51

II PROPRIÉTÉS FORMELLES DES GRAMMAIRES 59

1. Les automates abstraits 61
   1.1 Représentation de la compétence linguistique 61
   1.2 Automates strictement finis 66
   1.3 Automates linéaires bornés 75
   1.4 Automates à pile 76
   1.5 Transducteurs finis 84
   1.6 Transduction et automate à pile 86
   1.7 Autres types d'automates infinis-limités 91
   1.8 Machines de Turing 92
   1.9 Algorithmes et décidabilité 93
2. Systèmes de réécriture non limités 97

3. Grammaires dépendantes du contexte 101
4. Grammaires indépendantes du contexte 109
4.1 Classes spéciales de grammaires indépendantes du contexte 112
4.2 Grammaires indépendantes du contexte et automates infinis limités 115
4.3 Propriétés de fermeture 126
4.4 Propriétés indécidables des grammaires indépendantes du contexte 128
4.5 Ambiguité structurale 135
4.6 Grammaires indépendantes du contexte et automates finis 138
4.7 Définition possible des langages par des systèmes d'équations 151
4.8 Langages de programmation 161
5. Les grammaires catégorielles 162
Bibliographie 169

I

INTRODUCTION À L'ANALYSE FORMELLE DES LANGUES NATURELLES

Le langage et la communication jouent dans la vie humaine un rôle particulier et fondamental; savants et hommes de sciences de toutes origines ont médité et discuté ces questions. Par contre, les psychologues ont apporté une faible contribution à cet effort général. Aussi, afin de donner une vue comparée du problème plus vaste auquel se heurte un psychologue mathématicien quand il dirige son attention sur les questions de comportement verbal, cette étude devra sortir des limites traditionnelles de la psychologie.

Le fait fondamental en face duquel on se trouve placé dans toute recherche sur le langage et le comportement linguistique est le suivant: tout individu est capable de comprendre un nombre immense de phrases exprimées dans sa langue maternelle, phrases qu'il entend pour la première fois. Il a la possibilité, au moment voulu, de s'exprimer verbalement par de nouvelles expressions que d'autres possesseurs de cette langue vont comprendre de la même façon. On se posera alors les questions fondamentales suivantes:

1. Quelle est la nature exacte de cette aptitude?
2. Comment est-elle mise en œuvre?
3. Comment se constitue-t-elle dans l'individu?

On a essayé, de diverses façons, de formuler des questions de ce genre, sous une forme précise et explicite, et de construire des modèles représentant certains aspects de ces réalisations du sujet parlant (*native speaker*). Une fois construits des modèles suffisamment simples, il devient possible d'entreprendre certaines études purement abstraites portant sur leur caractère intrinsèque et leurs propriétés générales. De telles études en sont encore à leurs premiers balbutiements; rares sont les aspects du langage ou de la communication suffisamment formalisés pour que des recherches

de ce genre soient seulement concevables. Néanmoins, les résultats prometteurs sont de plus en plus nombreux. Nous donnerons ici un aperçu de certains de ces résultats et chercherons à montrer comment de telles études peuvent contribuer à notre compréhension de la nature et des fonctions du langage.

La première des trois questions fondamentales que nous nous poserons concerne la nature du langage lui-même. Afin de pouvoir y répondre, il nous faut rendre explicite la structure sous jacente commune à tous les langages naturels. L'étude de ce problème a ses origines en logique et en linguistique; dans les dernières années, l'attention s'est concentrée sur le concept, qui est d'importance critique, de grammaire. Exposer ce travail dans le présent traité s'explique par le désir de rendre les psychologues conscients, de façon plus réaliste, de ce que réalise un individu quand il a appris à parler et à comprendre un langage naturel. L'association de réponses vocales à des stimuli visuels — trait sur lequel s'est concentrée l'attention de nombreux psychologues — ne représente qu'un aspect limité du processus d'ensemble de l'apprentissage d'une langue.

Pour répondre à notre seconde question, il faut essayer de donner une caractérisation formelle, un modèle, de l'utilisation des langues naturelles. Les psychologues, dont on aurait escompté qu'ils aborderaient cette question dans le cadre de leur étude générale du comportement, n'ont jusqu'à présent apporté que des réponses extrêmement schématiques (et souvent peu plausibles). Quelques idées valables sur ce sujet sont venues des ingénieurs de la communication; leurs implications psychologiques étaient relativement claires et on les a rapidement assimilées. Cependant, les concepts venus des ingénieurs étaient essentiellement d'ordre statistique; ils ont peu de rapports avec ce que l'on sait de la structure interne du langage.

En présentant les questions 1 et 2 comme distinctes, nous rejetons de façon explicite l'opinion courante selon laquelle un langage n'est rien de plus qu'un ensemble de réponses verbales. Dire que telle règle de grammaire particulière vaut pour telle langue naturelle ne veut pas dire que les gens qui parlent cette langue suivront la règle systématiquement. Définir le langage est une chose. Caractériser celui qui l'emploie en est une autre. Les deux problèmes sont évidemment reliés l'un à l'autre, mais ne sont pas identiques.

Notre troisième question n'est pas moins importante que les deux premières. Jusqu'ici cependant, on a obtenu beaucoup moins de résultats en tentant de la formuler de façon telle qu'elle puisse servir de base à une recherche abstraite. L'étude de ce qui se passe quand un enfant commence

à parler dépasse les possibilités de nos modèles mathématiques. Mentionnons seulement l'aspect génétique et regrettons que ce problème soit relativement peu évoqué dans les pages suivantes.

1. LES LIMITES DE LA DISCUSSION

L'étude mathématique du langage et de la communication est un vaste sujet. Les intentions qui sont à l'origine de cet article nous obligent à nous limiter; l'exposé des restrictions que nous nous sommes imposées facilitera peut-être l'orientation du lecteur.

Nous nous sommes bornés à l'étude générale de ce qu'on appelle les langages naturels. Il existe naturellement de nombreux langages formels, développés par les logiciens et les mathématiciens: l'étude de tels langages est une des préoccupations centrales de la logique moderne. Dans les pages qui suivent, néanmoins, nous nous sommes efforcés de nous restreindre à l'étude formelle des langages naturels et nous ignorons dans une large mesure l'étude des langages formels. Il est quelquefois commode d'utiliser des langages miniatures, artificiels, dans le but d'illustrer une propriété particulière, dans un contexte simplifié. En ce sens, les langages mis au point par les spécialistes des machines à calculer dans un but de programmation, présentent souvent un intérêt particulier. Néanmoins, nos recherches seront axées, ici, sur les langages naturels.

Nous avons négligé, de plus, toute étude importante des systèmes continus. Le signal acoustique produit par un locuteur, est une fonction continue du temps et on la représente ordinairement par la somme d'une série de Fourier. La représentation de Fourier est particulièrement commode, quand on étudie l'effet des transformations linéaires continues (filtres). Heureusement, les mathématiciens et les ingénieurs des communications ont fréquemment exploré ce domaine important et l'ont traité à fond. (Son absence ici a donc peu d'importance).

Les systèmes de communications peuvent être envisagés comme discontinus par suite de l'existence de ce que les spécialistes des moyens de communication ont parfois appelé un critère de fidélité (Shannon, 1949). Un critère de fidélité détermine la manière d'obtenir une partition de l'ensemble de tous les signaux possibles pendant une durée de temps finie, en sous-ensembles de signaux équivalents — équivalents pour celui qui reçoit ces signaux. Un système de communications peut bien transmettre des signaux continus avec précision, mais si le récepteur ne peut (ou ne veut pas) faire attention à la subtilité des distinctions qu'autorise

le système, il ne sert à rien à la liaison d'être fidèle. Ainsi c'est le récepteur qui établit le critère d'acceptabilité du système. Plus élevé est ce critère, plus grand est le nombre de sous-ensembles distincts de signaux que le système de communications est capable de distinguer et de transmettre.

Le récepteur que nous désirons étudier est naturellement l'homme en tant qu'auditeur. Le critère de fidélité dépend de ses possibilités, de sa formation, de ses intérêts. Par suite, à partir de ses seuils différentiels de perception, nous pouvons établir un ensemble fini de catégories que nous utilisons comme symboles discontinus. Ces ensembles peuvent être des alphabets, des syllabaires, ou des vocabulaires; les éléments discontinus de ces ensembles sont les atomes indivisibles à partir desquels de plus longs messages peuvent être construits. La perception par un auditeur de ces unités discontinues pose naturellement un important problème psychologique; les aspects psycho-physiques formels des problèmes posés par la détection et la reconnaissance de tels signaux ont été discutés ailleurs[1] et nous ne les exposerons pas de nouveau ici. Cependant, on mentionnera rapidement, au § 6 (consacré à la structure phonique), un certain nombre de considérations qui sont peut-être spécifiques de la perception de la parole. Voir aussi Miller et Chomsky (1963), où on étudie de quelle manière la connaissance de la grammaire pourrait servir à organiser notre perception de la parole. Comme nous le verrons, la description précise d'un critère de fidélité de l'auditeur humain, relativement à la parole, est une chose complexe; mais pour le moment le point important est que les sujets parlant divisent les sons de la parole en sous-ensembles équivalents. Une notation discontinue est donc justifiée.

En outre, dans le domaine des systèmes discontinus, nous nous limiterons aux systèmes de *concaténation*, à leurs structures algébriques plus élaborées et à leurs inter-relations. En particulier, nous représentons le flux de la parole comme une suite d'atomes discrets qui sont juxtaposés immédiatement l'un après l'autre, c'est-à-dire concaténés. Aussi simple que cette limitation puisse sembler, elle implique un certain nombre de conséquences qu'il est intéressant de noter.

Soit L l'ensemble de toutes les séquences finies (y compris la séquence de longueur 0) pouvant être formé à partir des éléments d'un quelconque ensemble arbitraire fini V. Soient maintenant, $\phi, \chi \in L$. Si $\phi^\frown\chi$ représente le résultat de la concaténation des éléments ϕ et χ de telle sorte qu'ils forment une nouvelle séquence ψ, alors ψ appartient à L; c'est-à-dire que L est fermé sous l'opération binaire de concaténation. Par suite, la concaténation est associative,

1. *Handbook of Mathematical Psychology*, vol. I, ch. 3.

$$(\phi^\frown\chi)^\frown\psi = \phi^\frown(\chi^\frown\psi)$$

et la suite vide joue le rôle d'un élément neutre unique. Un ensemble qui comprend un élément neutre et est fermé sur une loi associative de composition est appelé un *monoïde*. Un monoïde satisfaisant 3 des 4 lois d'un groupe est quelquefois appelé *semi-groupe*. Un *groupe* est un monoïde dont tous les éléments ont des inverses.

C'est nécessairement en recourant à l'opération associative de concaténation que nous construirons les énoncés parlés. Cependant, il faut bien prendre soin de formuler ce point convenablement. Considérons, par exemple, la phrase anglaise ambiguë: *They are flying planes*, qui représente en fait deux phrases différentes:

$$They^\frown(are^\frown(flying^\frown planes)) \qquad (1a)$$
$$They^\frown((are^\frown flying)^\frown planes). \qquad (1b)$$

Si nous ne pensons qu'à l'orthographe ou à la prononciation des mots, l'exemple 1*a* est identique à l'exemple 1*b* et la concaténation simple n'offre pas de difficulté. Mais si nous pensons à la structure grammaticale ou à la signification, alors les exemples 1*a* et 1*b* présentent des différences significatives que n'indiquent habituellement ni la phonétique nil'écriture.

Les linguistes, habituellement, traitent de tels problèmes en déclarant qu'un langage naturel présente divers niveaux distincts. Dans les chapitres suivants, nous considérons chaque niveau comme un système de concaténation séparé ayant ses propres éléments et ses propres lois. La structure d'un niveau inférieur est spécifiée par la manière selon laquelle les éléments de ce niveau sont liés au niveau immédiatement supérieur. Dans le but de préserver le loi d'associativité, par suite, nous introduisons divers systèmes de concaténation et étudions les relations qui existent entre eux.

Considérons deux opérations différentes, qui peuvent être réalisées sur un texte écrit. La première opération applique une séquence de caractères écrits dans une séquence de signaux acoustiques: nous l'appellerons prononciation. Les prononciations des segments d'un message sont (approximativement) les segments de la prononciation de ce message. Ainsi la prononciation possède, relativement à ce message, une sorte de linéarité (cf. Sec. 6.3)

$$\text{pron}\,(x)^\frown\text{pron}\,(y) = \text{pron}\,(x^\frown y). \qquad (2)$$

Bien que l'équation 2 ne soit pas exacte (par exemple, elle ignore l'intonation et les transitions articulatoires entre segments successifs) elle est plus

proche de la vérité que la déclaration correspondante relative à l'opération suivante.

Cette opération applique la séquence des symboles dans une représentation de sa signification subjective; nous l'appellerons compréhension. Il est, pourtant, rarement possible d'identifier les significations des segments d'un message aux segments de la signification du message. Même si nous déclarons que les significations peuvent, d'une manière ou d'une autre, être purement et simplement concaténatées, dans beaucoup de cas nous trouverions probablement, sous réserve de quelqu'interprétation raisonnable de ces notions, que

$$\text{comp}\,(x)^\frown\text{comp}\,(y) \leqslant \text{comp}\,(x^\frown y). \tag{3}$$

C'est peut-être là une manière d'interpréter la proposition gestaltiste, selon laquelle la signification d'un tout est supérieure à la somme linéaire des significations de ses parties. A moins d'être un partisan impénitent de l'assotiationisme dans la tradition de James Mill, il n'est pas évident que parler de la concaténation de deux compréhensions ait un sens. On ne voit pas non plus clairement comment une telle opération pourrait se réaliser. En comparaison, les opérations représentées par l'équation 2 semblent bient définies.

Néanmoins, si l'on veut introduire le processus de compréhension, on arrive à de nombreux et difficiles développements. Récemment, il y a eu quelques propositions intéressantes relatives à l'étude abstraite de certains aspects dénotatifs (Wallace, 1961) et connotatifs (Osgood, Suci et Tannenbaum, 1957) des lexiques naturels. Aussi important que soit ce sujet pour toute théorie psycholinguistique générale, nous en dirons peu de choses dans cet article. Nous formons l'espoir, néanmoins, qu'en apportant quelques lumières dans le domaine des problèmes de syntaxe, nous aurons aidé à clarifier le problème sémantique au moins en indiquant certains des sens qu'il faut refuser au mot signification.

Finalement, comme nous l'avons déjà noté, ces chapitres concernent peu les processus de l'apprentissage du langage. Bien qu'il soit possible de donner une description formelle de certains aspects du langage, et bien que divers modèles mathématiques des processus d'apprentissage aient été développés, la conjonction de ces deux efforts théoriques reste si faible que l'on en est surpris.

Le problème de savoir comment un enfant peut, sans être soumis à une éducation spéciale, parvenir si rapidement à une parfaite maîtrise d'une langue, est un défi lancé aux théoriciens de l'apprentissage. Un adulte intelligent peut naturellement atteindre un certain degré de maîtrise d'une

nouvelle langue en utilisant avec persévérance une grammaire traditionnelle et un dictionnaire: mais un petit enfant atteint une maîtrise parfaite avec une facilité incomparablement plus grande et sans aucune instruction explicite. Une éducation attentive et une programmation précise des occasions de renforcement ne semblent pas nécessaires ... Il suffit apparemment d'une exposition remarquablement courte pour qu'un enfant normal acquière la compétence d'un sujet parlant.

On peut sans doute apporter quelques éclaircissements en ces domaines théoriques en imaginant qu'il nous faut construire un mécanisme qui serait une réplique de l'apprentissage de l'enfant (Chomsky, 1962 a). Une des parties de ce mécanisme devrait, à partir d'un échantillon d'énoncés grammaticaux, (avec peut-être quelques restrictions sur l'ordre de présentation) produire une grammaire de la langue (ainsi qu'un lexique). La description d'un mécanisme de ce type serait une hypothèse au sujet du bagage intellectuel inné, qu'un enfant met en jeu, en apprenant une langue. Naturellement, d'autres données de base peuvent jouer un rôle essentiel dans l'apprentissage du langage. Par exemple, les corrections par la communauté linguistique sont probablement importantes. Une correction indique qu'une certaine expression linguistique n'est pas une phrase. Ainsi, le mécanisme devrait admettre un ensemble de non-phrases, tout autant qu'un ensemble de phrases, comme données de base. De plus, il peut être prévu des indications selon lesquelles tel terme doit être considéré comme la répétition d'un autre et peut-être d'autres conseils et suggestions. Une importante question pour la recherche empirique consiste, évidemment, à savoir quels autres éléments de base il faut prévoir.

Il est tout aussi important, cependant, de préciser les propriétés de la grammaire que notre mécanisme universel d'apprentissage du langage est censé construire à la sortie. Cette grammaire est destinée à représenter certaines des aptitudes d'un sujet parlant adulte. En premier lieu, elle devrait indiquer comment il est capable de désigner une phrase correctement construite et, deuxièmement, elle devrait nous renseigner sur les dispositions des unités à l'intérieur de structures plus larges. Le mécanisme d'apprentissage du langage doit, par exemple, être amené à comprendre les différences qui existent entre les exemples 1*a* et 1*b*.

Comment déterminer les caractéristiques d'une grammaire qui énumère de façon explicite des phrases grammaticales, chacune étant décrite selon sa structure propre, tel est le problème fondamental posé par cet article. Nous voulons obtenir une grammaire formalisée, qui spécifiera les descriptions structurales correctes, à partir d'un nombre relativement faible de principes généraux de formation des phrases; cette grammaire

devra dépendre étroitement d'une théorie de la structure linguistique qui justifiera le choix de cette grammaire parmi d'autres grammaires possibles. Une des tâches du linguiste professionnel consiste, en un sens, à rendre explicite le processus que chaque enfant normal réalise implicitement.

Un mécanisme pratique d'apprentissage du langage devrait inclure des hypothèses restreignant sévèrement la classe des grammaires potentielles que peut avoir une langue naturelle. Vraisemblablement, le mécanisme possèderait à l'avance une description détaillée de la forme générale que peut prendre une grammaire ainsi que certaines procédures permettant de décider à partir des données de base si une grammaire particulière est meilleure qu'une autre. De plus, il devrait posséder certaines capacités phonétiques pour reconnaître et produire des phrases. Enfin, il devrait posséder une méthode qui, une fois donnée l'une des grammaires permises, déterminerait la description structurale d'une phrase quelconque. Si on s'imagine qu'on arrivera à choisir une grammaire adéquate parmi le nombre infini de grammaires possibles, au moyen de processus de pure induction s'exerçant sur un corpus fini d'énoncés, on se trompe complètement sur l'ampleur réelle du problème.

Le processus d'apprentissage, selon nous, consisterait en une évaluation des diverses grammaires possibles dans le but de trouver la meilleure, compatible avec les données de base. Le mécanisme s'efforcerait de trouver une grammaire qui énumère toutes les phrases et aucune des non-phrases et leur assigne des descriptions structurales de telle manière que les non-répétitions diffèrent aux points appropriés. Naturellement nous devrions associer l'appareil d'apprentissage du langage avec certains types de principes heuristiques qui lui permettraient, à partir des données de base ainsi que d'une série de grammaires possibles, de faire un choix rapide parmi les quelques grammaires qui semblent prometteuses; celles-ci seraient alors soumises à un processus d'évaluation. Ou encore, ces principes heuristiques permettraient d'évaluer certaines caractéristiques de la grammaire avant d'autres. On pourrait cependant simplifier les procédures heuristiques nécessaires en spécifiant à l'avance, de la manière la plus étroite possible, la classe des grammaires potentielles. La division du travail adéquate entre les méthodes heuristiques et la spécification de la forme reste à décider, naturellement, mais il ne faudrait pas faire trop confiance à la puissance de l'induction, même quand elle est aidée par une heuristique intelligente, pour découvrir la grammaire correcte. Après tout, les gens les plus stupides apprennent à parler, mais même le singe le plus brillant n'y parvient pas.

2. QUELQUES ASPECTS ALGÉBRIQUES DU CODAGE

L'application d'un monoïde dans un autre est une opération qui se trouve partout dans les systèmes de communication. Nous l'appellerons, avec quelque imprécision, *codage*, — comprenant sous ce terme les divers processus de codage, recodage, décodage et transmission. Afin de rendre cette discussion préliminaire plus concrète, considérons un monoïde qui consiste en toutes les suites qui peuvent être formées avec les caractères d'un alphabet fini A et un autre monoïde consistant en toutes les suites qui peuvent être formées avec les mots d'un vocabulaire fini V. Dans ce chapitre, par conséquent, nous considérons certaines propriétés abstraites du système de concaténation en général, propriétés qui s'appliquent de la même façon aux codes artificiels et aux codes naturels.

Un code C est une application 1 : 1, θ, des suites de V dans les suites de A telle que si v_i, v_j sont des suites de V, alors $\theta(v_i \frown v_j = \theta(v_i) \frown \theta(v_j)$. θ est un isomorphisme entre les suites de V et un sous-ensemble des suites de A; les suites de A fournissent les orthographes des suites de V. Dans ce qui suit, s'il n'y a pas de danger de confusion, nous simplifierons notre notation en supprimant le symbole de concaténation $\frown$, adoptant ainsi la convention normale des systèmes orthographiques. Considérons l'exemple simple d'un code que nous appellerons C_1. Soit $A = \{0, 1\}$ et $V = \{v_1, \ldots, v_4\}$. Définissons une application θ de la manière suivante:

$$\theta(v_1) = 1$$
$$\theta(v_2) = 011$$
$$\theta(v_3) = 010$$
$$\theta(v_4) = 00$$

Cette application particulière peut être représentée par un graphe en arbre, comme dans la figure 1. (Pour une discussion théorique des graphes en arbre voir, par exemple, Berge, 1958). Les nœuds représentent des points de choix; un chemin partant d'un nœud et allant vers le bas et à gauche représente le choix de 1 dans A et un chemin allant vers le bas et à droite représente le choix de 0. Chaque mot a une orthographe unique, indiquée par une branche unique dans le codage en arbre. Quand on atteint l'extrémité d'une branche et qu'un mot entier a été épelé, le système revient au sommet du graphe, prêt à épeler le mot suivant.

Afin de pouvoir décoder le message, il est naturellement essentiel de maintenir une synchronisation. Par exemple la suite de mots: $v_4v_1v_4v_1v_1$ est épelée 0010011, mais si la première lettre de cette orthographe manque, le résultat du décodage apparaîtra comme v_3v_2. Nous utilisons le symbole

au commencement d'une suite de lettres pour indiquer que l'on sait qu'il représente le commencement du message total; autrement, on utilise une suite de points (...).

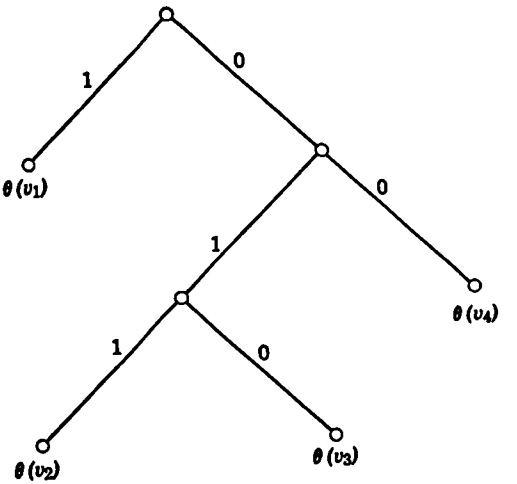

Figure 1

A chaque endroit particulier d'une suite de lettres qui épelle un message acceptable quelconque, il y aura un ensemble déterminé de prolongements possibles se terminant à la fin d'un mot. De plus, des suites initiales différentes peuvent permettre exactement les mêmes prolongements. En C_1, par exemple, les deux messages qui commencent par #000 ... et #10 ... peuvent se terminer par 0#, ... 10#, ... 11#, ou par un de ces groupes de lettres suivi d'autres mots. Nous disons qu'entre deux suites initiales quelconques il existe la relation R lorsque ces suites autorisent le même prolongement. Nous voyons immédiatement que R doit être réflexive, symétrique et transitive, et que c'est donc une relation d'équivalence. Deux suites initiales de caractères qui permettent exactement le même ensemble de prolongements sont appelées équivalentes à droite. A partir de cette relation, nous pouvons définir le concept important d'*état: l'ensemble de toutes les suites équivalentes à droite constitue un état du système de codage.*

L'état d'un système de codage constitue sa mémoire de ce qui est déjà survenu. Chaque fois qu'une lettre s'ajoute à la suite codée, le système passe à un nouvel état. En C_1 il y a trois états: (1) S_0 quand un mot complet vient juste d'être épelé, (2) S_1 après #0 ..., et (3) S_2 après #01. Ces états correspondent aux trois nœuds non terminaux dans l'arbre de la figure 1.

Nous inspirant de Schützenberger (1956), nous pouvons résumer les transitions entre états par des matrices, chacune étant définie pour chaque suite de lettres. Les lignes représentent alors l'état après n lettres et les colonnes représentent l'état après $n+1$ lettres. Si une transition est possible entre deux états on inscrit 1 à l'intersection de la ligne et de la colonne qui les représentent, sinon 0. Dans le cas du codage C_1 nous avons besoin de deux matrices: une pour représenter l'effet de l'addition de la lettre 0, l'autre pour représenter l'effet de l'addition de la lettre 1. (En général ceci correspond à un partition de l'arbre de codage en sous-graphes, un sous-graphe pour chaque lettre dans A). A chaque suite x nous associons la matrice M_x formée d'éléments m_{ij} donnant le nombre de chemins entre les états S_i et S_j quand la suite x apparaît dans les messages codés. Pour C_1 les matrices associées avec les suites élémentaires 0 et 1 sont

$$M_0 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{et} \quad M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Pour une suite plus longue, la matrice est le produit ordonné des matrices par les lettres de la suite. La matrice-produit $\mathscr{A}\mathscr{B}$ s'interprête de la façon suivante: le système se déplace de l'état S_i à l'état S_j selon les transitions permises par $\mathscr{A}$, puis se déplace de l'état S_j à l'état S_k selon les transitions permises par $\mathscr{B}$. Le nombre des chemins distincts de S_i à S_j à S_k est $a_{ij}b_{jk}$. Le nombre total de chemins de S_i à S_k, par sommation sur tous les états intermédiaires S_j, est $\sum_j a_{ij}b_{jk}$ produit ligne par colonne donné par les éléments de $\mathscr{A}\mathscr{B}$. Dans le cas où une lettre particulière ne peut apparaître dans un état donné, la ligne de sa matrice correspondant à cet état sera formée entièrement de 0. Toute matrice correspondant à une suite de A qui n'épèle aucune partie d'une suite de V sera une matrice 0. En général, il n'est pas nécessaire que les matrices possèdent des inverses; elles ne forment pas un groupe, mais elles forment un isomorphisme avec les éléments d'un semi-groupe.

Si la fonction: application de V dans A n'est pas bi-univoque, il y aura plus d'une entrée dans les matrices ou leurs produits. Ceci signifie qu'une suite unique de n lettres épelle plus d'une suite de mots. Quand ceci se produit, le message reçu est ambigu et ne peut être compris même s'il est reçu sans faute ni distorsion. Il n'est pas possible de savoir quelle suite de mots, parmi les suites possibles, était prévue.

Etant donné qu'il n'y a pas d'indicateur de frontière phonétique entre les mots successifs dans le flux normal de la parole, de telles ambiguités

se produisent facilement dans les langues naturelles. Miller (1958) donne l'exemple anglais suivant:

The good candy came anyway.
The good can decay many ways.

La suite d'éléments phonétiques peut se prononcer de manière relativement neutre, de sorte que l'expérience qu'en a celui qui écoute est en gros comparable à celle qui est ressentie dans les phénomènes visuels de perspectives réversibles. Les ambiguités de segmentation sont peut-être même plus courantes en français; le couplet suivant est un exemple bien connu de rime complète:

Gal, amant de la Reine, alla (tour magnanime),
Galamment de l'arène à la Tour Magne, à Nîmes.

En considérant des exemples de ce genre, on se rend compte que la production ou la perception de la parole impliquent bien plus de choses que la simple production ou la simple identification de qualités phonétiques successives, et que différents types de traitement de l'information doivent jouer à divers niveaux d'organisation. Ces problèmes sont définis de façon plus adéquate pour les langues naturelles au § 6.

Les difficultés de segmentation peuvent évidemment être évitées. Pour ce faire, on est amené à faire une classification simple des différents types de codes (Schützenberger, communication personnelle). Les *codes généraux* comprennent tous les codes qui font toujours correspondre à deux suites de mots différentes deux orthographes différentes. Un sous-ensemble particulier des codes généraux sont les *codes en arbres* dont les règles d'orthographe peuvent se représenter graphiquement comme dans la figure 1, l'orthographe de chaque mot se terminant à la fin d'une branche distincte. Tout code en arbre doit être de l'un des deux types suivants: les *codes en arbre à gauche* sont ceux dans lesquels l'orthographe d'aucun mot n'est le segment initial (segment de gauche) de l'orthographe d'aucun autre mot; les *codes en arbre à droite* sont, de la même façon, ceux dans lesquels aucun mot ne forme de segment terminal (segment de droite) d'aucun autre mot. (Les codes en arbre à droite peuvent être formés en inversant l'orthographe de chaque mot d'un code en arbre gauche). Un cas spécial est représenté par la classe des codes qui sont à la fois et simultanément des codes en arbre à droite, et des codes en arbre à gauche; Schützenberger les a appelés *codes anagrammatiques.*

Le sous-ensemble le plus simple formé de codes anagrammatiques est