ENCYCLOPÉDIE FRANÇAISE

CAHIERS
D'ACTUALITÉ ET DE SYNTHÈSE

LA CYBERNÉTIQUE

par L. COUFFIGNAL et M.P. SCHÜTZENBERGER

CONSTRUCTION SÉMANTIQUE DE LA LOGIQUE

par E. W. BETH

MATHÉMATIQUES

mises à jour détaillées rédigées sous la direction de Paul MONTEL

I

SOCIÉTÉ NOUVELLE DE L'ENCYCLOPÉDIE FRANÇAISE
Dépositaire Général : Librairie Larousse *13-21, rue du Montparnasse, PARIS VI*^e^

2. - La théorie de l'information

INTRODUCTION

Il est classique de comparer à l'énergétique la théorie de l'information. La première aurait pour but l'étude des systèmes dans lesquels une certaine quantité d'énergie se produit ou se transforme — la seconde aurait le même rôle, mais relativement à une nouvelle grandeur : l'*information*. A l'énergétique appliquée correspondraient les machines qui ont été le triomphe de la technologie du XIXe siècle : locomotive ou alternateur électrique. A la théorie de l'information reviendrait de faire les plans de ces mécanismes nouveaux que sont les calculatrices électroniques, les appareils de télévision, etc.

On peut ainsi poursuivre ces analogies assez loin, jusqu'au point même où les deux théories se rencontrent parce que l'on doit discuter les rapports mutuels entre les quantités d'énergie et d'information simultanément mises en jeu dans certains processus. Cependant une différence irréductible subsiste. Les deux théories correspondent à des idéalisations opposées de la réalité et comme nous le montrerons en détail plus loin, la seconde commence à ce point précis où sont postulées négligeables les quantités qui étaient l'objet même des recherches de la première.

Dans une perspective entièrement différente, des auteurs, moins techniciens que philosophes peut-être, ont voulu trouver dans la théorie nouvelle les fondements d'une science radicalement originale qui livrerait les clés de la signification et on a tenté ainsi, plus ou moins sérieusement, de rattacher la sémantique aux résultats des spécialistes en télécommunications.

Entre ces deux points de vue extrêmes, sinon dans leur but, tout au moins dans leur dépassement des possibilités actuelles, il nous semble préférable (*) de ne conserver de la première analogie que le principe des démarches de la physique théorique et de ne garder de la seconde que la présence d'entités nouvelles et notamment de « *sujets* » dont il faudra avant toute chose élucider les comportements.

Enfin, on ne saurait oublier que la théorie des communications, en tant que discipline appliquée, a une fonction normative : non contente de théoriser ce qui est ou ce qui peut être, elle doit fournir à l'ingénieur des critères raisonnables de décision qui, par nature, sont partiellement absolus, mais aussi partiellement arbitraires. En ceci encore, elle se distinguera donc des autres théories physiques auxquelles nous emprunterons le schéma initial d'analyse :

— décomposition de l'objet étudié en une série d'éléments plus simples ;

— fixation des principes de fonctionnement de chacune de ces parties ;

— calcul d'approximation permettant la synthèse des éléments constitutifs relativement à un principe extrémal d'optimalité.

LES ELEMENTS D'UN DISPOSITIF DE COMMUNICATION

Prenons, pour fixer les idées, l'exemple le plus courant d'un réseau télégraphique :

En suivant, depuis le début, la marche des opérations, nous y rencontrerons les articulations suivantes :

Emission

1) *L'expéditeur* compose le texte du télégramme.

Codage

2) L'agent du poste émetteur transforme la suite des lettres et signes typographiques, en une suite d'impulsions électriques qui sont envoyées sur un circuit électrique (matériel ou non).

Transmission

3) Les impulsions cheminant sur ce circuit sont soumises à des perturbations diverses (distorsion, « fading », parasites atmosphériques ou autres).

Décodage

4) Un second agent déchiffre les indications reçues et recompose en caractères typographiques usuels un texte qui est remis au destinataire.

Réception

5) Au reçu du télégramme, ce dernier prend une certaine décision, telle qu'aller attendre son correspondant à la gare à une certaine heure, effectuer telle ou telle opération commerciale, etc.

Enfin :

Bilan

6) Un bilan est établi qui chiffre d'une part le coût du télégramme, d'autre part le bénéfice ou la perte qu'a fait réaliser effectivement sa transmission plus ou moins rapide et correcte.

Si on le voulait, on pourrait voir, dans ce schéma un « modèle canonique de communication » et, en effet, les étapes que l'on vient d'énumérer se retrouvent quoique avec une importance variable dans tout processus de communications concevable. Nous conviendrons en outre de décomposer tout échange d'information (un dialogue par exemple) en une série de tels cycles dans lesquels le même individu peut d'ailleurs changer de rôle d'un cycle à un autre. L'identification est au demeurant toujours assez facile encore que l'opération « bilan » puisse paraître assez artificielle. C'est pourtant elle, comme nous le verrons plus loin, qui est la clé de voûte de toute la théorie.

1) L'émission. — Par hypothèse, un « émetteur » choisit un message à expédier dans une certaine liste de messages possibles. L'émetteur est le premier « sujet » que nous ayons à introduire. Par ce terme nous entendons d'ailleurs pour le moment une propriété négative : le comportement de l'émetteur ne *peut* pas être prévu exactement, car en effet si l'on savait quel message va être choisi il n'y aurait pas besoin de communication du tout. Ceci dit, l'émetteur peut être soit *passif*, c'est-à-dire considéré comme choisissant au hasard selon des probabilités fixes et connues, les différents messages, soit *actif*, c'est-à-dire agissant de façon entièrement autonome et imprévisible. Le cas du sujet actif sera traité plus en détails quand nous parlerons des principes de comportement.

Le deuxième aspect important de l'émission est l'existence d'une liste finie ou infinie, mais en tous cas *connue* de messages possibles : c'était l'ensemble des suites de signes typographiques composant un texte français dans notre exemple initial, mais ce serait aussi bien :

— l'ensemble des plages lumineuses plus ou moins éclairées formant une image susceptible d'être vue par une caméra de télévision ;

— l'ensemble de 3 lettres et 4 chiffres composant le numéro d'appel à un standard téléphonique ;

— une valeur numérique dans ce processus informationnel très spécial qu'est la mesure d'une grandeur physique (longueur, poids, voltage, etc.) ;

— un symbole dichotomique (« oui » ou « non ») fourni par un dispositif avertisseur par exemple.

L'existence de cette liste ne semble pas une restriction a priori puisque la liste peut être aussi longue et aussi complexe que l'on veut, mais elle a cependant une conséquence fondamentale : nous pouvons, dès le début, assigner à chacun des messages possibles un numéro d'ordre (au sens usuel

(*) Le point de vue qui sera adopté ici est celui qui avait guidé un séminaire sur la théorie des informations organisé en 1954-1955 dans le cadre de la chaire de Calcul des Probabilités de M. le Pr. G. DARMOIS par B. MANDELBROT et M. P. SCHÜTZENBERGER. Bien que ce dernier soit seul responsable de la forme donnée ici à cet exposé, son contenu reflète un point de vue commun développé au cours de longues et amicales discussions.

ou dans un sens généralisé) et considérer que la communication ne porte que sur ce numéro d'ordre, à l'exclusion de tout contenu sémantique du message qu'il repère. Ainsi — et c'est là qu'est la restriction essentielle — la théorie de l'information ne porte-t-elle que sur les « *signifiants* » et non sur les « *signifiés* » pour reprendre la distinction fameuse de SAUSSURE. Tout comme le nombre « 245 » désigne les propriétés collectives de tout ensemble formé de 245 objets distincts indépendamment de leur nature propre, le système de communication dont l'ensemble des messages possibles est « oui » ou « non » peut être étudié en soi sans tenir compte de ce que signifient ce oui et ce non.

Du même coup, il apparaît que la notion d'information n'existe qu'en relation avec la structure explicite de cette liste. Ce dernier point méritera toute notre attention, car son oubli est la source de bien des erreurs d'interprétation, dont la théorie a été l'objet.

2) Codage. — Même sous la forme d'un numéro d'ordre, le message n'est pas en général propre à être transmis tel quel :

Les images lumineuses doivent être transformées en modulations électriques avant d'être radiodiffusées, la pensée doit se traduire en mots, en gestes avant d'être perçue par autrui, le discours doit être écrit pour que la communication en soit possible au-delà du cercle des auditeurs, etc.

Le codage a pour but de faire correspondre à chaque message un autre symbole encore plus abstrait mais appartenant cette fois à la liste des *messages susceptibles d'être transmis* ou « *signaux* ».

Naturellement dans la pratique, ce codage est en réalité une série de codages successifs (de la pensée à la phrase, aux mots, aux phonèmes, aux ondes sonores, etc.), bien que nous l'envisagions ici comme une opération unique, dont le caractère essentiel est seulement d'être effectué selon des règles fixes, connues, et qui permettent le décodage ultérieur.

3) La ligne de transmission. — Ici encore, il s'agit d'une idéalisation à laquelle peuvent s'identifier les réalités les plus diverses depuis la page imprimée jusqu'à l'onde hertzienne. Son trait distinctif est d'être le lieu où s'introduit l'incertitude dans la communication.

Par une généralisation du langage des radiotransmissions, on appelle « *bruit* » toute perturbation au hasard du message transmis : par exemple sera réputé paradoxalement « bruit » le fait que le circuit téléphonique soit coupé pendant une conversation à condition que cette interruption soit :

1) imprévisible de façon autre que statistique ;

2) indépendante du message initial transmis.

(La censure n'est pas de la nature d'un « bruit » à proprement parler !)

Par conséquent, les distorsions qui peuvent en principe être rigoureusement corrigées (par exemple un affaiblissement de certaines fréquences) ne sont pas du bruit, car elles ne se produisent pas au hasard.

Elles ne sont pas « aléatoires » ou « stochastiques » (ces deux expressions étant pratiquement synonymes).

C'est en général une des données du problème que la structure du bruit et celle-ci s'exprime par des probabilités indiquant quelles sont les chances d'une transmission correcte ou incorrecte des différents signaux.

Observons d'ailleurs que le bruit peut dans certains cas être nul ou négligeable : c'est le cas le plus simple et sa théorie nous occupera tout spécialement.

4), 5), 6). — Décodeur. Récepteur. Bilan.

Il n'y a que peu de choses à dire pour le moment du décodeur et du récepteur, dont les fonctions sont évidentes et dont les principes d'action seront discutés plus loin. La notion de bilan, par contre, mérite un examen détaillé : supposons donc que nous soyons dans un cas où des évaluations monétaires précises aient un sens et voyons quels sont les « postes » qui doivent figurer à ce bilan.

Nous trouvons :

1) *Les frais d'équipement :* pour réaliser le codage, la transmission et le décodage, un certain matériel a dû être mis en place quel que soit l'usage ultérieur qui en sera fait. C'est donc l'une des tâches de théoriciens des communications que d'indiquer comment les mécanismes doivent être réalisés de la façon la plus simple pour des niveaux d'exactitude et de rapidité donnés.

2) *Les frais de fonctionnement :* ceux-ci dépendent à la fois du message émis et des opérations ultérieures. Il nous faut distinguer encore. D'une part il y a les coûts de codage, de transmission et de décodage qui varient selon le degré de minutie avec lequel ces opérations sont réalisées et, d'autre part se trouve le « *coût d'erreur* » qui est la clé de voûte de tout l'édifice. Par cette fonction, nous entendons ce qu'il en coûte au récepteur pour chaque message émis d'avoir pris sa décision en fonction d'un message plus ou moins tronqué ou incorrect.

Il n'est pas facile d'ailleurs de trouver des exemples naturels simples où le calcul de ce bilan soit tant soit peu rigoureux : en dehors d'opérations industrielles très précises il est hautement arbitraire de « taxer » monétairement les inconvénients de tel ou tel type d'erreur. Aussi avons-nous là comme une sorte de contradiction dans la théorie :

D'un côté, sans notion de bilan celle-ci devient triviale, car si les erreurs étaient sans importance, il serait inutile de communiquer quoi que ce soit, et de même si la transmission ne coûtait rien, il suffirait de répéter un très grand nombre de fois pour être à peu près sûr du résultat. Mais, d'autre part, dans chaque application, l'établissement de ce bilan est une affaire d'appréciations personnelles impossibles à chiffrer : comment trouver une commune mesure entre le coût en temps et en argent d'un long télégramme et le risque que son contenu trop elliptique soit incompréhensible pour le destinataire?

Comment comparer l'inconvénient d'une image télévisée plus ou moins floue avec l'avantage d'un prix de revient moindre d'un appareillage simplifié?

C'est donc d'abord à ce niveau que la théorie devra être normative en choisissant arbitrairement des critères raisonnables de coût, avec l'espoir souvent justifiable que la contradiction sera levée par ce fait mathématique que si les opérations sont répétées un très grand nombre de fois, le résultat pratique et peu sensible à des changements de critères tant que ceux-ci ont à peu près les mêmes propriétés qualitatives.

Voici des exemples de normes parmi les plus fréquemment utilisées :

Coût d'installation. — On dispose d'un stock de pièces détachées, toutes censées avoir le même prix et il s'agit d'en faire la synthèse de dispositif, en utilisant le plus petit nombre d'unités.

Coût de fonctionnement. — Un petit intervalle de temps est pris comme une unité à laquelle on ramène toutes les opérations que l'on cherche évidement à effectuer avec des délais minimum. Ou bien : l'énergie mise en jeu sur la ligne est prise comme unité de coût. Ou bien : cette même énergie est taxée à un tarif infime tant qu'elle est en dessous d'un certain palier et à un tarif infiniment élevé au-dessus.

Fonction d'erreur. — Les messages possibles étant en nombre fini, une « unité de gain » est enregistrée si et seulement si le message émis est correctement identifié, les erreurs étant toutes considérées comme égales en coût. Ou bien : les messages possibles étant restreints à deux (*oui* ou *non*), une taxation élevée oblige à rendre très faible la fréquence des « oui » reçus, comme « non » et l'on minimise la fréquence des « non » reçus comme « oui ». [1]

Les messages possibles étant des grandeurs (intensité, voltage, distance, etc.), on mesure la précision par une fonction simple de la différence $|\xi - \hat{x}|$ entre la grandeur émise ξ et la grandeur reçue $\hat{x}$ [2].

Ce choix fait pour chacune des parties, il sera le plus souvent suffisant au stade de l'application pratique de fixer des limites à tous les coûts, sauf un sur lequel on fera porter l'effort d'optimalisation : quitte d'ailleurs à reposer le problème sous un autre angle par la suite. Ainsi par exemple, dans le développement du téléphone, a-t-on pu voir successivement les ingénieurs s'efforcer de diminuer les coûts d'erreurs (l'objectif premier étant obtention d'une reproduction correcte de la voix parlée), puis les coûts d'installation (l'objectif étant la simplification des appareillages), puis maintenant en fonction d'exigences spéciales, les coûts de fonctionnement (la réduction des « largeurs de bande »).

Enfin, au-delà des techniques, l'introduction du bilan et de la fonction de coût d'erreur a une signification profonde pour toute la théorie, car c'est elle qui restitue aux messages émis ou reçus leur valeur concrète : le message initial était, nous l'avons vu, un simple numéro d'ordre sur une liste : au plus avait-il comme caractéristique sa probabilité plus ou moins grande d'être émis. La fonction d'erreur nous oblige, sinon à tenir compte intégralement de sa signification, tout au moins à considérer l'aspect de celle-ci qui retentit sur le bilan de la transmission : par exemple, si le bruit est tel que les messages dont les numéros d'ordre sont voisins ont plus de chance d'être confondus les uns avec les autres, il sera désirable d'éloigner par le codage les messages entre lesquels l'erreur est la plus grave. C'est là, si l'on veut, une possibilité de définition intrinsèque du degré de synonymie : deux messages sont d'autant plus synonymes que la confusion de l'un avec l'autre a un coût plus faible.

On observera d'ailleurs, à ce propos, que la relation de synonymie n'est pas nécessairement symétrique : confondre A et B n'a pas toujours la même importance, selon que c'est A ou B qui a été émis (cf. les dispositifs d'alarme évoqués plus haut).

LES PRINCIPES DE COMPORTEMENT

Ayant mis en place les acteurs et les décors, il nous faut maintenant fournir aux personnages des mobiles d'actions.

Le codeur et le décodeur sont par hypothèses de simples figurants : agissant selon des règles fixes connues de tous, ils sont mécanisables et mécanisés par principe.

Le bruit, lui est un acteur indépendant, mais nous avons convenu de le restreindre à n'agir que les yeux bandés sans savoir quels sont les messages qu'il déforme ou qu'il mutile.

Par contre, l'émetteur et le récepteur peuvent avoir des personnalités plus riches. Il suffit en ce qui concerne le premier d'évoquer le cas où l'émetteur est un avion ennemi ; les « messages émis » étant sa position et sa direction instantanée de vol. Le dispositif de communication a alors pour but d'informer le récepteur — qui est

(1) Par exemple un dispositif avertisseur d'alarme : on augmente d'abord sa sensibilité jusqu'à ce que les chances qu'il manque à fonctionner en cas d'accidents soient infimes (plus petites que 1/100 000 disons), puis dans un deuxième temps, on s'efforce d'empêcher au maximum qu'il donne l'alarme sans cause.

(2) Le plus souvent, le carré $|\xi - \hat{x}|^2$ de cette différence : c'est la méthode des « moindres carrés » connue des astronomes et des géodésiens depuis un siècle et demi et le « RMS criterium » des électroniciens modernes.

une batterie de D. C. A. — des éléments nécessaires au réglage du tir.

Il n'est évidemment pas possible de considérer que l'avion se déplace au hasard, car tout au contraire il dirige son vol de façon à rendre le plus difficile possible le pointage des pièces ennemies : il est *adversaire du récepteur* et il nous faut trouver une formulation mathématique de cette possibilité.

C'est la théorie des jeux — autre branche de la cybernétique — qui nous fournira les outils analytiques indispensables pour cette étude.

Pour l'instant, il nous suffit de savoir que l'émetteur adversaire peut dans des cas très généraux être considéré comme un sujet passif (c'est-à-dire choisissant au hasard), à condition que les probabilités en cause soient convenablement calculées.

Or, cette situation est beaucoup plus générale qu'il ne paraît : si, au lieu d'un adversaire subjectivement animé d'intentions hostiles, le récepteur avait en face de soi un sujet passif, mais dont il ignore les normes de conduite, ce pourrait être une politique raisonnable que d'agir comme à l'égard d'un ennemi déclaré.

Pour la seconde fois, la théorie des communications doit faire apparaître un principe normatif : en présence d'un émetteur dont il ignore les caractéristiques, le récepteur se conformera au principe « Minimax », c'est-à-dire qu'il manœuvrera comme si ces caractéristiques inconnues étaient les plus défavorables pour lui. Ceci nous amène alors à conclure que l'attitude du récepteur n'est en aucune façon différente de celle d'un statisticien. Pour ce dernier, en effet, une certaine grandeur inconnue a été « choisie » par la nature et ne lui est révélée qu'à travers un système de prises d'échantillons dont l'élément aléatoire peut être assimilé à un bruit. Par exemple le problème type suivant de la statistique peut être retraduit en termes de théorie des communications :

La longueur ξ comprise entre 99 et 101 cm d'un certain étalon est inconnue et non mesurable. On a reproduit 10 copies de cet étalon dont les mesures sont $x_1, x_2 \ldots x_{10}$ et l'on sait que des écarts — inévitables — dans le processus de copies sont distribués selon une loi connue Φ. Que peut-on dire de ξ connaissant les x_i ?

TRADUCTION

La liste des messages possibles est l'ensemble de tous les nombres entre 99 et 101. Le message émis ξ est codé en envoyant 10 fois un courant d'une intensité égale à ξ. Il existe un bruit qui ajoute ou soustrait à chacun des signaux une petite quantité distribuée selon une loi connue.

On pourrait donc considérer si l'on voulait que toute la statistique mathématique n'est autre chose que la théorie du récepteur dans les communications avec bruit si cette formulation ne subordonnait pas de façon un peu artificielle une discipline ancienne et largement développée à une autre encore dans l'enfance. Il n'en reste pas moins que la théorie de l'information ne se conçoit pas sans la statistique mathématique et il est assez intéressant du point de vue de l'histoire des sciences de constater que les ingénieurs des communications ont redécouvert en toute innocence des résultats et des principes bien connus depuis longtemps dans l'autre domaine.

LES GRANDS PROBLÈMES DE LA THÉORIE DE L'INFORMATION

La statistique mathématique ayant pour domaine propre le comportement du récepteur, il est clair que les résultats les plus nouveaux de la théorie de l'information se rencontrent à l'autre extrémité du cycle et tout spécialement dans l'étude du codage [3].

COMMUNICATIONS SANS BRUIT

Nous partirons du cas le plus simple et le plus typique où sont réunies les particularités suivantes :

i) la liste des messages élémentaires $M_1, M_2 \ldots M_n$ est *finie* et l'émetteur se borne à les choisir au hasard avec des probabilités fixes connues ;

ii) la ligne est exempte de bruit ;

iii) les signaux élémentaires qui peuvent être transmis sont en nombre fini et ont tous le même coût (calculé en unité de temps).

iv) aucune erreur n'est tolérée. (C'est-à-dire : le coût d'une erreur est infini !)

Sans grande perte de généralité on pourra supposer qu'il existe seulement deux signaux élémentaires, disons « + » ou « — », et le problème d'optimalité se formulera donc ainsi :

Comment faire correspondre à chacun des M_1, $M_2 \ldots$ une suite de + ou de —, de façon telle que cette correspondance permette un décodage sans erreur et qu'en moyenne le nombre des signaux élémentaires utilisés (« la longueur moyenne des messages du code ») soit le plus faible possible? Par exemple les messages à transmettre sont les suivants : M_1 « en avant », M_2 « en arrière », M_3 « à gauche », M_4 « à droite », M_5 « stop » avec des fréquences respectives de 1/8, 1/8, 1/8, 1/8 et 1/2. Comme nous n'avons à notre disposition

(3) On observera que l'opération appelée ici « codage » n'est pas inconnue de la statistique. Le soin de préparer les questions, c'est-à-dire les expériences ou les observations, de telle sorte que la « nature » y réponde le plus explicitement possible constitue ce que l'on appelle le « design of experiment » et tend à être un des chapitres les plus vivants de cette discipline. Il existe cependant des différences essentielles quant à la souplesse possible du codage qui est considérablement plus grande en théorie des communications.

que 2 signaux élémentaires + et – , il faut coder les M_i par des suites de + et de –. Supposons que l'on ait pris :

$$M_1 = + \; ; \; M_2 = - \; + \; ; \; M_3 = - - + \; ;$$
$$M_4 = - - - + \; ; \; M_5 = - - - -$$

Ceci est un code correct (sans ambiguïté à la lecture). Voyons quel serait son coût moyen. On trouve :

$$1 \times 1/8 + 2 \times 1/8 + 3 \times 1/8 + 4 \times 1/8 + 4 \times 1/2 = 26/8 = 3{,}25.$$

Naturellement il semblerait plus logique d'attribuer au message « stop » qui est de loin le plus fréquent la suite la plus courte de signaux élémentaires et nous pouvons essayer, par exemple, le code suivant :

$$M_1 = + + + \; ; \; M_2 = + + - \; ; \; M_3 = + - + \; ;$$
$$M_4 = + - - \; ; \; M_5 = -$$

de coût moyen :

$$3/8 + 3/8 + 3/8 + 3/8 + 1/2 = 16/8 = 2$$

donc beaucoup plus économique.

La question se pose d'elle-même de savoir si l'on ne pourrait encore faire mieux et c'est là l'objet du premier théorème de la théorie de l'information que nous formulerons ainsi :

« Dans les conditions décrites plus haut, le nombre moyen des signaux élémentaires ne peut pas descendre au-dessous de la valeur H, qui a l'expression suivante :

$$H = \Sigma p_i \operatorname{Log}_2 1/p_i$$

(p_i est la probabilité a priori du i-ème message).

$\operatorname{Log}_2 1/p_i$ est le logarithme de base de 2 de $1/p_i$, c'est-à-dire le nombre x tel que $(1/2)^x = p_i$).

Dans notre exemple : $p_1 = p_2 = p_3 = p_4 = 1/8$ et $\operatorname{Log}_2 8 = 3$; $p_5 = 1/2$ et $\log_2 2 = 1$, donc $H = 1/8 \times 3 + 1/8 \times 3 + 1/8 \times 3 + 1/8 \times 3 + 1/2 \times 1 = 2$.

D'autre part, un second théorème — plus profond que le précédent d'ailleurs — affirme que :

Il est possible de trouver au moins un codage tel que le nombre moyen de signaux élémentaires soit plus petit que H + 1.

Ces deux énoncés corrélatifs montrent que la quantité H a une liaison essentielle avec le problème. Nous l'appellerons la *quantité d'information de Hartley* et nous discuterons plus en détail sa signification dans les paragraphes suivants.

L'INFORMATION DE HARTLEY

La signification de cette quantité sera plus claire sans doute si nous considérons ce que devient H dans deux cas limites très simples.

D'abord quand une seule des probabilités n'est pas nulle (et est donc égale à 1). On trouve que H est nul et qu'il est différent de zéro pour tout autre système de valeur — ce que nous pouvons résumer par : l'information de HARTLEY est nulle si, et seulement si le message était connu à l'avance. Voyons maintenant l'autre cas extrême, celui où il existe N messages, tous également probables : un calcul facile montre que dans ce cas H a la valeur $\operatorname{Log}_2 N$ et que c'est pour ce choix des probabilités que celle-ci est maximum. A priori ce résultat n'a rien que de très intuitif : supposons pour simplifier que N soit une puissance de 2 — ($2, 2^2 = 4, 2^3 = 8, 2^4 = 16$ ... etc.), soit 2^h. D'après la définition qui a été donnée, H est précisément égal à h. Or il existe exactement 2^h — suites différentes formées de h symboles consécutifs + ou —. Si donc nous voulons faire correspondre à chaque suite un seul message, il faudra donc employer toutes ces suites et leur longueur moyenne (qui est leur longueur commune) sera bien h. D'autre part, nous ne saurions rien gagner en utilisant des suites de longueurs différentes, comme on peut le vérifier sans peine.

Le fait que pour un nombre donné N de messages, la quantité H atteigne un maximum quand ceux-ci sont tous aussi probables les uns que les autres correspond d'ailleurs bien à l'intuition : que ce cas est celui où l'incertitude est la plus grande possible. — C'est d'ailleurs sur des considérations de cette nature que s'était basé HARTLEY dès 1923 pour étudier cette question avant la formulation plus abstraite et plus systématique qu'en a donné C. SHANNON en 1949 dans un mémoire qui à la fois fonde et popularise la notion d'information.

Nous aurons cependant avantage à utiliser une deuxième interprétation de H qui en montre plus directement l'usage et nous prendront l'exemple schématique suivant : un objet appartient à l'une des 5 catégories C_1, C_2, C_3, C_4, C_5 dont les fréquences respectives sont 1/8, 1/8, 1/8, 1/8 et 1/2 et qui sont caractérisées par les propriétés suivantes :

	C_1	C_2	C_3	C_4	C_5
Taille	grand	grand	petit	petit	petit
Forme	long	court	long	long	long
Couleur	vert	vert	vert	vert	blanc
Poids	lourd	lourd	léger	lourd	léger

Supposant que toutes les observations ont le même coût, comment organiser au mieux le diagnostic (c'est-à-dire l'identification de la classe à laquelle appartient un objet)?

Nous empruntons aux systématiciens botanistes ou zoologistes leurs « clefs de diagnose » et nous construisons le schéma suivant :

Catégorie

$$\text{COULEUR}\begin{cases}\text{vert : TAILLE}\begin{cases}\text{grand : FORME}\begin{cases}\text{long : } C_1\\ \text{court : } C_2\end{cases}\\ \text{petit : POIDS}\begin{cases}\text{léger : } C_3\\ \text{lourd : } C_4\end{cases}\end{cases}\\ \text{blanc : } G_5\end{cases}$$

Celui-ci conduit à 2,00 observations en moyenne. Serait-il possible de faire mieux?

La théorie nous permet déjà de répondre *non :* en effet, le problème est rigoureusement le même que celui discuté plus haut une fois que l'on a introduit par exemple le code suivant :

1er symbole : + = vert ; — = blanc ;

2e symbole : + = grand ; — = petit ;

3e symbole : + = long ou léger ; — = court ou long

et un diagnostic plus rapide en moyenne signifierait un codage plus efficace, ce qui est impossible pour les valeurs données des fréquences.

Mais nous pouvons aller plus loin :

H correspondant aux valeurs (1/8, 1/8, 1/8, 1/8, 1/2) nous donne une sorte de mesure de ce que nous savons a priori sur l'objet en nous fournissant une estimation du nombre minimum d'indices qu'il faut recueillir pour l'identifier. Si nous savions que l'objet est « petit », les catégories C_1 et C_2 seraient exclues et les probabilités des autres catégories deviendraient 1/6, 1/6 et 4/6. A cet ensemble de valeurs correspond une quantité d'informations :

$$H = 1/6 \operatorname{Log}_2 6 + 1/6 \operatorname{Log} 6 + 4/6 \operatorname{Log} 6/4 = 1.2526...$$

On peut donc considérer que la connaissance du fait que « l'objet est petit » apporte une quantité d'information chiffrée par la différence :

$$2{,}00 - 1{,}252 \ldots \# 0{,}748 \ldots$$

Si au contraire nous savions que l'objet est grand, les seules catégories possibles seraient C_1 et C_2 avec une valeur de H correspondante égale à $H'' = 1/2 \operatorname{Log}_2 + 1/2 \operatorname{Log}_2 2 = 1{,}00$, soit une différence de $2{,}00 - 1{,}00 = 1{,}00$.

Finalement à l'observation portant sur la taille, il est logique d'associer la valeur moyenne de ces différences, soit :

$$1/4\,(H - H'') + 3/4\,(H - H') \# 0{,}811 = \bar{H}$$

que nous appelons la *valeur moyenne de l'information apportée par l'observation de la taille.*

La signification de $\bar{H}$ est claire : cette quantité exprime de façon approchée la réduction du nombre moyen d'observation restant à faire quand la taille est connue par rapport à la valeur correspondante quand la taille est inconnue.

C'est ici qu'intervient un phénomène mathématique de la plus haute importance dû au choix particulier qui a été fait pour l'expression analytique de H :

$\bar{H}$ est précisément la quantité d'information attachée au simple diagnostic de l'objet entre les deux groupes de catégories « grand » ou « petit ». Autrement dit, nous aurions pu faire l'économie de tout le calcul précédent et évaluer $\bar{H}$ simplement comme

$\bar{H} = 1/4 \operatorname{Log}_2 4 + 3/4 \operatorname{Log}_2 4/3 = 0{,}811\ldots$ (probabilité $C_1 + C_2$ (grand) $= 1/8 + 1/8$ probabilité $C_3 + C_4 + C_5$ (petit) $= 1/8 + 1/8 + 1/2$).

En particulier nous voyons qu'en première approximation une observation isolée sera d'autant meilleure que la valeur de $\bar{H}$ correspondante sera plus élevée. D'après ce que nous avons dit au début, ceci rejoint parfaitement l'intuition, car la question la plus « informative » est celle à laquelle toutes les réponses sont le plus également probables a priori.

Les remarques précédentes ont sans doute montré l'utilité de l'information de HARTLEY, mais il est nécessaire de ne pas perdre de vue ses limitations et surtout le fait qu'elle dépend des probabilités a priori des différentes catégories. Ainsi, si nous voulions savoir quelle est la quantité d'information apportée par une lettre de l'alphabet, nous ne pourrions donner de réponse que dans le cadre d'un système de référence précis qui pourrait être par exemple :

soit l'ignorance complète sur cette lettre (comme ce serait le cas si elle était l'indicatif de série d'un billet de loterie). On trouve alors :

$$H_1 = 26 \times 1/26 \times \operatorname{Log}_2 26 = 4{,}70,$$

soit la seule connaissance du fait que cette lettre appartient à un texte français. Ici :

$$H_2 = \text{à peu près } 2{,}5$$

d'après les données statistiques établies empiriquement,

soit la connaissance quelle est la première lettre d'un verbe français rimant avec « rouge » :

$$H = O$$

car la seule possibilité est b (de « bouge » ; cf. A. DE MUSSET). Si ces probabilités sont inconnues ou trop mal connues, l'information de HARTLEY ne sert plus guère qu'à formuler cette trivialité dont abusent les vulgarisateurs, à savoir que pour reconnaître un objet entre 2^h catégories ou plus, il faut au moins h observations dichotomiques.

La même mise en garde vaut pour la réalisation effective d'un programme de diagnostic : si les probabilités a priori des différentes catégories sont connues et si leur nombre est faible, il existe une méthode directe (due à HUFFMAN) permettant de construire un programme optimal sans utiliser H qui n'est en définitive qu'une fonction d'approximation. Par exemple, on présente parfois la solution du célèbre « problème de 12 pièces » (trouver en trois pesées une pièce fausse parmi 12) comme une application de la théorie de l'information et c'est là le type d'une prétention un peu exagérée.

La théorie permet bien de calculer qu'a priori il n'y a pas d'impossibilité en ce qui la concerne, mais la possibilité effective est de nature purement combinatoire. Ainsi si l'on savait qu'une pièce dans un lot de trois est plus lourde que les deux autres, la théorie de l'information montrerait qu'il n'est possible de la reconnaître par une seule observation qu'à condition que cette dernière soit susceptible de donner 3 réponses différentes : cette condition est nécessaire, mais non suffisante. Par exemple, l'opération est évidemment irréalisable avec une balance de type habituel.

Cependant, et c'est là l'aspect important, répétons-le, s'il s'agissait de reconnaître n pièces plus lourdes parmi 3n pièces, la deuxième partie du théorème montrerait que sous des conditions très larges on pourrait s'approcher proportionnellement autant que l'on veut de l'optimum à condition que n soit assez grand.

C'est d'ailleurs parce que H est une approximation et non pas une expression rigoureusement égale à la fonction de coût de fonctionnement que son efficacité est si grande comme nous le verrons plus loin.

PROBLEMES DE SCANSION

Nous n'avons envisagé jusqu'ici qu'un seul message. En général il s'agit plutôt de suites indéfinies de messages élémentaires, même si la liste de ceux-ci est finie. Par exemple, on doit considérer le cas où l'émetteur guiderait constamment un engin en dictant sans interruption une suite d'ordres tels que : en avant, à gauche, stop, stop, etc., etc., et un nouveau problème, algébrique celui-ci, se pose à l'ingénieur. Comment le décodeur va-t-il reconnaître que tel signal + ou — est un début ou une fin de *mot* (4) à moins de précautions spéciales. Par exemple, si le code avait été composé des mots suivants :

M_1 = + + + ; M_2 = — + + ; M_3 = + — + ;
M_4 = — — + ; M_5 = —

(c'est-à-dire l'inverse de celui donné plus haut). La suite d'ordres « droite arrière stop » (M_4 M_2 M_5) serait inutilisable car elle serait transmise sous la forme — — + — + + — et avant le dernier signe le récepteur ne pourrait savoir s'il ne s'agit pas de la suite « stop stop gauche... » (M_5 M_5 M_3) qui commence de la même manière. Ce problème n'est d'ailleurs pas purement théorique et il suffit d'avoir essayé de lire, par exemple, un texte classique javanais où aucun signe n'indique la fin des mots, pour apprécier ce genre de difficultés.

A ceci, plusieurs remèdes sont possibles : le plus naturel consiste à réserver un symbole spécial à la séparation des messages élémentaires successifs : c'est dans notre typographie ce signe spécial qu'est l'absence de tout autre signe et B. MANDELBROT en a fait la théorie dès 1952. Une autre technique consiste à n'employer que des unités de longueur fixe, les débuts de mots se retrouvant alors régulièrement.

Enfin, il est des classes de codes pour lesquels le délai d'attente entre la réception d'un symbole et celle de contexte nécessaire pour l'interpréter est réduite au minimum : ce sont les codes *unitaires.*

Dans le cas général, le problème de trouver les fins de mots, de *scander* le message conduit à l'introduction de notions intéressantes, notamment celle d'équivalence syntaxiques : (M. P. SCHÜTZENBERGER) les suites de symboles élémentaires a et a' sont syntaxiquement équivalentes si, quelles que soient les suites x et y, les suites composées xay et xa'y sont en même temps des messages corrects, c'est-à-dire composés d'une suite de mots.

Il y a là une analogie avec les parties du discours : (deux mots sont équivalents si on peut les substituer l'un à l'autre sans altérer la correction de la phrase, par exemple : deux adjectifs masculins singuliers, en français, ou en anglais : deux adjectifs quelconques) dont on a ainsi une sorte d'image algébrique formelle.

On observera au passage que la théorie des communications apporte une contribution à la discussion sur le sens des termes « équivalent » ou « échangeable » en linguistique . Nous avons parlé plus haut de synonymie (= équivalence du point de vue de la fonction de coût d'erreur) — l'équivalence syntaxique définie a rapport à la structure et non pas au contenu du message et en est donc qualitativement différente.

COMMUNICATIONS AVEC BRUIT

Nous conservons les mêmes hypothèses que plus haut, à cela près que nous ne supposons plus que la ligne de transmission soit exempte de bruit. La fonction d'erreur sera le nombre des identifications correctes toutes les erreurs étant également pénalisées. Prenons l'exemple très simple suivant :

Les messages à transmettre M_1 M_2 M_3 M_4 ont les mêmes probabilités a priori — 1/4. La ligne admet deux signaux + et —, mais pour chacun d'eux il y a une chance sur 10 pour que la transmission l'altère à tel point qu'il devienne impossible à reconnaître :

Ainsi ayant codé de la façon la plus économique, les messages par le tableau ci-dessous :

M_1 = + +
M_2 = + —
M_3 = — +
M_4 = — —

nous pouvons calculer que si M_1 est émis, il y a :

81 chances pour cent pour qu'il arrive correctement (+ +) ;

18 chances pour que l'un des symboles soit ininterprétable (le message reçu est ? + ou + ?) ;

1 chance pour cent pour que les deux symboles soient brouillés (? ?).

Dans les cas douteux, le récepteur pourra tirer au sort sa décision entre les deux possibilités qui existent (par exemple, s'il a reçu + ? c'est que soit + + = M_1, soit + — = M_2 avaient été émis.

(4) On appelle « mot » la suite de signaux élémentaires correspondant à un seul message élémentaire.

En moyenne on trouve que le résultat final sera correct dans :

$$\frac{81}{100} + \frac{1}{2}\,\frac{18}{100} + \frac{1}{4}\,\frac{1}{100} = 90{,}25\% \text{ des cas ou plus.}$$

Naturellement si un code plus long était utilisé, les résultats seraient meilleurs : par exemple avec

$$M_1 = + + + + \; ; \; M_2 = + + - - \; ;$$
$$M_3 = - - + + \; ; \; M_4 = - - - -$$

les chances sont beaucoup plus faibles pour que le message reçu soit ambigu : + + ? +, par exemple, ou même + ? ? + ne peuvent provenir que de M_1. Le même calcul que plus haut conduit à prédire un résultat correct dans 99,0025 des cas.

Il est intuitif qu'à condition d'augmenter suffisamment le nombre des symboles employés n'importe quel degré de certitude prescrit à l'avance pourrait être atteint et ceci quelle que soit d'ailleurs l'intensité du bruit à corriger.

Cependant ceci nous oblige à employer des codes parfois très longs et en tous cas ne réalisant pas le critère d'économie maximum permis par le théorème fondamental en l'absence de bruit. Les codes « antibruit » sont comme on dit *redondants* en ce sens qu'ils sont en moyenne plus longs que les codes optimaux correspondants.

Le problème se pose donc de savoir ce qui subsiste de la théorie dans ce cas et ceci est l'objet d'un théorème énoncé initialement par SHANNON puis démontré rigoureusement par FEINSTIEN, théorème qui constitue pour l'instant le résultat le plus important de toute la théorie de l'information.

Nous ne saurions ici en donner une expression rigoureuse, mais il est possible de montrer simplement en quoi il consiste :

Si nous revenons à l'exemple de tout à l'heure, nous constatons que quand 4 symboles sont consacrés à chaque message, il serait possible de transmettre non pas seulement 4 messages, mais 8 sans diminuer beaucoup la sécurité : par exemple le code :

$$M_1 = + + + + \; ; \; M_2 = + + - - \; ;$$
$$M_3 = + - + - \; ; \; M_4 = + - - + \; ;$$
$$M_5 = - + + - \; ; \; M_6 = - + - + \; ;$$
$$M_7 = - - + - \; ; \; M_8 = - - - -$$

donne 97,2 % de transmission correcte soit avec différence de 2 % à peine avec le code précédent alors que le nombre de messages utilisables est double.

Il y a donc très généralement une balance entre la sécurité permise par un code et le nombre de messages distincts qu'il contient ; balance dont les exemples courants sont nombreux. Typiquement : toute chose égale d'ailleurs, une augmentation de la sensibilité d'une émulsion photographique se paye par une diminution de la finesse des images ; il s'agit donc de trouver une quantité ne dépendant que de la ligne, c'est-à-dire du bruit, qui puisse servir en quelque sorte, de coefficient à cette équation et c'est là le rôle de la *capacité*.

LA CAPACITE

Nous avons caractérisé la ligne et le bruit par les probabilités avec lesquelles les signaux élémentaires sont transmis correctement ou, au contraire, sont altérés au point d'être confondus les uns avec les autres. Il est clair que si ces probabilités sont différentes, il y aura avantage à employer proportionnellement plus fréquemment ceux des symboles que le bruit affecte le moins, c'est-à-dire d'adapter en probabilités le codage à la ligne. La capacité est précisément la valeur moyenne du gain d'information réalisé quand cette adaptation est effectuée de façon optimale. Un calcul relativement simple permet d'en trouver la valeur et du même coup de fixer une limite supérieure au rendement de tout codage concevable pour un niveau de bruit donné.

C'est là l'extension du premier théorème que nous avons donné dans la section précédente. La seconde partie, elle aussi, est beaucoup plus profonde : elle indique qu'il existe sûrement un code qui permet à la limite (c'est-à-dire pour des messages extrêmement longs) d'atteindre cet optimum (c'est-à-dire de transmettre cette quantité d'information).

Le phénomène mathématique curieux est que le théorème affirme l'existence d'un tel code sans donner aucune indication sur la manière de le construire [5]. D'ailleurs jusqu'à une date très récente on ne connaissait aucun code qui approchât même d'assez loin cet optimum théorique.

Revenant à la première partie du théorème, il est impossible de ne pas mentionner son application au cas très important dans la pratique où les messages sont codés par la modulation d'un système périodique (modulation en fréquence ou en amplitude) : une théorie spéciale permet de fixer une limite infranchissable à la multiplicité des signaux reconnaissables pour une largeur de bande ou une énergie données et naturellement sur le plan pratique l'existence de cette limite est un guide précis pour la construction des appareillages.

C'est même à ce chapitre de la théorie que sont consacrés le plus grand nombre des travaux qui se publient chaque jour, mais leur nature essentiellement technique nous interdit d'y consacrer plus que ces brèves remarques.

INFORMATION DE WALD

A l'opposé en quelque sorte de la théorie précédente se situe le cas où deux messages seulement sont à distinguer, mais où l'accent est mis

(5) La situation est analogue à la suivante : nous pesons ensemble 50 caisses censées contenir chacune 100 kg. Le poids total est 4 950 kg, donc l'une des caisses n'est pas complètement remplie ! mais nous ne savons pas évidemment de laquelle il s'agit.

tout spécialement sur l'économie de codage des signaux et ceci nous donnera l'occasion de parler d'un autre type d'information, *l'information de Wald.*

Si le décodage des signaux élémentaires est très coûteux, il est certain que le récepteur pourra procéder de façon plus économique en arrêtant celui-ci aussitôt qu'il s'estimera assez sûr du résultat.

Par exemple, dans le premier code discuté plus haut, si le récepteur a la chance que les trois premiers symboles aient été transmis sans brouillage (+ + +, par exemple), il n'a nul besoin du quatrième pour être sûr que le message émis était M_1. Le quatrième symbole ne sert qu'au cas où l'un des signaux antérieurs aurait fourni une indication douteuse. Ceci est très typiquement la démarche même que nous employons chaque jour où, pour vérifier une hypothèse, nous renouvelons les observations et les expériences jusqu'à ce que la balance penche indiscutablement d'un côté ou d'un autre.

L'emploi systématique de cette méthode constitue ce que l'on appelle « l'analyse séquentielle » et cette théorie, due au mathématicien A. WALD, constitue un des développements les plus brillants de la statistique dans les dernières décades. Son application à la théorie des communications est encore dans l'enfance, mais il ncus semble important de la mentionner ici, car elle repose sur l'emploi d'une certaine quantité W (6) qui généralise l'information H de HARTLEY et qui permet a priori de limiter inférieurement le nombre minimum des signaux à décoder.

PROBLEMES DE TRANSMISSION CONTINUE. INFORMATION DE FISHER

Jusqu'ici nous avons toujours supposé que les messages pouvaient être arrangés sur une liste finie ou tout au moins que l'on pouvait leur donner un numéro d'ordre 1, 2... n.

Les considérations précédentes deviennent inapplicables si les messages sont des grandeurs continues comme cela se rencontre souvent dans la pratique.

Une technique très simple employée par exemple dans le Pulse Code Modulation consisterait à se ramener au cas précédent en *quantifiant* ces messages, c'est-à-dire en faisant choix d'une unité assez petite et en négligeant de transmettre les fractions de cette unité. C'est ce que font les chauffeurs de taxis qui tarifent leurs courses en *hectomètres* et non pas de façon exactement proportionnelle au chemin parcouru.

Cependant, sur le plan théorique aussi bien que pratique, il n'est pas sans intérêt de discuter en lui-même le cas d'un message continu et les statisticiens, dans les cinquante dernières années, ont élaboré des techniques analytiques très raffinées pour traiter ce problème.

Ici encore, selon le schéma logique, qui semble de règle, nous rencontrerons :

1) Une quantité ayant les propriétés formelles d'une information, c'est-à-dire telle qu'on puisse aussi bien soit la calculer globalement a priori, soit déterminer étape par étape le gain apporté par chaque observation. (C'est ici l'information de FISHER (7).)

2) Un théorème fixant une limite inférieure au nombre minimum d'observations nécessaires pour obtenir une précision donnée. (C'est l'objet du théorème dit de CRAMER-FRECHET-DARMOIS-RAO d'après les noms des auteurs qui l'ont successivement généralisé.)

3) Un second théorème montrant que si des observations très nombreuses sont permises, il est possible de les combiner de telle sorte que cette limite soit approchée d'aussi près que l'on veut.

Ici encore le sujet est essentiellement mathématique et il nous serait impossible d'entrer dans plus de détails sans recourir à un appareil analytique compliqué.

On mentionnera seulement que les « diverses quantités informatives » qui ont été introduites successivement, peuvent être fondées dans une large mesure à partir de principes axiomatiques très généraux sans faire appel à leur fonction spécifique (M. P. SCHÜTZENBERGER).

LES APPLICATIONS DE LA THÉORIE DE L'INFORMATION

THEORIE DE L'INFORMATION ET PHYSIQUE

L'expression formelle de l'information H de HARTLEY est identique à celle d'une des grandeurs fondamentales de la thermodynamique, à savoir l'entropie.

Il a donc semblé très vite qu'une liaison profonde devait exister entre ces deux notions ; liaison à laquelle d'ailleurs paraissait convier directement leur contenu intuitif :

d'une part une mesure de notre connaissance sur le système,

(6) Soient p_i et p'_i les probabilités respectives pour que le signal élémentaire i ait été reçu selon que M ou M' a été émis.
Soit $W = \Sigma p_i \operatorname{Log} p_i/p'_i$.
Alors, si M a été émis, il faudra décoder en moyenne au moins K/W signaux élémentaires par où K ne dépend pas de la ligne, mais seulement du niveau de sécurité que l'on désire atteindre.

(7) Si $f(x_i, \theta)$ est la probabilité que l'observation donne le résultat x_i quand le message a la valeur θ, on a :

$$F = \Sigma_i \quad f(x_i : \theta) \quad \partial^2/\partial\theta^2 \operatorname{Log} f(x_i : \theta).$$

d'autre part, puisque c'est là l'interprétation habituelle de l'entropie, une mesure de degré de désordre de ce même système.

Et, en effet, dans un travail assez peu remarqué à son époque, le physicien SZILARD avait déjà étudié en 1923 une relation analogue avant même que fût introduit le concept d'information. Pour exposer cette question, il nous faut ouvrir une brève parenthèse et rappeler ce qu'est un «démon» de MAXWELL :

C'est un postulat fondamental de la physique que l'impossibilité du mouvement perpétuel. Supposons cependant qu'une boîte remplie de gaz soit divisée en deux parties par une cloison munie d'une porte.

Toujours d'après les principes fondamentaux, les molécules de gaz sont en mouvement et leur vitesse est inégale, la valeur moyenne seule de celle-ci restant constante et caractérisant la température.

Par conséquent, si un « démon » posté près de la porte ouvrait et fermait judicieusement celle-ci, ce qu'il pourrait faire en ne dépensant qu'une énergie très faible, il parviendrait petit à petit à trier les molécules en ne laissant par exemple entrer dans l'une des parties que les molécules en mouvement rapide et dans l'autre celles en mouvement lent.

Au bout d'un certain temps, il aurait donc réalisé une différence de température entre ces deux parties, différence susceptible d'être convertie en énergie en violation apparente du principe de l'impossibilité d'un « perpetuum mobile ».

Naturellement ce paradoxe n'a de valeur que théorique, mais sa réfutation a suscité de nombreux travaux et notamment l'application la plus importante de la théorie de l'information à la physique :

Quel est en effet le résultat de l'action du «démon»: en classant les molécules il diminue leur désordre, c'est-à-dire qu'il diminue l'entropie. Or, ceci n'a été possible que parce qu'une certaine information a été obtenue sur les molécules.

Par conséquent, si nous identifions (au signe $+$ ou $-$ près) l'information et l'entropie comme le suggère leur analogie formelle, nous retrouvons bien un bilan total nul ainsi que le veut la théorie : le «démon» n'a pu acquérir de l'information sur les molécules qu'aux dépens d'une certaine augmentation d'entropie qui doit compenser exactement la diminution de cette même quantité qu'il a provoquée dans la boîte.

En développant plus en détails les calculs, BRILLOUIN a réussi à montrer qu'effectivement un semblable «démon» est impossible et à exorciser ainsi le paradoxe de MAXWELL.

Cette identification peut d'ailleurs être poussée plus loin et l'on retrouve en théorie de l'information des formules parallèles aux célèbres relations d'incertitude de HEISENBERG qui dominent la physique quantique.

Il y a là tout un domaine de recherches nouveau étudié notamment par GABOR et dont les principes mêmes sont encore loin d'être éclaircis, car étant donné que les énoncés purement mathématiques, sur lesquels se basent les deux développements, sont rigoureusement les mêmes, on peut se demander si cette assimilation de l'information à l'entropie est plus qu'un parallélisme formel. Peut-être d'ailleurs que cette question elle-même est vide de sens, puisqu'en toute rigueur les deux théories correspondent au départ à deux idéalisations opposées.

THEORIE DE L'INFORMATION ET SCIENCES HUMAINES

Il était naturel que le nom seul de théorie de l'information suscite un grand enthousiasme parmi les spécialistes des sciences humaines, comme contenant la promesse d'un instrument mathématique rigoureux pour traiter de concepts jusque-là rebelles à l'analyse numérique.

Il faut cependant admettre que les résultats n'ont peut-être pas encore égalé les espérances dans tous les domaines envisagés, et notre exposé sera par nécessité moins cohérent qu'on ne le souhaiterait.

Applications à la sémantique et à la psychologie. — CARNAP et d'autres logiciens comme MACKAY ont cherché à appliquer l'information de HARTLEY (ou à introduire des expressions analogues) afin d'étudier des cas plus généraux : jusqu'ici les résultats ont été surtout formels puisque comme nous l'avons vu, il n'a d'intérêt que si des probabilités explicites figurent au départ. Il n'en reste pas moins qu'une lacune grave existe : il n'existe actuellement aucune façon qui ne soit pas artificielle d'évaluer le « gain d'information » (au sens large comme opposé au sens étroit utilisé dans les chapitres précédents) réalisé par l'accomplissement d'un calcul ou d'un raisonnement logique : si par exemple nous voulons savoir quel est le plus petit nombre x tel que $x^2 - 14x + 12 = 0$, c'est intuitivement un « gain d'information » que d'apprendre que x est égal à 7 moins la racine carrée de 37. Cependant, du point de vue de la théorie de l'information, il n'y a aucun gain de réalisé. De façon tout aussi négative, aucune mesure intéressante ne permet d'estimer le progrès réalisé à chacune des étapes intermédiaires d'un Sorite.

L'intérêt d'un concept mathématique applicable à ces cas serait cependant considérable puisqu'il permettrait sans doute de fixer des limites aux nombres d'opérations élémentaires à effectuer pour arriver au bout d'un calcul.

Les seuls résultats connus sont, soit purement formels, soit obtenus par des méthodes d'épuisement de tous les cas possibles, soit des approximations extrêmement sommaires.

Ceci souligne, une fois de plus, s'il en était besoin, à quel point est erronée l'opinion trop répandue qui considère la théorie comme traitant de l'« information » au sens courant du terme et non pas seulement d'une certaine quantité appelée « information » (de HARTLEY, de WALD, de FISHER, etc.) définie dans un contexte rigoureux, quantité qui est d'ailleurs, en effet, parfois identifiable partiellement à ce que suggère son nom, mais sans que pour autant les résultats analytiques de la théorie soient valables plus loin que ce contexte [8] étroit.

A un niveau plus concret la théorie de l'information a cependant apporté des résultats encore que fragmentaires :

Les plus remarquables sont sans doute ceux de MAC CALLOCH et PITTS qui, dans le cadre plus large de la physiologie nerveuse, ont systématiquement étudié le nerf comme un système destiné à transmettre de l'information. Dans la même ligne de recherches, d'autres travaux ont mis à profit ces méthodes pour essayer de donner une théorie des organes des sens et notamment de l'audition.

Sur le plan expérimental on ne peut pas négliger les recherches récentes de H. QUASTLER qui a cherché à mesurer quelle quantité d'information par seconde le cortex humain était capable d'utiliser : des résultats convergents basés aussi bien sur l'observation de dactylographes hautement qualifiées que de musiciens entraînés déchiffrant à leur vitesse maximum, ont permis de trouver un palier correspondant approximativement à une douzaine de choix dichotomiques par seconde. Ceci correspond — grossièrement — aussi à la quantité reçue dans une lecture rapide et on entrevoit des applications nombreuses de ces recherches en plein développement.

Applications à la Linguistique. — C'est ici certainement que les résultats les plus riches et les plus encourageants ont été obtenus, jusqu'à présent. Nous distinguerons parmi ceux-ci deux groupes : les résultats d'observation et les résultats théoriques.

1) *Résultats d'observation.* Il est difficile de savoir les rôles respectifs joués par la théorie de l'information et les simples nécessités de la pratique des communications.

Le fait est néanmoins que la phonétique ayant amené des esprits de formation plus scientifique que littéraire à reconsidérer les problèmes de langage, divers faits nouveaux ont été découverts qui n'avaient pas attiré l'attention des grammairiens.

Essentiellement il s'agit de la structure statistique du langage, étude qui encore une fois décolle du plan de la réalité de la signification pour s'intéresser exclusivement à la façon dont cette signification est transmise.

Comme nous l'avons vu déjà, toute étude d'un système de communication doit être précédée d'une analyse de la fréquence des divers messages :

Ainsi, de façon assez grossière le code Morse reflète la composition moyenne par lettres de l'anglais (e, la lettre la plus fréquente est le « mot » le plus bref : un *point ;* q, très rare, demande sept fois plus de temps : *trait trait point trait*).

Cependant cette seule liste de fréquence des lettres ne caractérise pas — et de loin — un langage : il existe ce que l'on appelle des corrélations, c'est-à-dire, très simplement, que la fréquence des lettres n'est pas indépendante de la séquence qui précède son apparition : en français « q » entraîne presque automatiquement u, « x » est surtout une lettre finale. Les probabilités d'apparition des différents couples de lettres (les fréquences de « bigrammes ») forment donc une description plus fine du langage. Allant plus loin on peut étudier les fréquences des trigrammes, des quadrigrammes, quoique le décompte devienne vite inextricable, pratiquement en raison de la multiplicité des cas à considérer [9].

D'après ce que nous avons dit plus haut, l'existence de ces liaisons en probabilités diminue le contenu informationnel des lettres successives : A la question : quelle est la lettre suivante ?, nous sommes déjà le plus souvent capables de donner une réponse approchée alors qu'un codage optimal (du seul point de vue de l'économie) voudrait que toutes les 27 réponses possibles fussent aussi probables les unes que les autres.

Naturellement cette redondance n'est pas sans avantage : un texte partiellement altéré ou tronqué reste en partie compréhensible grâce aux contextes. La langue est un code anti-bruit très efficace [10].

Cette analyse statistique a révélé des aspects curieux bien qu'anecdotiques des langues parlées : le niveau de redondance est celui-là même qui permet l'existence de mots croisés ou encore est tel qu'un codage optimal n'utilisant que *deux* signes typographiques, permettrait de traduire à peu près tout texte dans un autre de même longueur. Une manière pittoresque de présenter ce dernier résultat est de dire qu'une phrase de n signes peut — en général — être devinée au moyen de n questions bien choisies dont la réponse est limitée à oui ou non.

(8) Comme l'extrapolation abusive qui voudrait appliquer le théorème des « forces vives » à l'« énergie » d'un général d'armée.

(9) Voici des exemples de ce que l'on peut obtenir en tirant au sort les lettres successives selon les fréquences des trigrammes : « PROSEJOURS DE MAIS LE QUE DONNENT TROIRE A BLEMER » (français), « UALLIS MINIET/LETAUDATORUM NUOS ET BINIBUS NONEMAE RESCENT » (Latin).

(10) Des recherches expérimentales ont d'ailleurs montré l'étendue étonnante des distorsions ou des mutilations que supporte un discours sans cesser d'être intelligible alors qu'assez vite disparaît la possibilité de reconnaître l'identité de l'orateur : il y a une marge extrêmement grande entre reproduction parfaite et reproduction utilisable dont la technique la plus avancée cherche à tirer profit pour simplifier les appareillages ou réduire les largeurs de bande utilisées.

De façon plus sérieuse, des considérations analogues peuvent être développées au sujet de l'analyse des phénomènes : combien d'oppositions élémentaires (vocalisé/non vocalisé ; nasal/non nasal, etc.) sont nécessaires pour identifier un phonème d'une langue donnée? Une équipe de chercheurs autour de JAKOBSON travaille activement cette question qui a des rapports évidents avec la réalisation d'une machine capable de traduire en un texte *écrit* un message *parlé*.

D'autres recherches expérimentales ont étudié la répartition des accents d'intensité, la longueur des phrases, etc. et il y a là comme nous le disions plus haut, un point de vue nouveau qui ne peut que féconder le développement de la linguistique.

2) *Recherches théoriques*. Bien qu'elles soient en marge de la théorie de l'information au sens strict, il est impossible de ne pas citer ici les recherches suscitées par le problème de la mécanisation des opérations de traductions.

Si l'on veut dépasser le stade primitif d'une simple mécanisation de dictionnaire, il est certain que les structures grammaticales doivent être interprétées dans un esprit nouveau : au lieu d'une aide à la *compréhension* d'une langue où d'un recueil de principes énonçant les règles de la civilité verbale et écrite, la grammaire opérationnelle doit être un manuel rigide prescrivant les démarches à effectuer pour identifier les parties de discours et leurs accidents ou pour faire l'inverse.

La rédaction de semblables grammaires est en train pour les langues principales, et déjà on commence à apercevoir des possibilités d'analyse structurelle radicalement différentes de celles jusqu'ici classiques, notamment l'existence d'« unités sémantiques » moins complexes que les phrases, mais plus abstraites que les mots.

Ici encore, les travaux en cours sont extrêmement prometteurs.

Beaucoup plus liée à l'aspect mathématique est par contre la théorie de B. MANDELBROT sur laquelle nous nous étendrons plus longuement.

Nous avons mentionné plus haut ce fait que si aucun signe spécial n'articulait les fins de mots, la lecture du message serait beaucoup plus délicate, voire même impossible en cas de perturbations altérant un ou plusieurs symboles.

Plus généralement cette division en unités sémantiques, cette quantification de la description du monde, que permet l'existence de ces atomes de discours que sont les mots, se révèle une nécessité à peu près absolue quand le langage doit être plus qu'un simple code référant à un système fixe de messages étroitement délimités à l'avance.

Mais si l'on admet ce postulat, il s'ensuit que les méthodes usuelles d'optimalisation du codage ne sont plus applicables brutalement et le problème se pose de savoir quelles devraient être la longueur et la fréquence relative des différents mots pour que ce codage envisagé reste le plus économique parmi ceux qui appartiennent au même temps.

En résolvant mathématiquement cette question, on découvre qu'il doit exister une relation très simple entre, d'une part : la fréquence relative de chacun des mots et, d'autre part : le nombre des mots plus fréquents que celui considéré.

Or, il se trouve que ces répartitions statistiques avaient été déjà compilées empiriquement par ZIPF sur des échantillons très longs de langues aussi diverses que l'anglais de T. S. ELIOTT, celui de James JOYCE, l'hébreu moderne, certaines formes de latin d'église, etc., sans parler des langues classiques.

L'accord entre la théorie et les observations est excellent et les cas exceptionnels s'interprètent d'une façon qui confirme la validité générale de la théorie.

Nous avons donc là un des premiers exemples d'application complète des méthodes des sciences physiques aux sciences humaines : un modèle mathématique est construit sur la base d'un principe d'optimalité et les calculs à eux seuls permettent à la fois de retrouver les régularités empiriques et d'expliquer les cas de non-concordance.

Mais il y a plus. Les structures mathématiques employées relèvent d'une théorie plus profonde dont B. MANDELBROT a étudié d'autres applications : on retrouve là des problèmes thermodynamiques spéciaux — ceux-là mêmes que la théorie habituelle néglige — comme ne correspondant pas aux approximations naturelles de l'interprétation énergétique classique, mais dont l'intérêt n'est pas pour autant moins grand : par exemple les nombres de genres et d'espèces dans lesquels se subdivisent des groupes aussi variés et aussi étendus que les algues, les coléoptères, etc., suivant certaines relations empiriques (« Loi de WILLIS ») de façon à peu près vigoureuse, relations qui déterminent statistiquement la fréquence relative des genres possédant *n* espèces pour toutes les valeurs de *n*. B. MANDELBROT a montré que cette loi pouvait précisément s'interpréter comme une conséquence d'un principe.

M. P. SCHÜTZENBERGER.