Structuration linguistique et mesure de l'information

Gabriel G. Bès

Structuration linguistique et mesure de l'information

Gabriel G. Bès
Groupe de recherches sur la condensation de l'information en langue naturelle (CILN)
Université Blaise-Pascal, Clermont II

Linguistique fonctionnelle. Débats et perspectives, présentés par M. Mahmoudian.
Presses universitaires de France, Paris, 1979, p. 129-141.

Résumé

Cet article s'intéresse au problème de la réduction des symboles binaires utilisés pour coder les messages ; il se propose de compléter l'optique mathématique de la théorie de l'information par une vision linguistique du problème de la codification. Cette vision devrait permettre d'appliquer le formalisme mathématique non pas à un langage décrit d'une manière non structurée, mais à des suites riches d'une description qualitative dépassant la description des agencements linéaires des symboles. Les résultats atteints dans quelques applications expérimentales à l'espagnol suggèrent que la notion de choix, et d'autres notions fondamentales de la linguistique fonctionnelle inspirées par A. Martinet, peuvent jouer un rôle capital dans l'application de la linguistique au domaine de la transmission des messages en langue naturelle.

Voir aussi

Gabriel G. Bès. *Identités et différences dans les unités de deuxième articulation.* Thèse. Université René Descartes - Paris V, 1972. https://hal.archives-ouvertes.fr/tel-01095229

Gabriel G. Bès. « Un théorème sur l'équivalence de la valeur H entre langages ». *Condenser*, Adosa, Clermont-Ferrand, février 1980, n° 1, p. 97-100. https://hal.archives-ouvertes.fr/hal-01100224

Gabriel G. Bès. « Description phonologique et codification économique du langage ». *Cahiers du Centre Interdisciplinaire des Sciences du Langage*, 1980, n° 2, p. 117-123. https://hal.archives-ouvertes.fr/hal-01100220

STRUCTURATION LINGUISTIQUE ET MESURE DE L'INFORMATION*

par Gabriel G. Bès

Le problème essentiel de celui qui a produit un message est de le faire parvenir au destinataire. Si le producteur du message est éloigné de son destinataire et si l'on souhaite que le message parvienne rapidement à celui-ci, on doit faire intervenir un ensemble d'opérations fort complexes : le message original *M* est codifié en un message formalisé *M'* lequel est, d'abord, transmis et, ensuite, décodifié, de manière à retrouver le message original M. Les opérations de codification, de transmission et de décodification sont compliquées et coûteuses. Elles mettent en œuvre, d'une part, un appareillage matériel; d'autre part, elles nécessitent un code permettant la double conversion du message original en message formalisé et du message formalisé en message original.

L'élaboration d'un système de codification efficace a été un des soucis majeurs des sciences et des techniques appliquées à la communication. Comment réduire autant que possible le temps de transmission d'un message sans le distorsionner ? Cette question, si l'on admet que le message formalisé doit être constitué de symboles binaires (désormais *sb*) peut se reformuler ainsi : comment

* Une partie des résultats et de la discussion ici présentés ont été exposés en janvier-février 1975, dans des conférences prononcées au département des Mathématiques pures de l'Université de Clermont et au département des Langues modernes de l'Université de Simon-Fraser (Burnaby, Canada). Dans la préparation de ce travail, j'ai largement profité des observations d'A. Hurtado, d'E. W. Roberts, et, tout particulièrement, de longues discussions avec Michel Chambreuil, qui m'ont permis, d'une part, de mieux comprendre la théorie de l'information dans ses aspects mathématiques et, d'autre part, d'achever la mise au point du théorème évoqué à la n. 4. M. Gentil, de l'IUT de Montluçon, a eu l'amabilité de faire le programme permettant le calcul des résultats du *Code C*, dont les données statistiques ont été obtenues grâce à la patience encourageante d'Irène Hackner, qui a dépouillé « à la main » les résultats obtenus sur un corpus d'espagnol.

réduire le nombre des *sb* utilisés pour coder les messages ? Dans le cadre de la théorie de l'information, inspirée des travaux de Shannon, ce problème a été abordé à partir de ce que l'on peut appeler l'optique mathématique de la codification. Dans cette contribution, nous nous proposons d'esquisser une autre possibilité : l'optique mathématique devrait pouvoir être complétée par une vision linguistique du problème de la codification. Le langage artificiel, qui suit, volontairement simplifié, illustre cette vision mathématique du problème et permet de la situer dans un cadre linguistique. Les résultats déjà atteints dans quelques applications expérimentales à l'espagnol, présentés par la suite, suggèrent que la notion de choix, et d'autres notions fondamentales de la linguistique fonctionnelle inspirées par André Martinet, peuvent jouer un rôle capital dans l'application de la linguistique au domaine de la transmission des messages en langue naturelle.

Soit les suites du langage L :

R^A^R^T^A^B^O^D^A^B^E^B^T^A
T^A^R^E^B^A^R^R^A^B^A^B^D^O
R^A^B^T^E^D^O^R^A^R^T^O^B^A^R^D^E^R
D^O^T^A^B^A^R^T^O^R^T^A^R^A^B^T^E^D^E^B^T^A^B

On connaît un certain nombre de caractéristiques de *L*, aussi bien qualitatives que quantitatives. Sur le plan qualitatif, on sait que :

a) L'inventaire de symboles de l'alphabet de *L* est { R, B, T, D, A, E, O }, auxquels on doit ajouter le symbole initial et final « # ».

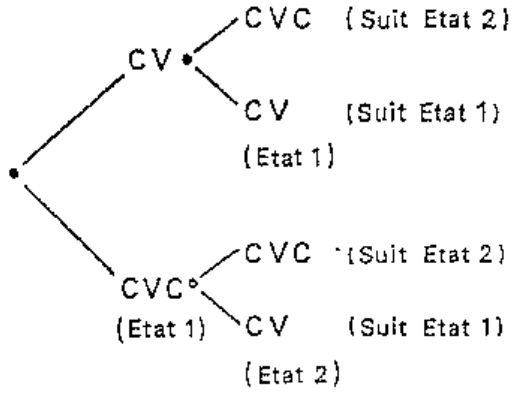
b) Les symboles de *L* se groupent en deux classes : { C } (consonnes) et { V } (voyelles). La classe { V } est constituée des symboles *A*, *E*, *O* et la classe { C } des symboles *R*, *B*, *T*, *D*.

c) Les éléments des classes { C } et { V } s'organisent en syllabes, dont il existe deux types : *CVC* et *CV*.

d) On définit trois positions dans la syllabe : position du noyau vocalique, occupée par le système { A, E, O }; position du prénoyau, occupée par des symboles appartenant à { C }; position du postnoyau, occupée aussi par des symboles appartenant à { C }. Si le prénoyau suit une syllabe du type *CV*, la classe { C } est représentée par le système { R, B, T, D }; si le prénoyau suit une syllabe du type *CVC*, la classe { C } est représentée par le sys-

tème { R, T, D }. La position du postnoyau est occupée par le système { R, B }.

Tenant compte des rapports entre les types syllabiques et l'existence de différents systèmes (consonantiques et vocalique), on a la combinatoire représentée dans le schéma 1 qui suit :



...	C	V	C	V	C	C	V	...
	R	A	R	A	R	R	A	
	B	E	B	E	B	T	E	
	T	O	T	O		D	O	
	D		D					

SCHÉMA 1

L'état 1 est constitué par les syllabes susceptibles d'apparaître à la suite de # ou d'une syllabe de type *CV*, l'état 2 par les syllabes susceptibles d'apparaître à la suite d'une syllabe de type *CVC*, les lignes pleines indiquant les possibilités de transition à partir d'un point donné.

Sur le plan quantitatif, on sait que :

e) Chaque type syllabique a une fréquence propre (f(CV) $\neq f$(CVC)), mais la fréquence de chaque type est la même dans les deux états (f(CV) dans l'état 1 $= f$(CV) dans l'état 2).

f) Chaque symbole vocalique possède une fréquence propre, qui reste cependant la même dans les différents types syllabiques : f(A) $\neq f$(E) $\neq f$(O); mais, par exemple, f(A) dans une syllabe de type *CV* $= f$(A) dans une syllabe de type *CVC*.

g) Chaque symbole de la classe { C } a une fréquence propre, qui est par ailleurs différente dans chaque système; par exemple, f(R) $\neq f$(B) $\neq f$(T) $\neq f$(D) dans le système { R, B, T, D }; f(R) dans le système { R, B, T, D } $\neq f$(R) dans le système { R, B } $\neq f$(R) dans le système { R, T, D }.

h) Dans les limites fixées par les contraintes qualitatives indiquées ci-dessus, la combinatoire des entités qui se suivent est quantitativement libre, dans ce sens qu'il n'existe pas des f conditionnelles entre une entité et l'entité qui suit (cf. n. 1).

Supposons qu'il soit nécessaire de transmettre les suites de

L : une suite S_i de L doit donc être codifiée en une succession de *sb*; celle-ci doit être transmise et, par la suite, décodifiée de manière univoque en la suite S_i. Dans le cadre de l'optique mathématique, la question de la codification des suites de L a été abordée à partir des points fondamentaux qui suivent[1] :

a) La codification minimale d'un langage mesurée en nombre de *sb* par symbole codifié, peut s'approcher autant qu'on veut de la valeur de H (information ou entropie) de la source (ensemble de symboles avec leurs transitions), mais elle ne peut pas être inférieure à cette valeur.

b) La valeur de H dépend du nombre d'unités de la source et de leur fréquence. Dans celle-ci, il faut considérer, d'une part, la fréquence de chaque unité et, d'autre part, la fréquence conditionnelle d'une unité à partir de telle autre. La valeur de H augmente avec le nombre d'unités de la source; à nombre égal d'unités de la source, la valeur de H augmente si les f de toutes les unités sont identiques et si la combinatoire entre les unités est libre.

c) La formulation mathématique susceptible de calculer la valeur de H ne s'applique qu'aux f conditionnelles mettant en rapport deux positions, l'une à la suite de l'autre. Le calcul de H pour un langage donné se fait donc par approximations successives. On calcule ainsi, d'abord, la valeur de H sans tenir compte de f de chaque symbole (approximation d'ordre zéro); par la suite, on considère f de chaque symbole mais sans tenir compte des f conditionnelles (approximation d'ordre 1). Dans le pas suivant (approximation d'ordre 2), on tient compte des f conditionnelles, dans les digrammes, entre chacun des symboles et tous les autres symboles. L'approximation suivante (d'ordre 3) considère les f conditionnelles dans les trigrammes. Pour ce faire, un trigramme est décomposé en un digramme suivi d'un symbole.

1. La technicité mathématique a été réduite à ce qui est strictement nécessaire pour suivre la discussion de fond, à savoir la possibilité d'incorporer les connaissances linguistiques sur la structuration du langage dans l'obtention des codes efficaces. Dans le texte, nous faisons allusion à Claude E. Shannon et Warren Weaver, *The mathematical theory of communication*, Urbana, Illinois, 1963 (1re éd. de 1949), en particulier le chap. Ier, et à Claude E. Shannon, Prediction and Entropy of Printed English, *Bell System Tech. J.*, 30 (1951), p. 50-64. On trouvera une présentation plus accessible et plus intuitive de la théorie de l'information dans Henri Atlan, *L'organisation biologique et la théorie de l'information*, Paris, Hermann, 1972, chap. 1 à 4, et dans J. R. Pierce, *Symbols, Signals and Noise*, New York, 1961. Les notations « $f(X)$ » et « $f(X \mid A)$ doivent se lire, respectivement, « la fréquence de X » et « la fréquence de X à la suite de A ». On se rappellera que la fréquence d'une suite $X^\frown Y$, dans le cas d'une combinatoire libre, est égale à $f(X)$ multiplié par $f(Y)$; en revanche, dans le cas des fréquences conditionnelles,

$$f(X^\frown Y) = f(X) \,.\, f(Y \mid X).$$

La valeur de H dans cette approximation d'ordre 3, dépend par conséquent des f conditionnelles entre chacun des blocs de deux symboles et tous les symboles qui les suivent. Au fur et à mesure que l'on augmente l'ordre d'approximation, on considère des f conditionnelles entre blocs de plus en plus grands et les symboles qui suivent. La valeur de H dans chaque approximation est égale ou inférieure à celle de l'approximation précédente. Rappelons, à titre d'illustration, les valeurs obtenues par Shannon pour un texte anglais, considérant 26 lettres et l'espace :

Approx. 0	Approx. 1	Approx. 2	Approx. 3
4,76	4,03	3,32	3,1

L'optique mathématique de la codification, caractérisée par les points *a)* à *c)* que l'on vient d'évoquer, conçoit le langage comme étant faiblement structuré sur le plan qualitatif : on ne considère que des unités qui s'alignent les unes derrière les autres et dont il s'agit de connaître les caractéristiques quantitatives, sans discuter la manière de les décrire et de les organiser. Des notions telles que contexte, paradigme et opposition, de syllabe et d'organisation hiérarchique, n'y sont pas utilisées. Aussi, il n'est pas étonnant que celui qui, selon cette optique, aborderait le problème de la codification de L, ne se soucierait pas de connaître les caractéristiques de L présentées ci-dessus.

Pourtant, si l'on tient compte de ces caractéristiques, il est possible de faire des prévisions intéressantes :

a) La combinatoire entre les symboles n'est pas libre. Du fait des caractéristiques qualitatives de L, la suite R⌢B, par exemple, n'est pas admise; par conséquent, $f(\mathrm{R}^\frown\mathrm{B}) = \mathrm{O} \neq f(\mathrm{R}).f(\mathrm{B})$.

b) Les fréquences conditionnelles existent non seulement à l'intérieur des digrammes — comme le montre l'exemple ci-dessus mais aussi dans les trigrammes. En effet, soit le trigramme R⌢R⌢A. A la suite de R⌢R on a comme possibilité le système { A, E, O }; en revanche, à la suite de R, on a comme possibilités le système { A, E, O } — lorsque R apparaît dans la position du prénoyau vocalique — et aussi le système { R, T, D }, lorsque R apparaît dans la position du postnoyau vocalique. Par conséquent, les prévisions que l'on est susceptible de faire à la suite de R⌢R sont différentes (et, en quelque sorte, plus « fines ») que celles que l'on peut formuler à la suite de R. Le résultat en est que $f(\mathrm{R}^\frown\mathrm{R}^\frown\mathrm{A}) \neq f(\mathrm{R}^\frown\mathrm{R}).f(\mathrm{A} \mid \mathrm{R})$; autrement dit, on ne peut pas cal-

culer la fréquence des trigrammes à partir de celle des digrammes, car il existe des contraintes qui vont au-delà d'une succession de deux lettres.

c) Il n'existe pas de contraintes qui aillent au-delà des trois lettres. Ceci implique que même si l'on considère des suites de plus de trois lettres, les prévisions que l'on est susceptible de formuler restent les mêmes que celles à l'intérieur des trigrammes. Soit, par exemple, la suite D⌢O⌢B⌢O⌢T⌢E⌢R⌢R⌢A. Les prévisions sur A que l'on peut formuler à la suite de R⌢R sont les mêmes que celles que l'on peut formuler à la suite de E⌢R⌢R, de T⌢E⌢R⌢R, de O⌢T⌢E⌢R⌢R, etc.

A partir de l'optique mathématique on peut, certes, aboutir à des résultats qui vont vérifier les prévisions précédentes. Ces résultats seront obtenus moyennant les calculs suivants :

I. — f des symboles dans une approximation de premier ordre, soit 7 valeurs.

II. — f des digrammes dans une approximation de deuxième ordre, soit 49 valeurs possibles (7^2) et 30 effectifs (toutes les combinaisons de deux lettres n'étant pas attestées).

III. — f des trigrammes dans une approximation de troisième ordre, soit 343 possibles (7^3), si l'on ne tient pas compte des contraintes déjà existantes au niveau des digrammes; ou 210 possibles (30 effectifs au niveau des digrammes, multipliés par les 7 symboles de l'alphabet), ou 118 effectifs.

IV. — f des quadrigrammes dans une approximation de quatrième ordre, soit 2 401 possibles (7^4), si l'on ne tient pas compte des contraintes au niveau des trigrammes, ou 826 possibles, si l'on ne tient compte que des trigrammes effectifs.

Les calculs I et II vont ratifier la prévision *a)* précédente; ceux de type III, la prévision *b)* et ceux de type IV, la prévision *c)*. L'optique mathématique conduira finalement à un système de codification où chaque symbole de l'alphabet aura une valeur particulière, calculée à la suite de chacun des symboles que ce symbole peut suivre, soit en tout 118 valeurs[2]. Désormais, ces

2. Nous rappelons que la valeur de H résulte de l'addition de ces 118 valeurs, pondérées par leur fréquence, chacune de ces valeurs étant le $\log_2$ d'une f déterminée. Ainsi, parmi les 118 valeurs, il y aura

$$f(R⌢B⌢A).\log_2 f(A \mid R⌢B);\ f(R⌢R⌢A).\log_2 f(A \mid R⌢R);\ \text{etc.}$$

(La formule générale est $H = -\sum_{i,j} f(X_i⌢X_j).\log_2 f(X_j \mid X_i)$.) Par ailleurs, il faut

valeurs, dont il faut tenir compte dans les attributions de *sb* aux symboles de l'alphabet, seront appelées *symboles codifiants*.

L'optique mathématique est, à notre connaissance, la seule qui a été employée pour traiter du problème de la codification des textes en langues naturelles. La vision linguistique peut la compléter. Elle devrait permettre d'appliquer le formalisme mathématique non pas à un langage décrit d'une manière non structurée, mais à des suites riches d'une description qualitative dépassant la description des agencements linéaires des symboles.

Revenons à *L* et aux caractéristiques connues de *L*. La vision linguistique de *L* nous dit qu'au-delà de l'alphabet de *L*, il existe plusieurs systèmes, déterminés par les symboles qui les composent et par la position vocalique. Chaque symbole de L n'est donc pas le résultat d'un choix sur l'ensemble total des symboles de l'alphabet, mais d'un choix qui porte sur les possibilités offertes par chaque système. Par exemple, dans la suite R⁀A⁀R⁀R⁀E, le premier R est un symbole appartenant à un ensemble de 4 possibilités (le système { R, B, T, D }; le deuxième R est un symbole d'un ensemble de 2 possibilités (le système { R, B }); le troisième R est un symbole d'un ensemble de 3 possibilités (le système { R, T, D }). Et il devrait être possible d'incorporer ces observations à un système de codification.

Les contextes qui déterminent les systèmes de *L* ne peuvent pas se définir toujours par la seule observation des symboles qui précèdent, à gauche, un symbole déterminé. Soit, par exemple, la suite B⁀A; après le symbole A on peut aussi bien trouver les symboles R ou B dans la position de postnoyau vocalique d'une syllabe de type *CVC*, que le symbole R ou B ou T ou D dans la position de prénoyau vocalique d'une syllabe *CV*. En revanche, si la suite B⁀A est suivie de, par exemple, R⁀A, on sait qu'il existe une limite syllabique après B⁀A. Autrement dit, pour déterminer l'appartenance à un système d'un symbole X qui apparaît dans une suite de *L*, il faut parfois tenir compte des symboles qui apparaissent *après* X.

Ceci pose un problème à la codification de X. En effet, supposons que l'on ait proposé les codifications suivantes (les

noter que dans tout le travail, nous avons identifié de manière simplificatrice, « codifier » (ou attribuer une succession de *sb* à un symbole déterminé) à « mesurer l'information » (par la formule ci-dessus évoquée).

chiffres tiennent lieu des successions de symboles binaires) : { R = 1; B = 2 }; { T = 1; D = 2; R = 3; B = 4 }. Soit la suite R⌢A⌢B⌢T⌢A. On sait que B, dans cette suite, appartient au système du postnoyau vocalique; il devrait par conséquent être codifié par « 2 ». Au moment de la décodification, lorsqu'on aboutira à ce « 2 », on aura par conséquent R⌢A 2... Ce « 2 » sera-t-il traité comme appartenant au système du postnoyau — il correspondrait dans ce cas à « B » — ou comme appartenant au système du prénoyau — il correspondrait dans ce cas à « D » ? En fait, on ne peut pas trancher, le système de codification ne permettant pas une décodification univoque.

La codification d'une unité quelconque doit, par conséquent, tenir compte, d'une part, du système auquel cette unité appartient, et d'autre part, cette appartenance doit être indiquée par des indices qui, dans le texte, doivent apparaître *avant* le symbole à codifier. Pour résoudre ce type de problème, il est nécessaire d'obtenir, par règle, un langage L' à partir de L; dans L' les indices qui déterminent l'appartenance des symboles à un système donné doivent être organisés de gauche à droite. En l'occurrence, les règles qui suivent permettent d'obtenir les suites de L' à partir de celles de L :

Symbole de L	Dans le contexte	Se réécrit dans L' comme :
R	$\begin{Bmatrix} V - & V \\ \# - & V \end{Bmatrix}$	R'_1
B	$\begin{Bmatrix} V - & V \\ \# - & V \end{Bmatrix}$	B'_1
T	$\begin{Bmatrix} V - & V \\ \# - & V \end{Bmatrix}$	T'_1
D	$\begin{Bmatrix} V - & V \\ \# - & V \end{Bmatrix}$	D'_1
R	C -	R'_2
T	C -	T_2
D	C -	D_2
R	$\begin{Bmatrix} -C \\ - \# \end{Bmatrix}$	R'_3
B	$\begin{Bmatrix} -C \\ - \# \end{Bmatrix}$	B'_3

Symbole de L	Dans le contexte	Se réécrit dans L′ comme :
A	- CV - #	A'_1
E	- CV - #	E'_1
O	- CV - #	O'_1
A	- CC - C#	A'_2
E	- CC - C#	E'_2
O	- CC - C#	O'_2
#		#

Les quatre suites présentées au début de ce travail deviennent, par application de ces règles :

$R'_1{}^\frown A'_2{}^\frown R'_3{}^\frown T'_2{}^\frown A'_1{}^\frown B'_1{}^\frown O'_1{}^\frown D'_1{}^\frown A'_1{}^\frown B'_1{}^\frown E'_2{}^\frown B'_3{}^\frown T'_2{}^\frown A'_1$

$T'_1{}^\frown A'_1{}^\frown R'_1{}^\frown E'_1{}^\frown B'_1{}^\frown A'_2{}^\frown R'_3{}^\frown R'_2{}^\frown A'_1{}^\frown B'_1{}^\frown A'_2{}^\frown B'_3{}^\frown D'_2{}^\frown O'_1$

$R'_1{}^\frown A'_2{}^\frown B'_3{}^\frown T'_2{}^\frown E'_1{}^\frown D'_1{}^\frown O'_1{}^\frown R'_1{}^\frown A'_2{}^\frown R'_3{}^\frown T'_2{}^\frown O'_1{}^\frown B'_1{}^\frown A'_2{}^\frown R'_3{}^\frown D'_2{}^\frown E'_2{}^\frown R'_3$

$D'_1{}^\frown O'_1{}^\frown T'_1{}^\frown A'_1{}^\frown B'_1{}^\frown A'_2{}^\frown R'_3{}^\frown T'_2{}^\frown O'_2{}^\frown R'_3{}^\frown T'_2{}^\frown A'_1{}^\frown R'_1{}^\frown A'_2{}^\frown B'_3{}^\frown T'_2{}^\frown E'_1{}^\frown D'_1{}^\frown E'_2$
$B'_3{}^\frown T'_2{}^\frown A'_2{}^\frown B'_3$ #

Le langage *L′* ainsi obtenu est susceptible d'être spécifié par une succession de systèmes, tels qu'ils sont indiqués sur le schéma 2 qui suit.

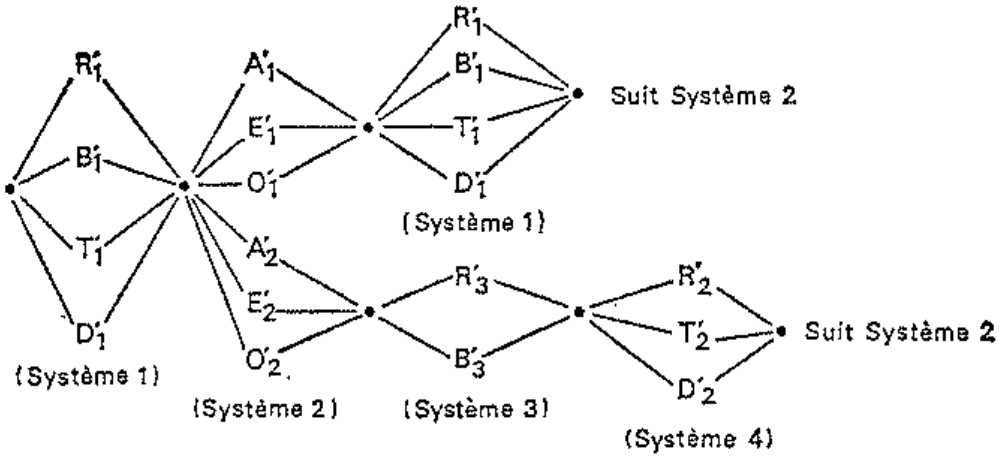


SCHÉMA 2

Les symboles de l'alphabet de L′ (soit R'_1, R'_2, R'_3, B'_1, B'_3, T'_1, T'_2, D'_1, D'_2, A'_1, A'_2, E'_1, E'_2, O'_1, O'_2) se groupent donc dans les 4 systèmes qui suivent :

S-1 : $\{ R'_1, B'_1, T'_1, D'_1 \}$
S-2 : $\{ A'_1, E'_1, O'_1, A'_2, E'_2, O'_2 \}$
S-3 : $\{ R'_3, B'_3 \}$
S-4 : $\{ R'_2, T'_2, D'_2 \}$.

Ces systèmes, étant donné les caractéristiques de *L* et les règles de passage qui ont permis l'obtention de *L′*, ont des caractéristiques intéressantes. D'une part, étant donné un symbole et son appartenance à un système, on sait quel est le système qui doit suivre; par exemple, à partir de A'_1 du système 2, ne peut suivre que le système 1. D'autre part, si les *f* de chaque symbole dans chaque système sont différentes de celles des autres symboles, il n'existe pas de *f* conditionnelles entre les symboles d'un système et les symboles du système qui suit. Par exemple, R'_3 et B'_3 dans le système 3 et R'_2, T'_2 et D'_2 dans le système 4 qui suit, ont chacun une *f* propre, mais *f* de R'_2 à partir de R'_3 est la même que *f* à partir de B'_3 ($f(R'_2 \mid R'_3 = f(R'_2 \mid B'_3)$). Ceci implique que l'on peut aboutir à la codification optimale de *L′*, c'est-à-dire la codification permettant l'attribution du nombre plus réduit possible de *sb* à chaque symbole, à partir de la connaissance de *f* de chaque symbole dans chaque système, soit en tout, dans le cas considéré, 15 valeurs ou symboles codifiants[3].

Le langage *L′* est donc susceptible d'être codifié de manière optimale à partir de la connaissance de 15 valeurs. Comme *L′* est obtenu par règle à partir de *L*, que d'autres règles permettent d'obtenir *L* à partir de *L′* (par exemple, R'_1 se réécrit R), il est possible de transmettre *L′*, d'appliquer ces règles de passage à sens contraire et de réobtenir *L* au point de destination des messages. Ceci implique que l'on peut obtenir la codification optimale de *L* à partir de la connaissance des 15 valeurs de *L′*.

La valeur optimale de la codification de *L* obtenue moyennant le passage par *L′* est la même que la valeur optimale de la codi-

3. La formule permettant le calcul de H de chaque système est $-\sum_{i=1}^{i=n} f(X_i) \log_2 f(X_i)$, où on retrouve les éléments fondamentaux de celle qui a été rappelée à la note 2, à ceci près qu'ici les *f* ne sont pas conditionnelles.

fication de L obtenue selon l'optique mathématique : on aboutira, dans les deux cas, aux mêmes valeurs en *sb* par symbole dans les suites de L[4]. La différence — et elle est grande — entre les deux optiques résulte cependant du « coût » dans le fonctionnement et la constitution de deux systèmes de codification. Dans le cas de l'optique mathématique, le système de fonctionnement devrait tenir compte de la valeur d'information apportée par 118 symboles codifiants en autant de trigrammes différents, alors que le système de fonctionnement dans l'optique linguistique n'implique que 15 symboles codifiants groupés en 4 systèmes. La constitution du système de codification, dans l'optique mathématique, implique l'obtention, au préalable, de presque un millier de magnitudes sur le comportement statistique des symboles et des blocs de symboles dans L. La constitution du système de codification dans l'optique linguistique ne suppose que l'obtention de 15 magnitudes.

Il est clair que le langage L est une illustration simplifiée et idéalisée des situations que l'on rencontre dans les langues naturelles. Les mêmes inconvénients dans l'optique mathématique de codification à propos de L et les mêmes avantages de l'optique linguistique doivent donc apparaître à propos des textes en langues naturelles. Les inconvénients ont déjà été remarqués à plusieurs reprises; signalons que pour obtenir la valeur de 3,1 *sb* par lettre et espace dans un texte anglais, on a dû opérer sur des trigrammes, soit environ 20 000 unités (27^3). Le système de codification devient ainsi vite irréaliste, l'ordre de chaque nouvelle approximation étant la puissance à laquelle il faut élever le nombre d'unités dans l'alphabet pour calculer le nombre d'unités possibles sur lesquelles il faut opérer (une approximation d'ordre 6, par exemple, implique 27^6 hexagrammes possibles, ce qui est déjà une magnitude impraticable). Ainsi on a pu dire, dans un contexte quelque peu différent, que l'optimum dans la codification du langage naturel — c'est-à-dire le nombre le plus réduit par lettre

4. Il est possible de prouver sur un plan général cette affirmation (cf. Gabriel G. Bès, *Un théorème sur l'équivalence de la valeur H entre langages*, Clermont, Institut de Linguistique, 1977, dactylographié). Plus précisément, H(L). Nombre de symboles de L = H(L'). Nombre de symboles de L', à condition que les règles de passage de L à L' établissent une relation bijective entre ces deux langages. En revanche, s'il est vrai que *dans le cas de l'exemple,* il n'existe pas des *f* conditionnelles entre les symboles appartenant aux systèmes de L', les conditions générales de ce fait n'ont pas été établies. On ne connaît donc pas exactement si dans le cas des langues naturelles il existe des conditions entièrement analogues à celles de l'exemple factice sur les rapports de L avec L', bien que les résultats empiriques déjà obtenus signalent l'existence des conditions très proches.

et espace permettant une transmission sûre — est un Eldorado mathématique : on sait qu'il existe, mais on ne peut pas l'atteindre[5].

Aujourd'hui nous sommes en mesure d'affirmer que, si ce n'est pas dans les proportions de l'exemple factice du langage L, les avantages de l'optique linguistique de codification existent pourtant aussi pour les textes en langue naturelle, en particulier pour des textes en espagnol, où elle a été appliquée de manière expérimentale. Suivant la même démarche que celle qui a été illustrée à propos du langage *L*, on a obtenu un système de codification permettant d'assigner entre 3,03 et 3,08 *sb* à chaque lettre et espace du texte original. Ce résultat a été acquis moyennant 273 symboles codifiants différents, groupés en 14 systèmes.

Le *Code de réduction alphabétique C*, qui a permis l'obtention de ces valeurs, a été construit de manière entièrement analogue à celle qui a été illustrée pour le langage *L*. Le point de départ a été la description des systèmes des segments phoniques de l'espagnol[6]; par la suite, on s'est donné des règles de passage rendant possible l'obtention d'un langage espagnol *bis*, où les indices permettant de reconnaître les différents systèmes étaient organisés de gauche à droite. Le point central de ces règles de passage est l'éclatement des lettres qui correspondent aux segments vocaliques en deux types : lettres qui apparaissent dans les syllabes ouvertes et lettres qui apparaissent dans les syllabes fermées. Par exemple, le « o » de *tostada* et le « o » de *coraza* deviennent deux symboles différents. Ceci permet de traiter séparément les systèmes consonantiques qui apparaissent dans la position de postnoyau vocalique de ceux qui apparaissent dans la position de prénoyau.

Les résultats obtenus demandent à être, d'une part, vérifiés, et d'autre part, élargis. Le *Code C*, dont nous faisons état ici, n'a été appliqué que de manière manuelle sur un corpus d'environ

5. Cf. Shannon et Weaver (*The mathematical theory...*, p. 108); Pierce (*Symbols, Signals and Noise*, chap. IV, V, VII); George A. Miller et Noam Chomsky, Finitary models of language users, dans R. Duncan Luce, Robert R. Bush et Eugene Galanter [Eds.], *Handbook of Mathematical Psychology*, New York..., 1963, vol. II, p. 427-430, p. 450-456; Jagjt Singh, *Ideas fundamentales sobre la teoría de la información, del lenguaje y de la cibernética*, traduction en espagnol de A. J. Garriga Trillo, Madrid, 1966, chap. 4.

6. Cette description est faite conformément aux propositions présentées dans Gabriel G. Bès, *Identités et différences dans les unités de deuxième articulation* (thèse, Paris, 1972), chap. X, où ont été développées des idées avancées par André Martinet, notamment dans Substance phonique et traits distinctifs, *BSL* 53 (1958), p. 72-85. Le *Code de réduction alphabétique C* (Clermont, Institut de Linguistique, 1977, dactylographié), fait suite aux *Codes de réduction alphabétique A et B*, préparés à Mendoza (1974), ceux-ci en collaboration avec D. E. Guillot.

3 000 lettres et espaces; étant donné les résultats précédents, il est peu probable que la vérification nécessaire sur des corpus plus larges puisse modifier de manière significative les résultats acquis. Mais ces résultats méritent d'être élargis, dans ce sens qu'ils ne constituent nullement la valeur optimale que l'on peut obtenir moyennant une optique linguistique. En effet, les codes de réduction alphabétique déjà préparés n'opèrent qu'en tenant compte des contraintes de la structure phonologique des langues naturelles (constitution des syllabes, des systèmes de phonèmes dans les différents contextes, etc.; cf. les caractéristiques présentées au début sur le langage *L*), et sans les avoir épuisées entièrement. Or, tous les systèmes de la langue (l'organisation des monèmes lexicaux et grammaticaux, la syntaxe, etc.) introduisent des facteurs qualitatifs de redondance. Analyser ces facteurs qualitatifs pour pouvoir opérer sur eux de manière efficace dans la constitution et le fonctionnement de systèmes de codification, ne paraît pas une utopie irréaliste : tout porte à croire que si les données ici présentées — rudimentaires certes, mais, nous les croyons déjà significatives — venaient à se confirmer, un pas important aura été franchi vers l'Eldorado de la codification optimale.

Université de Clermont II.