CONTRIBUTIONS À L'ÉTUDE DE CERTAINES INTERACTIONS INFORMATIONNELLES DANS LE ROUMAIN ÉCRIT

L'INFLUENCE À DISTANCE DES LETTRES

PAR

MIHAI DINU

1. INTRODUCTION

La forme particulière d'interdépendance syntagmatique connue sous le nom d'« influence à distance » s'avère, à une analyse plus attentive, comme un mode de manifestation de la redondance linguistique.

Si l'occurrence d'une unité linguistique entraîne des changements dans le champ de probabilité des unités qui lui succèdent, en limitant les possibilités d'apparition de celles-ci par une sélection avec prééminence de certaines combinaisons privilégiées, il résulte qu'elle nous communique aussi d'une manière implicite quelques informations relatives aux unités qu'elle précède. Ces dernières sont « anticipées », ce qui signifie que leur apparition devient prévisible dans une certaine mesure et comme telles elles présentent une redondance partielle.

Un cas-limite qui nous montre que l'influence à distance est due au caractère redondant de toute expression linguistique est la situation où la possibilité de supprimer l'élément « anticipé » sans affecter le contenu de la communication met en évidence que la détermination syntagmatique stricte est conditionnée par la redondance stricte du terme déterminé. Dans l'anglais écrit par exemple, l'occurrence de la lettre u après un q fournit une information nulle, parce que imposée par une règle d'orthographe sans exception. Dans le contexte q — la présence de la lettre u est donc réclamée avec nécessité, ce qui fait qu'elle soit de surcroît dans la communication.

On remarque que la dépendance informationnelle d'une unité linguistique par rapport à une unité précédente dérive d'un certain gaspillage produit au cours de la codification du message. Dans une langue idéale, d'une parfaite économicité, la notion d'influence à distance perdrait son sens.

La constatation, à la portée de tous, qu'en précisant un terme d'une communication inconnue on facilite la déduction du terme suivant, a pu trouver une expression quantitative seulement à partir de l'introduction dans l'étude de la langue de la notion d'entropie informationnelle due à C. Shannon.

Par analogie avec la thermodynamique où elle exprime le degré de désordre de la cinématique d'un ensemble de molécules, on a défini comme entropie dans la théorie de l'information la mesure quantitative de l'indétermination d'un système d'événements. A tout champ de probabilité on associe un nombre appelé conventionnellement entropie, dont

C.L.T.A., 9, fasc. 2, 211–240, Bucarest, 1972

la valeur dépend exclusivement des probabilités des événements qui composent le champ respectif. En notant les événements par a, b, c,... z, ... et leurs probabilités par p(a), p(b), ..., p(z)..., l'expression de l'entropie sera :

$$H = -p(a) \log p(a) - p(b) \log p(b) - \ldots - p(z) \log p(z)...$$

La valeur de l'entropie est proportionnelle à la difficulté de deviner lequel des événements possibles se produira à un moment donné.

Le roumain écrit, dont nous nous occuperons en ce qui suit, se caractérise par une distribution spécifique des fréquences des lettres. Comme la probabilité de l'apparition d'une lettre dans un texte peut être assimilée à la fréquence relative de la lettre respective dans un grand nombre de textes, à condition que la longueur de ceux-ci soit suffisante pour éliminer les effets perturbateurs de certaines déviations locales par rapport aux lois statistiques de la langue, il résulte qu'on pourra connaître en explorant des documents de langue écrite, le champ de probabilité associé à l'alphabet latin dans son hypostase particulière de système séméiotique destiné à fixer par écrit des messages de langue roumaine.

Une certaine valeur H_1 de l'entropie, appelée *entropie de premier ordre*, est propre à ce champ de probabilité. Si on nous demande de deviner la n-ième lettre d'un texte totalement inconnu, nous aurons à affronter une difficulté dont l'expression quantitative est H_1.

Mais si on reçoit l'indication supplémentaire que la (n—1)-ième lettre est, par exemple, un V, la situation change.

L'occurrence d'une série de lettres (b, c, d, f ...) après un v est incompatible avec les lois de la langue roumaine, et par suite un regroupement des probabilités des lettres a lieu, ce qui se traduit par une diminution de l'entropie parce que les variantes entre lesquelles on a à choisir ont été réduites et les chances d'un pronostic correcte ont augmenté sensiblement.

On peut s'attendre que la (n+1)-ième lettre souffre aussi en quelque sorte l'influence de la lettre connue. Si c'est ainsi, l'entropie qui caractérise cette position différera à son tour par rapport à H_1.

Le nombre total de positions affectées par la connaissance d'une lettre définira le rayon d'action de celle-ci.

On est à présent en mesure de donner une définition plus rigoureuse de l'influence à distance :

On entend par influence à distance d'une unité linguistique l'ensemble des modifications produites par la connaissance de cette unité dans le champ de probabilité des unités environnantes.

Dans la définition ci-dessus n'entre pas la condition que le terme connu précède les termes influencés. En effet, il est à attendre qu'une unité linguistique exerce aussi en sens contraire vers « l'amont » une action semblable à celle manifestée de gauche à droite vers « l'aval ». En reprenant l'exemple précédent, l'entropie de la position qui précède un v connu sera elle aussi plus petite que H_1, étant données les possibilités d'op-

tion plus restreintes, car avant un v, des lettres telles que b, f, j,... ne pourront pas apparaître.

Il résulte qu'en précisant une lettre on produit une dépression d'une longueur plus on moins grande dans le niveau général de l'entropie moyenne d'un texte inconnu, selon que la lettre est plus ou moins « influente ». La redondance de la langue explique, ainsi que l'on a déjà montré, l'existence de cet « entonnoir » informationnel.

2. MÉTHODES «CLASSIQUES» D'ÉTUDE DE L'INFLUENCE À DISTANCE

2.1. La méthode statistique

C. Shannon, le créateur de la théorie de l'information, a saisi dès le début la possibilité d'utiliser le concept d'entropie dans l'analyse de l'influence à distance. On lui doit l'étude qui a posé les fondements de la recherche dans ce domaine [11].

Shannon définit les entropies d'ordre supérieur, conçues comme entropies conditionnées d'événements composés, ainsi qu'il suit :

— l'entropie d'ordre 2 :

$$H_2^T = H_{\alpha 1}(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) = -p(aa)\log p(aa) - p(ab)\log p(ab) - p(ac)\log p(ac) - \ldots - p(az)\log p(az) - p(ba)\log p(ba) - p(bb)\log p(bb) - \ldots - p(zz)\log p(zz) + p(a)\log p(a) + p(b)\log p(b) + \ldots + p(z)\log p(z)$$

— l'entropie d'ordre 3 :

$$H_3 = H_{\alpha 1\alpha 2}(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2) = -p(aaa)\log p(aaa) - p(aab)\log p(aab) - \ldots - p(aaz)\log p(aaz) - p(aba)\log p(aba) - p(abb)\log p(abb) - \ldots - p(abz)\log p(abz) - \ldots - p(zzz)\log p(zzz) + p(aa)\log p(aa) + p(ab)\log p(ab) + \ldots + p(zz)\log p(zz)$$

. . . . . .

— l'entropie d'ordre n :

$$H_n = H_{\alpha 1\alpha 2 \ldots \alpha n-1}(\alpha_n) = H(\alpha_1\alpha_2 \ldots \alpha_{n-1}\alpha_n) - H(\alpha_1\alpha_2 \ldots \alpha_{n-1}) =$$
$$= -p(\underbrace{aa \ldots a}_{n \text{ fois}})\log p(\underbrace{aa \ldots a}_{n \text{ fois}}) - p(\underbrace{aa \ldots ab}_{n-1 \text{ fois}})\log p(\underbrace{aa \ldots ab}_{n-1 \text{ fois}}) - \ldots$$
$$- p(\underbrace{zz \ldots z}_{n \text{ fois}})\log p(\underbrace{zz \ldots z}_{n \text{ fois}}) + p(\underbrace{aa \ldots a}_{n-1 \text{ fois}})\log p(\underbrace{aa \ldots a}_{n-1 \text{ fois}}) +$$
$$+ p(\underbrace{aa \ldots ab}_{n-2 \text{ fois}})\log p(\underbrace{aa \ldots ab}_{n-2 \text{ fois}}) + \ldots p(\underbrace{zz \ldots z}_{n-1 \text{ fois}})\log p(\underbrace{zz \ldots z}_{n-1 \text{ fois}})$$

La signification linguistique de l'entropie H_k pour les lettres est la suivante : l'entropie d'ordre k reflète la difficulté de deviner la lettre qui continue un texte alors qu'on connaît la fréquence dans la langue écrite de toutes les combinaisons de 2, 3, ... k−1, k lettres.

Il est evident que la rangée des entropies d'ordre supérieur H_1, H_2, ..., H_k, ..., H_n, ... ne peut être que non-croissante, parce que la connaissance de plusieurs restrictions combinatoires, concernant les séquences de lettres, constitue une aide supplémentaire dans l'indication correcte de l'élément inconnu et entraîne donc après soi une diminution (ou tout au plus le maintien constant) de l'indétermination. On aura :

$$H_1 \geqslant H_2 \geqslant H_3 \geqslant \ldots \geqslant H_n \geqslant \ldots$$

D'autre part, pour n'importe quel n, H_n ne peut être que positif et non-nul ($H_n > 0$), parce que même si on connaissait toutes les lois statistiques d'une langue, on ne pourrait jamais prévoir la suite de n'importe quelle phrase possible de la langue respective.

Or, conformément à un théorème connu, une rangée infinie, monotone, non-croissante, supérieurement bornée est convergente. On aura donc un n tel que pour une valeur ε quelque petite qu'elle soit : $H_{n+1} - H \leqslant \varepsilon$. Au-delà de cet n l'indétermination ne diminue plus, ce qui signifie que l'information fournie par la connaissance d'une lettre n'est plus en mesure de mener à une diminution de l'entropie, même si on connaît les probabilités des combinaisons de plus de n lettres. On obtient ainsi la valeur moyenne n du rayon d'influence d'une lettre.

Mais l'application de cette méthode se heurte à la difficulté du calcul des entropies d'ordre supérieur, opération qui exige l'inventaire d'un nombre immense de combinaisons (digrammes, trigrammes, ... n-grammes), nombre qui augmente exponentiellement avec l'ordre de l'entropie recherchée.

Fondé sur des statistiques existantes, Shannon a réussi à déterminer pour l'anglais les valeurs $H_2 = 3{,}32$ et $H_3 = 3{,}10$ (ces deux valeurs ainsi que toutes celles auxquelles nous nous rapporterons plus loin sont exprimées en bits = unités binaires d'information) et, en utilisant certaines inventaires des fréquences des mots, il a donné une évaluation à caractère plutôt estimatif des entropies d'ordre 5 et 8. Comme $H_5 \neq H_8$, il résulte qu'en anglais le rayon moyen d'action d'une lettre est en tout cas plus grand de 5 lettres.

Ce résultat représente à peu près tout ce qu'on peut obtenir par la méthode statistique. Malgré les avantages indiscutables de l'automatisation du travail d'inventaire et de classement des combinaisons de lettres, on entrevoit peu de chances de pouvoir calculer, même dans une perspective éloignée, les valeurs des entropies d'ordre plus grand que 8. Il ne faut pas oublier aussi que si le nombre des séquences qu'on doit inventorier est pour de telles entropies au moins de l'ordre des trillions, les textes utilisés doivent contenir, pour donner une image quelque fidèle qu'elle soit du phénomène, un nombre de signes incomparablement plus grand.

On peut donc considérer que la méthode statistique dans la forme conçue par Shannon a épuisé ses possibilités dès son apparition et n'a

pas des chances de résoudre convenablement dans un proche avenir le problème de l'influence à distance.

2.2. La méthode des tests de prédiction

La poursuite des recherches dans ce domaine était conditionnée par l'obtention d'un moyen d'éluder la statistique. Il fallait élaborer une méthode qui permette le calcul des entropies d'ordre supérieur, sans déterminer au préalable les probabilités de toutes les combinaisons de lettres.

S'il n'est pas possible, pour le moment, d'expliciter toutes les lois statistiques qui gouvernent l'enchaînement des lettres dans un texte, on peut cependant utiliser la connaissance implicite de ces lois, propre à ceux qui parlent la langue respective.

L'idée appartient toujours à Shannon qui, dans la même étude consacrée à l'entropie de la langue anglaise [11], a mis les bases d'une méthode qui ne fait pas appel à la statistique, mais à la compétence linguistique de ceux qui parlent.

La méthode est fondée sur le test suivant : on met à la disposition du sujet de l'expérience un fragment de texte contenant $(n-1)$ lettres et on lui demande de deviner la n-ième lettre. L'épreuve se répète jusqu'à ce qu'il indique correctement cette lettre. Ensuite le test continue avec l'essai d'identifier la $(n+1)$-ième lettre et ainsi de suite. A mesure que la longueur du texte connu augmente, il est évident que les chances d'une prédiction correcte grandissent parce que la quantité d'information dont dispose le sujet est aussi plus grande et elle lui permet de s'orienter plus facilement quant à l'évolution ultérieure du texte. L'indétermination diminue et avec elle décroît aussi l'entropie de l'expérience. Mais ce processus ne peut pas continuer indéfiniment, car alors il existerait la possibilité qu'en lisant la première moitié d'un livre on puisse prévoir avec exactitude tout le texte de la seconde moitié, ce qui assurément n'arrive jamais. Il existe donc une frontière à partir de laquelle l'entropie de l'expérience reste pratiquement constante, limite qui marque la longueur moyenne du rayon d'influence de cette lettre dans la langue respective.

A partir des probabilités $q_i^{(n)}$ d'indication correcte de la n-ième lettre à la i-ième tentative, on peut approximer les entropies d'ordre supérieur grâce aux inégalités suivantes, établies par Shannon :

$$2(q_2^{(n)} - q_3^{(n)}) \log_2 2 + 3\,(q_3^{(n)} - q_4^{(n)}) \log_2 3 + \ldots + (k-1)$$

$$(q_{k-1}^{(n)} - q_k^{(n)}) \log_2 (k-1) + k q_k^{(n)} \log_2 k \leqslant H_k \leqslant - q_1^{(n)} \log_2 q_1^{(n)} -$$

$$- q_2^{(n)} \log_2 q_2^{(n)} - \ldots - q_k^{(n)} \log_2 q_k^{(n)}$$

Ainsi que le montrent les formules ci-dessus, ce qu'on obtient n'est pas la valeur exacte de l'entropie mais une limitation inférieure et supérieure du domaine auquel elle appartient. L'intervalle ainsi déterminé correspond à la réalité seulement dans l'hypothèse où celui qui devine indique toujours la lettre qui devrait immédiatement suivre en tenant compte de toutes les lois statistiques de la langue. Toute déviation par

rapport à ces lois mène à une extension artificielle de la limite du domaine.

Par la méthode des tests de prédiction, Shannon a calculé les entropies H_1, H_2, ..., H_{15} et H_{100}, ce qui lui a permis d'apprécier que l'entropie limite H_∞ est située, pour l'anglais, entre 0,66 et 1,33 bits, d'où on peut déduire que la redondance $R = 1 - \frac{H_\infty}{\log_2 27}$ a l'ordre de grandeur d'à peu près 80%.

Burton et Licklieder [2], en reprenant les tests de type Shannon sur un matériel plus riche, ont réussi à trouver des valeurs d'une précision supérieure pour les entropies H_1, H_2, H_4, H_8, H_{16}, H_{32}, H_{64}, H_{128} et $H_{10\,000}$. Ils ont constaté que $H_{32} \approx H_{64} \approx H_{128} \approx H_{10\,000}$, en tirant la conclusion que le rayon moyen d'action d'une lettre doit être d'environ 30 lettres. L'entropie limite conduit à une valeur de la redondance située entre 66% et 80%, du même ordre de grandeur que celle établie antérieurement pour l'allemand par K. Kupfmüller [7] et que celle qu'on a calculée ultérieurement pour le suédois par H. Hansson [5]. A partir de ces dates I.M. Iaglom et A.M. Iaglom [6] ont émis l'hypothèse qu'au moins pour les langues qui utilisent l'alphabet latin, le rayon moyen d'influence d'une lettre doit être d'environ 30 lettres.

Une contribution intéressante en vue d'augmenter la précision des tests de prédiction a été apportée par le mathématicien Kolmogorov qui a élaboré une méthodologie de travail qui permet le calcul de l'erreur commise inévitablement dans la détermination par voie expérimentale des entropies d'ordre supérieur [6] .

Le problème de la précision des résultats une fois résolu de manière satisfaisante, le seul côté faible de la méthode des tests de prédiction reste son caractère global. Elle nous donne une image d'ensemble de l'influence à distance mais elle ne peut pas offrir en même temps une description du mécanisme de la transmission de l'information linguistique le long du texte.

Comme on l'a montré dans l'introduction, à l'origine de ce transport d'information se trouvent les contraintes imposées par la langue et par l'orthographe à la concaténation des lettres en digrammes, trigrammes, tétragrammes, etc.

L'existence de ces restrictions à caractère combinatoire facilite la prédiction des lettres inconnues. Chaque combinaison a sa contribution dans la propagation à distance de l'information.

Si deux lettres quelconques x_i et x_j apparaissent dans la langue écrite avec les probabilités $p(x_i)$ et $p(x_j)$ et si la probabilité d'apparition du digramme $x_i\,x_j$ est $p(x_ix_j) = p(x_i) \cdot p(x_j)$, c'est-à-dire si l'événement composé x_ix_j se comporte comme si les événements qui le constituent seraient indépendants, il est évident que le digramme x_ix_j ne contribue d'aucune manière à la transmission à distance de l'information fournie par la lettre x_i, parce qu'il n'introduit aucune condition supplémentaire qui aide à deviner les lettres suivantes. Dans un pareil cas, si l'influence de la lettre x_i se manifeste pourtant plus loin dans le texte, le véhicule de cette information n'est pas le digramme x_ix_j mais bien une autre combinaison de lettres grevée par une certaine restriction linguistique. Mais pour éclairer l'apport spécifique des différents types de séquences de let-

tres dans la propagation de l'information, les tests de prédiction ne peuvent être d'aucune utilité. Quand le sujet de l'expérience indique correctement la n-ième lettre d'un texte dont (n−1) lettres sont connues, ni lui et ni l'expérimentateur ne sont en mesure de nous dire en quelle proportion cette réussite est due à l'existence de certaines formations privilégiées de 3, 4, 5, ... n, ... lettres, de la probabilité d'apparition desquelles il est à supposer que le sujet a tenu compte inconsciemment.

Pour suppléer à cette lacune, on propose en ce qui suit un modèle mathématique capable de saisir la contribution séparée des digrammes, des trigrammes, ..., des n-grammes, dans la transmission à distance de l'influence d'une lettre.

Ce modèle permet d'évidencier certaines particularités de la communication par l'intermédiaire de la langue, qui ont échappé jusqu'à présent aux recherches effectuées à l'aide des méthodes « traditionnelles ».

3. UN MODÈLE MARKOVIEN POUR L'ÉTUDE DE L'INFLUENCE À DISTANCE

3.1. Processus et chaînes Markov

On se limite ici seulement aux définitions des éléments indispensables pour comprendre le modèle mathématique proposé, en sélectant de la théorie très ample des processus Markov uniquement les aspects qui intéressent strictement la présente étude. Pour des détails et des précisions supplémentaires on peut consulter, par exemple [4].

Soit un système S avec n états possibles : S_1, S_2, ... S_n. A des moments déterminés le système S passe brusquement d'un état à l'autre. On doit préciser que l'ordre dans lequel se succèdent les états n'a aucune liaison avec leur numérotage qui est absolument arbitraire.

Le long de la « vie » du système tout état peut succéder à n'importe quel état. Mais la probabilité du passage de l'état S_i à l'état S_j a une valeur strictement déterminée p_{ij}. Dans le cas particulier $p_{ij} = 0$, les deux états ne peuvent pas succéder directement l'un à l'autre dans l'ordre S_i, S_j, mais l'ordre inverse peut être valable (au cas où $p_{ji} \neq 0$).

Une image d'ensemble des possibilités du système nous est donnée par la représentation des probabilités de passage d'un état à l'autre sous la forme d'un tableau rectangulaire que nous appellerons *matrice de passage* et qui a l'aspect suivant :

$$[\tau] = \left\| \begin{array}{llll} p_{11} & p_{12} & p_{13} \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} \cdots & p_{2n} \\ p_{31} & p_{32} & p_{33} \cdots & p_{3n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ p_{i1} & p_{i2} & p_{i3} \cdots & p_{in} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ p_{n1} & p_{n2} & p_{n3} \cdots & p_{nn} \end{array} \right\|$$

On remarque que le premier indice de tous les éléments d'une ligne est le même. La i-ième ligne reflète les chances que de l'état S_i le système passe à n'importe quel autre état S_1, S_2, ..., S_n. Comme l'un de ces n événements doit en tout cas se produire, la somme des éléments de chaque ligne sera nécessairement égale à 1 :

$$(1) \qquad \sum_{j=1}^{n} p_{ij} = p_{i1} + p_{i2} + \ldots\ldots + p_{in} = 1$$

La succession des états parcourus par le système S le long d'une période déterminée constitue une chaîne Markov. Si la matrice de passage ne souffre pas de modifications dans le temps, on dit que la chaîne Markov résultée est *stationnaire*.

Supposons qu'à un moment donné t, il existe la probabilité $p_1(t)$ que le système se trouve dans l'état S_1, la probabilité $p_2(t)$ qu'il se trouve dans l'état S_2, la probabilité $p_n(t)$ qu'il se trouve dans l'état S_n. On dira que la phase respective du processus est caractérisée par le vecteur :

$$V_0 = | \, p_1(t) \;\; p_2(t) \;\; p_3(t) \;\ldots\; p_n \, (t) |$$

Il va de soi qu'on aura de même :

$$\sum_{i=1}^{n} p_i(t) = p_1(t) + p_2(t) + \ldots + p_n(t) = 1$$

parce que le système S doit se trouver en tout cas dans l'un des n états possibles.

Au moment suivant (t+1), la situation du système sera décrite par un autre vecteur, différent en général du premier :

$$V_1 = |p_1(t+1) \quad p_2(t+1) \quad p_3(t+1) \ldots p_n(t+1)|$$

Les éléments de ce nouveau vecteur s'obtiennent par la multiplication conformément aux règles de l'analyse matricielle [1] du vecteur V_0 par la matrice $[\tau]$.

$$V_1 = V_0 \cdot [\tau]$$

De manière semblable on calcule le vecteur qui caractérise le moment (t+2) :

$$V_2 = V_1 \cdot [\tau] = V_0 \cdot [\tau]^2$$

et ainsi de suite.

Certains processus Markov jouissent de la propriété de tendre après un temps suffisamment long vers une situation probabiliste stable nommée *régime permanent* ou *régime limite*, pour laquelle les vecteurs qui correspondent à deux moments successifs deviennent identiques. Il résulte qu'à partir d'un certain seuil, le développement du processus ne dépend plus

de l'état initial du système. Cette propriété qui appartient en fait à la matrice de passage, s'appelle *ergodicité.*

La notion d'ergodicité est fondamentale pour notre étude, parce que l'extinction le long du texte de l'onde d'information générée par la connaissance d'une lettre est un phénomène typiquement ergodique.

3.2. La langue en tant que processus Markov

Imaginons à présent un processus Markov dans lequel les états du système soient les lettres de l'alphabet (le blanc, c'est-à-dire l'intervalle entre les mots est traité à son tour comme une lettre.

Tout texte écrit à l'aide de cet alphabet pourra alors être considéré comme une chaîne Markov. Le mot *lac* représentera, par exemple, le résultat d'un processus où le système S a passé successivement de l'état *l* à l'état *a* et de l'état *a* à l'état *c*.

Envisagés sous ce point de vue tous les mots, toutes les phrases et tous les livres écrits dans des langues qui utilisent le même alphabet ne sont que des chaînes Markov générées par des systèmes dont l'ensemble des états est le même. Ce qui diffère d'une langue à l'autre c'est seulement l'aspect de la matrice de passage. Celle-ci est constituée par les probabilités d'apparition dans la langue respective de tous les digrammes, parce que le passage de l'état x_i à l'état x_j équivaut à l'occurrence dans le texte de la paire de lettres $x_i x_j$.

Supposons qu'un texte commence par la lettre *c*. Le vecteur correspondant au moment initial du processus contiendra les éléments

$$p_a(0) = 0,\ p_b(0) = 0,\ p_c(0) = 1,\ p_d(0) = 0,\ p_e(0) = 0\ ,\ldots,\ p_z(0) = 0$$

$p_{\text{blanc}}(0) = 0$; il aura donc la forme :

$$V_0 = |0\ 0\ 1\ 0\ 0\ \ldots\ldots\ .0\ 0\ |$$

On remarque que, étant donné que l'événement *c* se produit avec certitude, la probabilité de tous les autres événements est nulle.

Quelle lettre pourra succéder à ce *c* initial?

Si on dispose d'une statistique des digrammes de la langue respective il ne reste qu'à multiplier le vecteur V_0 par la matrice de passage $[\tau]$ construite en partant de cette statistique et on obtiendra :

$$V_1 = V_0 \cdot [\tau] = |\,p_a(1)\quad p_b(1)\quad p_c(1)\ \ldots p_z(1)\,p_{\text{blanc}}(1)\,|$$

qui exprime la manière dont peut continuer le texte respectif, avec la probabilité de chacune des variantes possibles.

Par l'application répétée du même procédé on pourra s'éloigner indéfiniment de la position initiale. Il est à attendre qu'à un moment donné, dû au caractère ergodique de la langue, on ait pour un i suffisamment grand :

$$V_i = V_{i+1}$$

ce qui signifie que les probabilités d'apparition des lettres dans la position i ne dépendent plus de la lettre initiale. La valeur i est justement le rayon d'action de cette lettre.

On obtient une idée plus précise concernant le mécanisme de la transmission à distance de l'information fournie par la lettre connue si on calcule à chaque pas l'entropie du processus. Les modifications survenues dans la structure des vecteurs (constitués pour l'alphabet latin par presque 30 éléments) ne peuvent être que difficilement suivies et interprétées. En échange, la valeur de l'entropie propre à chaque vecteur offre une image synthétique et suggestive en même temps de la manière dont l'information se propage le long du texte.

L'entropie initiale $H(V_0)$ est nulle, parce qu'elle se rapporte à une situation parfaitement déterminée. On s'attend que $H(V_1) > H(V_0)$, puisqu'on ne connaît pas la lettre finale avec certitude mais on la suppose seulement, fondé sur la fréquence des paires des lettres. Il paraît normal que l'entropie augmente d'autant plus qu'on s'éloigne de la lettre connue et qu'elle tende, avec l'extinction de l'information initiale, vers une valeur limite qui n'est autre que l'entropie d'ordre 1 de la langue $H(V_\infty) = H_1$. On verra ci-dessous que cette variation de l'entropie n'est pas monotone (il existe aussi des moments où l'entropie diminue) et justement cette constatation représente un des résultats importants obtenus à l'aide de la méthode que nous proposons.

On a mentionné sous (1.) qu'il existe aussi un transport d'information dirigé vers « l'amont » qui s'accomplit de manière semblable à la propagation de l'influence à distance vers « l'aval » de la lettre considérée. Il est facile à remarquer que pour l'analyse de cet aspect du problème, il sera suffisant d'utiliser une matrice de passage obtenue à partir de la première par l'échange réciproque des lignes et des colonnes. Mais il est nécessaire que les lignes nouvellement obtenues soient modifiées de sorte qu'on réalise la condition (1) du point (3.1.). A cette fin, chaque élément de la matrice devra être divisé par la somme de la ligne qui le contient.

3.3. Quelques résultats concernant le roumain écrit

L'application à la langue roumaine écrite du modèle préconisé s'est heurtée à une difficulté : dans les statistiques existantes, résultées de l'analyse de certains textes plus riches, le dénombrement des digrammes ou des paires de phonèmes n'est pas complet.

Ainsi Octavian Tocaciu [12], qui a analysé des textes comprenant 442.730 lettres, a élaboré une matrice des fréquences des digrammes à l'intérieur des mots, qui ne tient pas compte de l'occurrence du blanc et n'est donc pas utilisable de notre point de vue. Alexandra Roceric-Alexandrescu dans son consistant travail de phonostatistique de la langue roumaine [10] prend en considération parmi les paires de phonèmes seulement les groupes vocaliques et consonantiques, sans examiner aussi les combinaisons mixtes voyelle-consonne et consonne-voyelle.

C'est pourquoi nous avons été obligés de faire appel à une statistique à dimensions plus modestes, visant certains textes poétiques emprun-

tés à l'œuvre de M. Eminescu et T. Arghezi, travail dû à E. Nicolau, C. Sala et Al. Roceric [9]. Ces auteurs ont dressé trois tableaux comprenant les probabilités d'apparition des digrammes dans :

1. — les premières poésies de M. Eminescu (parues en 1866) ;
2. — le poème « Hypérion » du même auteur ;
3. — quelques poésies (les titres ne sont pas précisés dans l'étude) de T. Arghezi.

Estimant que la langue des poésies de T. Arghezi, chronologiquement plus rapprochées de nous, reflète mieux la situation actuelle de la langue roumaine, nous avons opté pour l'utilisation de la troisième matrice. Mais en vérifiant le caractère stochastique de cette matrice on a constaté que dans la forme publiée se sont glissées certaines erreurs. Les lignes qui correspondent aux lettres b, î, n, ş, u, v et au blanc donnent des sommes différentes par rapport à 1 et elles sont donc incorrectes.

Nous avons préféré, au lieu de modifier arbitrairement certains éléments pour satisfaire la condition imposée, de remplacer entièrement les lignes en cause par les lignes homonymes de la matrice dressée pour le poème « Hypérion », étant donné que, faute de mieux, les probabilités de passage présentes là-bas respectent en quelque mesure les lois de fréquence de la langue. Mais pour la lettre b et pour le blanc, la matrice de « Hypérion » s'est avérée elle-même incorrecte et il a fallu recourir à la première statistique, celle des poésies publiées par M. Eminescu en 1866.

Toutes ces modifications ont altéré bien sûr le résultat de l'étude. Pourtant nous montrerons plus loin que si les conclusions qui découlent de l'analyse entreprise ne sont plus totalement valables pour aucun des textes qui ont servi à l'élaboration des matrices de passage considérées, en échange, elles ne contredisent pas l'esprit de la langue et concordent en bonne partie avec ce que l'on savait par des investigations antérieures concernant les propriétés statistiques de la langue roumaine.

Le point de départ a été donc une matrice à 26 lignes et tout autant de colonnes, qui correspondent aux 26 lettres rencontrées dans les textes, à savoir : a, ă, b, c, d, e, f, g, h, i, î, j, l, m, n , o, p, r, s, ş, t, ţ, u, v, z et le blanc. (On remarque que les textes analysés ne comprennent pas les lettres k, w, x et y, présentes pourtant, il est vrai avec des fréquences très réduites, dans l'inventaire de la langue roumaine écrite).

Pour effectuer les calculs, dont le volume est considérable on a utilisé un ordinateur.

L'influence de chaque lettre prise à part a été étudiée selon le schéma exposé. On prenait comme point de départ un vecteur stochastique avec tous les éléments nuls, exception faite pour celui afférent à la lettre choisie.

Au moyen des multiplications successives par la matrice de passage, on déterminait les phases suivantes du processus, en calculant aussi l'entropie qui leur correspond. Pour contrôler le bon fonctionnement de l'algorithme, on vérifiait pour chaque pas de calcul le caractère stochastique du vecteur résulté.

Le tableau 1 synthétise les résultats des calculs effectués. Dans la première colonne se trouvent inscrites les lettres dont nous nous sommes

Tableau 1

	1	2	3	4	5	6	7	8	6	10
a	3,518	4,162	4,027	4,151	4,119	4,114	4,124	4,120	4,120	4,120
ă	2,729	4,355	3,973	4,093	4,152	4,106	4,122	4,123	4,119	4,120
b	3,142	3,873	4,193	4,116	4,106	4,129	4,119	4,120	4,121	4,120
c	3,132	3,724	4,198	4,075	4,122	4,127	4,116	4,121	4,121	4,120
d	2,353	3,548	4,292	4,047	4,107	4,140	4,112	4,121	4,122	4,120
e	3,008	4,358	4,016	4,078	4,154	4,108	4,120	4,123	4,119	4,121
f	3,078	3,677	4,247	4,092	4,101	4,135	4,116	4,120	4,122	4,120
g	3,164	3,872	4,176	4,118	4,106	4,128	4,119	4,120	4,121	4,120
h	2,441	3,509	4,269	4,076	4,099	4,138	4,114	4,120	4,122	4,120
i	3,000	4,285	4,087	4,068	4,148	4,114	4,118	4,123	4,120	4,120
î	1,205	3,420	4,149	4,124	4,103	4,129	4,119	4,120	4,121	4,120
j	1,923	3,816	4,165	4,059	4,138	4,122	4,116	4,123	4,120	4,120
l	3,177	4,089	4,169	4,053	4,140	4,121	4,116	4,123	4,120	4,120
m	3,288	3,836	4,181	4,072	4,126	4,126	4,116	4,122	4,121	4,120
n	3,232	4,149	4,108	4,112	4,128	4,118	4,120	4,121	4,120	4,120
o	3,709	4,070	4,029	4,160	4,114	4,115	4,124	4,119	4,120	4,121
p	3,339	3,772	4,224	4,106	4,102	4,132	4,117	4,120	4,121	4,120
r	3,557	3,747	4,243	4,080	4,109	4,133	4,115	4,121	4,121	4,120
s	3,319	3,957	4,212	4,095	4,114	4,129	4,117	4,121	4,121	4,120
ş	1,564	3,470	4,252	4,111	4,077	4,142	4,117	4,118	4,122	4,120
t	2,880	4,066	4,181	4,049	4,139	4,123	4,115	4,123	4,120	4,120
ţ	1,441	3,162	4,299	4,071	4,079	4,146	4,113	4,119	4,123	4,120
u	3,332	4,094	4,058	4,144	4,114	4,119	4,122	4,120	4,120	4,121
v	2,563	3,638	4,250	4,082	4,102	4,136	4,115	4,120	4,122	4,120
z	2,692	3,531	4,279	4,074	4,098	4,139	4,114	4,120	4,122	4,120
blanc	4,142	3,811	3,935	4,204	4,094	4,115	4,128	4,117	4,121	4,121

proposé d'étudier l'influence. Les colonnes 1÷10 comprennent les entropies qui caractérisent les positions suivantes.

Par exemple, l'entropie des lettres qui peuvent apparaître en première position après un j est de 1,923 bits. En seconde position de 3,816 bits, en troisième de 4,165 bits et ainsi de suite.

Les conclusions plus importantes, du point de vue linguistique, résultées de l'examen du tableau 1, pourraient être formulées comme il suit :

1° — L'ergodicité de la chaîne Markov est évidente. On remarque qu'indifféremment de l'état de départ, le régime limite s'établit après environ 10 pas. Autrement dit, une lettre influence les probabilités d'apparition des autres 9 lettres, mais à partir de la dixième lettre cette influence s'éteint.

On pourrait remarquer ici que la frontière à laquelle nous nous rapportons est tout à fait relative. Le régime limite étant une situation vers laquelle ont tend asymptotiquement, le moment où nous pouvons considérer qu'il a été atteint dépend de la précision que nous imposons à la coïncidence des entropies. Ainsi du tableau 1, où l'on a pris en considération seulement trois chiffres décimales, il résulterait que pour la lettre a on obtient le régime permanent après huit positions. Mais si nous nous intéressons aux premières six décimales les calculs montrent qu'on rencontre l'entropie limite du texte (4, 120444 bits) seulement dans la 26-ième

position. Considérée à ce niveau de précision, l'influence d'une lettre est donc considérablement plus étendue.

Il résulte qu'on ne peut parler du rayon d'action d'une lettre qu'en fonction d'un certain seuil de proximité par rapport à l'entropie du texte, seuil que nous devrons nous imposer. Si nous considérons, par exemple, que l'influence d'une lettre devient négligeable à partir de la position pour laquelle l'entropie diffère avec moins de 1% en plus ou en moins par rapport à l'entropie de premier ordre du texte (ces limites seraient dans notre cas, $4{,}1204 \pm 0{,}0412 = 4{,}0792 \div 4{,}1616$), les rayons d'action des lettres seront les suivants :

a	ă	b	c	d	e	f	g	h	i	î	j	l
2	3	3	4	4	4	3	3	4	4	2	4	4

m	n	o	p	r	s	ş	t	ţ	u	v	z	blanc
4	1	3	3	3	3	5	4	4	3	3	4	4

Les différences de comportement informationnel sont clairement marquées : par rapport à la lettre n qui n'influence sensiblement que la lettre suivante, le rayon d'action d'un ş est 5 fois plus grand.

2° — L'entropie de premier ordre du texte est d'environ 4,12 bits (voir la colonne 10 du tableau 1). Ce résultat a dans notre cas une importance particulière. Il indique que les substitutions de lignes quelque peu arbitraires que nous avons été obligés d'effectuer dans la matrice de passage choisie ne nous ont pas trop éloignés des répartitions de probabilité qui caractérisent la langue roumaine écrite, pour laquelle des recherches antérieures ont montré que l'entropie de premier ordre est d'environ 4,11 bits.

D'ailleurs les probabilités en régime limite diffèrent assez peu des fréquences moyennes établies par des chercheurs différents [5], [8] et en tout cas ne contredisent pas de manière flagrante l'esprit de la langue.

Elles sont les suivantes :

				t... 4,70%
a...7,17%	e...8,44%	î...2,11 %	o...3,17%	ţ... 0,84%
ă...3,26%	f...1,01%	j...0,24 %	p...2,70%	u... 5,89%
b...0,82%	g...0,72%	l...3,56 %	r...5,13%	v... 1,00%
c...4,50%	h...0,19%	m...2,80 %	s...3,09%	z... 0,52%
d...3,57%	i...8,71%	n...5,60 %	ş...1,46%	blanc...18,81%

La fréquence du blanc nous fournit aussi la longueur moyenne d'un mot. Elle est :

$$L_{mot} = \frac{100}{18{,}81} = 5{,}32 \text{ lettres}$$

3° — Un résultat plus difficile à prévoir à l'avance, et justement pour cela d'autant plus intéressant, se rapporte à l'allure de la courbe de variation de l'entropie en « aval » d'une lettre donnée.

(Pour s'exprimer plus rigoureusement il ne s'agit pas d'une courbe, mais d'un ensemble de points isolés, l'interpolation entre ces points étant dépourvue de sens).

On se serait attendu qu'à partir de l'endroit où se trouve la lettre connue et où l'indétermination du choix entre les variantes est évidemment nulle, l'entropie augmentât continuellement en se rapprochant indéfiniment de la valeur de l'entropie moyenne du texte. L'intuition nous dit que la connaissance précise d'une lettre dans un texte inconnu devrait faciliter la prévision des lettres suivantes jusqu'à une certaine distance, ce qui se traduit en termes informationnels par une baisse de l'entropie, baisse d'autant plus petite que la lettre cherchée est plus éloignée de celle connue.

Mais les résultats des calculs contredisent cette supposition. Ainsi qu'il ressort du tableau 1, après un petit nombre d'intervalles (1 pour le blanc ; 2 pour a, ă, e, i, n ; 3 pour b, c, d, f, g, h, î, j, l, m, p, r, s, ş, t, ţ, v, z ; 4 pour o), on enregistre des valeurs de l'entropie plus élevées que l'entropie moyenne du texte, après quoi on tend vers celle-ci par des oscillations amortisées.

Ici nous est révélé un paradoxe de la langue qui n'avait pas été mis jusqu'à présent en évidence, à savoir que lorsque l'on demande de deviner une lettre dans un texte au sujet duquel nous ne disposons d'aucune information, la connaissance d'une lettre rapprochée de celle recherchée ne facilite pas toujours sa découverte. Au contraire même, l'information reçue peut nous dérouter en rendant l'identification encore plus difficile.

Un exemple éclaircira mieux cet effet :

Si on nous demande de deviner quelle est la lettre qui occupe une certaine position dans un texte que nous ne connaissons pas, la chance de donner une réponse correcte correspond environ à une entropie de 4,12 bits, ce qui équivaut à un choix entre 17,5 variantes également probables. Mais si nous sommes « aidés » en recevant l'information supplémentaire suivante : « la lettre située à deux intervalles à gauche par rapport à la lettre recherchée est un e », nos chances de deviner la réponse exacte baissent.

L'entropie d'une lettre, placée dans la seconde position après un e, est de 4,358 bits (v. le tableau 1) et elle correspond au choix entre un peu plus de 20 variantes également probables.

La difficulté de la réponse a donc augmenté sensiblement.

En nous précisant la lettre auxiliaire, on nous a fourni en fait une « information négative ». La mesure de cette information négative et la différence entre l'entropie moyenne à priori du texte et l'entropie de la position, après avoir reçu l'information supplémentaire (4,120 — 4,358 = = — 0,238 bits). Le tableau 2 comprend les informations « positives » et « négatives » que nous fournit la connaissance d'une lettre pour les dix autres lettres suivantes.

En vue d'une illustration spectaculaire du paradoxe ci-dessus, nous pourrons, en partant du tableau 2, combiner des textes possédant des propriétés étranges.

Soit par exemple le texte roumain suivant, composé de dix lettres (texte que nous pourrions très bien rencontrer dans un article d'électronique) :

— şi — diode — ?

où par — on a noté le blanc.

Quelles sont nos chances pour indiquer correctement la lettre qui suit (la onzième)?

On sait que la lettre cherchée est située immédiatement après un blanc. Cette indication nous procure, comme il résulte du tableau 2, une information de — 0,022 bits. Elle se trouve aussi à deux intervalles après un e, ce qui, ainsi que nous l'avons vu antérieurement, nous donne une nouvelle information négative de —0,238 bits. En allant plus loin à gauche, on remarque qu'absolument toutes les lettres antérieures à celle cherchée ne font que rendre plus difficile l'identification de celle-ci.

Tableau 2

	1	2	3	4	5	6	7	8	9	10
a	0,602	−0,042	0,093	−0,031	0,001	0,006	−0,004	0	0	0
ă	1,391	−0,235	0,147	0,027	−0,032	0,014	−0,002	−0,003	0,001	0
b	0,978	0,247	−0,073	0,004	0,014	−0,009	0,001	0	−0.001	0
c	0,988	0,396	−0,078	0,045	−0,002	−0,007	0,004	−0,001	−0,001	0
d	1,767	0,572	−0,172	0,073	0,013	−0,020	0,008	−0,001	−0,002	0
e	1,112	−0,238	0,104	0,042	−0,034	0,012	0	−0,003	0,001	−0,001
f	1,042	0,443	−0,127	0,028	0,019	−0,015	0,004	0	−0,002	0
g	0,956	0,248	−0,056	0,002	0,014	−0,008	0,001	0	−0,001	0
h	1,679	−0,611	−0,149	0,044	0,021	−0,018	0,006	0	−0,002	0
i	1,120	−0,165	0,033	0,052	−0,028	0,006	0,002	−0,003	0	0
î	2,915	0,700	−0,029	−0,004	0,017	−0,009	0,001	0	−0,001	0
j	2,197	0,204	−0,045	0,061	−0,018	−0,002	0,004	−0,003	0	0
l	0,943	0,031	−0,049	0,067	−0,020	−0,001	0,004	−0,003	0	0
m	0,832	0,284	−0,061	0,048	−0,006	−0,006	0,004	−0,002	−0,001	0
n	0,888	−0,029	0,012	0,008	−0,008	0,002	0	−0,001	0	0
o	0,411	0,050	0,091	−0,040	0,006	0,005	−0,004	0,001	0	−0,001
p	0,781	0,228	−0,104	0,014	0,018	−0,012	0,003	0	−0,001	0
r	0,563	0,373	−0,123	0,040	0,011	−0,013	0,005	−0,001	−0,001	0
s	0,801	0,163	−0,092	0,025	0,006	−0,009	0,003	−0,001	−0,001	0
ş	2,556	0,650	−0,132	0,009	0,043	−0,022	0,003	0,002	−0,002	0
t	1,240	0,054	−0,061	0,071	−0,019	−0,003	0,005	−0,003	0	0
ţ	2,679	0,958	0,179	0,049	0,041	−0,026	0,007	0,001	−0,003	0
u	0,788	0,026	0,062	−0,024	0,006	0,001	−0,002	0	0	−0,001
v	1,557	0,482	−0,130	0,038	0,018	−0,016	0,005	0	−0,002	0
z	1,428	0,589	−0,159	0,046	0,022	−0,019	0,006	0	−0,002	0
blanc	−0,022	0,209	0,185	−0,084	0,026	0,005	−0,008	0,003	−0,001	−0,001

La quantité totale d'information fournie par les lettres précédentes est de — 0,534 bits, donc l'entropie de la lettre cherchée sera de 4,120 + +0,534 = 4,654 bits, ce qui correspond à l'option entre 25 variantes également probables.

Etant donné que l'alphabet utilisé a 25 lettres il résulte que les lettres dont on dispose ont des chances presque égales d'occuper la position respective.

Il s'est produit ainsi une étrange égalisation des probabilités d'apparition des lettres entre lesquelles il existe d'habitude des différences assez importantes (par exemple p(i) = 8,71%, tandis que p(h) = 0,19%, c'est-à-dire les chances de rencontrer un h sont 46 fois plus petites que celles de rencontrer un i).

Voilà combien trompeuse peut être la connaissance partielle d'un texte! Cet aspect que les amateurs de mots croisés ont sans doute remarqué mérite d'être étudié sérieusement parce qu'il intéresse directement certains domaines de recherche, tels que l'épigraphie où il pourrait servir à la réconstitution des textes partiellement détruits.

Considérons maintenant un autre texte, cette fois-ci de 8 lettres :

— deşi — îţ?

où se pose le problème d'identifier la neuvième lettre.

On reprend le raisonnement ci-dessus.

La dernière lettre qui précède celle cherchée est un ţ (on a donc une information de 2,679 bits), l'avant-dernière — un î (0,700 bits) l'antépénultième le blanc (0,185 bits) et ainsi de suite.

On constate que dans ce cas toutes les lettres connues concourent à faciliter notre tâche et elles nous mettent à la disposition une information totale de 3,685 bits. L'entropie de la lettre cherchée sera 4,120—3,685 = = 0,435 bits, c'est-à-dire plus petite même que l'entropie de l'option entre deux variantes également probables. Ce fait ne surprend aucun Roumain, qui remplacera facilement le signe d'interrogation par un i, la seule lettre admissible dans le contexte donné.

Il serait encore à noter que, bien que d'aprés notre connaissance jusqu'à présent le paradoxe décrit plus haut n'ait pas été mis en évidence dans une forme explicite, il existait pourtant un indice potentiel pour l'apercevoir. Il y a déjà quelques années on a constaté [10] qu'en roumain l'entropie des phonèmes au commencement des mots est plus grande que l'entropie moyenne de la langue, ce qui revient à dire que la présence du blanc immédiatement à gauche de la lettre cherchée augmente le désordre du système et rend plus difficile l'identification de celle-ci c'est-à-dire exactement ce qui est exprimé par la valeur 4,142 de la première cassette de la dernière ligne du tableau.

4° — Examinons quelques aspects mathématiques étroitement liés à l'objet de notre étude :

Soit la matrice stochastique $[\tau]$ dont les éléments seront notés ainsi :

$$[\tau] = \begin{Vmatrix} p'(aa)\ p'(ab) \ldots\ldots\ldots\ldots\ p'(az)\ p'(a\ \text{blanc}) \\ p'(ba)\ p'(bb)\ \ldots\ldots\ldots\ p'(bz)\ p'(b\ \text{blanc}) \\ \cdot \\ \cdot \\ \cdot \\ p'(\text{blanc a})\ p'(\text{blanc b})\ \ldots\ p'(\text{blanc z})\ p'\ (\text{blanc blanc}) \end{Vmatrix}$$

où

$$\sum_{\alpha_2=a}^{\text{blanc}} p'(\alpha_1\alpha_2) = 1 \text{ pour tout } \alpha_1 \tag{2}$$

L'occurrence de l'événement $\alpha_1\alpha_2$ suppose la réalisation simultanée de deux événements indépendants :

1. — l'apparition de la lettre α_1 ;
2. — l'occurrence sans intermédiaire de la lettre α_2 à la droite de la lettre α_1.

La probabilité du premier événement est $p(\alpha_1)$ = la probabilité d'apparition dans la langue écrite de la lettre α_1.

La probabilité du second événement est justement $p'(\alpha_1\alpha_2)$. Le théorème de la multiplication des probabilités des événements indépendants fournit :

$$p(\alpha_1\alpha_2) = p(\alpha_1) \cdot p'(\alpha_1\alpha_2) \tag{3}$$

pour tout α_1 et tout α_2.

L'entropie de deuxième ordre par lettre s'exprime par (v. 2.2.) :

$$\mathbf{H}_2 = \mathbf{H}(\alpha_1\alpha_2) - \mathbf{H}(\alpha_1) = -p(aa)\log_2 p(aa) - p(ab)\log_2 p(ab) - \ldots - p(a\ \text{blanc})\log_2 p(a\ \text{blanc}) - p(ba)\log_2 p(ba) - p(bb)\log_2 p(bb) - \ldots - p(b\ \text{blanc})\log_2 p(b\ \text{blanc}) - \ldots - p(\text{blanc blanc})\log_2 p(\text{blanc blanc}) + p(a)\log_2 p(a) + p(b)\log_2 p(b) + \ldots + p(\ \text{blanc})\log_2 p(\text{blanc}) \tag{4}$$

En introduisant l'expression (3) dans (4) on obtient :

$$\mathbf{H}_2 = -p(a)\,p'(aa)\log_2[p(a)\,p'(aa)] - p(a)\,p'(ab)\log_2[p(a)\,p'(ab)] - \ldots - p(a)\,p'(a\ \text{blanc})\log_2[p(a)\,p'(a\ \text{blanc})] - p(b)\,p'(ba)\log_2[p(b)\,p'(ba)] - \ldots - p(\text{blanc})\,p'(\text{blanc blanc})\log_2[p(\text{blanc})\,p'(\text{blanc blanc})] + p(a)\log_2 p(a) + p(b)\log_2 p(b) + \ldots + p(\text{blanc})\log_2 p(\text{blanc}) \tag{5}$$

Notons encore par $E'(\alpha)$ l'entropie de la position qui succède à la lettre connue (la valeur de cette entropie pour chaque α est inscrite dans la colonne 1 du tableau 1, à l'intersection avec la ligne afférente à la lettre α).

On aura :

$$E'(a) = -p'(aa)\log_2 p'(aa) - p'(ab)\log_2 p'(ab) - \ldots - p'(a\ \text{blanc})\log_2 p'(a\ \text{blanc})$$

$$E'(b) = -p'(ba)\log_2 p'(ba) - p'(bb)\log_2 p'(bb) - \ldots - p'(b\ \text{blanc})\log_2 p'(b\ \text{blanc})$$

(6) .
.
.
.

$E'(\text{blanc}) = -p'(\text{blanc a}) \log_2 p'(\text{blanc a}) - p'(\text{blanc b}) \log_2 p'(\text{blanc b}) - \ldots - p'(\text{blanc blanc}) \log_2 p'(\text{blanc blanc})$

Dans l'expression (5) isolons en facteurs communs les valeurs p(a), p(b), ... p(blanc) et décomposons les logarithmes des produits dans des sommes des logarithmes des facteurs composants :

(7) $H_2 = p(a)\,[-p'(aa) \log_2 p(a) - p'(aa) \log_2 p'(aa) - p'(ab) \log_2 p(a) - p'(ab) \log_2 p'(ab) - \ldots - p'(a\ \text{blanc}) \log_2 p(a) - p'(a\ \text{blanc}) \log_2 p'(a\ \text{blanc})] + p(b)\,[-p'(ba) \log_2 p(b) - p'(ba) \log_2 p'(ba) - p'(bb) \log_2 p(b) - p'(bb) \log_2 p'(bb) - \ldots - p'(\text{b blanc}) \log_2 p(b) - p'(\text{b blanc}) \log_2 p'(\text{b blanc})] + \ldots + p(\text{blanc})\,[-p'(\text{blanc a}) \log_2 p(\text{blanc}) - p'(\text{blanc a}) \log_2 p'(\text{blanc a}) - \ldots - p'(\text{blanc blanc}) \log_2 p(\text{blanc}) - p'(\text{blanc blanc}) \log_2 p'(\text{blanc bla}_n\text{c})] + p(a) \log_2 p(a) + p(b) \log_2 p(b) + \ldots + p(\text{blanc}) \log_2 p(\text{blanc})$

En comparant (5) avec (6), on remarque que les sommes des termes pairs situés entre les parenthèses rectangulaires représentent justement les entropies E'(a), E'(b), ... E'(blanc), tandis que depuis les termes impairs on peut isoler en facteurs communs $\log_2 p(a)$ et, respectivement, $\log_2 p(b)$, ... $\log_2 p(\text{blanc})$.

On obtient ainsi :

$$H_2 = p(a)\left[E'(a) - \log_2 p(a) \sum_{\beta=a}^{\text{blanc}} p'(a\beta)\right] + p(b)\left[E'(b) - \log_2 p(b) \sum_{\beta=a}^{\text{blanc}} p'(b\beta)\right] + \ldots + p(\text{blanc})\left[E'(\text{blanc}) - \log_2 p(\text{blanc}) \sum_{\beta=a}^{\text{blanc}} p'(\text{blanc}\ \beta)\right] + p(a) \log_2 p(a) + p(b) \log_2 p(b) + \ldots + p(\text{blanc}) \log_2 p(\text{blanc})$$

Mais, conformément à (2), toutes les sommes notées par Σ dans la formule ci-dessus sont égales à 1. En tenant compte de cette observation

et en décomposant les parenthèses rectangulaires, on trouve :

$$H_2 = p(a)\ E'(a) - p(a)\ \log_2 p(a) + p(b)\ E'(b) - p(b)\ \log_2 p\ (b) + \ldots + p(\text{blanc})\ E'\ (\text{blanc}) - p(\text{blanc})\ \log_2 p(\text{blanc}) + p(a)\ \log_2 p(a) + p(b)\ \log_2 p(b) + p(\text{blanc})\ \log_2 p(\text{blanc})$$

Après avoir effectué toutes les réductions :

$$H_2 = p(a)\ E'(a) + p(b)\ E'(b) + \ldots + p(\text{blanc})\ E'(\text{blanc})$$

L'expression obtenue présente une importance particulière pour notre étude. Elle nous montre que l'entropie d'ordre 2 peut être obtenue comme moyenne pondérée de la première colonne du tableau 1 en prenant comme poids les probabilités des lettres dans l'ordre de la succession normale des lignes. On aura donc :

$$H_2 = 0{,}0717\ .\ 3{,}518 + 0{,}03263\ .\ 2{,}729 + \ldots + 0{,}1881\ .\ 4{,}142 = \\ = 3{,}26 \text{ bits}$$

Evidemment, le résultat obtenu doit être regardé avec une certaine circonspection, étant donnée la manière dont a été construite la matrice $[\tau]$ utilisée plus haut. Cependant, parce que fondée sur un inventaire des digrammes de certains textes concrets de langue roumaine il est assez probable que l'entropie d'ordre 2 de la langue écrite ne s'éloigne pas trop de ce chiffre. L'écart réduit par rapport à la valeur correspondante de l'anglais écrit semble aussi confirmer cette supposition.

Avec les réserves mentionnées, notons que la valeur déterminée ici est pour le moment la seule dont on dispose pour l'entropie de deuxième ordre du roumain écrit. Une seconde valeur sera établie plus loin (3.5) en partant de l'analyse d'un texte de langue littéraire contemporaine qui appartient au poète V. Voiculescu. Ces deux chiffres peuvent constituer une base pour l'étude comparative de la langue roumaine et d'autres langues de circulation pour lesquelles il existe des dates concernant les entropies d'ordre supérieur.

En connaissant H_2 on peut aussi calculer la redondance de deuxième ordre du texte considéré, qui est dans notre cas :

$$R_2 = 1 - \frac{H_2}{\log_2 27} = 1 - \frac{3{,}26}{4{,}70} = 0{,}306 = 30{,}6\%$$

3.4. L'apport des combinaisons de plusieurs lettres dans la transmission à distance de l'information fournie par une lettre

On reprochait plus haut (2.2) à la méthode des tests de prédiction son caractère d'estimation globale de l'influence à distance, qui la rendait inapte de relever la contribution différenciée de chacune des multiples restrictions de nature combinatoire par l'intermédiaire desquelles se réa-

lise la propagation de l'onde d'information le long du texte. Comment écarter cet inconvénient?

Jusqu'ici on n'a pris en considération que l'un des canaux de transmission de l'influence à distance, à savoir l'existence de certaines limitations statistiques dans la combinaison des lettres en digrammes. Mais un rôle important revient aussi aux restrictions concernant les trigrammes, les tétragrammes, etc. dont nous n'avons pas tenu compte plus haut. Conformément à notre modèle, étant donné qu'en roumain des combinaisons telles que br (brad) et rt (cort) sont possibles, on arrive à la conclusion erronée qu'on pourrait enregistrer avec une certaine probabilité aussi l'occurrence du trigramme brt, évidemment inadmissible. Il y a donc des limitations dans la construction des trigrammes qui contribuent elles aussi à faciliter la prognose correcte des lettres suivantes.

C'est ainsi que s'explique aussi la différence considérable entre le rayon moyen d'action d'une lettre calculée plus haut ($\approx$10) et celui établi par Burton et Licklieder ($\approx$30).

Les contraintes imposées aux digrammes ne peuvent expliquer que l'influence exercée sur les premières 9—10 lettres. Plus loin l'information se propage par d'autres voies. Le modèle markovien proposé a la qualité de permettre aussi l'analyse de celles-ci.

Pour l'étude de la contribution des trigrammes, il est suffisant d'inventorier les paires de lettres séparées par une troisième. Une pareille statistique n'est point plus difficile que celle des digrammes et incomparablement plus simple que la statistique des trigrammes, dont le nombre est beaucoup plus grand.

A partir d'une telle statistique on obtient une matrice de passage dans laquelle l'élément p_{ij} représente la probabilité que si l'initiale d'un texte est la lettre i, on rencontre dans la troisième position du texte la lettre j. On définit ainsi un nouveau processus Markov à pas doubles par rapport à celui qu'on à étudié ci-dessus, parce qu'il parcourt les positions du texte deux par deux. Le rayon d'influence déterminé pour le nouveau processus tiendra compte aussi de l'apport des trigrammes.

Le même procédé permet de prendre en considération le rôle des tétragrammes, des pentagrammes et d'autres séquences plus longues. Au lieu de déterminer les fréquences de toutes ces combinaisons dont le nombre est énorme, on aura à inventorier seulement les paires de lettres situées cette fois-ci à des distances progressivement plus grandes les unes par rapport aux autres. Les processus Markov respectifs, à pas de plus en plus grands, fourniront des valeurs croissantes du rayon d'action d'une lettre. Quand la grandeur de ce rayon cesse d'augmenter, on peut avoir la certitude qu'on a atteint la limite de la zone d'influence recherchée.

Cette méthode assure la description détaillée du mécanisme de l'influence à distance en épuisant pratiquement tous les aspects du problème.

Le dernier paragraphe du présent travail comprend les résultats de l'application des procédés décrits plus haut, effectués pour le moment, à titre expérimental, sur un matériel assez restreint.

3.5. L'analyse sur un échantillon de langue roumaine écrite du rôle des digrammes, des trigrammes, des tétragrammes et des pentagrammes dans la propagation de l'influence à distance

Pour illustrer les procédés de travail décrits plus haut (3.4) on a choisi un échantillon de langue littéraire contemporaine d'une longueur de 1000 lettres (y compris les blancs) extrait du récit « La tête de bison » par V. Voiculescu (du volume « Récits I », Bucarest 1966). Le fragment est le suivant :

« Groaznicul uragan care a bîntuit la începutul iernii m-a ţinut ca de obicei cîteva zile în pat, pradă unor cumplite dureri, care mă dezechilibrează totdeauna pînă la anihilare.

Cunosc explicaţia ştiinţifică a mizeriei mele, dar asta nu mă consolează cu nimic. Ştiu că de vină sînt năprasnicele mase de aer ce se grămădesc îngheţate la pol, fierbinte la ecuator şi care, năvălind potrivnice una asupra alteia, prefac văzduhul într-o năzdrăvană uzină de energie. Atunci enormele tensiuni electromagnetice cu care se încarcă lumea ne zdrobesc muşchii şi ne schilodesc nervii.

Atunci zac fără să pot scri, nici citi, nici măcar gîndi. Telefonul stă cu botniţa pusă, clopotele soneriilor sînt trimise să dea de veste în altă parte, la odăile slugilor, nimeni şi nimic nu clinteşte în preajmă-mi. Abia dacă uneori cerc să răsfoiesc ciorne de poeme părăsite ori vechi scrisori.

Într-o astfel de zi malefică, împins de o vrajă răufăcătoare, am desfăcut sicriaşul unei casete cu amintiri de la mama. Din fundul lui s-a răsturnat o [.....] ».

Dans ce texte apparaissent 27 lettres (y compris les blancs) dont les fréquences absolues sont les suivantes :

lettre	a	ă	b	c	d	e	f	g	h	i	î	j	l
occurrences	70	42	7	59	27	94	11	8	7	91	15	2	39

m	n	o	p	r	s	ş	t	ţ	u	v	x	z	blanc
26	64	34	19	56	37	8	46	5	41	10	1	12	169

On a dressé d'abord 4 tableaux qui contiennent les fréquences absolues des paires de lettres :

— contiguës ;
— séparées par une lettre ;
— séparées par deux lettres ;
— séparées par trois lettres ;

A partir de ces tableaux, « stochastisés » en divisant chaque élément par la somme de la ligne à laquelle il appartient, on a obtenu les matrices de passage de quatre processus Markov destinés à l'étude séparée de l'influence des di- tri-, tétra- et pentagrammes.

Par rapport au premier processus qui parcourt le texte position par position, les autres effectuent des pas 2, 3 ou 4 fois plus grands.

Les calculs ont été effectués par un ordinateur.

Les résultats sont systématisés dans les tableaux 3÷10, construits d'après les mêmes principes que les tableaux 1 et 2. Ils expriment d'un côté l'entropie des positions successives après une lettre connue, dans l'hypothèse qu'on tient compte seulement de l'apport des digrammes (tableau 3) ou aussi de celui des trigrammes (tableau 5), des tétragrammes (tableau 7) et des pentagrammes (tableau 9), et d'un autre côté la quantité d'information que la connaissance de la lettre respective fournit relativement aux lettres des positions suivantes (tableaux 4, 6, 8 et 10).

Une première remarque se rapporte à la comparaison des entropies de premier et de deuxième ordre de l'échantillon considéré avec celles des textes précédemment étudiés (3.3). On observe que H_1 a une valeur très rapprochée de celle fournie par le texte « corrigé » des poésies d'Arghezi (4,14 par rapport à 4,12, donc une différence de seulement 0,5%). Les deux valeurs sont supérieures aux valeurs déterminées pour d'autres langues qui utilisent l'alphabet latin (anglais 4,03 ; allemand 4,10, français 3,98 ; espagnol 4,01 ; [8]), ce qui met en évidence le caractère plus équilibré du roumain sous l'aspect des fréquences. Une indétermination plus grande dénote des valeurs plus rapprochées entre elles des probabilités d'apparition des lettres, ce qui, ayant en vue le caractère « phonétique » de notre orthographe est aussi une preuve indirecte de l'harmonie du système phonologique de la langue roumaine.

L'entropie de deuxième ordre résulte : $H_2 = 3{,}16$ bits, d'environ 3% plus petite que celle déterminée précédemment (3,26 bits). Ces deux valeurs se situent sous la valeur correspondante de l'anglais écrit (3,32 bits). Cette inversion du rapport existant entre les entropies de premier ordre des mêmes langues ($H_1^{\text{roumain}} > H_1^{\text{anglais}}, H_2^{\text{roumain}} < H_2^{\text{anglais}}$) met en lumière l'importance informationnelle hors ligne des digrammes roumains. Ceux-ci sont grevés d'un nombre de restrictions combinatoires sensiblement plus grand qu'en anglais et, par conséquent, en connaissant les règles de construction des digrammes on diminue l'indétermination des lettres dans une mesure plus importante. La quantité moyenne d'information fournie par la connaissance des possibilités de combinaisons des lettres en digrammes est pour le roumain $H_1 - H_2 = 0{,}86 \div 0{,}98$ bits par rapport à 0,71 bits en anglais.

Toutes les quatre paires de tableaux confirment notre image antérieure concernant le caractère oscillatoire de la propagation de l'information le long du texte. Nouvelle par rapport aux constatations précédentes est seulement l'observation que l'interférence des ondes d'information produites par les combinaisons de longueur différente peut soit amplifier, soit annuler cet effet. Quelques exemples aideront à mieux comprendre la manière dont a lieu la superposition des oscillations :

1. — Examinons les possibilités de prognose correcte de la quatrième lettre d'un mot.

Si on ne dispose d'aucune date relativement à la fréquence des lettres du roumain écrit, on a à choisir entre 27 variantes également probables. La difficulté d'indiquer correctement la lettre respective est exprimée par la valeur H_0 de l'entropie.

$$H_0 = \log_2 27 = 4{,}76 \text{ bits}$$

Tableau 3

	1	2	3	4	5	6	7
a	3,33	4,19	4,03	4,17	4,14	4,14	4,14
ă	2,55	4,31	4,03	4,13	4,15	,,	,,
b	2,13	3,66	4,15	4,10	4,16	,,	,,
c	3,24	4,00	4,19	4,09	4,15	,,	,,
d	2,11	3,56	4,27	4,09	4.14	,,	,,
e	3,05	4,31	4,08	4,13	4,15	,,	,,
f	2,41	3,58	4,22	4,07	4,16	,,	,,
g	2,50	3,68	4,11	4,15	4,14	,,	,,
h	1,15	3,54	4,20	4,09	4,15	,,	,,
i	3,42	4,21	4,10	4,15	4,14	,,	,,
î	0,70	3,46	3,94	4,21	4,12	,,	,,
l	1,00	3,28	4,20	4,17	4,10	4,15	,,
j	2,64	3,99	4,18	4,08	4,16	4,14	,,
m	2,61	3,73	4,25	4,07	4,15	,,	,,
n	3,39	3,93	4,20	4,13	4,14	,,	,,
o	3,30	3,84	4.05	4,19	4,12	4,15	,,
p	2,71	3,89	4,12	4,13	4,15	4 14	,,
r	3,17	3,98	4,21	4,10	,,	,,	,,
s	3,30	3,92	4,18	4,14	4,13	,,	,,
ş	1,81	3,76	4,16	4,16	,,	,,	,,
t	2,94	4,01	4,20	4,11	4,15	,,	,,
ţ	0,97	3,61	4,21	4,07	4,16	,,	,,
u	3,36	3,97	4,04	4,18	4,13	,,	,,
v	2,52	3,66	4,26	4,10	4,14	4,15	,,
x	0	2,71	3,89	4,12	4,13	,,	,,
z	2,46	3,64	4,20	4,15	,,	4,14	,,
blanc	4,09	3,94	4,04	4,18	,,	,,	,,

Tableau 4

	1	2	3	4	5	6	7
a	+0,81	−0,05	+0,11	−0,03	0	0	0
ă	+1,59	−0,17	+0,11	+0,01	−0,01	0	0
b	+2,01	+0,48	−0,01	+0,04	−0,02	0	0
c	+0,90	+0,14	−0,05	+0,05	−0,01	0	0
d	+2,03	+0,58	−0,13	+0,05	0	0	0
e	+1,09	−0,17	+0,06	+0,01	−0,01	0	0
f	+1,73	+0,56	−0,08	+0,07	−0,02é	0	0
g	+1,64	+0,46	+0,03	−0,01	0	0	0
h	+2,99	+0,60	−0,06	+0,05	−0,01	0	0
i	+0,72	−0,07	+0,04	−0,01	0	0	0
î	+3,44	+0,68	+0,20	−0,07	+0,02	0	0
j	+3,14	+0,86	−0,06	−0,03	+0,04	−0,01	0
l	+1,50	+0,15	−0,04	+0,06	−0,02	0	0
m	+1,53	+0,41	−0,11	+0,07	−0,01	0	0
n	+0,75	+0,21	−0,06	+0,01	0	0	0
o	+0,84	+0,30	+0,09	−0,05	+0,02	−0,01	0
p	+0,43	+0,25	+0,02	+0,01	−0,01	0	0
r	+0,97	+0,16	−0,07	+0,04	−0,01	0	0
s	+0,84	+0,22	−0,04	0	+0,01	0	0
ş	+2,33	+0,38	−0,02	−0,02	+0,01	0	0
t	+1,20	+0,13	−0,06	+0,03	−0,01	0	0
ţ	+3,17	+0,53	−0,07	+0,07	−0,02	0	0
u	+0,78	+0,17	+0,10	−0,04	+0,01	0	0
v	+1,62	+0,48	−0,12	+0,04	0	−0,01	0
x	+4,14	+1,43	+0,25	+0,02	+0,01	−0,01	0
z	+1,68	+0,50	−0,06	−0,01	+0,01	0	0
blanc	+0,05	+0,20	+0,10	−0,04	+0,01	0	0

Tableau 5

	2	4	6	8	10
a	3,89	4,17	4,13	4,14	4,14
ă	3,79	4,06	4,12	,,	,,
b	2,52	3,98	4,12	4,15	,,
c	3,41	4,09	4,15	4,14	,,
d	3,02	4,08	4,17	,,	,,
e	4,24	4,09	4,13	,,	,,
f	2,22	3,91	4,10	4,15	,,
g	2,16	4,24	4,11	4,14	,,
h	1,66	3,97	,,	,,	,,
i	3,91	4,15	4,14	,,	,,
î	2,68	3,97	,,	,,	,,
j	1,00	4,05	,,	4,12	,,
l	3,76	4,08	4,15	4,14	,,
m	3,31	4,01	,,	,,	,,
n	3,51	4,09	,,	,,	,,
o	3,54	4,18	4,14	,,	,,
p	2,82	4,09	4,12	4,15	,,
r	3,63	4,09	4,16	4,14	,,
s	3,36	4,11	,,	,,	,,
ş	2,16	4,15	,,	4,13	,,
t	3,36	4,08	,,	4,14	,,
ţ	2,32	3,94	4,13	,,	,,
u	3,53	4,13	4,14	,,	,,
v	2,92	3,98	,,	,,	,,
x	0,00	3,76	4,08	4,15	,,
z	2,63	4,04	4,14	,,	,,
blanc	3,79	4,20	4,13	4,14	,,

Tableau 6

	2	4	6	8	10
a	+0,25	−0,03	+0,01	0	0
ă	+0,45	+0,08	+0,02	0	0
b	+1,64	+0,16	+0,02	−0,01	0
c	+0,73	+0,05	−0,01	0	0
d	+1,12	+0,06	−0,03	0	0
e	−0,10	+0,05	+0,01	0	0
f	+1,92	+0,23	+0,04	−0,01	0
g	+1,98	−0,10	+0,03	0	0
h	+2,48	+0,17	+0,03	0	0
i	+0,23	−0,01	0	0	0
î	+1,46	+0,17	0	0	0
j	+3,14	+0,09	0	+0,02	0
l	+0,38	+0,06	−0,01	0	0
m	+0,83	+0,13	−0,01	0	0
n	+0,63	+0,05	−0,01	0	0
o	+0,60	−0,04	0	0	0
p	+1,32	+0,05	+0,02	−0,01	0
r	+0,51	+0,05	−0,02	0	0
s	+0,78	+0,03	−0,02	0	0
ş	+1,98	−0,01	−0,02	+0,01	0
t	+0,78	+0,06	−0,02	0	0
ţ	+1,82	+0,20	+0,01	0	0
u	+0,61	+0,01	0	0	0
v	+1,22	+0,16	0	0	0
x	+4,14	+0,38	+0,06	−0,01	0
z	+1,51	+0,10	0	−0,01	0
blanc,	+0,35	−0,06	+0,01	0	0

Tableau 7

	3	6	9	12
a	3,83	4,13	4,15	4,14
ă	3,77	4,17	4,14	,,
b	2,52	4,14	4,15	,,
c	4,02	4,13	4,14	,,
d	3,61	4,17	,,	,,
e	3,63	4,12	,,	,,
f	2,37	4,13	4,17	,,
g	2,75	4,05	4,14	,,
h	2,24	4,00	4,13	,,
i	3,73	4,06	,,	,,
î	3,06	4,07	,,	,,
j	1,00	4,08	4,14	,,
l	3,69	4,13	,,	,,
m	2,98	4,03	,,	,,
n	3,89	4,14	,,	,,
o	3,59	4,08	,,	,,
p	3,37	4,15	4,16	,,
r	3,99	4,13	4,14	,,
s	3,45	4,15	4,13	,,
ş	2,25	4,07	4,12	,,
t	3,98	4,18	4,14	,,
ţ	2,32	4,00	4,13	,,
u	3,64	4,12	,,	,,
v	3,12	4,08	4,15	,,
x	0,00	3,73	4,06	,,
z	3,08	3,99	4,12	,,
blanc	3,76	4,14	4,14	,,

Tableau 8

	3	6	9	12
a	+0,31	+0,01	−0,01	0
ă	+0,43	−0,03	0	0
b	+1,62	0	−0,01	0
c	+0,12	+0,01	0	0
d	+0,53	−0,03	0	0
e	+0,51	+0,02	0	0
f	+1,77	+0,01	−0,03	0
g	+1,39	+0,09	0	0
h	+1,90	+0,14	+0,01	0
i	+0,41	+0,08	+0,01	0
î	+1,08	+0,07	+0,01	0
j	+3,14	+0,06	0	0
l	+0,45	+0,01	0	0
m	+1,16	+0,11	0	0
n	+0,25	0	0	0
o	+0,55	+0,06	0	0
p	+0,77	−0,01	−0,02	0
r	+0,15	+0,01	0	0
s	+0,69	−0,01	+0,01	0
ş	+1,89	+0,07	+0,02	0
t	+0,16	−0,04	0	0
ţ	+1,82	+0,14	+0,01	0
u	+0,50	+0,02	+0,01	0
v	+1,02	+0,06	−0,01	0
x	+4,14	+0,41	+0,08	0
z	+1,06	+0,15	+0,02	0
blanc	+0,38	0	0	0

Tableau 9

	4	8	12	16
a	3,66	4,13	4,14	4,14
ă	3,72	4,21	,,	,,
b	2,52	4,07	4,13	,,
c	3,90	4,13	4,14	,,
d	3,53	,,	,,	,,
e	3,91	4,17	,,	,,
f	2,85	3,96	4,13	,,
g	2,50	4,05	,,	,,
h	2,81	3,87	4,12	,,
i	3,88	4,10	4,14	,,
î	3,06	4,11	,,	,,
j	1,00	3,81	4,20	,,
l	3,66	4,10	4,14	,,
m	3,51	4,20	4,12	,,
n	3,85	4,12	4,14	,,
o	3,47	4,10	4,13	,,
p	3,29	4,14	4,14	,,
r	3,67	4,13	4,15	,,
s	3,51	4,10	,,	,,
ş	2,75	4,00	4,14	,,
t	3,12	4,11	4,13	,,
ţ	2,32	3,92	4,12	,,
u	3,38	4,15	4,14	,,
v	2,72	4,07	4,15	,,
x	0,00	3,90	4,13	,,
z	3,02	4,00	,,	,,
blanc	4,22	4,14	4,14	,,

Tableau 10

	4	8	12	16
a	+0,48	+0,01	0	0
ă	+0,42	−0,07	0	0
b	+1,62	+0,07	+0,01	0
c	+0,24	+0,01	0	0
d	+0,51	+0,01	0	0
e	+0,23	−0,03	0	0
f	+1,29	+0,18	+0,01	0
g	+1,64	+0,09	+0,01	0
h	+1,33	+0,27	+0,02	0
i	+0,26	+0,04	0	0
î	+1,08	+0,03	0	0
j	+3,14	+0,33	−0,06	0
l	+0,48	+0,04	0	0
m	+0,63	−0,06	+0,02	0
n	+0,29	+0,02	0	0
o	+0,67	+0,04	+0,01	0
p	+0,85	0	0	0
r	+0,47	+0,01	−0,01	0
s	+0,63	+0,04	−0,01	0
ş	+1,39	+0,14	0	0
t	+1,02	+0,03	+0,01	0
ţ	+1,82	+0,22	+0,02	0
u	+0,76	−0,01	0	0
v	+1,42	+0,07	−0,01	0
x	+4,14	+0,24	+0,01	0
z	−1,12	+0,14	+0,01	0
blanc	−0,08	0	0	0

Toute autre est la situation dans le cas où l'on connaît la probabilité d'occurrence dans la langue écrite de chaque lettre à part. Notre tâche est alors considérablement facilitée. L'entropie diminue à

$$H_1 = 4{,}14 \text{ bits}$$

Pour augmenter l'exactitude de la réponse, on cherche à obtenir encore des informations supplémentaires. Une classe de pareilles informations est représentée par l'inventaire des digrammes du roumain écrit. Il est à supposer qu'en connaissant la modalité dont les lettres se joignent en digrammes, l'incertitude concernant la lettre située dans la quatrième position après le blanc enregistrera une diminution.

Mais le tableau 3 montre le contraire. Si on tient compte aussi de la statistique des digrammes, la difficulté d'identifier la lettre cherchée ne diminue pas. L'entropie augmente à 4,18 bits (voir la quatrième cassette de la dernière ligne du tableau 3).

Pourtant, on ne renonce pas. On essaie d'obtenir d'autres indications auxiliaires en faisant appel aussi à la statistique des trigrammes. Le tableau 5 prend en considération tant le rôle des digrammes que celui des trigrammes. Il nous fournira donc plus d'éléments pour découvrir la réponse correcte. Mais une autre surprise : l'entropie de la quatrième lettre d'un mot n'a diminué cette fois-ci non plus. Bien au contraire : elle est à présent de 4,20 bits.

Il ne nous reste qu'à prendre en considération aussi les inventaires des tétragrammes et des pentagrammes. Le travail pour dresser une telle statistique est accablant. Seulement le nombre des pentagrammes doit être de l'ordre des millions. L'information obtenue ne peut être que substantielle. Et pourtant...

La réponse du tableau 9 est catégorique : l'inventaire de toutes les combinaisons possibles de 2, 3, 4 et 5 lettres s'est avéré inutile. L'entropie de la lettre cherchée n'a diminué tant soit peu. Malgré les multiples informations dont on dispose, on aura à affronter une incertitude de 4,22 bits, un peu plus grande que si on avait ignoré les restrictions concernant la juxtaposition des lettres.

L'identité de la lettre située dans la quatrième position après le blanc reste tout aussi « mystérieuse » qu'au début. C'est un cas typique de non-fonctionnement des véhicules de l'information. L'écho de la présence du blanc quatre intervalles à gauche par rapport à la lettre cherchée n'est arrivé jusqu'ici que dans une très petite mesure et au lieu de faciliter la prédiction il la rend plus difficile (l'indétermination a augmenté dans une certaine mesure).

2. — Parfois est bloqué seulement l'un des canaux de transmission. Quand il s'agit de deviner la quatrième lettre située après un *a* les digrammes transportent une information négative tandis que les tétra- et les pentagrammes des informations positives (v. les tableaux 4, 8, 10) ; seulement les trigrammes ne contribuent nullement à la propagation de l'information. En comparant les tableaux 4 et 6, on constate que la quatrième lettre située après un *a* est tout aussi difficile à déduire qu'on tienne ou non compte de la fréquence des trigrammes. On se trouve dans une situa-

tion où le groupement des lettres en séquences de trois respecte fidèlement les lois qui gouvernent la formation des digrammes sans aucune autre restriction supplémentaire, c'est-à-dire :

$$p(x_i\ x_j\ x_k) = p(x_i\ x_j)\ .\ p(x_j\ x_k).$$

3. — Enfin, il y a des cas (peut-être les plus fréquents) quand chaque type de séquences de lettres contribue à la réduction de l'incertitude. L'indétermination de la quatrième position, d'après un ţ diminue avec 0,07 bits si on connaît les probabilités des digrammes, avec 0,20 bits si on tient compte des trigrammes et avec 1,82 bits si on prend en considération les tétra- et les pentagrammes. On doit souligner surtout le rôle important de ces derniers pour l'identification de la lettre cherchée.

On rencontre pourtant ici une certaine déformation des résultats due aux dimensions réduites de l'échantillon linguistique soumis à l'investigation. Les combinaisons auxquelles participent des lettres à fréquence moindre laissent l'impression qu'elles sont les seules possibles dans la langue et transmettent comme telles une quantité exagérée d'information. Le rayon d'influence de ces lettres apparaît pour cela plus grand qu'il n'est en réalité. L'exemple le plus concluant est celui de la lettre *x* qui a une seule occurrence dans le texte. Il n'existe donc qu'un seul digramme, un seul trigramme, ... un seul n-gramme qui commence par la lettre *x* (n variant de 2 jusqu'à 818 !). Toutes les 817 lettres qui suivent dans le texte après un *x* sont donc strictement déterminées ! Les dimensions réduites du matériel exploré constituent, ainsi que l'on remarque, la source de la dilatation artificielle du rayon d'action de la lettre *x*. Il n'y a pas de doute que dans un texte cent fois plus long l'indépendance d'un *x* par rapport à la 817-ième lettre qui lui succède serait évidente.

On doit retenir que, dans notre exemple, surtout la contribution des séquences longues de lettres est surévaluée. On s'attendra donc qu'en étudiant exhaustivement la langue écrite on obtienne des valeurs plus petites des rayons d'influence que celles qu'on a trouvées ici.

Une telle conclusion se trouve en évidente contradiction avec les résultats de Burton et Licklieder selon lesquels l'influence d'une lettre s'étend en moyenne sur une distance de 30 lettres. On ne trouve qu'une seule explication de cette discrépance :

Dans les tests de prédiction initiés par Shannon et utilisés aussi par les auteurs cités, en dehors des relations d'affinité dans l'enchaînement des lettres (l'existence de certains groupements privilégiés et d'autres repoussés par le système de la langue), une importante contribution est apportée par certains phénomènes de nature morpho-syntaxique tel que l'accord. Compatibles avec les règles de concaténation des lettres en séquences sont toutes les formes flexionnelles appartenant au paradigme d'un mot. Des formes comme *préparé*, *préparée*, *préparés* et *préparées* sont également admises par la langue, mais dans le contexte : « L'étudiante que je rencontrais chaque jour à la Bibliothèque Centrale Universitaire, où j'allais conspecter des matériaux pour un travail de grammaire comparée des langues indo-européennes, s'est montrée à l'examen très bien préparée » l'occurrence de la forme *préparée* est la seule possible.

Cette restriction grammaticale fait que le rayon d'influence de la lettre *e* du mot étudiante acquiert une valeur impressionnante (183 lettres !) La dernière lettre du texte (*e* de *préparée*) est strictement déterminée par cet *e*. Et encore cet exemple ne représente pas une limite. La phrase ci-dessus peut être amplifiée indéfiniment en intercalant des constructions subordonnées, sans que la relation obligatoire entre les deux lettres perde sa valabilité. L'information peut être transportée ainsi à des distances quelque grandes qu'elles soient. Cet aspect n'échappe pas évidemment aux tests de prédiction parce qu'il est à supposer que le sujet interrogé connaît et applique les règles concernant l'accord entre le déterminant et le déterminé. Mais ce genre d'influence représente un cas à part ; il n'est pas typique pour le comportement moyen des lettres. L'accord se rapporte seulement aux grammatèmes. Les lettres qui appartiennent aux racines échappent à ce genre d'interdépendance.

Nous croyons par conséquent que la longueur exagérée du rayon d'influence des lettres fournie par les expériments de prédiction doit être mise sur le compte de certains phénomènes grammaticaux du type de l'accord. Il est même probable que la valeur établie par Burton et Licklieder se trouve dans un certain genre de corrélation avec la longueur moyenne de phrases appartenant aux textes utilisés par ceux-ci.

L'examen des dernières lignes des tableaux 4, 6, 8 et 10 conduit à une conclusion intéressante concernant la morphologie. La fréquence du blanc (169 occurrences) indique la longueur moyenne des mots qui est :

$$L_{mot} = \frac{1000}{169} = 5{,}92 \approx 6 \text{ lettres}$$

D'autre part, les tableaux montrent que l'influence du blanc s'éteint environ à la sixième lettre suivante, indifféremment si on tient compte seulement de l'apport des digrammes ou aussi de celui des combinaisons plus longues (tri-, tétra-, pentagrammes). Par conséquent, la finale d'un mot garde une indépendance relative envers le corps du mot. La connaissance des combinaisons de tout au plus 5 lettres qui succèdent au blanc (il s'agit donc de toutes les lettres à l'exception de la finale) ne facilite pas la prédiction de la sixième, parce que son identité est dictée par d'autres lois que celles de la compatibilité des lettres antérieures. Ici se révèle le traitement différent des désinences par rapport aux radicaux.

4. CONCLUSIONS

Formulons sommairement les résultats de l'étude :

1. — A la différence des méthodes antérieures, le modèle markovien proposé offre la possibilité d'une analyse détaillée de la manière dont a lieu la propagation à distance de l'influence d'une certaine unité linguistique. (On s'est rapporté ici aux lettres, mais les procédés exposés sont valables aussi pour les phonèmes, les morphèmes, les mots, les catégories morphologiques, les fonctions snytaxiques, etc.). Grâce à ce modèle, on a pu mettre en évidence les agents du transport de l'information le long du texte, en précisant la contribution différenciée de chacun en vue de faciliter la déduction des éléments inconnus du message.

2. — L'étude en sens inverse, de droite à gauche, des corrélations informationnelles entre les éléments de la communication devient aussi possible.

3. — Il n'est pas sans intérêt de constater que l'application de la méthode ne réclame pas de la part du chercheur (comme dans les tests de prédiction) la connaissance de la langue étudiée. On peut donc analyser complètement l'influence à distance dans des textes appartenant à un idiome totalement inconnu, ce qui élargit considérablement le champ des applications.

4. — Il a été relevé pour la première fois un paradoxe de la transmission des informations par l'intermédiaire de la langue, qui consiste dans le caractère ondulatoire de la variation de l'entropie après une unité connue.

5. — Le calcul de l'entropie de deuxième ordre du roumain écrit constitue à son tour une « première ». La comparaison avec la valeur homologue de l'anglais a mis en lumière l'importance informationnelle particulière des digrammes roumains.

6. — Cette étude met un signe d'interrogation, sinon précisément quant aux résultats de Burton et Licklieder concernant la longueur du rayon d'influence d'une lettre, mais en tout cas quant à l'opinion de I.M. Iaglom et A.M. Iaglom selon lesquels ces résultats peuvent être étendus à toutes les langues. L'investigation d'un matériel de langue anglaise pourrait confirmer ou non les valeurs des rayons d'action des lettres établies par les chercheurs américains.

Note: Quelques fragments de cette étude ont été présentés sous forme de communication au Colloque de Linguistique Mathématique tenu à Bratislava en février 1970 et comme tels ils sont en cours de parution dans le volume consacré aux travaux du colloque [3].

Juin 1972

Institut de recherches pour l'économie des eaux
Bucarest

BIBLIOGRAPHIE

1. Bellman, Richard, *Introduction to matrix analysis*, McGraw-Hill Book Company, Inc. New York, Toronto, Londres, 1960.
2. Burton, N.G.; Licklieder, J.C., *Long range Contraints in the Statistical Structure of printed English*, dans « American Journal of Psychology », vol. 68, nº 4, Baltimore, 1955.
3. Dinu, Mihai, *Un modèle markovien de l'influence à distance dans les langues naturelles*, dans « Recueil linguistique de Bratislava », nº 4, 1970.
4. Faure, F.; Kaufmann, A.; Denis-Papin, M., *Mathématiques nouvelles*, Dunod Paris, 1964.
5. Hanson, H., *The entropy of Swedish language*, dans « Proceedings of the 2-nd Prague Conference » Prague, 1960.
6. Iaglom, A.M.; Iaglom I. M., *Probabilitate și informație*, Ed. Tehnică, Bucarest, 1963.
7. Küpfmüller, K., *Die Entropie der Deutschen Sprache* dans « Fernmeldetechnische Zeitschrift » nº 6, 1954.
8. Marcus, Solomon; Nicolau, Edmond; Stati, Sorin, *Introducere în lingvistica matematică*, Editura științifică, Bucarest, 1966.
9. Nicolau, E.; Sala, C.; Roceric, Al., *Observații asupra entropiei limbii române*, dans « Studii și cercetări lingvistice », X, 1, 1959, p. 35—53.
10. Roceric-Alexandrescu, Alexandra, *Fonostatistica limbii române*, Editura Academiei R. S.R. București, 1968.
11. Shannon, C.E., *Prediction and Entropy in printed English*, dans « Bell System Technical Journal », vol. 30, 1951, p. 50—64.
12. Tocaciu, Octavian, *Unele date statistice privind frecvența literelor și digramelor în limba (scrisă) contemporană*, dans « Studii și cercetări lingvistice », XVI, 5, 1965, p. 683—722.