**Zeitschrift:** Revue de linguistique romane
**Band:** 38 (1974)
**Heft:** 149-152

**Artikel:** Une vérification de loi linguistique par corrélation
**Autor:** Guiter, Henri
**DOI:** https://doi.org/10.5169/seals-399567

# UNE VÉRIFICATION DE LOI LINGUISTIQUE PAR CORRÉLATION

A date récente (*Rev. ling. rom.*, 1971, 35, p. 335), M. Jean Séguy publiait une importante étude, abondamment documentée, sur « La relation entre la distance spatiale et la distance lexicale ».

Pratiquant des « coupes » sensiblement rectilignes à travers les domaines de nombreux atlas linguistiques, M. Séguy a recherché comment varie la distance lexicale entre un point quelconque du parcours et le point de départ, en fonction de la distance géographique de ces deux points. Cette distance lexicale est en fait un pourcentage d'étymons différents, puisque, dans chaque atlas, M. Séguy a opéré sur cent cartes [1].

1. C'est dans le train qui nous conduisait au colloque de Strasbourg, en mai 1971, que Jean Séguy nous a montré, pour la première fois, ses courbes représentatives de la distance lexicale en fonction de la distance spatiale, et nous a fait part des difficultés qu'il éprouvait à les mettre en équation. La chose nous semblait au contraire relativement facile.

En septembre 1971, il nous adressait des calques de toutes ses courbes ; les équations correspondantes sont publiées dans son article cité en référence.

Mais, dans l'application que nous faisions aux atlas à points clairsemés, de la relation établie pour les atlas exhaustifs, il y avait aussi, sur courte distance, variation d'une distance linguistique en fonction d'une distance spatiale.

Il semblait donc logique que la relation de Jean Séguy et celle que nous avions proposée au colloque de mai 71, fussent liées.

En novembre 71, nous adressions à Jean Séguy une première mouture du présent article (avec d'ailleurs une faute de calcul, dont nous nous aperçûmes après l'envoi). Il nous répondait aussitôt une longue lettre enthousiaste, d'où nous extrayons : « Vous devez absolument publier cette trouvaille, dès que l'article sur les distances lexicales sera sorti et aussi les actes du colloque de Strasbourg puisque la lecture de votre communication de Strasbourg est indispensable à l'intelligence de votre lettre du 13/11. »

En nous retournant la seconde rédaction, un mois plus tard il nous renouvelait son désir de voir publier cette étude.

Hélas, il n'aura pas été donné à Jean Séguy de voir cette publication, pas plus d'ailleurs que celle des actes du colloque de Strasbourg, qui la conditionnait.

Nous l'offrons en hommage attristé à la mémoire de ce grand savant et infatigable travailleur, qui était aussi le plus loyal des amis.

Il a représenté par des courbes la variation de la distance lexicale N en fonction de la distance spatiale $x$ exprimée en myriamètres, opération effectuée pour chaque coupe particulière, puis pour la moyenne des divers résultats obtenus. Toutes ces courbes offrent un air de famille : le pourcentage de différences lexicales est fonction de la racine carrée du logarithme de la distance spatiale, $\sqrt{\log (x + 1)}$, ce terme pouvant parfois se présenter à la puissance 2, 3 ou même 4, mais restant le plus souvent à la puissance 1. Nous retiendrons comme équation-type, celle de la courbe moyenne :

$$N = 36 \sqrt{\log (x + 1)} \tag{1}$$

Cette représentation est approchée, car en réalité, les courbes représentatives présentent des séries de degrés.

On pourrait s'inquiéter en constatant que, d'après l'équation (1), N croît indéfiniment avec $x$, et peut donc théoriquement dépasser la valeur 100. En fait, cette valeur 100 serait atteinte pour $\log (x + 1) = 9$, donc pour $x =$ dix milliards de kilomètres, c'est-à-dire plusieurs heures-lumière, ou, si l'on préfère, la dimension du système solaire ! Déjà, pour que N atteignît 70, il faudrait $\log (x + 1) = 4$ et $x = 100\,000$ km, soit plus de deux fois le tour de la Terre. Pour une valeur de $x$ plafonnant à 1 000 km, N ne peut guère dépasser 50 %. En conséquence, il est inutile d'apporter une restriction à l'emploi de l'équation proposée, lorsqu'on l'applique, comme c'est le cas à travers le domaine roman.

* * *

Par ailleurs, dans notre contribution, « Atlas et frontières linguistiques », au Colloque organisé par le C. N. R. S. à Strasbourg en mai 1971 (Actes N° 930, 1973, p. 61), nous avons présenté une formule correctrice permettant de déterminer l'importance de la frontière linguistique passant entre deux points d'un atlas exhaustif, à partir des données fournies par un atlas à points clairsemés :

$$\Delta N = N \sqrt{\frac{Do}{D} - 1} \left(\frac{N}{100}\right)^2 \left(\frac{100 - N}{100}\right)^2 \tag{2}$$

Dans cette relation, Do est la densité des points de l'atlas exhaustif ; D, celle de l'atlas clairsemé considéré ; $\left(\frac{Do}{D}\right.$ est donc le rapport des nombres de points occupant un même domaine dans les deux atlas$\left.\right)$ ; N, le pourcentage de différences constatées entre les deux points. Précisons bien que nous représentions par N le pourcentage de toutes les différences possibles entre deux points, qu'elles fussent d'ordre phonétique, morphologique, syntactique ou lexical, et pas seulement le fait de correspondre à des étymons différents. Nos valeurs de N sont donc d'un ordre de grandeur plus élevé que celles envisagées dans l'article de M. Séguy.

Notre idée première était d'introduire la distance $x$ séparant les points, et de poser que la variation relative d'écarts $\frac{\Delta N}{N}$ était proportionnelle à la variation relative de distance $\frac{\Delta x}{x}$.

Pour des raisons de commodité, d'une part, parce qu'il est extrêmement facile de faire le rapport des points d'enquête présentés par deux atlas différents sur un même domaine géographique, d'autre part, parce que l'expérience montrait que l'adoption d'un tel terme nous dispensait de l'introduction de tout coefficient, nous nous sommes résolu à remplacer $\frac{\Delta x}{x}$ par $\sqrt{\frac{Do}{D}} - 1$, qui, sans lui être égal, a toujours des variations de même sens que lui. Il est aisé de comparer ces deux grandeurs graphiquement (fig. 1). Le rapport des densités $\frac{Do}{D}$ est égal à l'inverse du rapport des surfaces moyennes correspondant à chaque point : $\frac{Do}{D} = \frac{S}{So}$. Et $\frac{S}{So} - 1 = \frac{S - So}{So}$. Sur le

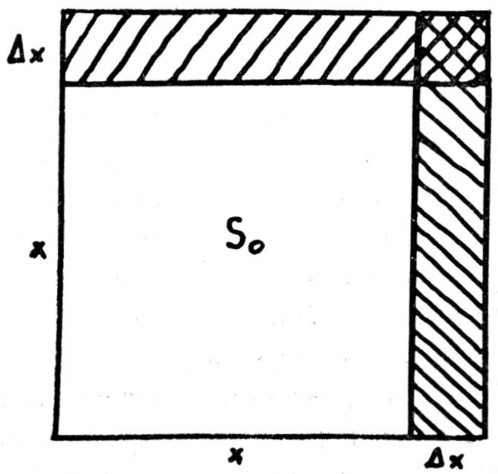



FIG. 1.

schéma, So est le carré blanc de côté $x$ ; S, le carré total de côté $(x+\Delta x)$ ; (S — So), la partie hachurée. Par conséquent, le carré du terme $\sqrt{\frac{\mathrm{Do}}{\mathrm{D}}}-1$ est représenté par le rapport de l'aire hachurée à l'aire du carré blanc. Comparons-lui le carré de $\frac{\Delta x}{x}$, soit $\frac{\Delta x^2}{x^2}$ : c'est le rapport de l'aire du petit carré quadrillé à l'aire du carré blanc. Les deux rapports varient donc toujours dans le même sens, bien qu'ils ne soient pas égaux.

L'équation (2), modifiée pour y introduire la variable $x$ au lieu de D, établirait elle aussi une relation entre une variation de distance linguistique et une variation de distance spatiale. Nous devons remarquer que la distance moyenne séparant deux points ne peut devenir inférieure à la maille $l$ de l'atlas exhaustif ; nous introduisons donc cette constante additive au dénominateur de $\Delta x$, qui ne peut ainsi tomber au-dessous de cette valeur. Par ailleurs, nous affecterons les distances d'un coefficient $k$, destiné à tenir compte de l'unité de longueur choisie et de la nature des faits linguistiques envisagés.

Mise sous forme différentielle, l'équation (2) devient :

$$\mathrm{dN} = \mathrm{N}\,\frac{k\,\mathrm{d}x}{k\,(x+l)}\left(\frac{\mathrm{N}}{100}\right)^2\left(\frac{100-\mathrm{N}}{100}\right)^2 \tag{3}$$

Si l'on admet qu'un atlas est exhaustif lorsqu'il comporte une enquête par commune, il est possible de se faire une idée de la valeur numérique moyenne de $l$.

Il y a en France 37 788 communes pour 550 986 km², soit 14,58 km² par commune, ce qui donne comme distance $l$ 3,8 km. (Le département des Pyrénées-Orientales, qui figure intégralement dans l'*ALPO* a 4 131 km² pour 231 points d'enquête, soit 17,88 km² par commune, et une distance de 4,2 km). Les communes italiennes sont plus grandes, 8 263 communes pour 286 648 km², soit 34,69 km² par commune, et une distance de 5,9 km ; et les communes espagnoles sont encore plus grandes, 9 287 communes pour 504 903 km², soit 54,36 km² par commune, et une distance de 7,3 km.

Pour l'élaboration de la courbe moyenne, M. Séguy a mis à profit 7 coupes d'atlas français, 3 d'atlas espagnols, et 2 d'atlas italiens ; ceci correspond à peu près exactement à une distance moyenne de 5 km entre les communes. (Nous manquons d'éléments pour apprécier les données relatives à l'atlas roumain).

***

Ceci posé, revenons à l'équation (3). En séparant les variables et en sommant, nous obtenons :

$$10^8 \int \frac{dN}{N^3 (100 - N)^2} = \int \frac{k\,dx}{k(x+l)} \tag{4}$$

Pour déterminer la fonction primitive du premier membre, nous avons intérêt à le transformer en une somme de monomes. Posons l'identité :

$$\frac{1}{N^3 (100-N)^2} = \frac{\alpha N^2 + \beta N + \gamma}{N^3} + \frac{\delta N + \varepsilon}{(100-N)^2}$$

Réduisons au même dénominateur le second membre, et identifions les numérateurs :

$$1 \equiv (\alpha + \delta) N^4 + (\varepsilon - 200\,\alpha) N^3 + (10\,000\,\alpha - 200\,\beta) N^2 + (10\,000\,\beta - 200\,\gamma) N + 10\,000\,\gamma$$

Il en résulte :

$$\alpha + \delta = 0 \qquad \varepsilon - 200\,\alpha = 0 \qquad 10\,000\,\alpha - 200\,\beta = 0$$
$$10\,000\,\beta - 200\,\gamma = 0 \qquad 10\,000\,\gamma = 1$$

D'où :

$$\gamma = \frac{1}{10\,000} \qquad \beta = \frac{2}{1\,000\,000} \qquad \alpha = \frac{4}{100\,000\,000}$$
$$\varepsilon = \frac{8}{1\,000\,000} \qquad \delta = -\frac{4}{100\,000\,000}$$

L'équation (4) devient :

$$\int \frac{4\,dN}{N} + \int 200 \frac{dN}{N^2} + \int 10\,000 \frac{dN}{N^3} + \int 400 \frac{dN}{(100-N)^2} + \int 4 \frac{dN}{(100-N)} = \int \frac{k\,dx}{k(x+l)} \tag{5}$$

Il est dès lors facile de passer aux fonctions primitives :

$$4 \operatorname{Log} N - \frac{200}{N} - \frac{5\,000}{N^2} + \frac{400}{100-N} - 4 \operatorname{Log}(100 - N) = \operatorname{Log} k(x+l) \tag{6}$$

Dans cette équation (6), il serait difficile d'expliciter N en fonction de $x$.

Mais nous remarquons que le troisième terme a un coefficient numérique très grand par rapport à ceux des autres termes du premier membre, ce qui peut permettre de négliger ces autres termes au moins pour les petites valeurs de N, et peut-être au-delà, s'il y a compensation entre ces autres termes. (Le symbole « Log » représente un logarithme népérien, qui est égal au logarithme décimal divisé par log e = 0,43429).

Groupons en un tableau les valeurs (arrondies) des divers termes, et leur total :

| N | 4 Log N | $-\frac{200}{N}$ | $-\frac{5\,000}{N^2}$ | $+\frac{400}{100-N}$ | $-4\,\mathrm{Log}\,(100-N)$ | TOTAL |
|---|---|---|---|---|---|---|
| 1 | 0 | — 200 | — *5 000* | + 4 | — 19 | — *5 215* |
| 10 | 9,2 | — 20 | — *50* | + 4,4 | — 18 | — *73* |
| 20 | 12 | — 10 | — *12* | + 5 | — 17,6 | — *22* |
| 40 | 14,8 | — 5 | — *3* | + 6,6 | — 16,4 | — *3* |
| 50 | 16 | — 4 | — *2* | + 8 | — 16 | + *2* |
| 60 | 16,4 | — 3,3 | — *1,3* | + 10 | — 14,8 | + *7* |
| 80 | 17,6 | — 2,5 | — *0,8* | + 20 | — 12 | + *22* |
| 90 | 18 | — 2,2 | — *0,6* | + 40 | — 9,2 | + *46* |
| 99 | 19 | — 2 | — *0,5* | + 400 | 0 | + *416* |

Un graphique exprimera encore mieux les résultats (fig. 2). Son examen montre que la courbe représentative de tout le premier membre, et celle représentative de son troisième terme, ont des allures et des parcours voisins jusqu'à N = 50 ; au-delà de N = 60, la divergence s'accuse et devient vite très importante.

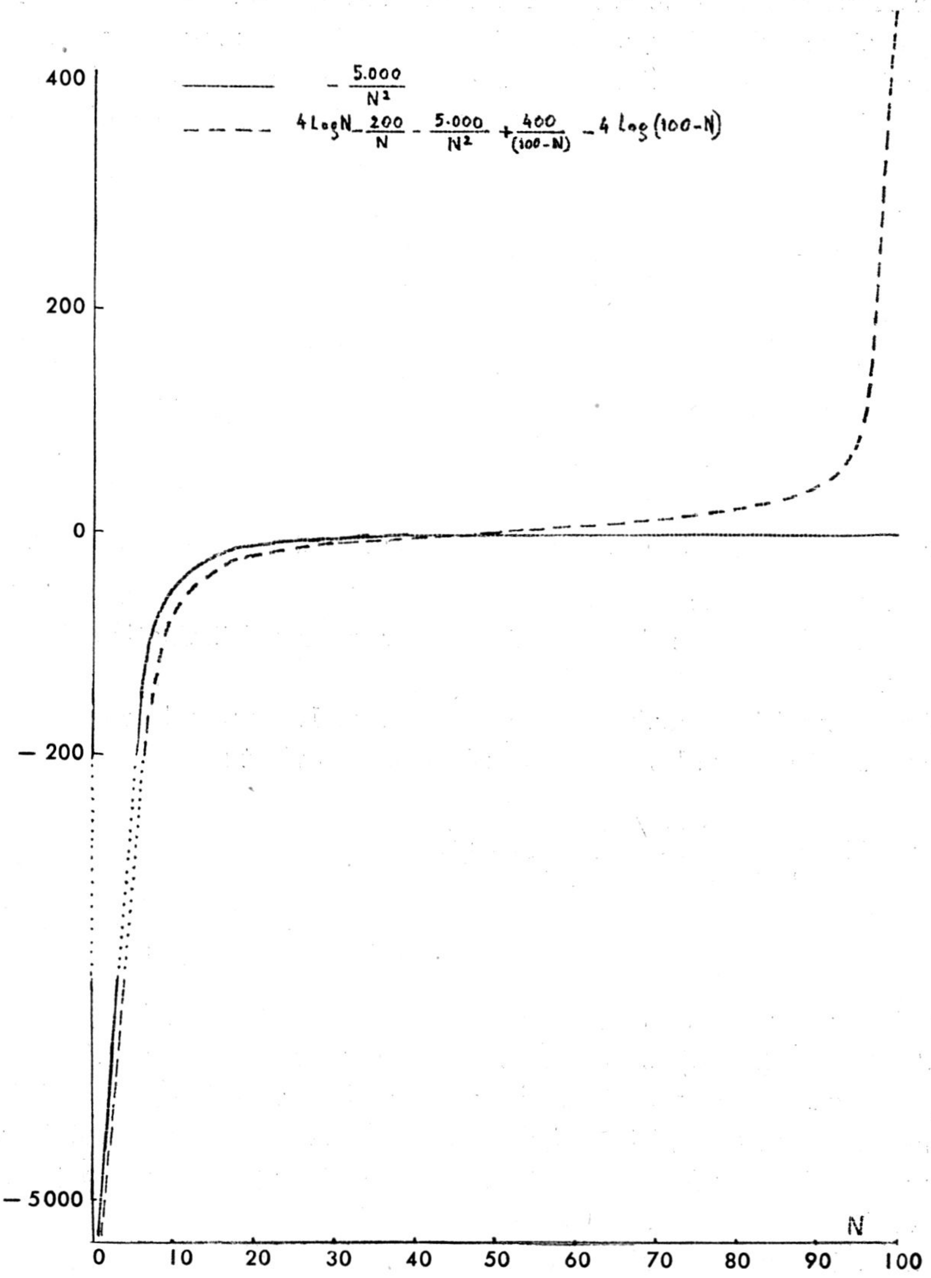
Étant donné que les courbes construites par M. Séguy plafonnent généralement à N = 50, et atteignent rarement N = 60, nous pouvons donc réduire le premier membre à son troisième terme, pour obtenir des résultats applicables à ces courbes.

Dans cette approximation, l'équation (6) se simplifie beaucoup :

$$(7) \qquad -\frac{5\,000}{N^2} = \mathrm{Log}\, k\,(x+l)$$

ou encore :

$$(8) \qquad \frac{N^2}{5\,000} = -\frac{1}{\mathrm{Log}\, k\,(x+l)}$$



400
200
0
− 200
− 5000
$-\frac{5.000}{N^2}$
$4\,\mathrm{Log}\,N - \frac{200}{N} - \frac{5.000}{N^2} + \frac{400}{(100-N)} - 4\,\mathrm{Log}\,(100-N)$
N
0 10 20 30 40 50 60 70 80 90 100


FIG. 2.

Mais l'équation résultant d'une intégration n'est exacte qu'à une constante additive près. Ici le fait est d'autant plus criard que les deux membres sont de signes contraires, si la constante reste fixée à la valeur zéro. Pour déterminer cette constante, remarquons que le second membre doit s'annuler pour $N = 0$ et $x = 0$. La constante satisfaisant à cette condition sera $\frac{1}{\operatorname{Log} kl}$ :

$$\frac{N^2}{5\,000} = \frac{1}{\operatorname{Log} kl} - \frac{1}{\operatorname{Log} k\,(x+l)} \tag{9}$$

Et, en réduisant au même dénominateur,

$$\frac{N^2}{5\,000} = \frac{\operatorname{Log} k\,(x+l) - \operatorname{Log} kl}{\operatorname{Log} kl \operatorname{Log} k\,(x+l)} = \frac{\operatorname{Log}\left(\frac{x}{l}+1\right)}{\operatorname{Log} kl \operatorname{Log} k\,(x+l)}$$

Le terme $\frac{x}{l} = \lambda$, qui s'introduit au numérateur du second membre, est la mesure de la distance $x$ en prenant comme unité de longueur la maille $l$ de l'atlas exhaustif, sorte d'unité naturelle entraînant l'élimination de $k$ au numérateur.

Si nous explicitons maintenant en fonction de N, et remplaçons les logarithmes népériens par les logarithmes décimaux, il vient :

$$\begin{aligned} N &= \frac{100\sqrt{\log e}}{\sqrt{2}\sqrt{\log kl \log k\,(x+l)}}\sqrt{\log(\lambda+1)} \\ &= \frac{46}{\sqrt{\log kl \log k\,(x+l)}}\sqrt{\log(\lambda+1)} \end{aligned} \tag{10}$$

équation dont la forme rappelle celle de l'équation (1). Cependant, le coefficient de $\sqrt{\log(\lambda+1)}$ n'est pas une constante, mais une fonction de $x$. Voyons avec quelle approximation nous pourrions considérer ce coefficient comme une constante ; ici aussi, nous allons prendre $l$ comme unité de longueur, ce qui amène le dénominateur à l'expression $\sqrt{\log k \log k\,(\lambda+1)}$.

La mise en équation de la courbe de M. Séguy en prenant $l$ (soit 5 km) comme unité, substituerait à l'équation (1) une équation :

$$N = 33\sqrt{\log(\lambda+1)} \tag{11}$$

Pour que le coefficient de l'équation (10) $\frac{46}{\sqrt{\log k \log k\,(\lambda+1)}}$ prenne la valeur 33, il faut que :

$$\sqrt{\log k \log k \, (\lambda + 1)} = \frac{46}{33} = 1{,}4$$

ou :

$$\log k \log k \, (\lambda + 1) \neq 2$$

expression qui peut se développer en :

$$\log^2 k + \log k \log (\lambda + 1) - 2 = 0$$

Lorsque $\lambda + 1 = 200$, cette équation admet comme solution $k = 4{,}8$ ; lorsque $\lambda + 1 = 20$, elle admet comme solution $k = 8$. Si nous adoptons pour $k$ une valeur moyenne telle que 6,5, les valeurs prises par le radical lorsque $\lambda$ varie sont :

| $\lambda + 1$ | 200 | 100 | 50 | 20 | 10 |
|---|---|---|---|---|---|
| $\sqrt{\log k \log k \, (\lambda + 1)}$ | 1,58 | 1,50 | 1,42 | 1,30 | 1,21 |

soit des écarts de 13 % au-dessus et au-dessous de la valeur moyenne 1,4 lorsque la distance spatiale varie entre 50 et 1 000 km.

Si nous nous contentons de cette nouvelle approximation, l'équation (10) devient identique à l'équation (11), qui est elle-même la forme prise par l'équation (1) lorsqu'on substitue au myriamètre, comme unité, la maille moyenne des atlas exhaustifs, soit 5 km.

Nous constatons donc qu'au prix de deux approximations,

— la première consistant à négliger quatre des termes de l'intégrale devant le terme $\frac{5\,000}{N^2}$, c'est-à-dire à confondre les deux courbes de la figure 2 lorsque N est inférieur à 60,

— la seconde consistant à adopter une valeur moyenne pour un coefficient qui varie en réalité de 13 % en-dessus ou en-dessous,

la courbe expérimentale établie par M. Séguy admet comme équation une intégrale première de l'expression de variation des différences linguistiques avec la distance spatiale, expression à laquelle nous avaient conduit précédemment des considérations de simple logique.

*
* *

On pourra s'étonner que nous n'ayons pas adopté une démarche inverse, en principe plus facile, c'est-à-dire dériver l'équation de M. Séguy pour essayer de retrouver la nôtre.

Une telle démarche aurait été vouée à l'échec pour la raison suivante : l'équation relativement simple de M. Séguy visait à rendre compte de l'allure de la courbe à des distances du point origine largement supérieures à la maille de l'atlas exhaustif ; notre équation s'appliquait au contraire à des mailles d'atlas, c'est-à-dire, à des distances immédiatement supérieures à la maille de l'atlas exhaustif.

Si nous différencions l'équation (11), nous obtenons :

$$(12) \qquad dN = \frac{33 \times 0{,}43}{2} \frac{d\lambda}{\sqrt{\log(\lambda+1)}\,(\lambda+1)} = 7{,}1 \frac{d\lambda}{\sqrt{\log(\lambda+1)}\,(\lambda+1)}$$

La variation de N fournie par l'équation (12) n'est pas modulée en fonction des valeurs de N, ce qui s'explique par les termes négligés en cours d'intégration. Les adaptations aux cas particuliers de l'équation (11) (ou de l'équation (1)), se traduisaient, dans l'article de M. Séguy, par des modifications de coefficients ou d'exposants.

Nous pouvons comparer les valeurs numériques de $\Delta N$ qui résultent de l'équation (12), et celles qui résultent de l'équation (2).

Pour établir $\Delta N$ avec l'équation (12), nous pouvons ajouter les valeurs successives de dN lorsque $d\lambda = 1$ (c'est-à-dire lorsque la distance spatiale croît de 5 en 5 km au delà de la maille de l'atlas exhaustif).

$$\lambda = 1 \qquad dN = 7{,}1 \frac{1}{0{,}55 \times 2} = 6{,}4 \qquad \Delta N = 6{,}4$$

$$\lambda = 2 \qquad dN = 7{,}1 \frac{1}{0{,}7 \times 3} = 3{,}4 \qquad \Delta N = 9{,}8$$

$$\lambda = 3 \qquad dN = 7{,}1 \frac{1}{0{,}77 \times 4} = 2{,}3 \qquad \Delta N = 12{,}1$$

$$\lambda = 4 \qquad dN = 7{,}1 \frac{1}{0{,}84 \times 5} = 1{,}7 \qquad \Delta N = 13{,}8$$

$$\lambda = 5 \qquad dN = 7{,}1 \frac{1}{0{,}88 \times 6} = 1{,}3 \qquad \Delta N = 15{,}1$$

En utilisant l'équation (2) nous devons nous fixer une valeur de N, 50 par exemple. Lorsque $(\lambda + d\lambda)$ prend les valeurs 2, 3, 4, 5, 6, etc., la maille de l'atlas est 2, 3, 4, 5, 6 fois plus grande, et la densité des points 4, 9, 16, 25, 36 fois plus faible. Les valeurs correspondantes de $\Delta N \left( = \frac{50}{16} \sqrt{\frac{Do}{D} - 1} = 3{,}1 \sqrt{\frac{Do}{D} - 1} \right)$ seront :

$$\frac{Do}{D} = 4 \qquad \Delta N = 3{,}1\sqrt{3} = 5{,}3 \text{ au lieu de } 6{,}4$$

$$\frac{Do}{D} = 9 \qquad \Delta N = 3{,}1\sqrt{8} = 8{,}8 \text{ au lieu de } 9{,}8$$

$$\frac{Do}{D} = 16 \qquad \Delta N = 3{,}1\sqrt{15} = 12 \text{ au lieu de } 12{,}1$$

$$\frac{Do}{D} = 25 \qquad \Delta N = 3{,}1\sqrt{24} = 15{,}2 \text{ au lieu de } 13{,}8$$

$$\frac{Do}{D} = 36 \qquad \Delta N = 3{,}1\sqrt{35} = 18{,}3 \text{ au lieu de } 15{,}1$$

Les rapports $\frac{Do}{D}$ ne dépassent pratiquement pas cet ordre de grandeur ; leurs valeurs sont de 38 pour l'ALF, 9 pour l'A. L. S. I., 18 pour l'A. L. C., 20 pour l'A. L. P. I., et sensiblement plus faibles pour les atlas régionaux.

La concordance des valeurs de $\Delta N$ obtenues en utilisant l'équation (2) et l'équation (12) est satisfaisante (fig. 3), si l'on considère qu'à trois reprises, des approximations ont été indispensables pour faciliter les calculs.

***

Voilà donc deux méthodes « dialectométriques » dont les genèses ont été complètement indépendantes, et qui visaient d'ailleurs des buts différents.

L'une mettait à profit toutes les différences linguistiques, et son rayon d'action n'excédait guère six fois la maille d'un atlas exhaustif, soit une trentaine de kilomètres.

L'autre se bornait aux différences lexicales, et la longueur de ses coupes pouvait atteindre 1 000 km.

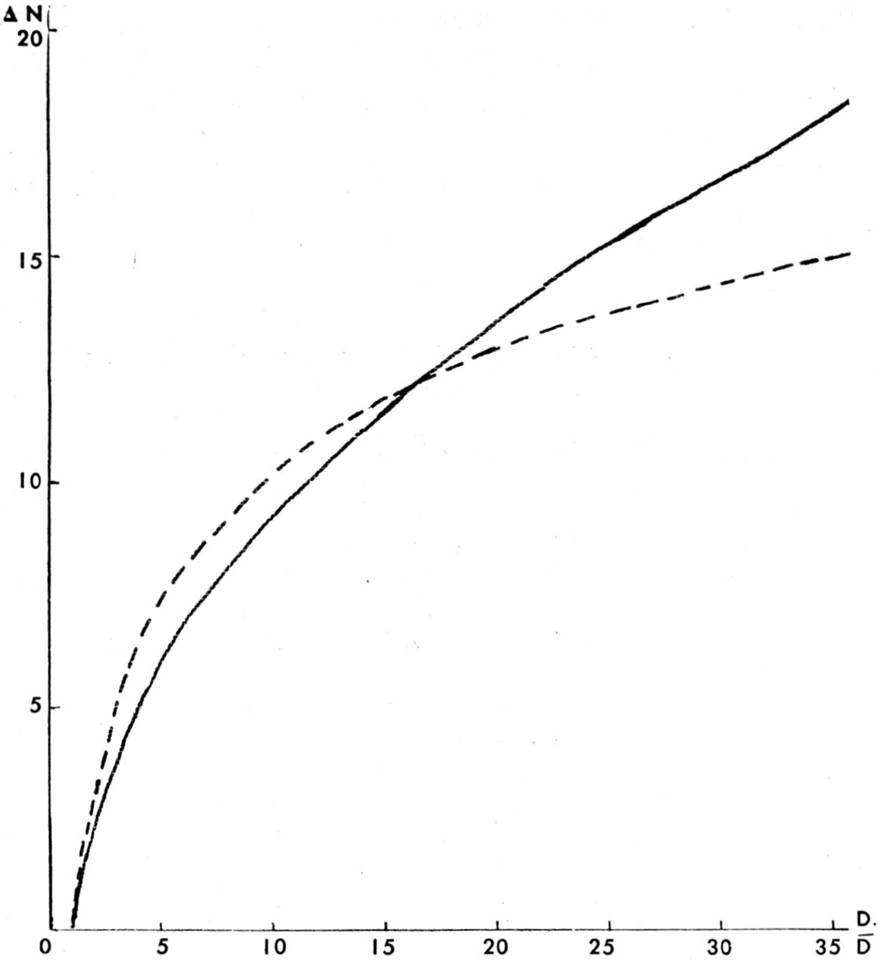



FIG. 3.

——— Équation 2
------ Équation 12

Or, il se trouve qu'à quelques approximations près, la forme de l'équation à laquelle aboutit la seconde, peut se déduire par une intégration, de l'équation par laquelle s'exprime la première.

Il y a là une justification réciproque des deux méthodes et de leurs expressions mathématiques, dont l'intérêt est manifeste.

Henri GUITER.