Appréciation de l'importance des écarts en géolinguistique

APPRÉCIATION DE L'IMPORTANCE DES ÉCARTS EN GÉOLINGUISTIQUE

Quand on étudie, en employant la méthode globale (1), une même région avec des atlas linguistiques différents, ou avec des choix différents de cartes prises dans un même atlas, il arrive fréquemment que l'on trouve des distances linguistiques différentes entre deux points d'enquête donnés.

Les études d'atlas portent très souvent sur 100 cartes, ce qui simplifie les calculs de pourcentage. Mais, ainsi que nous l'avons exposé dans le travail donné en référence (1), il serait dangereux de descendre au-dessous de ce nombre 100, et c'est vraiment le minimum qui offre quelques garanties.

A titre d'exemple, nous rappelons ci-dessous le début du tableau des pourcentages de différences, N, dans les interpoints des cartes 1 à 100 et 101 à 200 de l'ALPO, ainsi que l'écart du second comput par rapport au premier.

Points comparés	Cartes 1-100	Cartes 101-200	Ecarts	Points comparés	Cartes 1-100	Cartes 101-200	Ecarts
1/2	11	19	+ 8	1/3	6	8	+ 2
1/5	14	14	0	2/3	7	13	+ 6
2/4	11	15	+ 4	2/5	9	13	+ 4
2/6	10	6	— 4	2/10	16	14	— 2
3/4	8	12	+ 4	4/10	13	15	+ 2
4/17	9	11	+ 2	5/6	9	11	+ 2
5/8	12	16	+ 4	6/7	15	15	0
6/8	9	11	+ 2	6/9	15	17	+ 2
6/10	9	9	0	7/9	13	11	— 2
7/10	12	14	+ 2	7/11	11	9	— 2
7/31	96	92	— 4	7/33	96	94	— 2
8/9	15	15	0	8/27	94	94	0
9/27	96	96	0	9/29	96	94	— 2

(1) Henri Guiter. — Atlas et frontières linguistiques. Les dialectes romans de France à la lumière des atlas régionaux, Paris, C.N.R.S., 1973, pp. 61-109.

Points comparés	Cartes 1-100	Cartes 101-200	Ecarts	Points comparés	Cartes 1-100	Cartes 101-200	Ecarts
9/33	96	94	— 2	10/11	13	13	0
10/17	9	15	+ 6	10/18	16	20	+ 4
11/18	13	13	0	11/31	95	91	— 4
17/18	13	17	+ 4	17/19	7	7	0
17/20	14	14	0	18/20	9	9	0
18/31	96	92	— 4	19/20	7	11	+ 4
19/22	11	11	0	20/21	7	9	+ 2

Ce bref échantillon d'un long tableau, qui porte sur 1.049 interpoints, suffit à nous donner une idée de ce que l'on peut obtenir : la valeur — 4 de l'écart apparaît pour 4 interpoints ; la valeur — 2, pour 6 ; la valeur 0, pour 12 ; la valeur +2, pour 8 ; la valeur + 4, pour 7 ; la valeur + 6, pour 2 ; la valeur + 8, pour 1. Il s'ébauche une courbe en cloche avec son sommet pour l'écart 0 ; elle est dissymétrique parce que l'étude des 40 premières mesures est trop courte pour que puisse jouer une loi de grands nombres. Remarquons qu'avec 17 interpoints de plus (soit les 57 premiers), les nombres d'écarts + et — s'égalisent, et la courbe devient symétrique (Fig. 1).

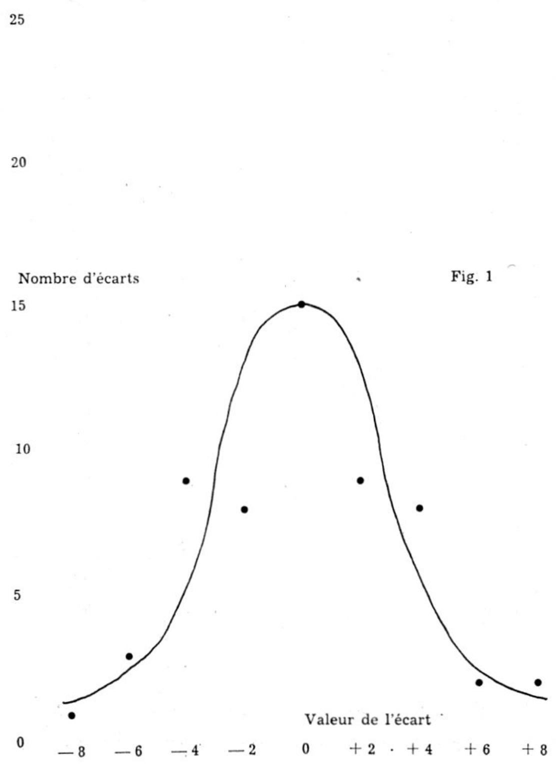

Fig. 1

Nous constatons la possibilité de mesures différentes pour un même interpoint, et l'on est amené à se demander si les écarts présentent un caractère aléatoire, ou si, au contraire, ils sont significatifs.

Bien entendu, ce n'est pas l'écart des pourcentages qui peut nous renseigner à ce sujet. Il faut recourir à une méthode statistique, et nous allons employer celle du X^2. Prenons deux exemples, l'un avec une faible valeur de N, l'autre avec une valeur de N élevée.

Interpoint 5/8	Cartes 1-100	Cartes 101-200	Totaux
Différences	12 (14)	16 (14)	28
Concordances	88 (86)	84 (86)	172
Totaux	100	100	200

Si la répartition était parfaitement aléatoire N serait égal à 14 dans chaque lot de 100 cartes.

$$X^2 = 2 \times \frac{4}{14} + 2 \times \frac{4}{86} = 0{,}68.$$

La probabilité de l'hypothèse nulle est de 40 % ; l'écart n'est pas significatif.

Interpoint 7/31	Cartes 1-100	Cartes 101-200	Totaux
Différences	96 (94)	92 (94)	188
Concordances	4 (6)	8 (6)	12
Totaux	100	100	200

Avec une répartition parfaitement aléatoire, N égalerait 94 dans chaque lot.

$$X^2 = 2 \times \frac{4}{94} + 2 \times \frac{4}{6} = 1{,}42.$$

La probabilité de l'hypothèse nulle est encore de 22 % ; l'écart n'est pas significatif.

Prenons maintenant l'interpoint 1/2 qui présente le plus grand écart.

	Cartes 1-100	Cartes 101-200	Totaux
Différences	11 (15)	19 (15)	30
Concordances	89 (85)	81 (85)	170
Totaux	100	100	200

$$X^2 = 2 \times \frac{16}{15} + 2 \times \frac{16}{85} = 2{,}50.$$

La probabilité de l'hypothèse nulle tombe maintenant à 12 %, mais c'est encore bien suffisant pour que l'écart ne soit pas significatif.

Il serait évidemment fastidieux de répéter 1.049 fois le calcul du X^2 pour tous les interpoints des cartes de l'ALPO. Il est plus commode de construire une abaque, sur laquelle une simple lecture nous permettra de déterminer si un écart est, ou n'est pas, significatif.

Sur la figure 2, nous avons porté en abscisse le nombre de différences relevées sur 100 cartes pour un interpoint, et en ordonnée les valeurs du X^2. Les traits en tirets parallèles à l'axe des abscisses ont pour ordonnées les valeurs du X^2 correspondant à des probabilités de distribution aléatoire de 5 % (au-dessous de laquelle la distribution devient significative), de 1 % et de 0,1 %. Ils nous permettront de mieux situer les courbes. Les quatre courbes que nous avons tracées, correspondent à des écarts en pourcentage de 5, 10, 15 et 20. Bien entendu,

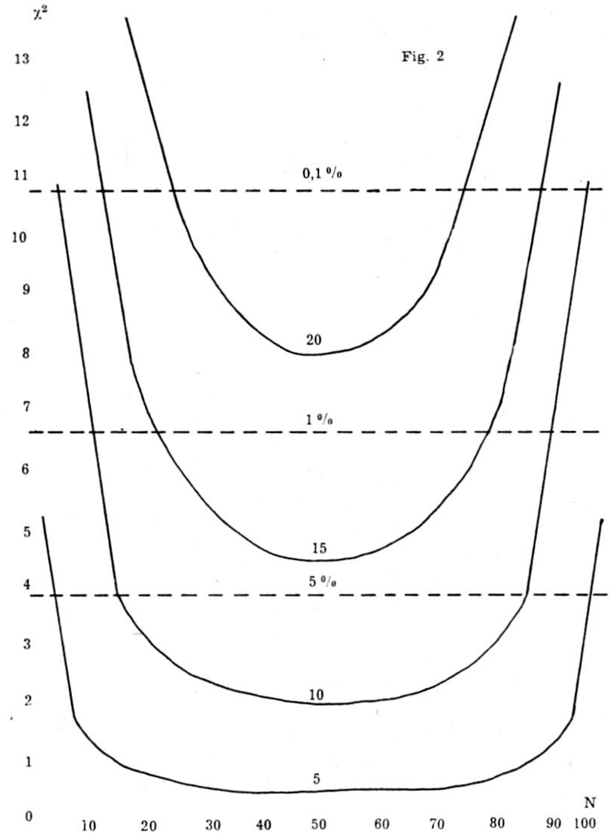

Fig. 2

on pourrait tracer une courbe pour chaque valeur d'écart, d'unité en unité, mais la figure présenterait moins de netteté.

Ces courbes ont été tracées par points, le X^2 étant porté en fonction de la valeur moyenne des deux différences comparées. Prenons, par exemple, la construction de la courbe correspondant à la valeur 10 de l'écart. Nous envisageons d'abord le cas où la différence aurait la valeur 0 (ou 100) d'après l'une des séries de cartes, et 10 (ou 90) d'après l'autre série.

	1re série	2e série	Totaux
Différences	0 (5)	10 (5)	10
Concordances	100 (95)	90 (95)	190
Totaux	100	100	200

$$X^2 = \frac{25}{5} \times 2 + \frac{25}{95} \times 2 = 10,9.$$

Les points d'ordonnée 10,9 correspondront aux abscisses 5 et 95.

Nous passons au point suivant :

	1re série	2e série	Totaux
Différences	10 (15)	20 (15)	30
Concordances	90 (85)	80 (85)	170
Totaux	100	100	200

$$X^2 = \frac{25}{15} \times 2 + \frac{25}{85} \times 2 = 3,9.$$

Les points d'ordonnée 3,9 correspondront aux abscisses 15 et 85. Et ainsi de suite.

Nous donnons les coordonnées des points ayant servi à construire les diverses courbes.

Valeur 5 de l'écart :

N	2,5 et 97,5	7,5 et 92,5	12,5 et 87,5	17,5 et 82,5	22,5 et 77,5
X^2	5,13	1,79	1,14	0,87	0,71
N	27,5 et 72,5	32,5 et 67,5	37,5 et 62,5	42,5 et 57,5	47,5 et 52,5
X^2	0,63	0,57	0,53	0,51	0,50

Valeur 10 de l'écart :

N	5 et 95	15 et 85	25 et 75	35 et 65	45 et 55
X^2	10,9	3,9	2,66	2,18	2,02

Valeur 15 de l'écart :

N	7,5 et 92,5	17,5 et 82,5	27,5 et 72,5	37,5 et 62,5	47,5 et 52,5
X^2	16,2	7,78	5,64	4,8	4,5

Valeur 20 de l'écart :

N	10 et 90	20 et 80	30 et 70	40 et 60	45 et 55
X^2	22,2	12,5	9,52	8,32	8,08

Un écart de 20 entre deux mesures est évidemment très improbable ; mais nous avons voulu montrer comment évoluait la famille de courbes en fonction du paramètre « écart des deux mesures d'un même interpoint ».

Nous constatons qu'entre les valeurs 4 et 96 de N, la courbe correspondant à l'écart 5 est constamment au-dessus du seuil de probabilité 5 %, c'est-à-dire qu'entre ces limites, un écart de 5 n'est pas significatif, et ne correspond qu'à des fluctuations aléatoires. Or, sur les 40 premières mesures, que nous avons prises comme échantillon, 37, c'est-à-dire 92,5 % présentent des écarts inférieurs à 5, et, par ailleurs, toutes les valeurs de N sont comprises entre 4 et 96.

La courbe correspondant à l'écart 10 n'est dans le domaine des fluctuations aléatoires que dans des limites un peu moins larges, entre les valeurs 15 et 85 de N. Aucun de nos 3 écarts supérieurs à 5, n'arrive à 10. L'écart 8 correspond à N = 15, et se trouve donc dans la zone aléatoire ; les deux écarts 6 correspondent à des valeurs 12 et 10 de N, et sont encore situés dans cette même zone.

Lorsque nous passons à la courbe correspondant à l'écart 15 (et il en serait déjà de même avec l'écart 14), nous constatons qu'elle ne coupe plus la droite représentative de 5 % de probabilité de l'hypothèse nulle. Tous les écarts égaux ou supérieurs à 14 ont donc valeur significative. Entre les valeurs 22 et 78 de N, les écarts 15 introduisent une probabilité d'hypothèse nulle comprise entre 1 et 5 % ; cette probabilité diminue beaucoup hors de ces limites.

Enfin la courbe correspondant à un écart égal à 20, nous montre qu'il introduit toujours un élément fortement significatif, la probabilité d'hypothèse nulle étant déjà inférieure à 0,5 % dans la partie centrale où N = 50, et s'abaissant à moins de 0,1 % hors de l'intervalle 25 - 75 de N.

Si les objets d'étude sont des atlas linguistiques non exhaustifs, c'est-à-dire pour lesquels chaque commune du domaine n'a pas constitué un point d'enquête, c'est la valeur corrigée N', qui devra être substituée à la mesure brute N. Il ne pourra guère s'agir que de mesures internes, comparant les diverses parties d'un même atlas, car deux atlas non exhaustifs différents ne choisissent pas forcément les mêmes communes comme points d'enquête.

Les fluctuations aléatoires s'amenuisent lorsque le nombre de cartes étudiées augmente. D'autre part, les valeurs des dénominateurs du X^2 s'accroissent, ce qui tend à diminuer celui-ci. Donnons-en quelques exemples.

Avec un écart de 20, la valeur minima du X^2, au milieu de la courbe, sera :

pour 100 cartes	1re série	2e série	Totaux
Différences	40 (50)	60 (50)	100
Concordances	60 (50)	40 (50)	100
Totaux	100	100	200

$$X^2 = 4 \times \frac{100}{50} = 8.$$

pour 200 cartes	1re série	2e série	Totaux
Différences	90 (100)	110 (100)	200
Concordances	110 (100)	90 (100)	200
Totaux	200	200	400

$$X^2 = 4 \times \frac{100}{100} = 4.$$

pour 400 cartes	1re série	2e série	Totaux
Différences	190 (200)	210 (200)	400
Concordances	210 (200)	190 (200)	400
Totaux	400	400	800

$$X^2 = 4 \times \frac{100}{200} = 2.$$

L'écart de 20 était très fortement significatif avec 100 cartes ; il ne l'est presque plus avec 200 cartes ; et il cesse de l'être avec 400 cartes.

Une recherche portant sur 400 cartes n'est pas invraisemblable : elle a été réalisée sur l'Atlas Lingüístic de Catalunya (2).

Le réseau de courbes de la figure 2, établi pour des études portant sur des tranches de 100 cartes, serait donc très fortement décalé vers le bas, si les études venaient à porter sur des tranches de 400 cartes : la courbe correspondant à l'écart 20, viendrait prendre la place de celle qui correspond ici à l'écart 10.

Les résultats obtenus sont valables, non seulement lorsque la méthode globale aura été appliquée, mais encore toutes les fois que d'autres méthodes, telles que la méthode dialectométrique de Jean Séguy (3) ou la méthode de l'indice général d'identité de Hans Goebl (4), mesurent, elles aussi, des différences linguistiques, ou simplement lexicales, entre points d'atlas linguistiques.

Chaque fois qu'une comparaison aura été établie entre des groupes différents de cartes linguistiques recouvrant un même domaine géographique, la présente étude permet d'apprécier si l'écart qui peut apparaître sur un interpoint, est, ou non, significatif.

Henri GUITER

---

(2) Antoni Griera. — Atlas lingüistic de Catalunya. Barcelona, 1924 et sq. Sor Anna Sardà i Enric Guiter. — L'Atlas lingüistic de Catalunya i la fragmentació dialectal del català. Miscellanea Barcinonensia XL, Barcelona, 1975, pp. 93-112.

(3) Jean Séguy. — La relation entre la distance spatiale et la distance lexicale. Revue de Linguistique Romane, 1971, pp. 335-357.

(4) Hans Goebl. — La dialectométrie appliquée à l'A.L.F. (Normandie). Actes du XIVe Congrès International de Linguistique Romane II, Naples, 1976, pp. 165-195.