PROGRAMMES D'ORDINATEUR POUR LA DETERMINATION ET LE TRAITEMENT STATISTIQUE DES FREQUENCES DE PHONEMES

On admet généralement que la fréquence des phénomènes linguistiques est stable. Cette affirmation signifie qu'il existe pour chaque langue une norme dont dépend la fréquence d'un phénomène; elle implique également que les variations dues à l'auteur ou au genre littéraire s'écartent peu de la norme*.

Cependant, il faut remarquer que cette double hypothèse n'a pas fait l'objet d'un nombre suffisant de vérifications; d'autre part, elle se borne à affirmer le peu d'importance des variations dues à l'auteur sans indiquer la limite des écarts admissibles.

Avant de définir la structure d'une langue au point de vue des sons, il est donc nécessaire de vérifier l'hypothèse de la stabilité des phénomènes. Mais une telle vérification suppose l'examen d'un très grand nombre de données qu'il serait difficile de rassembler manuellement; la machine apparaît ici comme un auxiliaire indispensable non seulement parce qu'elle permet au chercheur de disposer en un temps particulièrement court d'une

* G. HERDAN, *The Calculus of linguistic Observations*, La Haye, 1962, p. 18-19.

documentation qu'il aurait mis plusieurs mois à rassembler mais aussi parce que les résultats obtenus offrent une sécurité plus grande. Dans un même ordre d'idées, le calcul des écarts est fait par la machine de sorte que le philologue peut se consacrer davantage à l'analyse des résultats.

Notre premier objectif est de rendre possible l'établissement automatique de telles listes pour la langue latine. Nous avons réalisé un programme d'ordinateur qui permet d'effectuer rapidement des relevés sur un nombre important de phonèmes. La machine dénombre les différents phonèmes contenus dans un échantillon ainsi que le nombre total de phonèmes de cet échantillon. Nous avons prévu une seconde série de comptages : l'ordinateur établit une liste fréquentielle des phonèmes situés à l'initiale des mots et il compte le nombre de phonèmes initiaux; leur répartition peut en effet révéler certaines particularités stylistiques et permettre une étude détaillée du phénomène important de l'allitération. Après cette première recherche, et partant des résultats obtenus, nous avons mis au point un programme qui permet d'apprécier l'importance des écarts entre les différents échantillons ou l'écart d'un échantillon par rapport à la norme.

La plus grande partie de cet article sera consacrée au programme de comptage. Nous montrerons le déroulement de ses étapes principales et nous nous attarderons quelque peu sur certains problèmes posés par des cas particuliers. Dans une seconde partie, nous exposerons le fonctionnement général du programme de calcul.

Nos programmes ont été réalisés pour un ordinateur IBM 1620 auquel est connecté un lecteur-perforateur de cartes (IBM 1622).

Le support matériel des données est la carte mécanographique IBM-80 colonnes. Le programme de comptage est conçu pour l'exploitation des

fichiers du L.A.S.L.A. où le texte est perforé à raison d'un mot par carte dans une zone déterminée. Mais il est possible, en utilisant les indicateurs de la console, de changer les limites de cette zone et même de traiter un fichier perforé en continu; la seule exigence dans ce cas est qu'il y ait une colonne blanche entre les différents mots.

Comme nous l'avons dit, deux séries de comptages s'effectuent au cours d'un même traitement; la machine dénombre les occurrences de chacun des phonèmes de l'échantillon ainsi que le nombre total de phonèmes. De la même façon, on établit la fréquence de chacun des phonèmes à l'initiale d'un mot et le nombre total de phonèmes initiaux.

.- LES PHONEMES

La première difficulté que nous avons rencontrée a été la définition de l'unité de comptage. On peut opposer la lettre au son et se demander dans quelle mesure l'une reproduit l'autre. C'est là une question qui n'entre pas dans notre propos. En ce qui concerne le latin, langue morte, le problème est plus difficile encore : sa prononciation a soulevé plus d'une controverse.

Cependant, le comptage de lettres, tout comme celui de sons, présente un intérêt évident. Par ailleurs, la lettre isolée peut être définie comme la représentation écrite d'un phonème. C'est pourquoi il nous a semblé que, dans un domaine assez peu exploré, le choix d'une unité simple s'imposait; aussi, celle que nous avons retenue se situe au niveau le plus élémentaire, celui du son indivisible représenté dans la plupart des cas par un caractère alphabétique. Nous avons retenu 23 phonèmes "simples", à savoir les 21 signes que comportait l'alphabet latin au temps de César et de Cicéron auxquels nous avons ajouté Y et Z.

D'autre part, nous avons joint à cette liste les sons représentés par deux caractères : les diphtongues AE, AU, EU et OE et les consonnes aspirées CH, PH et TH.

Notre liste ainsi constituée comprend au total 30 phonèmes différents.

2.- LE DEROULEMENT DU PROGRAMME

Après avoir décrit les problèmes généraux que présentent les relevés automatiques de sons, nous montrerons comment nous avons adapté le traitement aux fichiers du L.A.S.L.A. dont on a trouvé une description dans cette Revue*.

Le comptage des phonèmes simples ne pose guère de difficultés : la machine compare le contenu d'une colonne de la carte traitée avec la liste des phonèmes jusqu'à ce qu'elle trouve l'identité entre les deux; à ce moment, elle procède à une addition de 1 dans le compteur correspondant, puis elle passe au contenu de la colonne suivante.

Cependant, l'automatisme de ce procédé exige une possibilité de contrôle; en effet, si par erreur un caractère non-valide (un chiffre, par exemple) s'est introduit dans le fichier, la machine ne peut trouver son correspondant dans la liste des phonèmes; elle risque alors de s'arrêter et de détruire les résultats enregistrés. Pour empêcher de tels accidents, nous avons prévu une intervention à la machine à écrire connectée à l'ordinateur; ce dernier,

* A. BODSON et S. GOVAERTS, *Codification d'un texte latin sur cartes mécanographiques 80 colonnes,* dans *Revue de l'Organisation Internationale pour l'Etude des Langues anciennes par Ordinateur,* 1966, 1, pp. 1-46.

lorsqu'il rencontre sur une carte un caractère non-valide, transmet par la machine à écrire le contenu de cette carte et s'arrête. L'opérateur peut à ce moment enregistrer en mémoire, par l'intermédiaire de la machine à écrire, la forme correcte et ensuite permettre à l'ordinateur de continuer le travail.

Par définition, la machine repère une lettre initiale lorsqu'elle traite la première colonne de la zone où se trouve la forme. L'ordinateur compte également comme lettre initiale le caractère précédé par une colonne sans perforation; ce système est utilisé pour les formes complexes de la conjugaison dont nous devrons parler dans la suite.

En ce qui concerne les diphtongues ainsi que les lettres doubles CH, PH et TH, nous avons adopté des systèmes de repérage différents. Pour CH, PH et TH, le repérage s'effectue à partir du second caractère. C'est lorsque H se présente à la machine que se fait l'examen de la colonne précédente pour savoir si elle contient C, P ou T. Nous avons préféré ce système pour accélérer le plus possible le traitement. H est un phonème relativement rare, beaucoup moins fréquent, en tous cas, que C, P ou T; il était, dès lors, plus économique d'effectuer le repérage de ces groupes à partir du son H.

Pour les diphtongues, c'est le premier élément qui entraîne le traitement de la colonne suivante : lorsque la machine rencontre A, elle vérifie si le caractère suivant est E ou U (V); le procédé est le même pour E (EU) et pour O (OE).

Le mode d'identification des phonèmes ainsi que le système de comptage que nous venons de décrire sont fort simples et permettent de compter de façon purement automatique un nombre élevé de phonèmes en peu de temps : l'ordinateur identifie et compte en dix minutes entre deux mille cinq cents et trois mille phonèmes.

Pour terminer cette description générale du programme, il nous reste à ajouter que, lorsque les comptages sont finis, l'ordinateur transmet les résultats de deux manières : il les imprime à la machine à écrire et les perfore sur cartes mécanographiques. Ces cartes, nous le verrons, seront utilisées dans des opérations ultérieures.

LES DIPHTONGUES

Nous avons déjà fait allusion aux difficultés que présentent les diphtongues dans un traitement automatique. Il faut y revenir.

En effet, deux voyelles successives ne forment pas nécessairement une diphtongue. Ainsi dans le mot AER, A et E sont deux phonèmes distincts, tandis que pour CAEDO, il s'agit d'une diphtongue.

Par ailleurs, il n'existe aucune règle qui permette à la machine de différencier les cas où deux voyelles forment diphtongue. Nous avons dû examiner chaque cas et voir dans quelle mesure il était possible d'énoncer des principes utilisables en machine.

Plusieurs cas n'ont pu être traités de façon automatique; lorsqu'ils se présentent, ils sont transmis à la machine à écrire pour que l'opérateur indique par un chiffre le nombre de syllabes du groupe vocalique.

AU

Ce groupe est le seul pour lequel on puisse se fonder sur le principe suivant : AU est diphtongue lorsqu'il n'est pas directement précédé ou suivi

d'une voyelle. Partout ailleurs, il s'agit de deux phonèmes distincts.

p. ex. aufero (au)
auis (a et u)
lautitia (au)
lauare (a et u)

AE

Le cas de AE est singulièrement plus complexe que le précédent. Nous devons distinguer AE en première syllabe et AE en finale.

1) En première syllabe

a.- A l'initiale

En position initiale, AE n'est pas traité automatiquement; il aurait fallu cataloguer les cas où il s'agit d'une diphtongue et ceux où on est en présence de mots du type AER. La difficulté a été résolue en prévoyant une intervention de l'opérateur à la machine à écrire. Toutefois, AER et ses dérivés, relativement fréquents dans la langue, ont été rangés en mémoire et sont traités automatiquement.

b.- En première syllabe après une consonne

Ici encore, pour un certain nombre de mots, l'opérateur doit fournir la solution à la machine; cependant, nous avons pu distinguer certains groupes (consonne + AE), particulièrement fréquents, dans lesquels AE est toujours diphtongue. Ils ont été placés en mémoire et sont comptés sans

intervention de l'opérateur. C'est ainsi que nous procédons pour CAE, MAE, PAE, PRAE et SAE.

2) AE a la finale d'un mot

Les mots du type DANAE, c'est-à-dire les mots empruntés au Grec, écartent toute possibilité d'un traitement automatique. Nous savons que ces mots sont assez peu nombreux et qu'ils n'ont pas une fréquence très élevée mais nous n'avons pas trouvé de relevés exhaustifs des mots dont la finale est de ce type. Nous avons dû prévoir ici encore une intervention de l'opérateur.

En revanche, le fait que, dans la technique du L.A.S.L.A., chaque mot porte une analyse morphologique codifiée nous a permis d'établir une règle partielle. Nous avons défini comme suit la présence de AE diphtongue en finale :
AE, en fin de mot, est diphtongue lorsque la forme est analysée comme génitif ou datif singulier et comme nominatif ou vocatif pluriel. Ce recours à l'analyse de la forme permet de traiter automatiquement presque tous les cas de AE en finale.

EU et OE

Ces groupes se rencontrant dans un assez petit nombre de mots, nous n'avons pas prévu de traitement particulier; chaque fois que l'ordinateur perçoit une de ces diphtongues, il attend une intervention de l'opérateur. Cependant, les exemples les plus fréquents de EU diphtongue (CEU, EHEU, HEU, NEU et SEU) ont été placés en mémoire et sont traités immédiatement par la machine.

Les interventions de l'opérateur

Nous avons utilisé notre programme sur plusieurs échantillons d'environ 25.000 phonèmes; il nous est apparu que l'opérateur devait intervenir entre cent et cent-cinquante fois pour fournir une solution à la machine. Ce résultat semble assez satisfaisant et n'entraîne pas un ralentissement excessif du déroulement des opérations. Cependant, lorsque nous aurons appliqué à un nombre suffisant de textes ce programme de comptage, nous pensons qu'il sera utile de prévoir un traitement automatique pour les mots qui se présentent le plus souvent. Nous avons prévu, à cet effet, la possibilité d'enrichir les listes qui se trouvent en mémoire centrale.

3.- LES ENCLITIQUES

Les fichiers du L.A.S.L.A., nous l'avons rappelé, portent un mot par carte. Pour des raisons évidentes d'analyse et de comptage, l'enclitique est séparé du mot sur lequel il s'appuie et est perforé sur la carte suivante. Or, dans les relevés d'initiales, on ne peut compter la première lettre de l'enclitique et comme, d'autre part, la machine identifie une lettre initiale lorsqu'elle traite le contenu de la première colonne réservée à la forme, il fallait prévoir une boucle de programme qui empêche le comptage des lettres initiales des enclitiques.

A cet effet, nous avons rangé en mémoire CUM, NE, QUE et VE. Dans le cas de QUE et de VE, le traitement est facile mais pour NE et CUM, il a fallu ajouter des tests supplémentaires.

NE enclitique se distingue de ses homographes par le fait qu'il est adverbe

interrogatif; pour le comptage de ce mot, nous avons fait intervenir les codes d'analyse du L.A.S.L.A.

Pour CUM, le problème est plus complexe; en effet, CUM préposition n'est enclitique que lorsqu'il régit un pronom personnel. Lorsque CUM se présente à la machine, celle-ci vérifie s'il s'agit d'une préposition; si c'est le cas, l'ordinateur va consulter une zone de réserve où il a placé le contenu de la carte précédente et il vérifie si cette carte porte un pronom personnel à l'ablatif.

4.- QUELQUES PARTICULARITES PROPRES AUX FICHIERS DU L.A.S.L.A.

Pour terminer cette première partie, nous dirons quelques mots des fichiers que nous traitons habituellement.

A.- Les formes composées de la conjugaison

Selon les conventions du L.A.S.L.A., ces expressions sont perforées deux fois. La première carte porte le lemme de l'élément significatif, la seconde celui de l'auxiliaire. Dès lors, prendre en considération les deux cartes équivaudrait à compter deux fois la même forme. C'est pourquoi l'ordinateur ne tient pas compte des formes reprises sous le lemme de l'auxiliaire, la forme complexe est traitée quand se présente la carte qui porte le lemme de l'élément significatif. En ce qui concerne le comptage des initiales, c'est la présence d'une colonne sans perforation entre les deux éléments qui permet à la machine de distinguer les deux mots.

D'autre part, lorsque, pour une de ces formes dont nous venons de parler, l'auxiliaire est sous-entendu, on le perfore à la suite de l'élément significatif en le faisant précéder et suivre d'une perforation symbolique.

L'auxiliaire ainsi restitué ne doit pas être pris en considération. C'est le code spécial dont l'auxiliaire est entouré qui permet de ne compter que l'élément significatif.

B.- Les expressions complexes

Les expressions du type NON SOLUM SED ETIAM sont entièrement perforées sur quatre cartes qui ne diffèrent entre elles que par le lemme. Il fallait dans ce cas trouver un système analogue à celui employé pour les formes de la conjugaison afin de ne compter qu'une seule fois chaque élément. Nous avons utilisé la solution suivante : les mots qui composent ces expressions sont tous invariables, leur forme est identique à leur lemme; par conséquent, lorsqu'une de ces expressions est traitée, le comptage se fait sur le lemme.

* * *

Nous ne nous attarderons guère au second programme; nous nous limiterons à en donner les principes généraux.

Les résultats fournis par les comptages ne peuvent être interprétés sans un recours au calcul des écarts par rapport à une norme ou à une moyenne.

Le test de χ^2 donne, à cet égard, des indications très intéressantes.

Pour un phénomène relativement stable, la fréquence observée dans un échantillon peut être utilisée comme approximation de probabilité (p'). Cette approximation est d'autant meilleure que l'échantillon est plus étendu. On désigne alors par q' l'approximation de la probabilité du contraire

c'est-à-dire la fréquence des expériences où se produit un phénomène autre que celui que l'on considère. Ainsi, on pourra par exemple étudier la fréquence de A et celle d'un phonème autre que A.

Le programme que nous avons réalisé consiste à calculer pour des échantillons différents la valeur de p' et de q', les effectifs théoriques, les χ^2 partiels et leur somme selon la formule :

$$\chi^2 = \Sigma \frac{(x_i - n_i p')^2}{n_i p' q'}$$

En parlant du programme de comptage, nous avons dit que, en fin de travail, les résultats étaient perforés sur cartes; ce sont ces cartes qui servent de données au programme de calcul. Elles portent les renseignements suivants :

1) l'indication du phonème considéré;
2) la fréquence de ce phonème dans l'échantillon;
3) le nombre total de phonèmes de l'échantillon.

Les mêmes renseignements sont fournis, par une seconde série de cartes, au sujet des phonèmes initiaux avec, à la place du nombre total de phonèmes, le nombre de phonèmes initiaux.

Pour utiliser le programme de calcul, il faut disposer des résultats de deux échantillons au moins (sans quoi aucune comparaison n'est possible). Il est nécessaire, avant de calculer les écarts, d'interclasser les cartes résultats des échantillons et de placer les unes à la suite des autres les cartes qui concernent un même phonème. Cette opération peut se faire à l'aide d'une trieuse.

Au début d'une série de calculs, l'opérateur indique par un chiffre le nombre d'échantillons à comparer; ce renseignement désigne à la machine la quantité de cartes qu'elle doit enregistrer avant de commencer les opérations.

Nous avons vu que la machine dispose de deux séries de chiffres : le nombre total de phonèmes et le nombre de phonèmes A, par exemple. Le calcul de la probabilité (p') s'obtient en divisant le total des phonèmes A par le nombre total de phonèmes.

L'effectif théorique d'un échantillon a comme valeur le produit de p' et du nombre total de phonèmes de cet échantillon.

Pour obtenir la valeur de χ^2 partiel, il suffit de diviser le carré de la différence entre les effectifs réels et les effectifs théoriques par le produit du nombre total de phonèmes de l'échantillon et p'q'.

En ce qui concerne le temps d'opération, l'ordinateur calcule trente fois, en moins de trois minutes, la somme des χ^2 de deux échantillons. Ce temps augmente à peine si le nombre de parties est plus élevé.

* * *

Les programmes dont nous venons de parler ont comme objectif de rendre possibles la description et l'étude phonétiques de la langue latine; ils présentent cet avantage de supprimer toute tâche de documentation : le philologue peut se consacrer entièrement à l'analyse et à l'interprétation des données. C'est là, nous semble-t-il, un résultat appréciable.

Joseph DENOOZ
Boursier du Patrimoine
Université de Liège