RECHERCHE STATISTIQUE SUR LA COMPOSITION PHONOLOGIQUE DE LA LANGUE ITALIENNE EXÉCUTÉE AVEC UN SYSTÈME IBM

A. ZAMPOLLI

Plusieurs observations empiriques ont montré la stabilité des fréquences relatives des formes linguistiques, fournissant les prémisses pour l'étude statistique du langage qui propose, entre autre, une interprétation en termes statistiques de la distinction saussurienne entre langue et parole. La langue est conçue non seulement comme l'ensemble, organisé en système des formes que l'on rencontre aux différents niveaux constitutifs du langage (niveau phonétique, phonématique, lexical, morphologique etc...), mais comme l'ensemble de ces unités avec, en plus, leur respective probabilité d'emploi. Le concept de forme linguistique est associé d'une façon inséparable avec sa fréquence d'apparition, et avec les fréquences de ses combinaisons, avec les autres formes du même niveau. La langue présente ainsi les attributs essentiels d'une population statistique; elle est caractérisée par les fréquences relatives et par les probabilités définies de la variable qui est représentée, dans le cas présent, par les formes de chaque niveau. Ces probabilités ne sont pas connues a priori, il est seulement possible de les évaluer a posteriori sur la base des fréquences relatives pouvant être observées dans les actes de parole concrets et individuels. Ces actes prennent pourtant, vis-à-vis de la langue, la fonction d'échantillons statistiques de la population. Donc la matière première indispensable pour la vérification de toute hypothèse, ou pour la formulation de toute loi statistique sur le langage, est l'ensemble des données statistiques sur sa composition. Leur absence est encore aujourd'hui le premier obstacle au développement de ces études.

Naturellement la situation est différente d'une langue à l'autre, et pour l'Italien on a fait peu de choses. Cette recherche exécutée comme thèse de lettres (A. Zampolli: Études de statistique linguistique exécutées avec des installations IBM Université de Padoue année 1959—1960), se propose de combler la lacune en ce qui concerne la composition phonétique de l'Italien.

La récolte de la documentation statistique demande, avant tout, l'abstraction ou formation des cas statistiques.

Pour arriver à la formation des cas statistiques, il est nécessaire de choisir les traits devant être relevés parmi les nombreux caractères reconnaissables dans la composition phonétique d'une langue. Les phonèmes selon la défi-

nition de l'école structuraliste répondent parfaitement au concept de cas statistiques et il est certain que la statistique, en principe, doit être statistique d'unités phonologiques. Les analogies qui résultent lorsqu'on compare la définition de phonème de Troubetzkoy (*Principes de phonologie*, trad. franç. de J. Cantineau, Paris 1949, pp. 39—40) et la définition de cas statistiques de Boldrini (*Statistica*, Teoria e Metodi, pp. 111—112) sont significatives.

A partir de cette considération, nous avons dû prendre conscience de problèmes de fond, de nature méthodologique, qui peuvent être ici résumés dans les 4 points suivants:

I. choix de la prononciation pour retranscrire le texte échantillon
II. liste des unités phonologiques dont il fallait compter les fréquences
III. choix des principales variantes dont il fallait relever les fréquences
IV. choix et dimensions de l'échantillon.

Nous les exposons brièvement:

I.

Il n'est pas possible de passer directement de la façon de parler individuelle, à un type unique de prononciation nationale. Dans ce domaine les études ont à peine commencé, toutefois il est désormais certain qu'on ne peut parler d'un seul système phonématique italien, mais seulement de différents systèmes phonématiques régionaux. Les différences consistent dans les phonèmes employés et dans la différente distribution des phonèmes communs dans les mots. Par exemple l'opposition entre la variété fermée et ouverte du *e* et du *o* est connue dans de nombreux types d'Italien, toutefois il est facile d'établir une liste de morphèmes qui dans une région sont prononcés avec *e* ou *o* fermés, dans une autre région avec *e* ou *o* ouverts et vice-versa.

Les florentins prononcent: *bène, piède, cèntro, accapatòio* et les vénitiens *béne, piéde, céntro, accapatóio* (G. Lepsky).

Le choix d'un système régional de prononciation plutôt qu'un autre conditionne deux moments importants de l'enquête statistique: il fait varier le nombre et l'inventaire des phonèmes (cas statistiques) dont on recherche la fréquence, et il fait varier la transcription phonématique, qui détermine la distribution des phonèmes dans le texte, c'est-à-dire la composition de l'échantillon choisi. Nous avons choisi le Florentin cultivé débarrassé des caractéristiques et des nuances de la prononciation dialectale, car il offre les avantages suivants aux fins du relévé statistique.

Avant tout, la majeure partie des études de phonématique ou de phonétique italienne prennent comme base l'italien florentin; ainsi, la majeure partie des tableaux ou des listes de phonèmes de l'italien représentent en réalité le système florentin; en second lieu les dictionnaires donnent généralement la prononciation des lemmes selon l'usage florentin; en plus celui-ci

est probablement le type d'italien le plus connu auprès des specialistes étrangers. Tout ceci sans considérer que, sur la base d'une série de considérations historiques et de constatations de fait, le florentin est proposé comme modèle orthoépique et présenté comme prononciation nationale. Citons à ce propos les propositions très claires du Dictionnaire Encyclopédique Italien (p. XIII—XIV).

II.

Les spécialistes qui ont décrit le système phonématique de l'italien littéraire, c'est-à-dire sans adjectifs, florentin ou toscan cultivé, ne sont pas d'accord sur le contenu des listes de phonèmes qui le composent. Le nombre des phonèmes établis varie d'un minimum de 27 (G. Porru et d'autres) à un maximum de 54 (Castellani). Cette situation pose naturellement un grand problème d'une importance décisive pour l'organisation et les limites de *validité* de toute la recherche statistique.

Les divergences parmi les systèmes phonologiques proposés, excluent-elles la possibilité de mener la recherche de façon que les résultats obtenus soient utilisés par le spécialiste quel que soit son système phonologique? En d'autre terme, est-il inévitable de conduire les opérations statistiques selon un système phonologique déterminé, ayant pour conséquence que, à la fin des opérations, les données relevées (fréquences, combinaisons, etc...) ne sont rigoureusement valables que dans la perspective particulière du système phonématique adopté? *Les oscillations du nombre des phonèmes s'expliquent par la constatation de quelques points controversés*, pour chacun desquels on donne deux ou plusieurs solutions différentes. L'examen de ces divergences nous a amenés à cette conclusion: il est possible d'établir une liste d'unités phoniques telles que la fréquence des phonèmes de n'importe lequel des systèmes proposés

— ou bien coincide avec la fréquence d'une des unités phoniques,
— ou bien est égale à la somme des fréquences de deux ou plusieurs unités phoniques opportunément choisies.

Puisque la vérification de ces affirmations serait trop longue, nous nous limitons à établir la liste des points controversés, et la liste des unités correspondantes dont nous avons compté les fréquences.

Voyelles et semi-voyelles

Le système italien, selon l'opinion la plus commune, emploie phonologiquement deux des propriétés caractéristiques des voyelles: le degré d'ouverture orale et la localisation, ou zone d'articulation. La position plus répandue est probablement celle qui reconnaît 7 phonèmes vocaliques et deux semi-vocaliques: respectivement: *u*, *o̩*, *o̧*, *a*, *e̩*, *ȩ*, *i* et *w*, *y*. Cependant, différents systèmes sont proposés, puisque on interprète différemment le rôle de l'accent, l'opposition entre les deux variétés, ouverte et fermée, de *e* et de *o*,

et l'opposition *voyelles: semi-voyelles*. Le nombre oscille entre 5, 7, 9, 10, 12 phonèmes vocaliques, avec, ou sans, addition de deux phonèmes semi-vocaliques. Dans notre statistique, on compte les fréquences de 14 unités: *a, à, e, é, è, o, ò, ó, i, ì, y, u, ù, w.*

Accent. On pose le problème s'il faut considérer l'accent comme un trait pertinent devant être ajouté aux autres traits distinctifs des voyelles, c'est-à-dire, le point d'articulation et le degré d'ouverture. En ce cas, 5 des 7 voyelles citées: *i*, *ɐ*, *a*, *ɑ*, *u* (mais, non *ɐ*, *ɑ* qui n'apparaissent jamais en position atone) seraient dédoublées chacune en deux phonèmes autonomes, l'un atone et l'autre tonique: ainsi, la liste des phonèmes vocaux s'élèverait à 12 unités (*a, à, e, é, è, o, ó, ò, i, ì, u, ù*), ou bien à 10, si on ne considère pas phonématiques les oppositions *é : è*, *ó : ò*. Il est évident que, sans prendre position en faveur d'une de ces alternatives, il est possible de fournir des résultats utilisables, quelle que soit l'alternative choisie. Il suffit en effet de compter séparément, pour chaque voyelle, la fréquence en position tonique et en position atone. Celui qui ne met pas l'accent entre les caractéristiques phonématiques, peut obtenir la fréquence de chaque phonème vocalique en additionnant la fréquence de ses apparitions en position atone et la fréquence de ses apparitions en positions toniques.

Opposition ẹ : ɐ, ọ : ɑ

Certains nient l'importance phonématique des oppositions en question, et les mettent parmi celles qui „nella lingua italiana hanno importanza non tanto per l'essenziale differenziazione semantica, quanto per l'eleganza della dizione" (W. Belardi). Des affirmations de ce genre doivent probablement être attribuées aux différences considérables des usages locaux. Les deux variétés de *e* et de *o* sont connues et employées dans la majeure partie du domaine linguistique italien, mais leur distribution entre les racines et les morphèmes, comme on a vu, est très differente d'une région à l'autre.

Des paires minimales existent probablement dans chaque type d'Italien, mais elles sont différentes d'un type à l'autre, et certainement elles ne sont pas très nombreuses. En outre les différentes prononciations sont généralement tolérées. Ceux qui cherchent à emprunter leur propre prononciation au modèle orthoépique toscan, le font généralement pour de motifs *professionels* de correction, de style, d'élégance de diction: la plupart de ces personnes sont des speakers, des acteurs, des enseignants, etc.

Cependant, il est certain que dans le système phonématique florentin sont présentes des paires minimes. Ces oppositions sont neutralisées en position atone où apparaît l'archiphonème, qui habituellement est considéré identique à *ẹ*, *ọ* en position tonique. En résumant ce qui concerne *e* et *o*, on pose deux problèmes presque indépendants l'un de l'autre, chacun avec deux possibilités alternatives:

a) traiter l'accentuation comme trait pertinent des voyelles, ou considérer l'accent à part comme un fait prosodique?
b) attribuer, ou non, la relevance phonématique à l'opposition entre les deux variétés, fermée et ouverte?

On aura naturellement les diverses combinaisons pour chaque solution de *a)*, avec chaque solution de *b)*, ce qui donne 4 hypothèses. Il est possible d'établir la fréquence des phonèmes considérés par chacune d'entre elles, quand on connaît la fréquence des 6 unités vocaliques suivantes:

e	en position atone
é	fermé en position tonique
è	ouvert en position tonique
o	en position atone
ó	fermé en position tonique
ò	ouvert en position tonique.

Semi-voyelles

Il n'y a aucun doute sur la nature particulière, articulatoire et acoustique des semi-voyelles *y* et *w*, si bien que certains phonéticiens les classent parmi les sons *consonantiques*.

Cependant, doivent-elles être considérées comme des phonèmes autonomes, ou plutôt comme simples variantes *combinatoires* des phonèmes respectifs *i* et *u*?

Pour *i* et *u* on a donc, comme pour *e* et *o*, deux problèmes, dont chacun admet deux solutions:

a) traiter, ou non, l'accentuation comme *trait pertinent* des voyelles,
b) considérer, ou non, pertinente l'opposition voyelle: semivoyelle.

Naturallement on a, à partir des combinaisons de chaque solution de a) avec chaque solution de b), 4 hypothèses possibles:

1. on considère deux phonèmes: *i, u*
2. on considère quatre phonèmes: *i, í, u, ú*
3. on considère quatre phonèmes: *i, y, u, w*
4. on considère six phonèmes: *i, í, y, u, ú, w*

Il est immédiatement évident que la solution de compter les fréquences de 6 unités, *i, í, y, u, ú, w* permet de calculer les fréquences des phonèmes considérés par chacune des 4 hypothèses citées.

Consonnes

En exécutant les statistiques nous avons compté les fréquences de 40 unités consonantiques, *p,p|, b, b|, t, t|, d, d|, č, č|, ǧ, ǧ|, k, k|, g, g|, f, f|, v, v|, s, s|, š, S, l, l|, r, r|, l', n, m, ⁺m, m̨, ⁺n, n, n̨, m|, n|, z, z'*. Nous écrivons les consonnes géminées avec le symbole de la simple plus/. Citons également, comme pour les voyelles, les points controversés.

Nasales

La position la plus commune consiste à considérer en plus des *n'* (palatalisée) et, éventuellement, des géminées *m* (et *n*), deux phonèmes nasaux: *m* et *n*. On dit du *m* qu'il peut apparaître devant voyelle ou consonne labiale *p* et *b*; de *n*, on distingue au moins 3 variétés: *m̨* qui apparaît devant les labiodentales *f* et *v*; *n̨* qui apparaît devant les velaires *k* et *g*; et enfin *n* qui apparaît devant toutes les voyelles et toutes les consonnes sauf *p, b, f, v, k, g*. On ressent aussi le problème si la variété *m̨* doit être considérée une variante combinatoire du phonème *m* plutôt que du phonème *n*.

Une troisième solution serait de considérer pertinente l'opposition *m* : *n* seulement devant voyelle, tandis qu'elle serait neutralisée devant consonne, position où la réalisation de l'archiphonème serait conditionnée par la consonne suivante. Pour répondre aux exigences de ces trois solutions différentes il est nécessaire de compter les fréquences de 9 unités:

m cons. nasale bilabiale devant voyelle
⁺m cons. nasale bilabiale devant p — b
m̨ cons. nasale labiodentale devant f, v
n̨ cons. nasale velaire devant k, g
⁺n cons. nasale dentale devant consonne sauf p, b, f, v, k, g,
m/ cons. nasale bilabiale géminée (seulement) prévocalique
n/ cons. nasale dentale géminée (seulement) prévocalique
n' cons. nasale palatale (seulement) prévocalique

Opposition sourde: sonore pour les consonnes s et z

Certains auteurs ne retiennent pas pertinente l'opposition *s* : S, exemple: *fuso* — (le fuseau): *fuSo* (part. de fondre), et éliminent S de la liste des phonèmes consonantiques.

Dans notre statistique nous avons compté séparément la fréquence de *s* sonore et de *s* sourde. Si quelqu'un ne retient pas pertinente cette opposition, il peut additionner les deux fréquences et les attribuer à un seul phonème, réalisé tantôt sourd, tantôt sonore.

Géminées

En ce qui concerne les lettres appelées doubles, se pose le problème suivant: savoir si les doubles (appelées aussi géminées, fortes, longues) doivent être considerées, ou non, comme des phonèmes autonomes, distincts des simples (faibles, brefs) correspondants. Chez les auteurs, l'on trouve trois solutions:

1. Les géminées sont interprétées comme biphonématiques, c'est-à-dire répétitions d'un même phonème.
   L'opposition *pala: palla* serait du type *pala : palma*. Dans *pala*, on compte 4 phonèmes; dans *palma* et dans *palla*, 5.

2. Les géminées, à qui on a reconnu des caractéristiques monophonématiques, sont opposées aux simples en tant que phonèmes autonomes. L'opposition *pala : palla* serait du type *pala : paga*. Dans *palla*, on compte 4 phonèmes comme dans *pala* et dans *paga*. 15 consonnes (*p, b, t, d, č, ǧ, k, g, f, v, s, l, r, m, n*) donennt alors 30 phonèmes: (*p : p|; b : b|; t : t|;* etc...) Restent exclus de l'opposition *géminées: simple*, les consonnes *l'*, *n'*, *š*, *z*, *z'*, car toujours longues, et le phonème S car toujours bref, ainsi que les variétés nasales ^{+}m, m_{ξ}, ^{+}n, n_{ξ}.
3. Aux géminées, est reconnue une nature monophonématique, mais elles ne sont pas opposées aux simples comme phonèmes autonomes; simple et géminée correspondante sont simplement considérées comme degrés différents d'articulation (respectivement faible et fort) d'un même phonème.

Dans notre statistique on a compté séparément les fréquences des 30 unités proposées par la 2ème solution.

Ceux qui soutiennent la 3ème solution, doivent seulement additionner entre elles la fréquence de la simple et de la géminée; au contraire, ceux qui acceptent la 1ère solution doivent ajouter à la fréquence de la simple la fréquence, rédoublée, de la géminée correspondante.

III.

Le nombre des variantes est pratiquement illimité, c'est pourquoi on se demande si on doit compter les fréquences des variantes combinatoires et de quelles variantes. Comme on a vu, les phonèmes considérés selon la définition structuraliste, répondent parfaitement au concept de cas statistique et il n'y a aucun doute qu'une statistique du matériel phonique d'une langue doit être, en principe, statistique des unités phonologiques distinctives.

Toutefois, ce n'est pas seulement en vue de l'utilisation en phoniatrie et pour la théorie de l'information qu'il est intéressant de relever les fréquences des variantes; de nombreuses études récentes en soulignent l'importance lingustique. *Ou bien sur un plan synchronique*, comme éléments qui, étant communs à tout le langage d'une communauté, constituent une sorte d'accompagnement constant, si bien qu'ils contribuent à l'identification et à la reconnaissance du phonème ou d'un groupe de phonèmes; *Ou bien sur le plan diachronique*, comme facteurs déterminants dans de nombreux processus d'évolution phonétique.

C'est pourquoi, en plus des fréquences de certains types de variantes combinatoires, particulièrement soulignées par les spécialistes de phonétique italienne, nous avons relevé aussi toutes les combinaisons de 2 ou 3 phonèmes avec leur fréquence d'occurence.

Durée des voyelles

La différence entre voyelle longue et voyelle brève est considérable en italien, même si elle ne semble pas phonologiquement pertinente. Le facteur essentiel est la durée proportionnelle. Nous pouvons parler sur des temps différents, et la durée absolue de chacune des articulations en sera profondément changée, sans que les rapports de longueur entre les articulations successives soient beaucoup changés.

Cette durée proportionnelle dépend aussi du milieu phonétique et particulièrement de la structure de la syllabe. Les auteurs sont pratiquement d'accord en considérant longues les voyelles en syllabe tonique ouverte, et brèves toutes les autres (Jossely, Battisti). Nos tableaux comportent pour chaque voyelle tonique les fréquences de ses apparitions:

a) *en syllabe ouverte*,
   ou suivie par une consonne sourde,
   ou suivie par une consonne sonore;
b) *en syllabe fermée*,
   ou suivie par une géminée,
   ou suivie par une nasale,
   ou suivie par *l*.

Voyelles nasalisées

Les voyelles nasales sont généralement produites par le fait qu'une consonne nasale nasalise plus ou moins la voyelle contiguë. Parfois ce phénomène d'assimilation arrive jusqu'à la chute complète de la consonne nasale (Panconcelli). Sur le plan diachronique l'assimilation joue un rôle considérable dans l'évolution phonétique de plusieurs langues. Les *Prof. Croatto* et ces qui ont fait des études particulières sur la nasalisation des voyelles italiennes, soulignent les reflets importants de ce phénomène dans certains troubles de la parole. La nasalisation, généralement régressive, mais qui peut être aussi progressive, se fait normalement dans l'unité syllabique d'une façon particulière, quand la voyelle est suivie par les groupes *mf*, *ns*, *nl*, *nr*. On a ainsi compté les fréquences des voyelles:
1. précédées d'une nasale — 2. précédées d'une nasale dans la même syllabe — 3. suivies par une nasale — 4. suivies par une nasale dans la même syllabe — 5. précédées et suivies en même temps de nasales.

Degrés d'intensité des consonnes

Certains spécialistes de phonétique distinguent 3 degrés d'intensité (faible, moyen et fort) pour les consonnes *p; b; t; d; č; ǧ, k, g, f, v; s; l; r; m; n*. Pour le degré fort, ce que nous avons dit pour les géminées est valable. Le contraste *faible: moyen* est certainement sans valeur phonologique. Cepen-

dant nous avons compté distinctement les fréquences des dites consonnes en position intervocalique ou suivies par *l*, *r* (degré faible), et les fréquences en début de phrase phonétique, ou précédées par consonne, ou suivies par consonne différente de *l*, *r* (degré moyen). Il serait trop long d'expliquer ici le traitement réservé à *š*, *n'*, *l'*, *z*, *z'*. Pour le renforcement syntaxique, nous verrons plus loin.

Phonétique syntaxique

Les phénomènes qui arrivent entre phonème final de parole et phonème initial de parole suivante, quand les deux paroles appartiennent à la même phrase phonétique, constituent un type particulier de variante. La transcription de ces phénomènes, dits de phonétique syntaxique, pose le difficile problème de déterminer les pauses qui servent d'indices ou de frontières de la phrase phonétique. Dans notre statistique, on a compté les fréquences des phénomènes suivants:

a) entre consonnes finales et consonnes initiales de paroles voisines:
   1. nasales finales qui deviennent homoorganiques à la consonne initiale suivante: *con pari = compari*
   2. *l*, *r*, *n*, terminales de parole qui se fondent avec *l*, *r*, *n* initiales de la parole suivante, en donnant lieu à *l/*, *n/*. Par ex.: *nel lago = nel/ago;*

b) entre voyelles terminales et voyelles initiales de paroles voisines:
   1. *i* non accentué devient *y*, si, en position initiale ou terminale de parole, il est respectivement précédé ou suivi par un mot qui termine ou commence avec voyelle (excepté *i*);
   2. *u* initial non accentué devient *w* si la parole précédente se termine par voyelle;
   3. *i* final tombe dans les digrammes et trigrammes *ci*, *gi*, *cci*, *ggi*, *gli*, *gni*, *sci* + voyelle initiale de mot suivant, sauf *i*.
      Ex.: *amici amorosi = amičamorosi*. En effet, ce n'est pas un phonème, mais un simple signe graphique;

c) entre parole »qui renforce« et consonne non intense au début du mot suivant. A la différence des autres, tel phénomène dont les raisons résident dans l'évolution historique de l'Italien, n'est pas senti en Italie du nord, mais il est »très net et très pur dans la prononciation toscane« où il a une très grande influence parce qu'il empêche le passage de *k* à *h*, de *c* à *x* et de *g* à *γ*; il existe aussi, bien que moins nettement, dans la prononciation romaine (Tagliavini).

IV.

Le texte que nous avons analysé est *Veglia d'Armi, pièce en 2 actes avec intermède de D. Fabbri, Vallecchi Editore, Firenze 1957.*
Le texte contient 200 pages, dont 34, d'introduction, n'ont pas été prises en considération. Naturellement nous avons seulement examiné les paroles que les acteurs prononcent; toutes les autres ont été laissées de côté. Les paroles ainsi reconnues sont 18.970, constituées par 83.098 phonèmes et par 37.117 syllabes. Les formes différentes sont 4.916, se réduisant à 2.451 lemmes. Le mot le plus long est de 19 lettres (internazionalizzato).

Le choix de l'échantillon présente des difficultés considérables, parce qu'il faut s'assurer que l'échantillon observé résulte suffisant et capable de donner une idée de tous les cas possibles, ou, comme on dit, qu'il soit représentatif de son univers (Boldrini).
En pratique on se demande:

1. Quelle doit être la longueur de l'échantillon. *Guiraud* dit: »Des décomptes rapides portant sur 500 ou 1.000 phonèmes sont [...] à peu près sans valeur; en dessous de 10.000 phonèmes un dénombrement de ce genre ne peut tout au plus que constituer une indication«. *Bocca, Lafon* et d'autres sont d'accord, en proposant une longueur idéale de texte de 100.000 phonèmes. Les données de *Bocca* et de *Manfrino* et les nôtres démontrent toutefois que les variations des fréquences relatives ne peuvent être relevées au-dessus de 5.000 phonèmes.
   Nous avons divisé *Veglia d'Armi* en 17 sections et pour chacune nous avons calculé les fréquences des différentes unités phonologiques. Les 17 séries de fréquence ont été comparées entre elles et avec les fréquences globales du texte moyennant l'indice χ^2 de *K. Pearson*, proposé à cette fin par *M. Boldrini* et *G. Herdan*. Les valeurs de l'indice ainsi obtenues ne sont pas significatives, c'est-à-dire les distributions des phonèmes dans les 17 séries du texte (qui peuvent être considérées comme 17 échantillons du texte considéré comme univers) diffèrent très probablement entre elles par pur hasard.
2. Dans quelles conditions et dans quelle mesure le style de l'auteur ou les particularités psychologiques, sociales, culturelles des personnes qui parlent influent sur la normalité du texte examiné.
   Veglia d'Armi présente une variété d'arguments: les personnages de cette œuvre passent pratiquement en revue tous les secteurs de la vie et de la culture, sans employer cependant une terminologie specialisée. Par exemple il parlent de la science moderne et de ses réalisations techniques en termes que l'homme de la rue emploie habituellement. Le dialogue est serré, avec de nombreuses interruptions, anacoluthes, exclamations, ex-

pressions, et peu de descriptions. C'est, en somme, une conversation, comme on peut en entendre parmi des personnes de culture moyenne. Ces caractères intrinsèques sont confirmés par certaines données statistiques obtenues. Le text de χ^2 appliqué à la distribution des phonèmes dans la statistique déjà citée de Boldrini et à la distribution des phonèmes de notre texte, s'est révélé significatif. C'est-à-dire que les différences entre les résultats de Boldrini et les nôtres ne sont pas le résultat du seul hasard. Au contraire le valeur de χ^2 en comparaison du texte de D. Fabbri et de quelques articles de journaux (d'environ 15.000 phonèmes) dépouillés dans ce but ne donne pas un résultat significatif. Ceci paraît confirmer la validité de l'énchatillon choisi: en effet les statistiques de Boldrini ont été faites sur des textes littéraires et poétiques.

PRINCIPALES DONNÉES STATISTIQUES OBTENUES

1. Fréquences absolues et pourcentages des unités phonologiques:
   a) Dans leur ensemble;
   b) Réparties par catégories grammaticales;
   c) Réparties entre racines, affixes et désinences des mots;
   d) Réparties par paroles de longueur différente et selon la position occupée dans le mot.
2. Fréquences absolues et pourcentage des principales variantes combinatoires et des phénomènes de phonétique syntaxique.
3. Fréquences des diverses syllabes et des différentes structures syllabiques (consonne - voyelle - consonne; voyelle - liquide; etc.).
4. Fréquences des différentes combinaisons de 2 ou 3 unités phonologiques.
5. Fréquences des lettres et de la ponctuation.
6. Fréquences des diverses catégories grammaticales et morphologiques.
7. Fréquences et listes des mots et des lemmes du texte.

Les résultats obtenus sont extrêmement intéressants.

Ces résultats intéressent non seulement la composition de tests phoniatriques, audiologiques, et psychologiques en langue italienne et les interprétations linguistiques, mais aussi les physiciens et les ingénieurs pour leurs études dans le champ de la théorie de l'information, des transmissions et des télécommunications. On a eu, naturellement, confirmation que peu d'éléments couvrent la majeure partie d'un texte parlé: par exemple, des 54 unités phonologiques les 5 plus fréquentes couvrent la moitié des paroles et les 10 plus fréquentes 73 %.

Des 96 structures syllabiques différentes, les 10 plus fréquentes constituent 75 % environ du texte. Les 25 paroles plus fréquentes composent 35 % du texte; les 100 plus fréquentes 60 %. En plus des tableaux de

fréquence et des listes des matériaux phonologiques du texte dont nous venons de parler, nous pouvons énumérer parmi les résultats fournis par l'ordinateur, les séries de fiches qui restent à la disposition pour les recherches suivantes:

1. Une fiche pour chaque mot du texte
2. Une fiche pour chaque forme graphique
3. Une fiche pour chaque forme phonologique
4. Une fiche pour chaque lemme
5. Une fiche pour chaque phonème
6. Une fiche pour chaque triphonème
7. Une fiche pour chaque syllabe.

PRÉPARATION ET EXÉCUTION AUTOMATIQUE DES STATISTIQUES PAR DES INSTALLATIONS MÉCANOGRAPHIQUES ET ÉLECTRONIQUES

Il est évident comment l'utilisation des installations mécanographiques et des systèmes de *electronic data processing* facilite les opérations de depouillement, de comptage de pourcentage et de composition des données, etc.

Nous pouvons, pour plus de facilité, réserver à chaque unité phonologique relevée dans l'échantillon une fiche ou un *record* logique, dans lequel sont enregistrées, outre l'unité elle-même, opportunément codifiée, d'autres données qui la décrivent différemment. Par exemple l'énumération de ses traits pertinents, exprimés en termes *articulatoires*. En outre, un numéro d'ordre de l'unité phonologique dans le texte échantillon, ou dans la phrase phonétique, ou dans le morphème. En outre, un code qui qualifie grammaticalement le morphème dans lequel apparaît l'unité. En outre encore: une seconde codification qui enregistre l'éventuelle modification subie par l'unité même dans la chaîne parlée: variantes etc.

Une fois exécuté cet enregistrement des données, la machine peut les regrouper au moyen des fonctions élémentaires de selection et de tabulation, en autant de classes homogènes que de caractères dont les modalités sont enregistrées, et elle peut aussi, en mettant en corrélation les informations existantes, créer de nouveaux paramètres pour l'analyse.

Il est moins évident au contraire si, et dans quelle mesure, on peut exécuter avec des moyens automatiques l'opérations de transcription du texte échantillon qui dans la procédure de documentation statistique correspond au moment de relèvement des données. On se demande s'il est nécessaire au specialiste de lire le texte pour le remettre à la perforation déjà codifié mot à mot, selon l'alphabet phonétique — et il n'y a personne qui ne voie le poids et des dangers d'une telle opération — ou s'il est au contraire pos-

sible de le confier à l'élaborateur dans l'alphabet traditionnel. Le problème présente, avant même un aspect technique, un aspect théorique.

Dans les traités de phonétique les exemples de transcriptions sont souvent précédés par des expressions du genre: „exécutés selon la prononciation de Monsieur...“ La prononciation individuelle peut, en effet, altérer la transcription d'un texte, mais, à notre avis, de telles altérations ne rentrent pas, sinon dans une certaine limite, dans la transcription qui doit être à la base d'une recherche statistique comme la nôtre. Cette transcription, comme nous l'avons dit, doit en premier lieu permettre de compter les unités phonologiques présentes dans le texte c'est-à-dire saisir le plan de la „*langue*“. En deuxième lieu, nous avons donné aussi la documentation de certaines variantes, mais les variantes que nous avons choisies se situent, par leur constance et généralité, au niveau que *Coseriu* appelle la „*norma*“. Et, en effet, nous avons divisé la documentation statistique en deux parties: l'une concerne exclusivement les unités phonologiques; l'autre considère certaines variantes de ces unités phonologiques.

Or, ce qui peut varier d'une prononciation à l'autre, c'est avant tout la subdivision en phrases phonétiques, ou, en d'autres mots, la distribution des pauses dans la chaîne parlée, dont dépend la position de l'accent sur certains monosyllabes et la réalisation de phénomènes de phonétique syntaxique dans certains points de la chaîne. Le relevé des phénomènes de phonétique syntaxique, exécuté par nous dans un but expérimental, dans les limites indiquées plus haut, a été gardé bien à part dans la recherche.

Il y a ensuite toute la gamme des variantes expressives et stylistiques (Troubetzkoy), dont certaines peuvent modifier essentiellement la réalisation d'un phonème: par ex. *le hiatus*. Il semble évident que des phénomènes de ce genre, très intéressants pour une étude phonostylistique, et qui par conséquent peuvent être notés dans une transcription phonétique, ne rentrent pas dans le cadre de notre recherche et ne doivent en tous cas pas être considérés, par définition, par une transcription *phonologique*, qui fait abstraction des faits contrastifs, prosodiques, stylistiques.

Une transcription phonématique est au contraire intéressée aux cas dans lesquels le système phonologique de la communauté du locuteur, admet deux prononciations différentes d'un même morphème, de façon qu'un phonème peut être substitué au phonème avec lequel il est en opposition. En italien cela se vérifie quelquefois par les oppositions *é:è;ó:ò; s:ʃ; z:z'; i:y; u:w*.

Les dictionnaires enregistrent quelques-unes de ces couples de prononciation: il est donc possible de résoudre le problème en décidant, par exemple, de faire adopter par la machine la prononciation que le dictionnaire choisi (dans notre cas le Cappuccini-Migliorini) indique comme la plus répandue. On confie ainsi au dictionnaire la cohérence de la prononciation

adoptée. Ceci étant établi, et ayant fait abstraction pour le moment de la phonétique syntaxique, examinons maintenant l'aspect technique du problème. La plus simple solution consisterait à confronter le texte avec un répertoire de formes déjà transcrites; le terme de la recherche serait évidemment constitué par la forme écrite dans l'alphabet traditionnel; comme fonctions, on comparaitrait le morphème analysé dans les phonèmes qui le composent, et eventuellement, d'autres informations du genre déjà illustrées plus haut par des exemples. Même dans la consultation de ce „dictionnaire de machine", on aurait des formes du texte non prévues par le répertoire, et des homographes (du genre *cápito* et *capíto*), à résoudre par un procédé analogue à celui opéré par la lemmatisation. (Cf. la communication de P. R. Busa S. J.)

A ce point-là, si on veut considérer aussi les phénomènes de phonétique syntaxique, on doit ranger les mots dans l'ordre de texte, et employer ensuite un programme fonctionnant, par exemple, comme le programme d'analyse que nous allons décrir. Pour la langue italienne il n'existe malheureusement pas de répertoire de cette espèce: cependant nous avons fait la transcription de „Veglia d'Armi" selon un procédé différent, qui permettrait, entre autres, de créer, à partir d'un dictionnaire de formes graphiques, un dictionnaire de formes transcrites phonologiquement.

Notre procédé peut être résumé en deux étapes.

La première a pour but d'introduire dans le texte ces informations que le texte ne contient pas explicitement, et qui pourraient être tirées d'un dictionnaire. En particulier la distinction de faits phonétiques non représentés avec cohérence, mais indispensables pour une transcription phonologique.

Ces faits phonétiques sont en italien:

— la position de l'accent nécessaire entre autre pour distinguer automatiquement des cas comme „*scalpiccio*" de cas comme „*fáccia*" transcrits respectivement *(skalpič/io)* et *(fáč/a)*;
— cas dans lesquels *i*, *u* atones à côté de voyelles ne sont pas des semivoyelles;
— les oppositions *ó:ò; é:è; s:S; z:z'*;
— quelques rares mots dans lequels le trigramme *gli* doit être transcrit *(gli)* et non pas *(l')*, et le trigramme *gn* doit être transcrit *(gn)* et non pas *(n')*; ex.: „*ganglio*", „*gnoseologia*".

On peut imaginer facilement les opérations de machine: on perfore le texte dans l'alphabet traditionnel, on le met en ordre alphabétique, on imprime le répertoire des formes graphiques; sur ce répertoire on applique éventuellement les classements grammaticaux, on marque la position de l'accent etc.

On transporte ces informations, par l'itermédiaire des formes, à toutes les occurrences du texte, qui à ce point-là est écrit avec un alphabet de 34 lettres. Aux 21 lettres traditionnelles se sont ajoutées en effet: *á; ó; ò; é; è; S; z'; í; ú; i*, et *u* qui ne deviennent pas des semi-voyelles; et en outre, un signe de séparation des morphèmes, et une de pauses phonétique.

A ce point-ci commence la deuxième phase entièrement automatique. Le programme analyse le texte, une lettre à la fois, en la comparant avec une table mémorisée, comme l'image l'illustre, dans laquel sont énumérées par ordre alphabétique toutes les 33 lettres citées avec, en face, la respective unité phonologique.

Pour quelques-unes de ces lettres, pourtant, une simple confrontation n'est pas suffisante, parce que leur valeur phonologique doit être définie au fur et à mesure sur la base des lettres qui suivent dans le texte. Pour celle-ci on doit prévoir tous les digrammes, trigrammes etc. dans lesquels elles prennent une valeur phonologique différente, et les énumérer, par ordre alphabétique, avec la correspondante unité phonologique qui les transcrit, dans une seconde section de la table, à la quelle la *routine* de recherche tabellaire a recours, en confrontant les digrammes, trigrammes etc. dans lesquels la lettre apparaît dans le texte.

Le fonctionnement de cette routine est simple.

Considérons le texte composé par la série de lettres n^0, n^1, n^2 etc. L'ordinateur examine la lettre n^0 et recherche la ligne correspondante dans la table. Un *zéro* dans la col. 2 indique que la lettre a une valeur phonologique définie; la lettre est ensuite transcrite et l'ordinateur recommence le cycle pour n^1.

Si, au contraire, la col. 2 contient *1*, l'ordinateur forme un digramme, en associant, au symbole de la col. 3, n^1, et recherche ce digramme parmi ceux énumérés dans la seconde section de la table. S'il ne le trouve pas, il transcrit n^0 avec l'unité phonologique indiquée par la première section; s'il le trouve, il analyse la col. 3 de la ligne correspondante de la table. Si cette dernière contient *zéro*, il assume pour la transcription l'unité phonologique de cette ligne; si elle contient 1, il poursuit la recherche dans cette même deuxième section avec un nouveau digramme formé par le symbole de la col. 3 et par n^2, et ainsi de suite. Le numéro *1*, *2*, *3*... de la col. 5 indique le nombre de lettres transcrites par l'unité phonologique.

En effet, dans le cas de *celo*, la valeur *(č)* est déterminée en analysant le digramme *ce*; mais l'unité *(č)* transcrit seulement la lettre *c* et le cycle doit être recommencé pour *e*; dans *bacino*, *(č)* transcrit seulement *c*, tandis que, dans *bacio*, *(č)* transcrit *ci*; mais pour déterminer si le digramme *ci* doit être transcrit *(č)*, ou bien *(či)*, il est nécessaire d'examiner même la deuxième lettre après *c*.

Illustrons ce procédé par un exemple. Supposons, pour simplicité, un

texte composé seulement de 3 lettres différentes: *p*, *a*, *á*, par exemple *páppa* (n^0 n^1 n^2 n^3 n^4).
á est transcrit toujours *(á)*
a est transcrit toujours *(a)*
p est transcrit *(p)* sauf quand il se trouve dans le digramme *pp* lequel est transcrit *(p/)*.

La table sera ainsi composé comme suit:

1	2	3	4	5
a	0		a	1
á	0		á	1
p	1	x	p	1
xp	0		p/	2

L'ordinateur examine n^0 *(p)*.

Sur la table *p* résulte indéfini: en effet il a *1* en col. 2.

Donc l'ordinateur forme le digramme avec le symbole *x* de la col. 3, et avec la lettre suivante (n^1), qui est *á*. Le digramme *xa* n'est pas présent sur la table: ainsi le premier *p* (n^0) est transcrit *p*. Au contraire *á* (n^1) est défini *(0* en col. 2) et il est immédiatement transcrit avec *(á)* . n^2 est encore *p*, mais cette fois le digramme formé est *xp* qui se trouve dans la table.

L'unité phonologique est *(p/)* qui transcrit deux lettres (col. 5) et ainsi le cycle successif aura pour objet n^4 *(a)* et non n^3, le troisième *p*, qui est déjà transcrit.

Cette routine a un caractère évidemment général *(utility)*, et elle a été par la suite employée aussi pour des transcriptions pas exactement phonologiques, comme celles d'un code de perforation à un code différent de mémorisation et d'impression, pour l',,Index Thomisticus", pour certains textes de la ,,Accademia della Crusca", etc.

Un procédé qui se base sur les mêmes considérations théoriques est possible même avec des machines traditionnelles à fiches perforées, sans utilisation d'un ordinateur électronique.

Il consiste dans le fait de reproduire, moyennant des passages successifs avec la machine reproductrice, chaque lettre du texte sur une fiche particulière, accompagnée du trigramme dont la lettre est au centre.

La machine trieuse, opérant selon un programme spécial de regroupement, subdivise les fiches-trigramme de façon que la lettre centrale de tous les trigrammes au même groupe soit transcrite avec la même unité phonologique, qui est multiperforé depuis matrice, jusqu'à tous les fiches trigrammes du même groupe.

Division en syllabes

En transcrivant le texte, l'on eut soin de mettre à côté de chaque unité phonologique un code numérique, que j'appellerai fonction-syllabique. Voici les symboles et les unités auxquelles chacun équivaut:

1 = a, á, ẹ, ẹ́, ẹ̀, ọ, ọ́, ọ̀, u, ú, i, í
2 = w, y
3 = p, b, t, d, č, ǧ, k, g, f, v
4 = m, ɱ, n, ɳ, l, r, s, z, ʒ, ʒ́,
5 = b/; č/; d/; f/; g/; ǧ/; k/; l/; m/; n/; p/; r/; s/; t/; v/; ʒ/; ʒ́/; ñ; š; l'
6 = pause, ou confins de phrase phonétique.

Le programme de division en syllabes, réalisé soit en IBM Cardatype, soit en IBM 1401, fut compilé en nous attenant aux règles exposées par A. Camilli.[1] Il se divise en 2 cycles.

1er cycle — la syllabe précédente soit déjà conclue; on commence une nouvelle syllabe. On écrit dans la nouvelle syllabe toutes les unités phonologiques jusqu'à la première voyelle (symbole 1) comprise.

2ème cycle — il y a en cours une syllabe dont l'unité vocalique centrale a déjà été écrite. Les phonèmes suivants sont répartis entre la syllabe en cours et la successive, suivant les combinaisons suivantes:

1 | 1	
1 | 2 1	
1 2 | 3	
1 2 | 4 1	
1 2 | 4 2	
1 2 4 | 3,4,5,6	
1 2 5 | 5	
1 2 6 |	

1 | 3	
1 3 6 |	
1 | 4 1	
1 | 4 2	
1 4 | 3,4,5	
1 4 6 |	
1 5 | 5	
1 6 |	

[1] A. Camilli, *Pronuncia e Grafia dell Italiano*, No. 73, 1947.