2

C. N. R. IST. DI LINGUISTICA COMPUTAZIONALE P I S A	E. 1. 1	fr	BIBLIOTECA
--	---------	----	------------

INTERNATIONAL ASSOCIATION
OF LOGOPEDICS AND PHONiatrics

Proceedings of the
XIIth International Speech and Voice Therapy Conference
Padua, 1962

Edited by L. CROATTO, M.D. and C. CROATTO-MARTINOLLI, M.D.

Separatum

CENTRO PER L'AUTOMAZIONE DELL'ANALISI LINGUISTICA
DI GALLARATE

CENTRO MEDICO CHIRURGICO DI FONIATRIA DI PADOVA
ISTITUTO DI GLOTTOLOGIA DELL'UNIVERSITA' DI PADOVA

UNA RICERCA STATISTICA SULLA COMPOSIZIONE
FONOLOGICA DELLA LINGUA ITALIANA PARLATA, ESEGUITA
CON UN SISTEMA IBM A SCHEDE PERFORATE

R. BUSA, C. CROATTO-MARTINOLLI, L. CROATTO, C. TAGLIAVINI,
A. ZAMPOLLI; Gallarate, Padova

Consiglio Nazionale delle Ricerche
ISTITUTO DI LINGUISTICA COMPUTAZIONALE
INVENTARIO N° 957/L
COLLOCAZIONE E. 1. 1. 85

Da più parti l'attenzione degli studiosi si rivolge agli studi quantitativi del linguaggio e ai procedimenti appropriati per condurli, in particolare all'impiego dei moderni strumenti di elaborazione elettronica dei dati.

La linguistica strutturale ha stabilito l'esistenza di un sistema qualitativo di forme in opposizione distintiva, definito dalla sua struttura, dal suo equilibrio, dal suo rendimento. Nel discorso, i fonemi ricorrono con frequenze stabili e costituiscono schemi numerici che permetterebbero di parlare di un sistema di frequenze, o sistema fonostatistico (*Guiraud*). Lo studio delle frequenze trova applicazione anche in linguistica diacronica: si spiega l'evoluzione fonetica attraverso una serie di spostamenti che mirano a ristabilire l'equilibrio compromesso dalla frequenza, anormalmente elevata rispetto a quella ammissibile dalla struttura del sistema, assunta da certe articolazioni sotto l'azione di agenti esterni (migrazioni, calchi, ecc.) od interni (estensione di certe desinenze, o categorie grammaticali, ecc.). L'interpretazione delle frequenze degli elementi fonologici è stata tentata da diversi punti di vista (*J. Van Ginneken, G.K. Zipf, Tomás Navarro, N. Troubetzkoy, E. Martinet, P. Guiraud*, ecc.). In particolare la frequenza dei fonemi, studiata alla luce della teoria dell'informazione, è vista in relazione al contenuto di informazione e al costo, cioè alla complessità, del fonema. Ma quest'ultimo valore appare, allo stato attuale delle conoscenze, come incognita: potrebbe essere la risultante di diversi componenti: l'energia impegnata nella produzione del suono, la facilità di articolazione, la facilità di combinazione con altri fonemi, il rendimento acustico, la percettibilità, ecc.

In questa direzione è fondamentale il contributo che la fonetica sperimentale e in particolare l'audiologia, la foniatria e la psicologia del linguaggio possono dare alla linguistica. Lo studio delle manifestazioni patologiche del linguaggio dovrebbe mettere in evidenza le linee di maggior difficoltà di un sistema linguistico. Sarebbe importante una comparazione, anche statistica, tra le frequenze normali e le frequenze patologiche, all'interno di un sistema linguistico e soprattutto tra sistemi linguistici diversi.

Non c'è bisogno di sottolineare l'importanza del test fonetico co-

me strumento indispensabile per l'audiologo che si propone di esaminare la funzione uditiva. E' necessario conoscere la costituzione quantitativa della lingua che il *test* deve rappresentare. La fedeltà del *test* dipende direttamente dal numero di parametri secondo i quali esso rispecchia la composizione della lingua parlata. Le varie ambientazioni di un fonema determinano diversi gradi nella sua udibilità e nella sua complessità articolatoria, cosicchè è indispensabile, nel comporre liste foneticamente equilibrate, tenere conto non solo delle frequenze pure e semplici dei fonemi, ma anche delle varianti combinatorie di ciascuno, specificate ulteriormente con la considerazione delle combinazioni in bifonemi e trifonemi, delle posizioni nella parola e in sillabe di struttura diversa. La struttura del sistema linguistico del paziente, l'importanza e la posizione di un fonema nel sistema, la sua funzione diacritica e il suo contenuto semantico, il suo rendimento funzionale: queste informazioni costituiscono il contributo che la linguistica, e la fonologia in particolare, possono dare all'audiologo, al foniatra e al logopedista.

Non solo dati statistici validi e sufficientemente dettagliati¹ sono indispensabili tanto alla linguistica quanto alle scienze foniatriche e audiologiche: ma i risultati della ricerca statistica non sono utilizzabili se non nella misura in cui nel rilevarli si è tenuto conto dei punti di vista, complementari, delle due scienze. Per questo lo studio statistico sulla composizione fonologica e fonetica della lingua italiana moderna che qui presentiamo è stato reso possibile dalla collaborazione tra l'*Istituto di Glottologia* del Prof. Carlo Tagliavini dell'Università di Padova e il *Centro Medico-Chirurgico di Foniatria* del Prof. Lucio Croatto. L'estensione del testo campione ed il numero di parametri dell'analisi non sarebbero stati possibili senza l'impiego degli *impianti elettronici della IBM Italia*, ed in particolare dell'elaboratore IBM 1401.

Nell'impostare questa ricerca eseguita come *tesi di laurea in lettere* (Antonio Zampolli - *Studi di statistica linguistica eseguiti con*

¹ Analisi statistiche delle frequenze nella lingua italiana erano già state fatte, per studi crittografici, stenografici, dattilografici (Valerio, Sacco). Queste ricerche sono state però effettuate sui grafemi e non sui fonemi. Invece le statistiche di Marcello Boldrini, in *Contributi del Laboratorio di Statistica*, vol. XXI, 1948 e di E. Bocca e A. Pellegrini in *Archivio Italiano di Otologia, Rinologia e Laringologia*, vol. LVI, 1950 sono basate su una trascrizione del testo e forniscono un termine di paragone molto interessante al nostro studio, anche se è minore il numero delle unità foniche e dei fenomeni di cui si rilevano le frequenze.

TABELLA I

FREQUENZE ASSOLUTE E PERCENTUALI DELLE UNITA' FONICHE
NELLE PAROLE E NELLE FORME

		nelle parole				nelle forme	
a	b	c	d	e	f	g	h
a	6	5410	6.432	6	1756	6.57	-0.14
á	10	3336	3.966	7	1309	4.90	-0.94
b	30	440	0.523	30	205	0.73	-0.21
b/	35	214	0.254	40	59	0.22	+0.03
ċ	26	651	0.774	22	264	0.98	-0.21
ċ/	45	74	0.087	44	46	0.16	-0.08
k	9	3456	4.109	11	863	3.23	+0.87
k/	38	173	0.205	34	98	0.36	-0.16
d	11	2787	3.313	12	784	2.93	+0.38
d/	49	14	0.016	49	9	0.03	-0.02
e	1	6908	8.214	1	2353	8.81	-0.60
é	15	2388	2.839	17	553	1.99	+0.84
è	17	1880	2.235	19	465	1.74	+0.49
f	25	694	0.825	21	278	1.04	-0.22
f/	46	50	0.059	46	33	0.12	-0.06
g	33	324	0.385	31	176	0.65	-0.27
g/	50	4	0.004	50	4	0.01	-0.006
ġ	34	324	0.385	33	145	0.54	-0.16
ġ/	44	86	0.102	45	42	0.15	-0.05
i	5	5470	6.504	4	2013	7.53	-1.03
í	20	1630	1.938	18	515	1.92	+0.01
y	19	1763	2.096	16	544	2.03	+0.06
l	12	2688	3.196	14	683	2.55	+0.64
l/	28	552	0.656	36	84	0.31	+0.34
l'	36	184	0.218	39	62	0.23	-0.02
m	13	2617	3.111	10	883	3.30	-0.19
m/	42	106	0.126	41	50	0.18	-0.06
n	3	6115	7.271	8	1134	4.24	+3.03
n/	43	103	0.122	42	49	0.18	-0.06

a	nelle parole				nelle forme			
	b	c	d	e	f	g	h	
n'	39	155	0.184	43	46	0.17	+0.01	
o	2	6728	8.000	3	2149	8.04	-0.04	
ó	16	1910	2.271	20	366	1.37	+0.90	
ò	21	1165	1.385	23	257	0.96	+0.42	
p	14	2507	2.981	13	754	2.82	+0.16	
p/	41	137	0.162	38	63	0.23	-0.07	
r	4	5750	6.837	2	2276	8.52	-1.69	
r/	40	145	0.172	37	74	0.27	-0.10	
s	8	3720	4.423	9	1113	4.16	+0.26	
s/	29	528	0.627	28	204	0.76	-0.14	
š	37	176	0.209	35	98	0.36	-0.15	
ŷ	32	332	0.394	32	154	0.57	-0.18	
t	7	4775	5.677	5	1913	7.16	-1.49	
t/	27	571	0.678	26	232	0.86	-0.19	
u	24	720	0.856	24	254	0.95	-0.10	
ú	22	1076	1.279	25	245	0.91	+0.37	
w	23	1000	1.189	29	199	0.74	+0.45	
v	18	1795	2.134	15	598	2.23	-0.10	
v/	47	43	0.051	47	29	0.10	-0.05	
z	31	407	0.483	27	206	0.77	-0.29	
z'	48	17	0.020	48	14	0.05	-0.03	
Totale		84098	99.977		26703	99.63		

impianti IBM - Università di Padova, Anno Accademico 1959-60), dovremmo risolvere problemi di fondo, di notevole importanza metodologica:

- I. scelta della pronuncia per trascrivere il testo campione,
- II. elenco delle unità fonologiche delle quali conteggiare le frequenze,
- III. scelta delle principali varianti delle quali rilevare le frequenze,
- IV. scelta e dimensioni del campione.

Li esponiamo brevemente:

I. Non è possibile passare direttamente dal modo di parlare individuale a un unico tipo di pronuncia nazionale: l'italiano ha varie pronunce regionali. Gli studi in questo campo sono appena iniziati, tuttavia è ormai accertato che non si potrebbe parlare di un unico sistema fonetico italiano, ma solo di diversi sistemi fonemati regionali. Le differenze consistono nei fonemi usati e nella diversa distribuzione dei fonemi comuni entro le parole. Ad esempio l'opposizione

tra le varietà chiusa e aperta di *e* e di *o* è conosciuta in molti tipi di Italiano, tuttavia è facile elencare una serie di morfemi che in una regione sono pronunciati con *e* ovvero *o* chiusa, in un'altra regione con *e* ovvero *o* aperta, e viceversa. I fiorentini pronunciano *bène, piède, cèntro, accappatòio*, e i veneziani *bène, piède, cèntro, accappatòio*. (Giulio Lepsky).

La scelta di un sistema regionale di pronuncia piuttosto che un altro condiziona dunque due momenti fondamentali dell'indagine statistica: fa variare il numero e l'inventario dei fonemi (casi statistici) dei quali si ricercano le frequenze, e fa variare la trascrizione fonematica, che determina la distribuzione dei fonemi del testo, cioè la composizione del campione prescelto. Scegliemmo il fiorentino colto, sfronato dalle caratteristiche e dalle sfumature della pronuncia vernacola, perchè offre i seguenti vantaggi ai fini del rilevamento statistico: la maggior parte degli studi sia di fonematica sia di fonetica italiana prende come base l'italiano-fiorentino, cosicchè la maggior parte delle *tabelle* o *elenchi* di fonemi dell'italiano rappresentano in realtà il sistema fiorentino; in secondo luogo i vocabolari attestano generalmente la pronuncia dei lemmi secondo l'uso fiorentino; inoltre questo è probabilmente il tipo d'Italiano meglio conosciuto presso gli studiosi stranieri. Tutto ciò senza considerare che, sulla base di una serie di considerazioni storiche e di constatazioni di fatto, esso viene proposto come modello ortoepico e presentato come pronuncia nazionale. Citiamo a questo proposito le chiarissime proposizioni del Dizionario Enciclopedico Italiano (p. XIII-XIV).

II. Gli studiosi che hanno descritto il sistema fonematico dell'italiano letterario, o senza aggettivi, ovvero fiorentino, o toscano colto che dir si voglia, non sono d'accordo nell'elencare i fonemi che lo costituiscono: il numero dei fonemi elencati varia da un minimo di 27 (G. Porru ed altri) a un massimo di 54 (A. Castellani e P. Fiorelli).

Questa situazione pone naturalmente un grande problema, di importanza decisiva per l'impostazione e per i limiti di validità di tutta la ricerca statistica sulla composizione fonetica dell'italiano: *le divergenze tra i sistemi fonologici proposti escludono la possibilità di condurre la ricerca in modo che i risultati ottenuti siano utilizzabili dallo studioso qualunque sia il sistema fonologico prescelto?* In altre parole, è inevitabile condurre le operazioni statistiche secondo un ben determinato sistema fonologico, con la conseguenza che alla fine delle operazioni i dati rilevati (frequenze, combinazioni, ecc.) sono validi, a rigore, solo nella particolare prospettiva del sistema fonematico adottato? Le oscillazioni nel numero dei fonemi si spie-

gano con la constatazione di alcuni punti controversi, per ciascuno dei quali si prospettano due o più soluzioni. L'esame di queste divergenze ci ha condotto alla conclusione che è possibile stabilire un elenco di unità foniche tali che, essendo note le loro frequenze, dato uno qualsiasi dei sistemi fonologici proposti, e dato uno qualsiasi dei fonemi che lo compongono, la frequenza del fonema, o coincide con la frequenza di una delle unità foniche, ovvero è pari alla somma delle frequenze di due o più unità foniche opportunamente scelte.

La verifica di queste affermazioni sarebbe troppo lunga e perciò fuori luogo in questa sede; ci limitiamo ad elencare i punti controversi con le diverse soluzioni proposte a ciascuno di essi, fornendo l'elenco delle unità foniche delle quali abbiamo conteggiato le frequenze per poter rispondere a tutte le soluzioni proposte.

Vocali e semivocali

Il sistema fonematico italiano, secondo l'opinione più comune, utilizza fonologicamente due delle proprietà caratteristiche delle vocali: il grado di apertura orale e la localizzazione, ovvero zona di articolazione. La posizione più diffusa è probabilmente quella che riconosce 7 fonemi vocalici e due semivocalici: rispettivamente /u, ɔ, o, a, e, ɛ, i/ ed /w, y/. Grazie però alle diverse interpretazioni del ruolo dell'accento e ai diversi atteggiamenti nei confronti della opposizione tra le due varietà, aperta e chiusa, di *e* e di *o* e della opposizione *vocali*: *semivocali*, vengono proposti sistemi diversi, nei quali il numero oscilla tra 5, 7, 9, 10, 12 fonemi vocalici, con o senza l'aggiunta di due fonemi semivocalici. Nella statistica si conteggiano le frequenze di 14 unità: *a, á, e, é, è, o, ó, ò, i, í, y, u, ú, w*.

Accento

Il problema è se l'accento sia da considerare un tratto pertinente da aggiungere agli altri distintivi delle vocali, cioè il punto di articolazione e il grado di apertura. In tal caso 5 delle 7 vocali citate *i, e, a, o, u* (ma non *ɛ, ɔ* che non appaiono mai fuori di accento) sarebbero sdoppiate ciascuna in due fonemi autonomi, l'uno atono e l'altro tonico, così che l'elenco dei fonemi vocalici salirebbe a 12 unità: (*a, á, e, é, è, o, ó, ò, i, í, u, ú*) ovvero a 10 se non si considerano fonematiche le opposizioni *é:è, ó:ò*. È evidente che, senza prendere posizione in favore di una di queste alternative, è possibile fornire dei risultati utilizzabili quale che sia l'alternativa scelta.

Basta infatti conteggiare separatamente, per ciascuna vocale, la

TABELLA 2

FREQUENZE PERCENTUALI DELLE UNITA' FONICHE
IN ORDINE DECRESCENTE

N. d'ord.	in forme			in parole		
	Unità fonica	Freq. %	Progress.	Unità fonica	Freq. %	Progress.
1	e	8,81	8,81	e	8,21	8,21
2	r	8,52	17,33	o	8,00	16,21
3	o	8,04	25,37	n	7,27	23,48
4	i	7,53	32,90	r	6,83	30,31
5	t	7,16	40,06	i	6,50	36,81
6	a	6,57	46,63	a	6,43	43,24
7	á	4,90	51,53	t	5,67	48,91
8	n	4,24	55,77	s	4,42	53,33
9	s	4,16	59,93	k	4,10	57,43
10	m	3,30	63,23	á	3,96	61,39
11	k	3,23	66,46	d	3,31	64,70
12	d	2,93	69,39	l	3,19	67,89
13	p	2,82	72,21	m	3,11	71,00
14	l	2,55	74,76	p	2,98	73,98
15	v	2,23	76,99	é	2,83	76,81
16	y	2,03	79,02	ó	2,27	79,08
17	é	1,99	81,01	è	2,23	81,31
18	í	1,92	82,93	v	2,13	83,44
19	è	1,74	84,67	y	2,09	85,53
20	ó	1,37	86,04	í	1,93	87,46
21	f	1,04	87,08	ò	1,38	88,84
22	č	0,98	88,06	ú	1,27	90,11
23	ò	0,96	89,02	w	1,18	91,29
24	u	0,95	89,97	u	0,85	92,14
25	ú	0,91	90,88	f	0,82	92,96
26	t/	0,86	91,74	č	0,77	93,73
27	ʒ	0,77	92,51	t/	0,67	94,40
28	s/	0,76	93,27	l/	0,65	95,05
29	w	0,74	94,01	s/	0,62	95,67
30	b	0,73	94,74	b	0,52	96,19
31	g	0,65	95,39	ʒ	0,48	96,67
32	ʃ	0,57	95,96	ʃ	0,39	97,06
33	ǰ	0,54	96,50	g	0,38	97,44
34	k/	0,36	96,86	ǰ	0,38	97,82
35	š	0,36	97,22	b/	0,25	98,07

N. d'ord.	in forme		in parole			
	Unità fonica	Freq. %	Freq. %	Unità Progress. fonica	Progress.	
36	l/	0,31	97,53	l'	0,21	98,28
37	r/	0,27	97,80	š	0,20	98,48
38	p/	0,23	98,03	k/	0,20	98,68
39	l'	0,23	98,26	n'	0,18	98,86
40	b/	0,22	98,48	r/	0,17	99,03
41	m/	0,18	98,66	p/	0,16	99,19
42	n/	0,18	98,84	m/	0,12	99,31
43	n'	0,17	99,01	n/	0,12	99,43
44	č/	0,16	99,17	ǰ/	0,10	99,53
45	ǰ/	0,15	99,32	č/	0,08	99,61
46	f/	0,12	99,44	f/	0,05	99,66
47	v/	0,10	99,54	v/	0,05	99,71
48	ʒ'	0,05	99,59	ʒ'	0,02	99,73
49	d/	0,03	99,62	d/	0,006	99,74
50	g/	0,01	99,63	g/	0,004	99,75
Totale		99,63			99,75	

frequenza in posizione tonica e in posizione atona: chi non pone l'accento tra le caratteristiche fonematiche, può ottenere la frequenza di ciascun fonema vocalico sommando assieme la frequenza delle sue apparizioni in posizione atona e la frequenza delle sue apparizioni in posizione tonica.

Opposizione *e:ɛ, o:ɔ*

Alcuni negano rilevanza fonematica alle opposizioni in questione, come *W. Belardi* che le pone tra quelle che «nella lingua italiana hanno importanza non tanto per l'essenziale differenziazione semantica, quanto per l'eleganza della dizione». Affermazioni di questo tipo sono da riferire probabilmente alla notevole diversità di usi locali.

Le due varietà di *e* e di *o* sono conosciute ed usate nella maggior parte del dominio linguistico italiano, ma la loro distribuzione entro le radici e i morfemi, come si è visto, è assai varia da regione a regione. Coppie minime esistono probabilmente in ogni tipo di Italiano, ma sono diverse da tipo a tipo e certamente non sono molto numerose.

Inoltre le diverse pronunce sono perfettamente tollerate: quanti cercano di adeguare la propria pronuncia al modello ortoepico toscano, lo fanno generalmente per motivi professionali di correttezza, di stile, di «eleganza della dizione»: sono per lo più annunciatori, attori, insegnanti, ecc.

Però sta di fatto che nel sistema fonemico fiorentino è documentata la presenza di coppie minime. Queste opposizioni sono neutralizzate in posizione atona, nella quale appare dunque l'arcifonema, che solitamente è considerato identico ad *é*, *ó* di posizione tonica.

Riassumendo, riguardo ad *e* ed *o*, si pongono due problemi, pressochè indipendenti l'uno dall'altro, ciascuno con due possibili alternative:

- a) trattare l'accentazione come tratto pertinente delle vocali, o considerare a parte l'accento quale fatto prosodico?
- b) attribuire o no rilevanza fonemica alla opposizione tra le due varietà, chiusa ed aperta?

Naturalmente si avranno le varie combinazioni di ogni soluzione di a. con ogni soluzione di b., le quali danno 4 ipotesi.

E' possibile determinare la frequenza dei fonemi considerati da una qualsiasi di queste 4 posizioni, qualora si conoscano le frequenze delle 6 unità vocaliche seguenti:

- *e* in posizione atona,
- *é* chiusa, in posizione tonica,
- *è* aperta, in posizione tonica,
- *o* in posizione atona,
- *ó* chiusa, in posizione tonica,
- *ò* aperta, in posizione tonica.

Semivocali

E' fuori dubbio la particolare natura articolatoria ed acustica delle cosiddette semivocali *y* ed *w*, tanto che da alcuni fonetisti sono classificate tra i suoni consonantici. Ma debbono essere considerati fonemi autonomi, o piuttosto semplici varianti combinatorie rispettivamente dei fonemi *i* ed *u*?

Per *i* ed *u* si hanno dunque, come per *e* ed *o*, due problemi, ciascuno dei quali ammette due soluzioni:

- a) trattare o no l'accentuazione come tratto pertinente delle vocali?
- b) considerare, o no, pertinente, l'opposizione vocale;semivocale?

Naturalmente si hanno dalle combinazioni di ogni soluzione di a) con ogni soluzione di b) 4 possibili ipotesi:

- 1) si considerano 2 fonemi: *i*, *u*.
- 2) si considerano 4 fonemi: *i*, *í*, *u*, *ú*.
- 3) si considerano 4 fonemi: *i*, *y*, *u*, *w*.
- 4) si considerano 6 fonemi: *i*, *í*, *y*, *u*, *ú*, *w*.

E' immediatamente evidente che la soluzione di conteggiare le

frequenze di 6 unità: *i, i, y, u, ú, w* permette di calcolare le frequenze dei fonemi considerati da ciascuna delle 4 ipotesi citate.

Consonanti

Nell'eseguire le statistiche abbiamo conteggiato le frequenze di 40 unità consonantiche: *p, p', b, b', t, t', d, d', č, č', ĝ, ĝ', k, k', g, g', f, f', v, v', j, s', s, š', l, l', l', r, r', m, +m, m̃, +n, n, ŋ, m', n', n', z, z'*. Elenchiamo analogamente a quanto fatto per le vocali i punti controversi.

Nasali.

La posizione più comune consiste nel considerare, oltre ad *n'* ed, eventualmente, alle geminate *m'* ed *n'*, due fonemi nasali *m* ed *n*. Di *m* si dice che può apparire dinanzi a vocale o alle consonanti labiali *p* e *b*; di *n* si distinguono almeno tre varietà: *m̃* che appare dinanzi alle labiodentali *f* e *v*; *ŋ*, che appare dinanzi alle velari *k* e *g*, ed infine *n* che appare dinanzi a vocale e a tutte le consonanti meno *p, b, f, v, k, g*.

E' però sentito il problema se la varietà *m̃* debba essere considerata o no variante combinatoria del fonema *m* anziché del fonema *n*. Una terza soluzione poi consisterebbe nel considerare pertinente l'opposizione *m:n* solo dinanzi a vocale, mentre sarebbe neutralizzata davanti a consonante, posizione nella quale la realizzazione dell'arcifonema sarebbe condizionata appunto combinatoriamente dalla consonante seguente. Per rispondere alle esigenze di queste tre diverse soluzioni è necessario conteggiare le frequenze di 9 unità:

<i>m</i>	cons. nasale	bilabiale	dinanzi a:	vocale
<i>+m</i>		bilabiale	»	<i>p, b</i>
<i>m̃</i>		labiodentale	»	<i>f, v</i>
<i>ŋ</i>		velare	»	<i>k, g</i>
<i>n</i>		dentale	»	vocale
<i>+n</i>		dentale	»	consonante, meno <i>p, b, f, v, k, g</i>
<i>mm</i>		bilabiale geminata	(solo)	prevocalica
<i>nn</i>		dentale geminata	(solo)	prevocalica
<i>ñ</i>		palatale	(solo)	prevocalica

Opposizione sorda:sonora per le consonanti esse e zeta

Alcuni autori non ritengono pertinente l'opposizione *s:f* (come *fúso*, (per filare): *fúfo* (participio di fondere), ed eliminano *f* dall'elenco dei fonemi consonantici.

Nella statistica abbiamo conteggiato separatamente la frequenza di esse sonora e di esse sorda.

Chi non ritenesse pertinente questa opposizione può sommare le due frequenze e attribuirle a un unico fonema, realizzato ora sordo ora sonoro. Analoghe considerazioni valgono per l'opposizione *zeta* sorda: *zeta* sonora.

Geminate

Riguardo alle consonanti cosiddette *doppie* si pone il problema se le doppie (dette anche geminate, forti, lunghe) siano da considerare, o no, fonemi autonomi, distinti dalle semplici (deboli, brevi) corrispondenti. Negli autori si trovano tre soluzioni:

- 1) Le geminate sono interpretate come nessi bifonematici, cioè ripetizioni di uno stesso fonema. L'opposizione *pala:palla* sarebbe del tipo *pala:palma*; in *pala* si contano quattro fonemi; in *palma* e in *palla* cinque.
- 2) Le geminate, cui vengono riconosciute caratteristiche monofonematiche, sono opposte alle semplici come fonemi autonomi. La opposizione *pala:palla* sarebbe del tipo *pala:paga*. In *palla* si contano quattro fonemi come in *pala* e in *paga*. 15 consonanti (p, b, t, d, č, ĝ, k, g, f, v, s, l, r, m, n) danno luogo allora a 30 fonemi p:p/; b:b/; t:t/; ecc.). Restano esclusi dalla opposizione geminata: semplice, i fonemi l', n', š, z, z', perchè sempre lunghe, e j perchè sempre breve, nonché le varietà nasali +m, ŋ, +n, ŷ.
- 3) Alle geminate viene riconosciuta natura monofonematica, ma non sono opposte alle semplici come fonemi autonomi; semplice e ge-

TABELLA 3 a

FREQUENZE DISTRIBUITE SECONDO IL PUNTO DI ARTICOLAZIONE DEI FONEMI

Punto di articolazione	Nelle parole			Nelle forme		
	Frequenza assoluta	Frequenza %		Frequenza assoluta	Frequenza %	
		Sul totale dei fonemi	Sul totale delle conson.		Sul totale dei fonemi	Sul totale delle conson.
Labiali	6011	14,090	7,159	2014	7,542	14,675
Bilabiali	2582	3,070	3,070	938	3,512	6,834
Dentali	28500	33,704	66,704	8929	33,438	65,061
Palatali	1650	1,961	3,861	703	2,632	5,122
Velari	3971	4,726	9,302	1141	4,272	8,313
Totale	42714	50,812	99,997	13724	51,399	100,005

minata corrispondente sono considerate semplicemente gradi diversi di articolazione (rispettivamente debole e forte) di uno stesso fonema.

Nella statistica si sono conteggiate separatamente le frequenze delle 30 unità proposte dalla seconda soluzione. Chi sostiene la terza soluzione deve solamente sommare tra loro la frequenza della sem-

TABELLA 3 b

FREQUENZE DISTRIBUITE SECONDO IL MODO DI ARTICOLAZIONE DEI FONEMI

Modo di articolazione	Nelle parole			Nelle forme		
	Frequenza assoluta	Frequenza %		Frequenza assoluta	Frequenza %	
		Sul totale dei fonemi	Sul totale delle conson.		Sul totale dei fonemi	Sul totale delle conson.
Occlusive	16537	19,663	38,715	5676	21,184	41,358
Nasali	9096	10,815	21,285	2162	8,096	15,753
Fricative	2582	3,070	6,040	938	3,512	6,834
Affricate	424	0,504	0,992	220	0,823	1,603
Sibilanti	4756	5,655	11,134	1596	5,875	11,450
Liquide	3424	4,071	8,017	829	3,104	6,049
Rotante	5895	7,009	13,801	2350	8,800	17,149
Totale	42714	50,787	99,994	13724	51,394	100,196

plice e della geminata, mentre chi accetta la prima soluzione deve aggiungere alla frequenza della semplice la frequenza raddoppiata della geminata corrispondente.

III. Varianti

Il numero delle varianti è praticamente illimitato: si pone perciò il problema se e di quali varianti combinatorie conteggiare le frequenze. Come si è visto, i fonemi, intesi secondo la definizione strutturalistica, rispondono perfettamente al concetto di *caso statistico* e non vi è dubbio che una statistica del materiale fonico di una lingua debba essere, in principio, statistica delle frequenze delle unità fonologiche distintive. Tuttavia non è solo in vista della utilizzazione in sede foniatrica e di teoria dell'informazione che interessa rilevare le frequenze delle varianti: molti studi recenti ne sottolineano l'importanza linguistica.

Sia sul piano sincronico, come elementi che, comuni in tutto il parlare di una comunità sono caratterizzanti e tipici di una lingua determinata, e, pur non rientrando tra le opposizioni significative del

sistema, costituiscono una specie di loro costante accompagnamento, cosicchè contribuiscono alla identificazione e al riconoscimento del fonema o di un gruppo di fonemi, quando non si sostituiscano con valore funzionale, soprattutto in certi modi rapidi di elocuzione, ai tratti distintivi pertinenti del sistema; anche nel campo diacronico come fattori determinanti in molti procedimenti di evoluzione fonetica.

Perciò oltre alle frequenze di alcuni tipi di varianti combinatorie, particolarmente sottolineati dagli studiosi di fonetica italiana, abbiamo rilevato anche tutte le combinazioni di due e tre fonemi, con la loro frequenza di apparizione.

Durata delle vocali

La differenza tra vocale lunga e vocale breve è notevole in italiano, anche se non pare fonologicamente pertinente. Il fattore essenziale è la durata proporzionale. Possiamo parlare a tempi diversi, cosicchè la durata assoluta delle singole articolazioni sarà profondamente mutata, senza che i rapporti di lunghezza fra le successive articolazioni ne siano molto cambiati. Questa durata proporzionale dipende anche dall'ambiente fonetico circostante e particolarmente dalla struttura della sillaba. Praticamente, gli autori concordano nel considerare lunghe le vocali in sillaba aperta tonica, brevi tutte le altre (*Josselyn, Battisti*). Le nostre tabelle recano per ogni vocale tonica:

le frequenze delle sue apparizioni a. In sillaba tonica: 1. seguite da consonante sorda, 2. seguite da consonante sonora. b. In sillaba chiusa: 3. seguita da geminata, 4. seguita da nasale, 5. seguita da *l*.

Vocali nasalizzate

Le vocali nasali sono generalmente prodotte dal fatto che una consonante nasale nasalizza più o meno la vocale attigua. Talora questo fenomeno di assimilazione giunge fino alla completa caduta della consonante nasale (*Panconcelli Calzia*). Sul piano diacronico la assimilazione giuoca un ruolo notevole nella evoluzione fonetica di varie lingue. I *Croatto*, che hanno condotto particolari studi sulla nasalizzazione delle vocali italiane, sottolineano gli importanti riflessi di questo fenomeno in alcune turbe della parola. La nasalizzazione, generalmente regressiva (ma può essere anche progressiva) normalmente si compie nell'unità sillabica, in particolar modo quando la vocale è seguita dai gruppi *mf, ns, nl, nr*.

Si sono pertanto conteggiate le frequenze delle vocali:

1. precedute da nasale, 2. precedute da nasale nella stessa sillaba, 3. seguite da nasale, 4. seguite da nasale nella stessa sillaba, 5. precedute e seguite contemporaneamente da nasali.

Gradi di intensità delle consonanti

Alcuni fonetisti distinguono tre gradi di intensità (tenue, medio e forte) per le consonanti *p, b, t, d, ċ, ġ, k, g, f, v, s, l, r, m, n*.

Per il grado forte vale quanto detto per le geminate. Il contrasto tenue: medio è invece certamente senza valore fonologico. Pertanto si sono conteggiate distintamente le frequenze delle suddette consonanti in posizione intervocalica o seguite da *l, r* (grado tenue) e le frequenze all'inizio di frase fonetica, o precedute da consonante, o seguite da consonante diversa da *l, r*. Sarebbe troppo lungo spiegare in questa sede il trattamento riservato ad *š, ñ, l', z, z'*. Per il rafforzamento sintattico si veda invece qui sotto.

Fonetica sintattica

Un particolare tipo di variante costituiscono i fenomeni che avvengono tra fonema finale di parola e fonema iniziale di parola seguente quando le due parole appartengono alla stessa frase fonetica. La trascrizione di questi fenomeni, detti di fonetica sintattica, pone il difficile problema di determinare le pause che fungono da indici o frontiere della frase fonetica.

Nelle statistiche sono state conteggiate le frequenze dei seguenti fenomeni:

- a) *tra consonanti finali e consonanti iniziali di parole attigue.*
 - 1) nasali finali che divengono omorganiche alla consonante iniziale seguente.
 - 2) *l, r* finali di parola che si fondono con *l, r* iniziali della parola seguente, dando luogo a *l', r'*. Es. *nel lago*—*nel/ago*.
- b) *tra vocale finale e vocale iniziale di parole attigue.*
 - 1) *i* non accentata, diviene *y* se, in posizione iniziale o finale di parola, è rispettivamente preceduta o seguita da parola che termina o inizia con vocale (eccetto *i*).
 - 2) *u* non accentata iniziale di parola diviene *w* se la parola precedente termina con vocale.
 - 3) *i* finale di parola cade nei digrammi e trigrammi *ci - gi - cci - ggi - gli - gni - sci* + vocale iniziale di parola seguente, meno *i*. Es. *amici amorosi*—*amičamorosi*. Infatti non è un fonema, ma un semplice segno grafico.
- c) *Tra parola rafforzativa e consonante non intensa iniziale di parola seguente.*

TABELLA 4

FREQUENZE PERCENTUALI, NELLE PAROLE, DELLE LETTERE
IN ORDINE DECRESCENTE

Rango	Lettera	Freq. %	Freq. Progressiva
1	e	12,510	12,510
2	o	10,973	23,483
3	i	10,515	33,998
4	a	9,790	43,788
5	n	7,249	51,037
6	r	6,762	57,799
7	t	6,624	64,423
8	s	6,093	70,516
9	c	4,652	75,168
10	l	4,450	79,618
11	m	3,166	82,784
12	d	3,151	85,935
13	u	3,128	89,063
14	p	3,113	92,176
15	v	2,105	94,281
16	h	1,496	95,777
17	g	1,394	97,171
18	b	0,971	98,142
19	f	0,888	99,030
20	q	0,695	99,725
21	z	0,535	100,260

Tale fenomeno, le cui ragioni risiedono nella evoluzione storica dell'italiano, a differenza degli altri non è avvertito nell'Italia del Nord, ma è «vivissimo e purissimo nella pronunzia toscana», ove ha grandissima influenza perchè impedisce il passaggio di *k* a *h*, di *č*: a *č* e di *ǰ* a *ǰ*, e sia pure in minor misura, in quella romana.

IV. Il testo da noi analizzato è *Veglia d'Armi*, rappresentazione in due parti ed un intermezzo di Diego Fabbri - Vallecchi Editore, Firenze 1957. Il testo conta 200 pagine, delle quali 34, occupate dall'introduzione, non sono state prese in considerazione. Naturalmente, sono state esaminate solo le parole che gli attori pronunciano: trascurammo tutte le altre. Le parole così individuate sono 18.970, costituite da 83.098 fonemi e da 37.117 sillabe. Le forme diverse sono 4.916, riducibili a 2.451 lemmi. La parola più lunga è di 19 lettere (internazionalizzato). La scelta del campione presenta notevoli difficoltà perchè occorre assicurarsi che il campione osservato risulti sufficiente e adatto per fornire una idea di tutti i casi possibili; o, come si dice, sia rappresentativo del suo universo (*Boldrini*).

In pratica ci si chiede:

- 1) Quanto deve essere esteso il testo campione. Dice *Guiraud*: 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* e gli altri autori concordano, ponendo verso i 100.000 fonemi la lunghezza ideale del testo. I dati di *Bocca*, di *Manfrino* e i nostri dimostrano tuttavia che le variazioni delle frequenze relative non sono rilevanti in brani al di sopra di 5.000 fonemi.

Abbiamo suddiviso *Veglia d'Armi* in 17 sezioni e per ciascuna abbiamo calcolato le frequenze delle diverse unità fonologiche. Le 17 serie di frequenze sono state comparate tra loro e con le frequenze complessive del testo mediante l'indice χ^2 di *K. Pearson*, proposto a tale scopo da *Marcello Boldrini* e da *Gustav Herdan*. I valori dell'indice così ottenuti non sono significativi, cioè le distribuzioni dei fonemi nelle 17 serie del testo (che possono essere viste come 17 campioni del testo considerato come universo) con grande probabilità differiscono tra loro per puro caso.

- 2) In che condizione e in che misura lo stile dell'autore o le particolarità psicologiche, sociali, culturali dei parlanti influiscano sulla normalità del testo esaminato.

TABELLA 5

1. Frequenza complessiva dei fonemi nelle diverse posizioni di parola		2. Lunghezza delle parole	
Posiz della parola	Frequenza fonemi	Lunghezza	Tot. Parole
17	2	17	2
16	5	16	3
15	11	15	6
14	31	14	20
13	110	13	79
12	233	12	123
11	462	11	229
10	917	10	455
9	1539	9	622
8	2553	8	1014
7	3886	7	1333
6	6043	6	2157
5	8062	5	2419
4	10633	4	2171
3	13123	3	2490
2	17232	2	4109
1	18887	1	1655
Totale	84729	Totale	18887

Veglia d'Armi presenta varietà di argomenti: i personaggi di quest'opera passano in rassegna, praticamente, tutti i settori della vita e della cultura, senza però usare mai una terminologia specializzata. Ad esempio parlano della scienza moderna e delle sue realizzazioni tecniche coi termini che di solito usa l'uomo comune.

Il dialogo è serrato, con frequenti interruzioni, anacoluti, esclamazioni, modi di dire, e poche sono le descrizioni. E' insomma una conversazione, come se ne possono sentire tra persone di cultura media, e questo suo carattere è confermato da alcuni dei dati statistici ottenuti. Il test di χ^2 applicato alla distribuzione dei fonemi nella statistica citata di *Boldrini* e alla distribuzione dei fonemi nel nostro testo, si è rivelato significativo. Cioè le differenze tra i risultati di *Boldrini* e i nostri non sono imputabili al solo caso. Invece il valore di χ^2 nel confronto tra il testo di *Diego Fabbri* e alcuni articoli di giornali (di circa 15.000 fonemi) spogliati a questo scopo, non è risultato significativo. Questo conferma la validità del campione scelto: infatti le statistiche di *Boldrini* sono state condotte su testi letterari e poetici.

Preparazione ed esecuzione automatica del rilevamento statistico mediante impianti meccanografici ed elettronici IBM

La classificazione vera e propria, cioè la riunione dei casi in categorie omogenee, secondo le modalità simili dei caratteri, si realizza mediante lo spoglio dei casi rilevati. Prima di valerci delle macchine IBM, sperimentammo quali sarebbero state le difficoltà nell'eseguire questo lavoro con carta e penna. Alcuni saggi ci convinsero della impossibilità pratica di compiere il lavoro in tempo utile, se non rinunciando a dati (per es. la distribuzione delle unità fonologiche nelle diverse parti della parola: radice, desinenza, ecc.) che sono necessari per la validità e la completezza della statistica.

Il Prof. *Bocca*, che condusse una ricerca analoga, ci disse che il conteggio manuale dei fonemi nei testi analizzati fu straordinariamente faticoso, senza contare che la mole delle operazioni manuali e la loro lunghezza aumentavano notevolmente la possibilità di errore e di confusione nella registrazione dei dati. L'uso delle macchine oltre ad assicurare una notevole economia di tempo e l'esattezza dei risultati ha permesso o meglio ha indotto a moltiplicare i parametri della rilevazione statistica dalla quale dipende la finezza dell'analisi dell'universo studiato e, a estendere l'ampiezza del campione, da cui dipende la validità dei dati. Una volta effettuata la registrazione dei dati elementari, cioè la trascrizione di ogni individuo o caso statistico, la macchina può, senza ulteriori trascrizioni, raggrupparli e conteggiarli in tante classi omogenee quanti sono i caratteri di cui sono registrate le modalità, e può anche, combinando secondo certe condizioni le informazioni esistenti, creare nuovi parametri per l'analisi.

Trascrizione fonologica automatica

Sarebbe stata possibile, ovviamente, una procedura come questa: ritrascrivere a mano l'intero testo nell'alfabeto fonologico e poi farlo perforare, impiegando una scheda per ogni fonema. Per risparmiare tempo, ed evitare errori, ci chiedemmo se e in che misura fosse possibile far perforare il testo nell'alfabeto normale ed istruire poi le macchine ad eseguire automaticamente la trascrizione dall'alfabeto storico a quello fonologico. La macchina, nel trascrivere il testo, deve:

- trascrivere una stessa lettera ora con un fonema, ora con un altro.

Esempio: e ————— e
 $\acute{e}$
 $\grave{e}$

- trascrivere due lettere diverse con lo stesso fonema.

Esempio: c ————— k
 q

- trascrivere digrammi o trigrammi con un solo fonema.

Esempio: sci ————— $\check{s}$

- non trascrivere mai certe lettere.

Esempio: h .

Si aggiunga che, per tenere conto dei fenomeni di fonetica sintattica, la macchina deve eseguire in modo particolare la trascrizione di lettere, digrammi o trigrammi che si trovano in posizione iniziale o finale di parola, differenziando inoltre il suo comportamento a seconda che le lettere, i digrammi o i trigrammi siano, o no, anche iniziali o finali di frase fonetica.

Si pensi a uno straniero che debba leggere un testo italiano: egli può apprendere la posizione di un accento o la apertura o chiusura di una *e* o di una *o*, la sonorità di una *s* o *z* solo con l'uso, e cioè memorizzando la conoscenza della singola parola come unità semantica. Ma per il resto egli riesce a leggere uno scritto applicando un numero limitato di regole. Ora, scrivere un programma per le macchine, altro non è che comunicare un elenco finito di regole; la macchina può trattare qualsiasi tipo di informazioni, purchè la struttura della elaborazione sia riconducibile a un algoritmo finito: dunque può trascrivere tutte quelle lettere il cui esatto valore fonologico può essere stabilito tenendo conto delle lettere immediatamente precedenti e delle lettere immediatamente seguenti nel testo. Lo *Zampolli* ha costruito, per ciascuna di queste lettere, una tabella in cui è definito il valore della lettera per ciascun trigramma di cui la lettera può essere al centro.

TABELLA 6

CONSONANTI NASALI

a) prospetto delle diverse varietà considerate

	Labiali	Labio-dentali	Dentali	Palatali	Velari
preconsonantiche	m	ɱ	n		ŋ
prevocal. o finali di parola	+m		+n	n'	
geminate	m/		n/		

b) frequenze delle diverse varietà di nasali

Unità fonica	Frequenza assoluta	% sul totale delle nasali	% sul totale dei fonemi
m	469	5,15	0,55
ɱ	110	1,20	0,13
n	2133	23,44	2,53
ŋ	367	4,03	0,43
+m	2148	23,62	2,48
+n	3501	38,50	4,16
m/	106	1,16	0,12
n/	103	1,16	0,12
n'	155	1,70	0,18
Totale cons. nasali	9092	99,96	11,72

Oltre ai prospetti di frequenza e alle liste dei materiali fonologici del testo di cui parleremo tra poco, possiamo annoverare, tra i risultati forniti dal calcolatore, le seguenti serie di schede, che restano a disposizione per successive ricerche:

1) Una scheda per ogni parola del testo. 2) Una scheda per ogni forma grafica. 3) Una scheda per ogni forma fonologica. 4) Una scheda per ogni lemma. 5) Una scheda per ogni fonema. 6) Una scheda per ogni trifonema. 7) Una scheda per ogni sillaba.

Principali dati statistici ottenuti

— Frequenze assolute e percentuali delle unità del sistema fonologico italiano: 1. complessive (Fig. 1, 2, 3). 2. Ripartite per categorie grammaticali. 3. Ripartite tra radici, affissi e desinenze delle parole. 4. Ripartite per parole di diversa lunghezza e a seconda della posizione occupata nella parola (Fig. 5).

- Frequenze assolute e percentuali delle principali varianti combinatorie e dei fenomeni di fonetica sintattica (Fig. 6, 7).
- Frequenze delle diverse sillabe e delle differenti strutture sillabiche (cons.-voc.; cons.-voc-con.; liquida-vocale; ecc).
- Frequenze delle diverse combinazioni di due e di tre unità fonologiche.
- Frequenze delle lettere (Fig. 4) e delle interpunzioni.
- Frequenze delle diverse categorie grammaticali e morfologiche.
- Frequenze e liste delle parole e dei lemmi del testo.

I risultati ottenuti sono estremamente interessanti. Oltre che per la composizione di *tests* foniatrici, audiologici e psicologici in lingua italiana e per le interpretazioni linguistiche, questi risultati interessano fisici e ingegneri per i loro studi nel campo della teoria dell'informazione, delle trasmissioni e delle telecomunicazioni.

Si è avuta un'altra conferma che pochi elementi coprono la maggior parte di un testo parlato: per esempio, delle 54 unità fonologiche, le 5 più frequenti coprono metà delle parole e le 10 più frequenti il 73 per cento.

Delle 96 strutture sillabiche diverse, le 10 più frequenti costituiscono il 75 per cento circa del testo. Le 25 parole più frequenti compongono il 35 per cento del testo; le 100 più frequenti il 60 per cento.

TABELLA 7

VARIAZIONI NELLE FREQUENZE DOVUTE A FONETICA SINTATTICA

Unità fonica	Freq. senza fonetica sintattica		Freq. con fonetica sintattica		Variazione	
	Assoluta	%	Assoluta	%	Assoluta	%
a	5410	6,43	5673	6,75	+263	+0,32
á	3336	3,96	3073	3,65	-263	-0,32
e	6908	8,21	7441	8,83	+533	+0,62
é	2388	2,83	1875	2,22	-523	-0,61
è	1880	2,23	1870	2,22	-10	-0,01
i	5470	6,50	4436	5,27	-1034	-1,23
í	1630	1,93	1483	1,76	-147	-0,17
y	1736	2,09	2817	3,36	+1081	+1,27
o	6728	8,00	6977	8,29	+249	+0,29
ó	1910	2,27	1726	2,05	-184	-0,22
ò	1165	1,38	1110	1,31	-65	-0,07
u	720	0,85	652	0,77	-68	-0,08
ú	1076	1,27	854	1,01	-222	-0,26
w	1000	1,18	1290	1,52	+290	+0,34

Riteniamo forse il più valido risultato del nostro lavoro il metodo messo a punto per l'automazione dell'analisi fonologica: in particolare quello per la trascrizione fonologica automatica e per la divisione automatica in sillabe.

Questi metodi, che abbiamo tradotto in un programma *utility* per il calcolatore IBM 1401, permetteranno ricerche fonostatistiche su qualsiasi corpus si voglia analizzare.