RENDICONTI
del
SEMINARIO MATEMATICO
della
UNIVERSITÀ DI PADOVA

C. MINNAJA
L. PACCAGNELLA

Variazioni dell'entropia linguistica dell'italiano scritto e calcolo di un'entropia fonematica

Rendiconti del Seminario Matematico della Università di Padova, tome 57 (1977), p. 247-265

<http://www.numdam.org/item?id=RSMUP_1977__57__247_0>

REND. SEM. MAT. UNIV. PADOVA, Vol. 57 (1977)

Variazioni dell'entropia linguistica dell'italiano scritto e calcolo di un'entropia fonematica

C. MINNAJA - L. PACCAGNELLA (*)

§ 1. - Introduzione.

L'idea del calcolo dell'entropia di un testo linguistico è già vecchia di almeno un quarto di secolo ([1] , [2]). Il calcolo effettivo è stato effettuato per alcune lingue (vd., ad es. [3]), sulla base di tabelle di frequenza dei grafemi elaborate per alcune lingue europee. È ben nota l'importanza che tali studi rivestono nella teoria matematica delle comunicazioni; in particolare, per effettuare prove di intelligibilità di un messaggio, si è ricorsi a modelli di sorgente dotati della proprietà di Markov, e simulabili mediante un elaboratore elettronico ([4]).

È invece più recente l'idea proposta da ingegneri e linguisti di considerare la lingua come un sistema isolato, e di applicare a tale sistema le considerazioni proprie dell'entropia (vd., ad es. [5] , a cui è annessa un'ampia bibliografia). La proposta è stata formulata in maniera assai dubitativa, anche perchè, nel caso preso sia pur superficialmente in considerazione in [5] , parrebbe che l'evoluzione di una lingua nel tempo portasse ad una diminuzione di entropia.

Evidentemente l'entropia calcolata a livello di grafemi varia a seconda del testo che si considera, ma tale variazione è in genere

---

(*) Indirizzo degli AA.: Istituto di Matematica Applicata, Università, Padova.

Lavoro eseguito mentre L. Paccagnella godeva di una borsa di studio del CNR, Bando n. 201/1/57 del 28/5/75.

piuttosto limitata e decresce con la lunghezza del testo scelto. L'entropia calcolata a livello di fonemi varia fortemente a seconda della trascrizione fonematica scelta. Entrambi i tipi di calcolo sono utili, a seconda se si vogliano inquadrare in uno studio per un simulatore di una sorgente markoviana o per un sintetizzatore della voce umana ([6] , [7]) .

Nel presente lavoro viene dapprima calcolata l'entropia di primo ordine su due testi giornalistici italiani, a livello grafematico, e il risultato viene confrontato con quelli di [3], per una verifica sperimentale dell'ipotesi espressa in [5]. Viene poi calcolata l'entropia degli stessi testi, ma con alfabeti più ampi, come proposta per modelli di sorgente markoviana più adeguati alla lingua italiana.

Successivamente viene calcolata l'entropia di primo ordine a livello fonematico, e infine viene proposta un'interpretazione diversa di un possibile calcolo dell'entropia a livello di parole.

§ 2. - Calcolo dell'entropia sull'alfabeto base.

La classica formula di Shannon

$$H_1(X) = - \sum_{x \in X} p(x) \lg_2 p(x)$$

fornisce l'entropia calcolata su un insieme X di simboli x, sul quale è stata introdotta una funzione di probabilità p . I risultati trovati da Manfrino ([3]) per H_1 sono ricavati dall'analisi di tre testi, ciascuno di 10.000 lettere dell'alfabeto italiano. I tre testi erano rispettivamente di carattere scientifico, storico e giornalistico, scritti rispettivamente nel 1955, 1940 e 1958. L'entropia media risultante era di 3,9433 bit/lettera.

La nostra analisi si esprime in un confronto tra due testi giornalistici, uno di politica interna e uno di politica estera. Entrambi i pezzi sono di 10.000 lettere, prescindendo dal fatto che queste siano soltanto le 21 lettere dell'alfabeto italiano classico. Compaiono infatti anche cifre arabe e romane (raggruppate in un unico elemento) e lettere di altri alfabeti.

Le occorrenze e le relative entropie sono riportate nella tabella I. Il testo n° 1 è preso dall'articolo di fondo del quotidiano « La Repubblica » del 24/3/1976, pagg. 1 e 3, dal titolo « La lunga notte della DC » ; esso si compone di due parti, a firma rispettivamente

TABELLA I

Lettera	Testo N° 1		Testo N° 2	
	occ./10^4	H_1	occ./10^4	H_1
a	0.1191	0.3656	0.1081	0.3470
b	0.0077	0.0541	0.0089	0.0606
c	0.0441	0.1986	0.0450	0.2013
d	0.0403	0.0398	0.1867	0.1851
e	0.1089	0.3484	0.1067	0.3445
f	0.0095	0.0638	0.0103	0.0680
g	0.0195	0.1108	0.0180	0.1043
h	0.0096	0.0643	0.0101	0.0670
i	0.1069	0.3448	0.1199	0.3669
l	0.0648	0.2558	0.0623	0.2495
m	0.0213	0.1183	0.0271	0.1411
n	0.0743	0.2787	0.0724	0.2742
o	0.0977	0.3278	0.0874	0.3073
p	0.0268	0.1399	0.0277	0.1433
q	0.0039	0.0312	0.0024	0.0209
r	0.0652	0.2568	0.0655	0.2576
s	0.0513	0.2198	0.0585	0.2396
t	0.0728	0.2752	0.0648	0.2558
u	0.0276	0.1429	0.0277	0.1433
v	0.0132	0.0824	0.0132	0.0824
z	0.0131	0.0819	0.0091	0.0617
j	0.0000	0.0000	0.0002	0.0025
k	0.0000	0.0000	0.0006	0.0064
w	0.0001	0.0013	0.0007	0.0073
x	0.0000	0.0000	0.0001	0.0013
y	0.0000	0.0000	0.0011	0.0108
cifre	0.0023	0.0202	0.0124	0.0785
Totali	1.0000	3.9693	1.0000	4.0282

di F. De Luca e G. Valentini. Il testo n° 2 è preso dalla rubrica « Affari esteri » del settimanale « Panorama » del 23/3/1976; esso si compone di due parti, una intitolata « Distensione proibita » a firma di M. Conti, l'altra (fino al completamento delle 10.000 lettere) intitolata « Se il rosso vince » a firma di S. Parone. In entrambi i testi non sono stati considerati nè titoli nè sottotitoli. In appendice,

nelle tabelle I_A e II_A sono riportate le occorrenze suddivise per singole migliaia di lettere; ad esse si riferiscono i due gruppi di grafici annessi. Questi dimostrano come siano fortemente differenziate le occorrenze in ciascun migliaio di lettere.

Si può notare che le singole entropie per i tre testi calcolate in [3] erano parecchio simili tra loro, rispettivamente: 3,9453; 3,9395; 3,9453. Invece la prima metà della tabella I fornisce già 3,9693, e la seconda metà addirittura 4,0282. Per quanto l'estensione dei testi in esame sia relativa, tali differenze potrebbero già portare un supporto alla tesi che lo stile, almeno quello giornalistico, in 18 anni si è evoluto verso una maggiore entropia. In particolare è da notare la differenziazione relativamente forte tra l'entropia dei due testi attuali entrambi giornalistici, uno di politica interna e l'altro di politica estera, superiore di circa dieci volte alle differenze registrate tra i testi di [3], che pure erano di tipo piuttosto differente tra loro.

§ 3. - Ampliamenti dell'alfabeto.

I calcoli effettuati per la tabella I tendevano a raffrontare le frequenze delle varie lettere con i testi di qualche decennio fa. Qualora si verificassero, con l'evolversi della lingua, significativi spostamenti di tali frequenze, potrebbe essere utile modificare il codice telegrafico in vista di una maggiore economia del costo di trasmissione.

Se aumentiamo il numero degli elementi dell'alfabeto X l'entropia in generale crescerà. Un ampliamento abbastanza naturale è quello di considerare anche i segni di interpunzione. Nel testo n° 1 sono stati trovati 12 diversi segni di interpunzione, che hanno portato il totale del corpus a 10.330 elementi; nel testo n° 2 ne sono stati trovati 14, per un totale di 10.393 elementi. Ciò ha portato alla tabella II, nella quale le frequenze sono state rapportate a 10.000. Si noti come sia sempre maggiore l'entropia del testo n° 2, per quanto la differenza percentuale resti pressochè costante.

Questo ampliamento dell'alfabeto comporta però alcune difficoltà. Infatti, adottando un codice binario, se l'alfabeto è composto da n elementi, bisogna usare k cifre per distinguere un elemento, dove k è legato ad n dalla relazione

$$2^{k-1} + 1 < n < 2^k$$

TABELLA II

Lettera	Testo N° 1		Testo N° 2	
	freq./10^4	H_1	freq./10^4	H_1
a	0.1153	0.3593	0.1040	0.3396
b	0.0075	0.0527	0.0085	0.0585
c	0.0427	0.1942	0.0432	0.1958
d	0.0390	0.1826	0.0382	0.1799
e	0.1054	0.3422	0.1026	0.3370
f	0.0092	0.0622	0.0099	0.0659
g	0.0189	0.1081	0.0173	0.1013
h	0.0093	0.0627	0.0097	0.0649
i	0.1061	0.3434	0.1153	0.3593
l	0.0627	0.2506	0.0599	0.2433
m	0.0206	0.1155	0.0260	0.1369
n	0.0719	0.2731	0.0696	0.2676
o	0.0946	0.3218	0.0840	0.3002
p	0.0259	0.1367	0.0266	0.1392
q	0.0038	0.0304	0.0023	0.0202
r	0.0631	0.2516	0.0630	0.2513
s	0.0497	0.2151	0.0562	0.2334
t	0.0705	0.2697	0.0623	0.2495
u	0.0267	0.1396	0.0266	0.1392
v	0.0128	0.0804	0.0127	0.0800
z	0.0127	0.0799	0.0087	0.0595
j	0.0000	0.0000	0.0001	0.0013
k	0.0000	0.0000	0.0005	0.0055
w	0.0001	0.0013	0.0006	0.0064
x	0.0000	0.0000	0.0001	0.0013
y	0.0000	0.0000	0.0010	0.0100
cifre	0.0022	0.0196	0.0119	0.0760
.	0.0070	0.0499	0.0071	0.0507
,	0.0130	0.0813	0.0152	0.0918
;	0.0001	0.0013	0.0000	0.0000
:	0.0013	0.0121	0.0016	0.0149
?	0.0003	0.0034	0.0004	0.0045
!	0.0002	0.0023	0.0000	0.0000
( )	0.0008	0.0080	0.0007	0.0073
»	0.0000	0.0000	0.0001	0.0013
« , »	0.0048	0.0372	0.0051	0.0388
—	0.0002	0.0023	0.0001	0.0013
...	0.0001	0.0013	0.0003	0.0035
+	0.0002	0.0023	0.0000	0.0000
'	0.0041	0.0323	0.0051	0.0388
%	0.0000	0.0000	0.0008	0.0082
(0),	0.0000	0.0000	0.0005	0.0055
Totali		4.1264		4.1896

ed in questo caso avremmo già bisogno di 6 cifre, che si adatterebbero ad alfabeti con un numero di elementi compreso tra 33 e 64.

Ancora maggiore è tuttavia l'entropia di un testo scritto quando si considerino diverse le maiuscole dalle minuscole. Nella nostra analisi sono state considerate maiuscole soltanto quelle lettere maiuscole che non comparivano in inizio di frase, in quanto una maiuscola dopo un punto è chiaramente ridondante. Ciò ha portato per il testo n° 1 ad un alfabeto di 40 simboli (22 minuscole, cifre, 17 maiuscole) e per il testo n° 2 ad un alfabeto di 46 simboli (25 minuscole, cifre, 20 maiuscole). Tali numeri sono ovviamente legati al corpus ; in linea generale si dovrebbero prevedere almeno 53 simboli (26 minuscole, 26 maiuscole, cifre). Si noti che nel testo n° 2 alcune lettere compaiono più frequentemente nella forma maiuscola che nella minuscola, il che mostra come la politica estera affidi una discreta parte della sua informazione ai nomi propri. Infatti le maiuscole sono 122 nel testo n° 1 e 146 nel testo n° 2. L'entropia totale, considerando maiuscole, minuscole e cifre, senza considerare gli altri segni tipografici, è 4.0136 per il primo testo e 4.1143 per il secondo.

Un ulteriore ampliamento dell'alfabeto, comprendente maiuscole, minuscole, cifre, segni di interpunzione porta ad un codice a 7 elementi, in quanto è da prevedere che l'alfabeto sia di $53 + 14 = 67$ simboli. Le tabelle III e IV portano rispettivamente le occorrenze di tutti questi elementi e le rispettive entropie, già calcolate rapportate a 10.000, per il testo n° 1 e per il testo n° 2 (*).

Potremmo considerare che questo ampliamento dell'alfabeto è il massimo effettivamente significativo, a meno che non si voglia distinguere una vocale scritta con l'accento da una senza accento, oppure una cifra dall'altra, il che è essenziale dal punto di vista dell'informazione, ma meno importante dal punto di vista dello stile linguistico. Un discorso a parte meriterebbe lo spazio come segno tipografico.

(*) Il calcolo relativo a queste tabelle è stato eseguito da G. Zilli e F. Degan sul calcolatore Olivetti 652.

TABELLA III

Simbolo	occ.	H_1	Simbolo	occ.	H_1
a	1183	0.3580	*r*	648	0.2509
A	8	0.0080	*R*	4	0.0044
b	67	0.0472	*s*	510	0.2146
B	10	0.0097	*S*	3	0.0034
c	427	0.1903	*t*	727	0.2698
C	14	0.0129	*T*	1	0.0013
d	389	0.1784	*u*	276	0.1398
D	14	0.0129	*U*	0	0.0000
e	1086	0.3420	*v*	131	0.0800
E	3	0.0034	*V*	1	0.0013
f	81	0.0549	*z*	111	0.0704
F	14	0.0129	*Z*	20	0.0175
g	183	0.1032	*w*	1	0.0013
G	12	0.0113	*W*	0	0.0000
h	96	0.0628	.	72	0.0499
H	0	0.0000	,	134	0.0813
i	1068	0.3388	;	1	0.0013
I	1	0.0013	:	13	0.0121
l	646	0.2504	?	3	0.0034
L	2	0.0024	!	2	0.0023
m	204	0.1119	( )	8	0.0080
M	9	0.0089	« »	48	0.0372
n	741	0.2730	—	2	0.0023
N	2	0.0024	...	1	0.0013
o	977	0.3221	—	2	0.0023
O	0	0.0000	'	41	0.0323
p	264	0.1354	cifre	22	0.0196
P	4	0.0044			
q	39	0.0304			
Q	0	0.0000	Totali	10.330	4.1973

TABELLA IV

Simbolo	occ.	H_1	Simbolo	occ.	H_1
a	1079	0.3393	*T*	1	0.0013
A	2	0.0024	*u*	257	0.1320
b	81	0.0546	*U*	20	0.0174
B	8	0.0080	*v*	129	0.0786
c	443	0.1941	*V*	3	0.0034
C	7	0.0071	*z*	91	0.0599
d	390	0.1777	*Z*	0	0.0000
D	8	0.0080	*j*	0	0.0000
e	1060	0.3360	*J*	2	0.0024
E	7	0.0071	*k*	2	0.0024
f	95	0.0619	*K*	4	0.0044
F	8	0.0080	*w*	1	0.0013
g	175	0.0992	*W*	6	0.0062
G	5	0.0053	*x*	1	0.0013
h	99	0.0640	*X*	0	0.0000
H	2	0.0024	*y*	11	0.0105
i	1197	0.3591	*Y*	0	0.0000
I	2	0.0024	Cifre	124	0.0762
l	623	0.2434	.	74	0.0508
L	0	0.0000	,	158	0.0918
m	248	0.1286	;	0	0.0000
M	23	0.0195	:	17	0.0151
n	722	0.2673	?	5	0.0053
N	2	0.0024	!	0	0.0000
o	873	0.3002	( )	8	0.0080
O	1	0.0013	»	2	0.0024
p	263	0.1342	« »	54	0.0394
P	14	0.0128	—	2	0.0024
q	24	0.0202	...	4	0.0044
Q	0	0.0000	'	54	0.0394
r	643	0.2484	%	9	0.0088
R	12	0.0113	(0),	6	0.0062
s	576	0.2313			
S	9	0.0088			
t	647	0.2494	Totali	10.393	4.2870

§ 4. - Entropia della lingua parlata.

Le percentuali dei fonemi calcolate in [8] ci permettono, sempre con l'uso della formula di Shannon

$$H(X) = - \sum_{x \in X} p(x) \lg_2 p\,(x),$$

di calcolare anche l'entropia dei singoli fonemi e quindi della lingua parlata. Usiamo qui lo stesso testo analizzato in [8] (la « Veglia d'armi » di Diego Fabbri), e la stessa trascrizione fonetica usata in [8]; tale trascrizione è molto precisa, nonostante non sia troppo convincente la motivazione per cui i fonemi z e z' vengono esclusi dall'opposizione « geminata : semplice ». Il fatto che lo spettro acustico presenti anche per z e z' semplici l'aspetto di una consonante lunga non autorizza, a nostro giudizio, a dichiarare trascurabile tale opposizione. Sono invece molto convincenti i motivi per i quali è stato scelto tale testo come esempio di lingua parlata : si tratta di un lavoro teatrale, quindi colloquiale, con argomenti di varia natura, per quanto sempre riferibile ad un linguaggio di un ceto medio-colto. Il testo conta 18.970 parole che sono state trascritte in 83.098 fonemi. La tabella V riporta i valori di H da noi ricavati e, per comodità, anche le frequenze rapportate a 10.000 come da [8]. I fonemi distinti sono 50, il che, dal punto di vista tecnico, ci porterebbe ad un codice a 6 cifre.

Come si vede, l'entropia calcolata sui fonemi è notevolmente superiore a quella calcolata sui grafemi, tuttavia non così superiore quanto potrebbe apparire considerando solo i grafemi registrati da [3].

§ 5. - Successivi livelli di entropia.

Per uno studio più avanzato si può calcolare l'entropia condizionata di livello r così definita

$$H_r(Y) = - \sum_{x,y \in Y} pb_x(y) \lg_2 =$$

$$= - \sum_{x,y \in Y} p(b_x, y) \lg_2 p(b_x, y) + \sum_{x \in X} p(b_x) \lg_2 p(b_x)$$

TABELLA V

Fonema	freq./10^4	H_1	Fonema	freq./10^4	H_1
a	0.06432	0.25461	*m/*	0.00126	0.01213
á	0.03966	0.18466	*n*	0.07271	0.27496
b	0.00523	0.03963	*n/*	0.00122	0.01180
b/	0.00254	0.02189	*n'*	0.00184	0.01671
č	0.00744	0.05428	*o*	0.08000	0.29150
č/	0.00087	0.00884	*ó*	0.02271	0.12400
k	0.04109	0.18922	*ò*	0.01385	0.08550
k/	0.00205	0.01830	*p*	0.02981	0.15107
d	0.03313	0.15386	*p/*	0.00162	0.01501
d/	0.00016	0.00201	*r*	0.06837	0.26462
e	0.08214	0.29617	*r/*	0.00172	0.01582
é	0.02839	0.14588	*s*	0.04423	0.19898
è	0.02235	0.12255	*s/*	0.00627	0.04588
f	0.00825	0.05710	*š*	0.00209	0.01860
f/	0.00059	0.00632	*ʃ*	0.00394	0.03147
g	0.00385	0.03088	*t*	0.05677	0.23495
g/	0.00004	0.00058	*t/*	0.00678	0.04884
ǧ	0.00385	0.03088	*ú*	0.01279	0.08043
ǧ/	0.00102	0.01013	*u*	0.00856	0.05879
i	0.06504	0.25642	*w*	0.01189	0.07602
í	0.01938	0.11025	*v*	0.02134	0.11844
y	0.02096	0.11687	*v/*	0.00051	0.00557
l	0.03196	0.15876	*z*	0.00483	0.37160
l/	0.00656	0.04757	*z'*	0.00020	0.00245
l'	0.00218	0.01927			
m	0.03111	0.15575	Entropia totale		5.04781

dove $p_{b_x}(y)$ è la probabilità di y condizionata dal gruppo b_x, formato da $r-1$ lettere, che precede y.

Le H_2 e H_3 per i grafemi della lingua italiana sono già state calcolate in [3]. Per i fonemi nulla finora è stato calcolato, per quanto per il calcolo di H_2 potrebbe servire di spunto la tabella II di [4], nonostante che la trascrizione sia meno fine di quella usata in [8], pur riferendosi allo stesso testo.

È interessante il calcolo dell'entropia a livello di parole. Per questo serve una valutazione piuttosto precisa della lunghezza media delle parole della lingua in questione. A livello grafematico la lunghezza media riportata in [3] è di 5,20 lettere/parola ; nei testi n° 1

e n° 2 oggetto del presente studio le parole erano sempre 1911, il che conduce ad una lunghezza media di 5,23 lettere/parola. A livello di fonemi il calcolo sui dati di [8] ci conduce a 4,380 fonemi/parola.

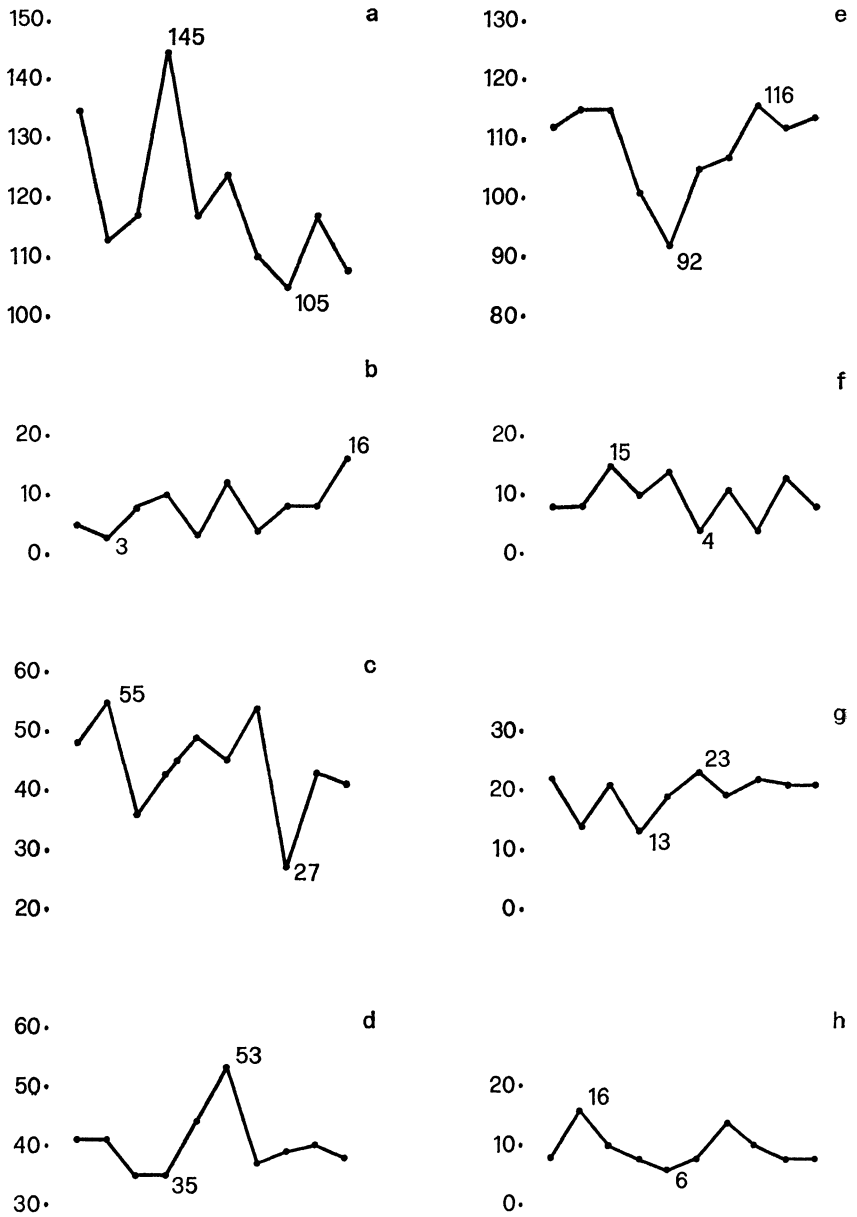
Resta aperto il problema del calcolo dell'entropia a livello di parole per un numero significativo di testi di varie epoche, per poter verificare una variazione di entropia con il tempo. Si potrebbe anche definire un vocabolario di concetti base da assumere come alfabeto: nella lingua italiana di uso comune non sono più di 700-800, nonostante che i lemmi siano molte migliaia. In base a tale alfabeto si potrebbero calcolare entropie di testi scritti e parlati, rilevando così con maggior rigore somiglianze e differenze di stile e di contenuto.

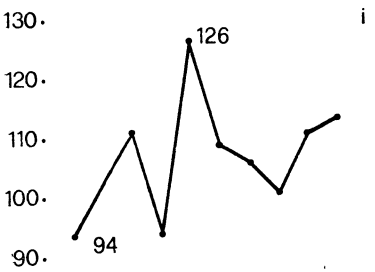
TABELLA I_A

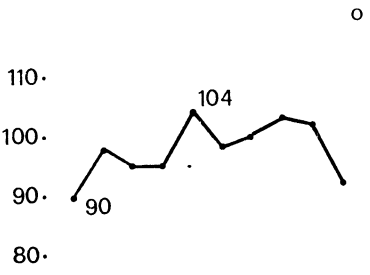
a	135	113	117	145	117	124	110	105	117	108
b	5	3	8	10	3	12	4	8	8	16
c	48	55	36	43	49	45	54	27	43	41
d	41	41	35	35	44	53	37	39	40	38
e	112	115	115	101	92	105	107	116	112	114
f	8	8	15	10	14	4	11	4	13	8
g	22	14	21	13	19	23	19	22	21	21
h	8	16	10	8	6	8	14	10	8	8
i	94	103	111	94	126	109	106	101	111	114
l	80	68	58	60	60	68	61	67	64	62
m	19	23	30	18	20	20	15	15	22	31
n	64	78	77	60	81	75	71	81	79	77
o	90	98	95	95	104	98	100	103	102	92
p	20	25	28	35	28	23	28	28	30	23
q	3	4	3	4	4	4	3	5	3	6
r	64	55	59	60	56	68	78	71	71	70
s	50	53	71	59	48	40	49	51	46	46
t	80	72	61	92	75	62	66	85	69	66
u	24	25	20	29	33	27	36	32	16	34
v	16	11	15	14	9	16	10	14	14	13
z	11	18	13	15	12	14	13	13	10	12
j	—	—	—	—	—	—	—	—	—	—
k	—	—	—	—	—	—	—	—	—	—
w	—	—	—	—	—	—	1	—	—	—
x	—	—	—	—	—	—	—	—	—	—
y	—	—	—	—	—	—	—	—	—	—

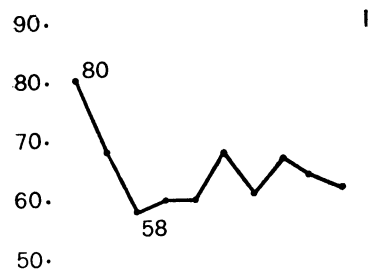
TABELLA II_A

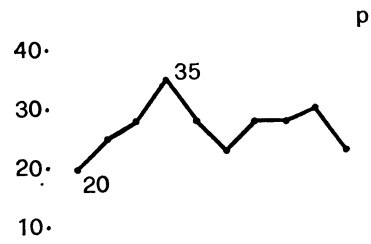
a	124	127	121	103	100	102	98	112	101	112
b	6	16	3	17	5	10	7	10	8	10
c	39	58	34	63	32	47	48	41	50	41
d	42	35	28	37	46	45	40	42	39	42
e	111	97	106	100	119	115	106	113	104	113
f	16	6	14	7	7	8	12	9	11	9
g	20	20	18	23	17	14	12	20	18	20
h	10	15	8	10	13	11	13	7	7	7
i	109	113	133	124	120	128	114	110	120	110
l	54	55	62	49	63	63	75	62	77	62
m	21	16	35	24	36	28	26	26	25	26
n	82	83	62	68	76	75	82	59	66	59
o	85	93	75	97	81	81	87	91	87	91
p	32	23	35	23	23	19	28	29	30	29
q	3	1	5	—	2	5	1	3	2	3
r	67	68	72	59	62	60	64	84	65	84
s	63	56	59	61	65	64	54	52	58	52
t	68	58	61	70	77	65	65	66	58	66
u	26	22	29	22	38	31	24	28	30	28
v	11	18	11	14	9	14	16	17	15	17
z	6	13	13	6	5	8	10	11	9	11
j	1	1	—	—	—	—	—	—	—	—
k	1	2	1	1	1	—	—	—	—	—
w	1	2	1	2	1	—	—	—	—	—
x	1	—	—	—	—	—	—	—	—	—
y	1	2	—	—	2	1	1	2	2	2

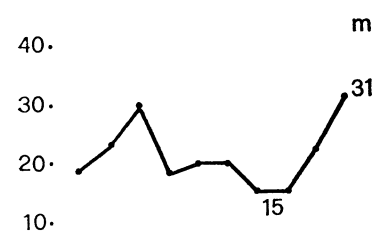

a
150.
140.
130.
120.
110.
100.
145
105
b
20.
10.
0.
16
3
c
60.
50.
40.
30.
20.
55
27
d
60.
50.
40.
30.
53
35
e
130.
120.
110.
100.
90.
80.
116
92
f
20.
10.
0.
15
4
g
30.
20.
10.
0.
23
13
h
20.
10.
0.
16
6


i
130
120
110
100
90
126
94

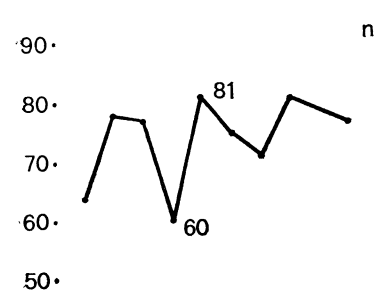

o
110
100
90
80
104
90

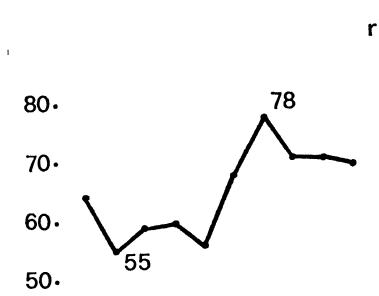

l
90
80
70
60
50
80
58

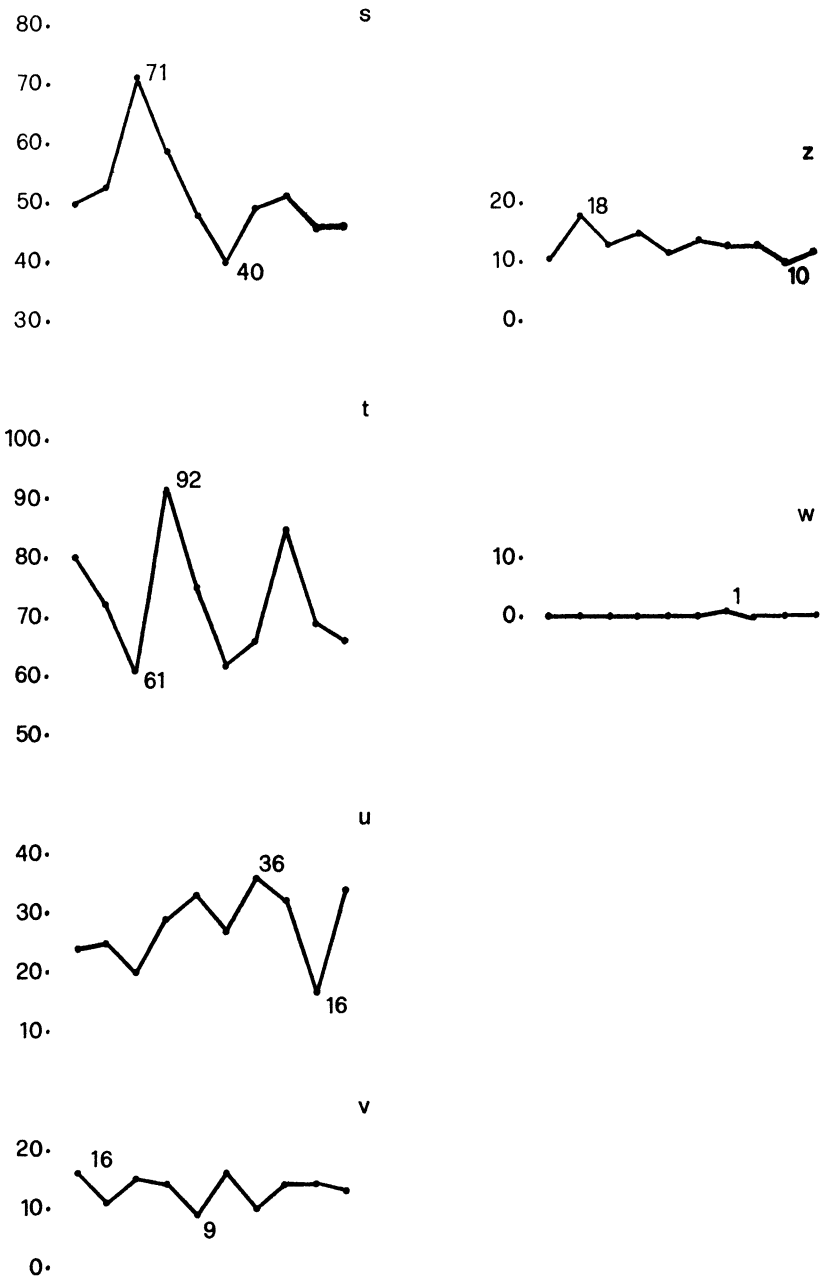

p
40
30
20
10
35
20

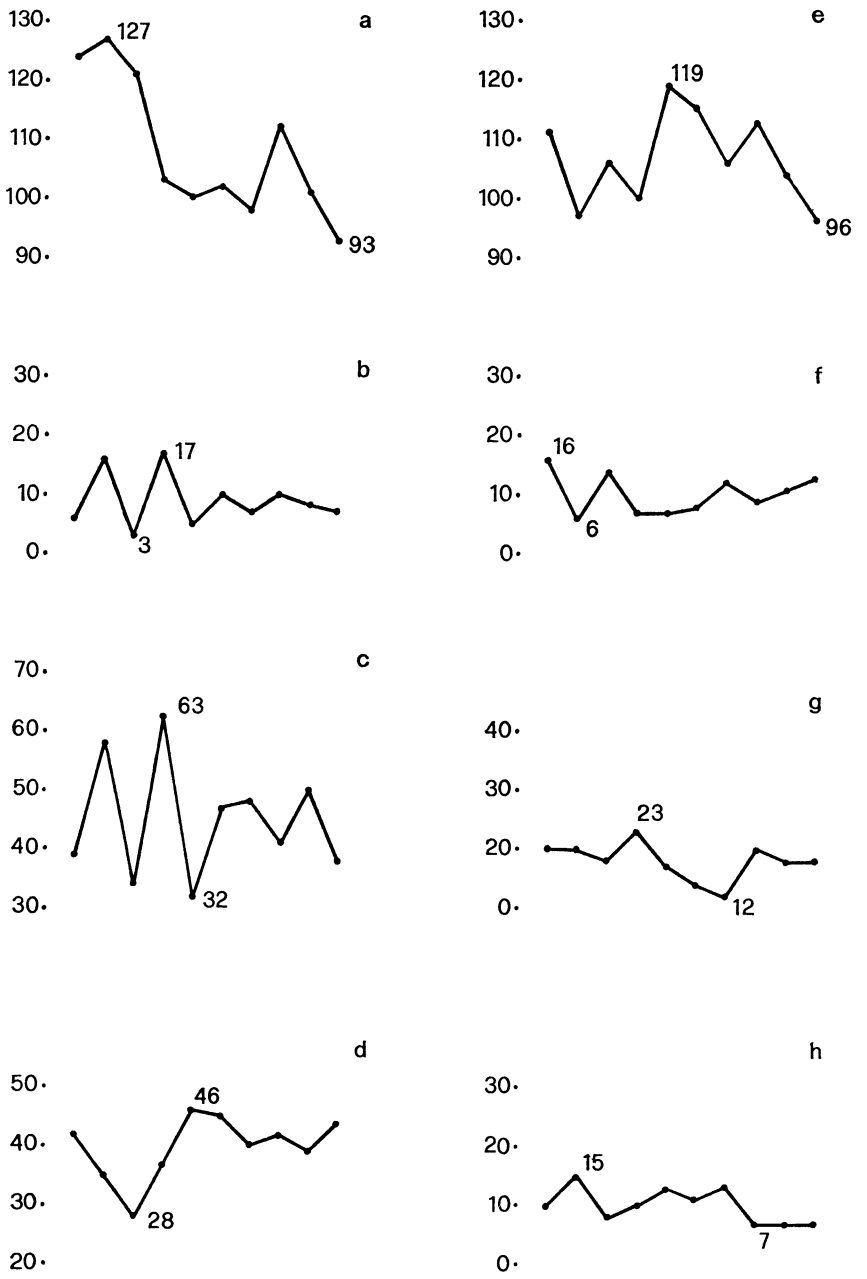

m
40
30
20
10
31
15

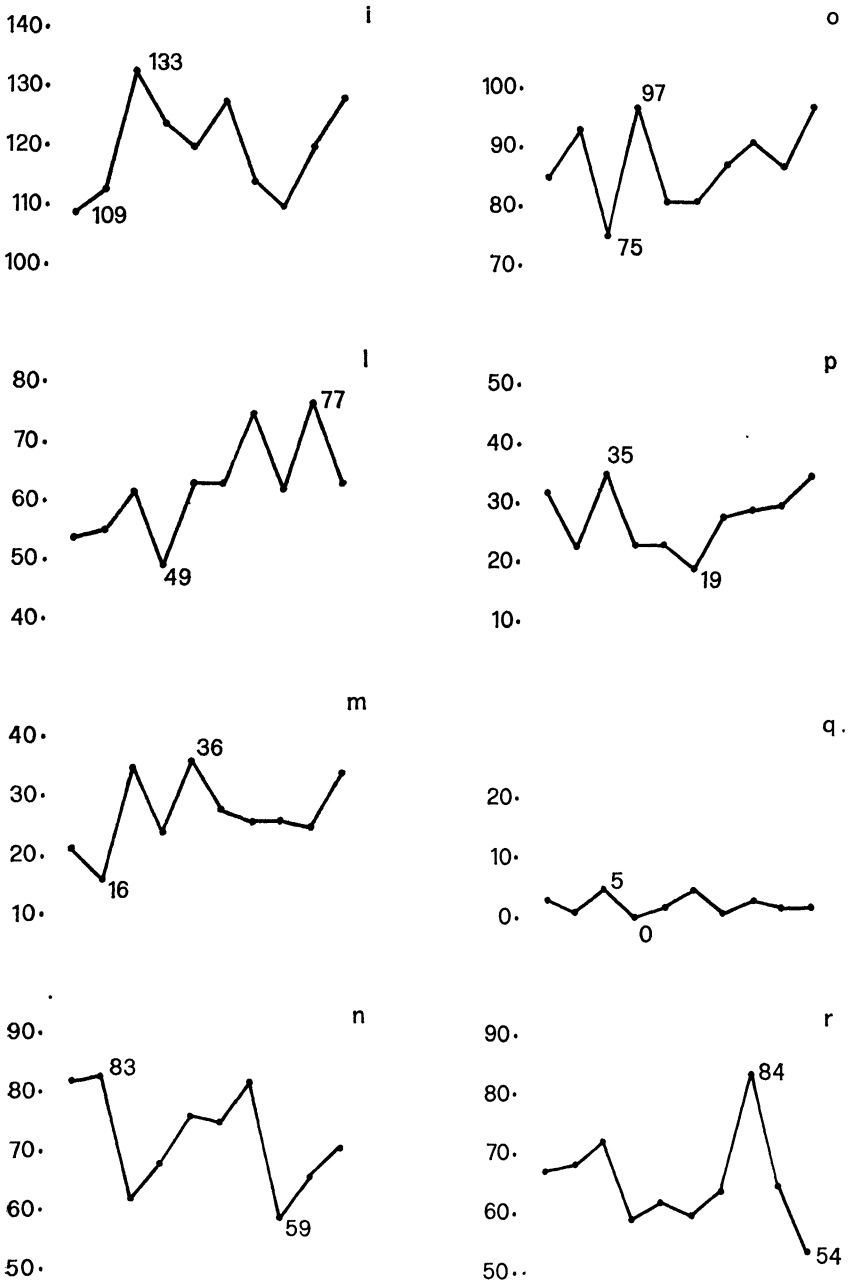

q
15
10
0
6
3

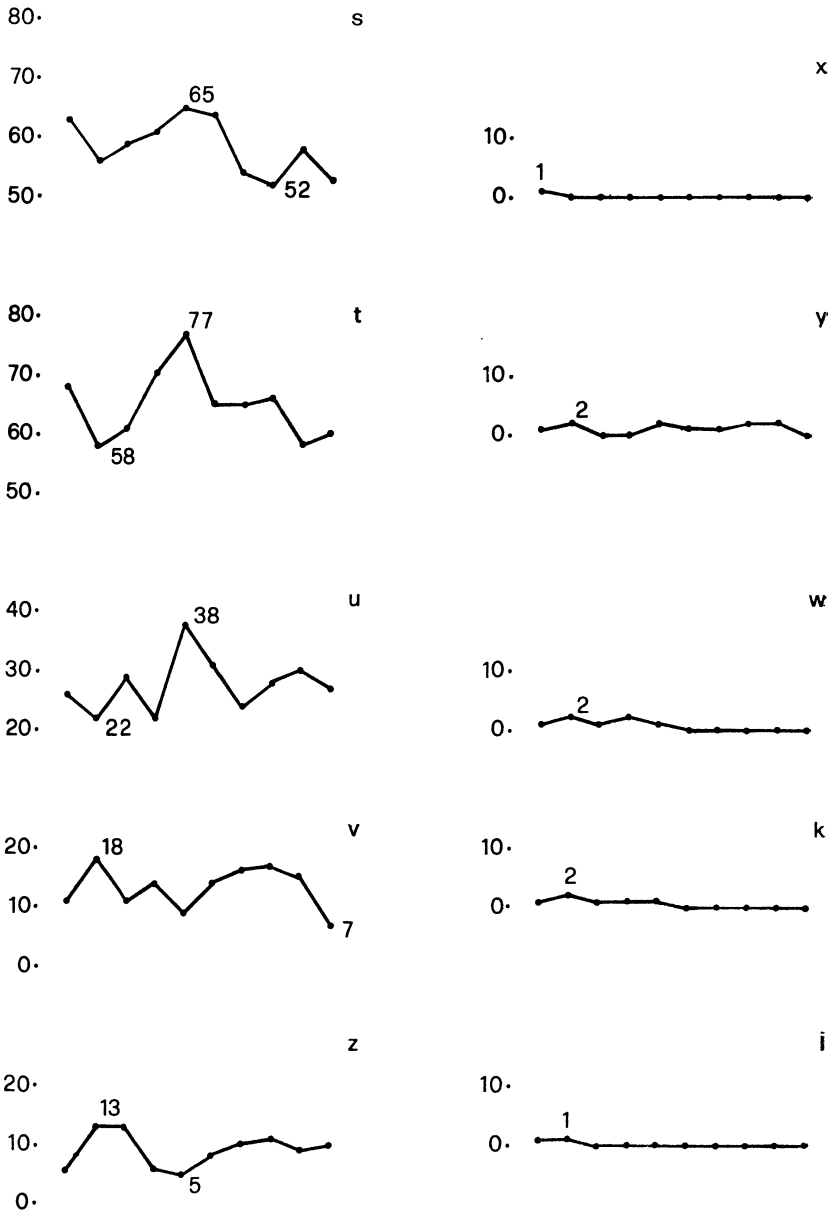

n
90
80
70
60
50
81
60


r
80
70
60
50
78
55


s
80.
70.
60.
50.
40.
30.
71
40
z
20.
10.
0.
18
10
t
100.
90.
80.
70.
60.
50.
92
61
w
10.
0.
1
u
40.
30.
20.
10.
36
16
v
20.
10.
0.
16
9


a
130
120
110
100
90
127
93
e
130
120
110
100
90
119
96
b
30
20
10
0
17
3
f
30
20
10
0
16
6
c
70
60
50
40
30
63
32
g
40
30
20
0
23
12
d
50
40
30
20
46
28
h
30
20
10
0
15
7


i
140·
130·
120·
110·
100·
133
109
o
100·
90·
80·
70·
97
75
l
80·
70·
60·
50·
40·
77
49
p
50·
40·
30·
20·
10·
35
19
m
40·
30·
20·
10·
36
16
q·
20·
10·
0·
5
0
n
90·
80·
70·
60·
50·
83
59
r
90·
80·
70·
60·
50··
84
54


s
80·
70·
60·
50·
65
52
x
10·
0·
1
t
80·
70·
60·
50·
77
58
y
10·
0·
2
u
40·
30·
20·
38
22
w
10·
0·
2
v
20·
10·
0·
18
7
k
10·
0·
2
z
20·
10·
0·
13
5
i
10·
0·
1

BIBLIOGRAFIA

[1] C. E. SHANNON, *Prediction and Entropy of Printed English*, Bell S.T.J., 1951, XXX, pp. 50-64.

[2] G. A. BARNARD, *Statistical Calculation of Word Entropies for Four Western Languages*, I.R.E. Trans. Information Theory, 1955, IT-I, p. 49.

[3] R. MANFRINO, *L'entropia della lingua italiana ed il suo calcolo*, « Alta frequenza », 1960, XXIX, pp. 4-29.

[4] G. B. DEBIASI - G. A. VALLI, *Sorgenti di messaggi per prove di intelligibilità della parola*, Mem. Acc. Pat. Sc. Lett. Arti, 1968, LXXX, pp. 293-314.

[5] E. A. AFENDRAS - N. S. TZANNES - J. G. TRÉPANIER, *Distance, Variation and Change in Phonology : Stochastic Aspects*, Folia linguistica, 1973, 6, pp. 1-27.

[6] G. B. DEBIASI - G. DE POLI - G. A. MIAN - C. MILDONIAN - C. OFFELLI, *Italian speech synthesis from unrestricted text for an automatic answerback system*, Proc. 8th Int. Congr. Acoustics, Londra, 1974, p. 296.

[7] G. B. DEBIASI - G. DE POLI - G. A. MIAN - C. OFFELLI, *Voce dagli elaboratori : prospettive di sviluppo e realtà di una applicazione*, Atti 3° Convegno di Cibernetica e Biofisica, S. Marino, 1974.

[8] R. BUSA - C. CROATTO-MARTINOLLI - L. CROATTO - C. TAGLIAVINI - A. ZAMPOLLI, *Una ricerca statistica sulla composizione fonologica della lingua italiana parlata eseguita con un sistema IBM a schede perforate*, Proc. 12th Int. Speech & Voice Therapy Conference, Padua, 1963, pp. 542-562.

Manoscritto pervenuto in redazione il 6 aprile 1977.