U. BORTOLINI • C. TAGLIAVINI • A. ZAMPOLLI

Lessico di frequenza della lingua italiana contemporanea

GARZANTI

Questo volume, che mette a disposizione di un più largo numero di lettori i risultati di ricerche prima riservate a pochi specialisti, è ricavato – per cortese concessione della IBM-Italia alla Casa Editrice Garzanti, che intende così arricchire la propria tradizione lessicografica – da un recente lavoro che come lo studio sulla Divina Commedia, non in commercio, è frutto dell'attività di ricerca linguistica e letteraria svolta dalla IBM-Italia in collaborazione con il Centro Nazionale Universitario di Calcolo Elettronico di Pisa (CNUCE).

Il presente lavoro, basato su uno spoglio di 500.000 parole della lingua italiana contemporanea, fornisce una prima raccolta di circa 5000 lemmi disposti sia in ordine alfabetico sia in ordine di frequenza decrescente e di altri parametri statistico-matematici, come vedremo in seguito. Esso non ha la pretesa di essere assolutamente paradigmatico poiché, soprattutto a un livello di frequenza inferiore a quello dei primi 1000/1500 lemmi, i dati sono più legati alla scelta dei campioni, talché una diversa campionatura potrebbe portare spostamenti e anche variazioni di lemmi. Con esso ci proponiamo di dare un utile strumento di lavoro per lo studio del lessico contemporaneo che, il più delle volte, nei dizionari normativi viene necessariamente quasi sommerso nel *mare magnum* di voci e forme tradizionali o già uscite dall'uso, o in via di sparizione, o limitate a lessici speciali o, anche se tuttora vitali, fornite di accezioni particolari.

D'altra parte, pur convinti della relatività dei dati da noi offerti, un elenco di questo genere non può non avere anche un notevole interesse pratico per lo studio della lingua italiana sia come lingua materna sia, e forse più, come lingua straniera.

1 PRECEDENTI STORICI

1.1.1 Per quanto l'esigenza di una considerazione quantitativa del linguaggio dal punto di vista scientifico sia relativamente recente e l'applicazione dei metodi statistici alla linguistica sia diversamente interpretata e valutata, bisogna riconoscere che, fin dall'epoca dei Greci e dei Romani, filologi e grammatici, più o meno consapevolmente, hanno tenuto conto del fattore quantitativo, per lo meno come opposizione fra voci rare e obsolete (o addirittura *hapax legomena*) e parole di particolare frequenza e uso. È ben noto che la grammatica classica, pressappoco come fu formulata dagli alessandrini, rappresenta un compromesso fra i principi degli *analogisti* e quelli degli *anomalisti*. L'analogia, cioè la regola, era formata da ciò che veniva indicato come *normale* e per ciò stesso più frequente, l'anomalia da ciò che era eccezionale e per ciò stesso più raro. Anche i filologi, per l'attribuzione dei testi adespoti ad un autore o ad un altro, tenevano conto dell'apparizione di certe parole o di certi stilemi con una maggiore o minore frequenza. Tale frequenza era determinata, più che altro, dal senso linguistico e dall'esperienza dei singoli filologi, finché questi non si poterono basare su lessici speciali che garantivano una certa completezza. Le prime concordanze furono fatte nel medioevo per i testi biblici: per ciascuna parola erano riportati tutti i contesti dove appariva. Nei secoli successivi le imprese di questo tipo si moltiplicarono e si ebbero concordanze di testi di autori classici e moderni, e lessici speciali dedicati a opere particolari di determinati autori.

Nel XIX secolo compaiono anche le prime liste di frequenza, compilate però per scopi extralinguistici. Una delle prime opere e senz'altro la più vasta realizzata con spogli manuali è il volume *Häufigkeitswörterbuch der deutschen Sprache* pubblicato a Berlino (Steglitz) nel 1898 a opera di F.W. Kaeding (e di moltissimi collaboratori): si basa su uno spoglio di circa 11.000.000 di parole con lo scopo di stabilire le frequenze dei singoli grafemi, delle singole

sillabe e, infine, delle parole ad uso degli stenografi del sistema Stolze-Schrey. È ancora in un'opera compilata per scopi stenografici che si procede ad un esame critico delle liste di frequenza: J.B. Estoup, nelle sue *Gammes sténographiques* (Paris 1907, ed edizioni successive), stabilì, infatti, l'importantissima relazione fra il numero di occorrenze di un termine e il suo rango nella serie delle parole ordinate per frequenza decrescente, che fornì uno dei punti di partenza della linguistica quantitativa. Gli indici e repertori di frequenza, lo spoglio dei testi hanno dunque una lunga storia, ma né i grammatici alessandrini né gli eruditi del medioevo e del rinascimento pensarono mai di utilizzarli per l'insegnamento delle lingue. Fu solo quando questo insegnamento ebbe preso maggiore estensione e ci si rese conto dell'impossibilità di dominare una lingua in qualche anno di studio, per poche ore la settimana, che si pensò di limitare l'insegnamento lessicale e di non ritenere, per insegnarle agli allievi, che le parole più frequenti (v. R. Michéa, *Les vocabulaires fondamentaux*).

S'iniziò allora, dapprima negli Stati Uniti, poi anche negli altri paesi, la compilazione di dizionari fondamentali soprattutto ad opera di pedagogisti e psicologi che volevano da una parte rendersi conto dell'estensione del vocabolario infantile a diverse età, dall'altra trarre indicazioni sull'apprendimento del lessico.

Vari testi vengono sottoposti all'analisi statistica: opere letterarie, libri scolastici, lettere personali ecc. Si compilano e si confrontano liste diverse per tentare di raggiungere risultati più generali e per ciò stesso più validi, nella convinzione che le parole più frequenti siano anche le più utili.

Ne ricordiamo le principali:[1]

per l'inglese:

E.L. THORNDIKE, *The teacher's word book*, New York, 1921;

E.L. THORNDIKE, I. LORGE, *The teacher's word book of 30.000 words*, New York, 1944;

per il francese:

G.E. VANDER BEKE, *French word book*, New York, 1929;

V.A.C. HENMON, *A french word book based on a count of 400.000 running words*, Madison, 1924;

per lo spagnolo:

M.A. BUCHANAN, *A graded spanish word book*, Toronto, Ont., 1927;

H. KENISTON, *A basic list of spanish words and idioms*, Chicago, 1933.

I vocabolari fondamentali compilati in quest'epoca non sono d'altronde tutti basati sulla frequenza: Ogden e Richard, autori del *Basic English* (*Basic = British American Scientific International Commercial*) (1923-1928), « partendo dalla constatazione che un piccolo numero di parole basta a definirne molte altre: cioè a realizzare un gran numero di *meanings* » (E. Arcaini: *Princìpi di linguistica applicata*, p. 397), hanno compilato una lista di 850 parole (di cui soltanto 18 verbi) che sono generalmente le più frequenti, ma che non sono state scelte in base a spogli di frequenza: la selezione, infatti, è stata guidata dall'intento di esprimere col minimo numero di parole il maggior numero di

1 Per informazioni bibliografiche più dettagliate rinviamo a P. Guiraud, *Bibliographie critique de la statistique linguistique*, Utrecht, 1954, ed a D. Harkin, *The history of word counts* in « Babel », 3, 1957, pp. 113-124, e (per lavori apparsi dopo questa data) a B. Malmberg, *Les nouvelles tendances de la linguistique*, Paris, 1966, p. 278 e sgg.

oggetti e concetti. Si è venuta così costituendo una lingua semiartificiale che non è l'inglese naturale, poiché questo sistema presuppone sempre una trasposizione analitica delle frasi che non è sovente né semplice né facile.[2]

Questi spogli dei pedagogisti attirarono l'attenzione anche di linguisti e statistici: G.K. Zipf fu uno dei primi a studiare i fenomeni di distribuzione delle parole e ne cercò l'origine nel principio del minimo sforzo (v. G.K. Zipf, *Human behavior and the principle of least effort*, Cambridge, Mass. 1949). Si studiarono anche i problemi posti dalla compilazione delle liste di frequenza, il loro uso e valore pedagogico; si vide così che certe frequenze risultavano da cause accidentali e passeggere. La frequenza perciò, come vedremo meglio in seguito, doveva essere controllata tramite la ripartizione contestuale, poiché una parola che appare in un piccolo numero di testi, ma è spesso ripetuta, è necessariamente legata al contenuto di questi testi e quindi a delle circostanze particolari (Vander Beke nel suo *French word book* distingue già l'uso delle parole e la loro distribuzione contestuale). Si notò inoltre che le liste di frequenza ottenute dallo spoglio dei testi contenevano soprattutto parole grammaticali, verbi, aggettivi e qualche sostantivo di carattere generale, ma pochi sostantivi concreti, che un vocabolario essenzialmente pratico e destinato alla conversazione nelle circostanze più comuni della vita quotidiana avrebbe dovuto contenere.

Fu per questo che più tardi G. Gougenheim, R. Michéa, P. Rivenc e A. Sauvageot, autori del *Français fondamental* del CREDIF, pensarono di analizzare la lingua parlata registrando conversazioni di persone poste nelle situazioni più comuni della vita di ogni giorno. Ma anche i dati così ottenuti non rendevano conto di tutto il « vocabolario disponibile » nella mente del parlante anche se non frequentemente usato e gli autori hanno cercato di determinarlo mediante serie di interrogazioni su vari « centri di interesse ».

1.1.2 L'italiano è stato un po' la lingua dimenticata in questi spogli di frequenza; finora sono stati fatti solo due brevi saggi, quello di T.M. Knease, *An italian word list from literary sources*, Toronto 1933 e quello di B. Migliorini, *Der grundlegende Wortschatz des Italienischen*, Marburg 1943.

Il primo rappresenta uno spoglio manuale su un totale di circa 4000 parole tratte da fonti esclusivamente letterarie, comprensive di opere in prosa e poesia che vanno dalla seconda metà dell'Ottocento ai primi decenni del Novecento. Sono state omesse quasi tutte le parole grammaticali e anche altre voci ritenute « particolarmente frequenti » e non sono stati distinti gli omografi. Da un raffronto sommario fatto con le nostre liste non ci risulta che parole di colorito prettamente letterario siano presenti nella Knease ed assenti nel nostro Lessico di frequenza della lingua italiana contemporanea (che d'ora in poi abbrevieremo LIF), come la diversità del campione poteva far prevedere.

Il lavoro del Migliorini è stato compilato unicamente sulla base del senso linguistico dell'autore, insigne storico della lingua italiana ed esperto lessicografo, che si proponeva di dare per esclusivi scopi didattici una lista delle

2 Per esempio: *to accelerate* viene reso con *to go faster*, *convinction* viene reso con *strong belief*.

1500 parole da lui ritenute fondamentali nella lingua italiana contemporanea e perciò più utili da proporre per lo studio della nostra lingua agli stranieri. Egli distingue le 500 parole più importanti stampandole in grassetto mentre lascia in caratteri normali le altre 1000. Un rapido raffronto con la nostra lista ci ha mostrato che, a parte alcune voci evidentemente legate alla situazione politico-sociale del momento come *balilla*, *corporativo*, *corporazione*, *aviere*, *duce*, le quali figuravano nel Migliorini fra le prime 500 e che non trovano posto nella nostra lista, ve ne sono altre che il Migliorini considera (*eroina* femminile di eroe, *voluttà, solere* ecc.), che rispecchiano aspetti lessicali e stilistici ormai in decadenza. Altre voci, invece, comprese dal Migliorini fra le 1500 compaiono nella nostra raccolta al di sotto della soglia d'uso che abbiamo scelto come limite della nostra lista, per esempio: *ardore, artiglieria, attirare, balzare, berretto, fisionomia* ecc. per un totale di 77 lemmi.[3]

1.2 Gli inizi del nostro secolo hanno dunque visto tutto un moltiplicarsi di ricerche vòlte alla compilazione di dizionari fondamentali, e tali erano lo slancio e il fervore delle attività, che R. Michéa (*Les vocabulaires fondamentaux*, p. 22) ha parlato di un'epoca eroica dei dizionari fondamentali.

Questo movimento, che raggiunge il suo apice negli anni immediatamente precedenti la seconda guerra mondiale, si accompagnò allo sviluppo generale della cosiddetta *linguistica applicata*, termine che, soprattutto negli Stati Uniti, fu per lungo tempo sinonimo di *glottodidattica*.

Il diffondersi, anche tra i cultori delle discipline umanistiche, delle tecniche di elaborazione automatica dei dati diede un nuovo impulso, negli anni dopo il 1950, ai progetti di dizionari fondamentali, i quali poterono beneficiare

3 Diamo l'elenco completo dei lemmi presenti nella lista della Knease e che non compaiono nel nostro LIF; le voci non contrassegnate da un asterisco compaiono nei testi spogliati, ma per il loro basso uso non sono state accolte nel LIF; mancano assolutamente quelle contrassegnate con asterisco. **Affogare, agonia, alito, allegramente, ambedue, ardore, *arrischiare, assiduo, astrarre, avvedere, balenare, balzare, berretto, borbottare, brontolare, celare, chiarore, *chioma, cipresso, collera, colto (-ó-), commento, contentare, cullare, curvare, dì, *dileguare, discorrere, divorare, erta, ferreo, fiero, *figgere, filosofico, filosofo, fioco, fisonomia, *florido, forestiere, *fosco, fremito, freschezza, fruscìo, *gaiezza, gaio, gemito, giovanile, gomito, gota, impacciare, impallidire, indifferenza, indugiare, infame, intervallo, invincibile, leggenda, lettore, *livido, mercante, oblio, *onde (avv.), pallore, palma, palpebra, palpitare, pendere, *pensoso, pentire, percorrere, *peregrino, pittura, placido, pranzare, quantunque, quercia, rammentare, rasserenare, rauco, ricamo, rimprovero, ripigliare, risvegliare, ritto, rondine, ruga, santità, sapienza, sboccare, scintillare, scrollare, sdegno, seguitare, serrare, servitore, sgomento, *sigaro, singhiozzare, smorto, socchiudere, solennità, solere, soprabito, sospiro, spasimo, spettro, spirare, stridere, sublime, svoltare, tempio, tenue, torbido, torpido, tranquillamente, *tremito, trionfare, tuono, turbamento, veglia, *velare, ventaglio, vergognoso, villaggio, viltà, virile, volto, *voluttà, vuotare.*

Ed ecco quelli presenti nell'elenco di B. Migliorini, che non compaiono nelle nostre liste: *accampamento, *affogare, *annuncio, ardore, artiglieria, atterrire, automobilista, aviazione, *aviere, avvedersi, *balilla, balzare, berretto, codesto, *costei, collera, compiacente, *compratore, conversare, convoglio, *corporativo, *corporazione, designare, dì, *dileguarsi, discorrere, duce, eroico, *eroina, federale, federazione, fiero, filosofo, fisionomia, forestiero, fremito, gaio, gomito, gota, incrociatore, indifferenza, insegnamento, modestia, naviglio, nobiltà, nono, novanta, *onde (avv.), palpebra, pendere, pentirsi, pilotare, pittura, principiare, *quantunque, *radioascoltatore, rammentare, ritto, scintillare, seguitare, serrare, *settimo, socchiudere, solere, sospiro, spasimo, turbamento, *velare, *ventesimo, villaggio, *voluttà, vuotare, zero.*

delle nuove possibilità di spoglio e di elaborazione statistica dei dati; nel Primo Congresso Internazionale di Linguistica Applicata (Nancy, 1964) venne riservato un posto di grande rilievo alla sezione dei dizionari fondamentali.

Questa seconda fase della storia dei dizionari fondamentali è caratterizzata da una maggiore preoccupazione per i problemi metodologici e teorici: non si dubita certo dell'utilità didattica di liste di parole particolarmente frequenti nella lingua da insegnare, ma si vogliono rendere espliciti i presupposti statistici e linguistici di questa prima intuizione, così da poter utilizzare, per un affinamento e un potenziamento delle procedure di compilazione dei dizionari di frequenza, i risultati e le tecniche recentemente sviluppati dalla statistica linguistica.

Spogli di testi condotti per i più diversi motivi hanno mostrato che un numero relativamente piccolo di vocaboli molto frequenti costituisce la maggior parte dell'intero testo. P. Guiraud in *Les caractères statistiques du vocabulaire* (p. 10) dice:

« Un très petit nombre de mots convenablement choisis couvrent la plus grande partie de n'importe quel texte, et il est possible d'établir une liste de mots telle que:

Les 100 premiers mots couvrent 60% de n'importe quel texte
Les 1000 premiers mots couvrent 85% de n'importe quel texte
Les 4000 premiers mots couvrent 97,5% de n'importe quel texte
Le reste (40 à 50.000 mots?) couvre 2,5% de n'importe quel texte. »

Si suppone non solo che questo tipo di distribuzione delle frequenze lessicali sia comune a tutti i testi, ma anche che le parole di maggiore frequenza siano pressappoco le stesse in tutti i testi. Ne deriverebbe, da un lato, che per identificarle basterebbe spogliare un numero adeguato di testi; dall'altro che, una volta elencate, converrebbe iniziare da questo elenco l'insegnamento del lessico di una lingua, perché apprendendo un numero relativamente limitato di parole, per esempio poche centinaia, l'allievo sarebbe posto subito in condizione di comprendere più dell'80% delle parole di qualsiasi testo.

Non è qui il luogo di discutere la proporzione tra numero di parole «capite» e comprensione del testo; questo ci porterebbe a discutere la relazione tra la frequenza di una unità linguistica e il suo valore informativo.[4] Lo studio di questo problema, condotto accuratamente dai teorici dell'informazione (nel senso introdotto da C. Shannon) e con risultati non trascurabili a livello grafematico e talora fonematico, è ancora agli inizi per quanto riguarda le unità linguistiche di altri livelli, in particolare del lessico e della sintassi.

È invece opportuno riferire brevemente i problemi teorico-pratici connessi al concetto della *frequenza* di una parola in una lingua e al *metodo* per rilevarla partendo da uno spoglio di testi.

4 « On établit, par exemple, des langues de base, qui sont des vocabulaires minimums, fondés sur cette observation que les mots les plus fréquents sont les mêmes dans la plupart des textes et que mille mots, convenablement choisis à partir d'une enquête statistique, constituent par leurs répétitions 85% de n'importe quel texte. Le calcul montre que ces mots, très fréquents, ont un faible contenu d'information et ne constituent que 50% de l'information du texte. » P. Guiraud, *Langage et théorie...*, 1968, pp. 160-161.

2.1 I testi sottoposti a spoglio si configurano, nelle intenzioni dei compilatori dei dizionari di frequenza, come campioni della lingua intesa quale universo statistico e di conseguenza pongono i problemi della campionatura per i quali la statistica ha elaborato tutta una serie di metodologie e di tecniche (v. M. Boldrini, *Statistica...*, 1968, p. 512 e segg.).

Da un lato occorre delimitare l'universo, dall'altro assicurarsi della rappresentatività del campione. Se si conoscono bene le caratteristiche dell'«universo», si può fabbricare un campione in cui ciascuna di queste caratteristiche sia distribuita secondo le stesse proporzioni, con una sottile stratificazione (Ch. Muller, *Initiation...*, 1968, p. 15)[5]. Che non sia questo il caso della lingua è facilmente intuibile.

Le frequenze si osservano in un testo, cioè in una realizzazione del sistema linguistico da parte di un individuo, realizzazione determinata da un lato dalle sue caratteristiche personali, dall'altro da quelle della comunità cui appartiene, caratterizzata sul piano sociale, geografico, storico, ecc. Se il testo è un testo letterario, nella sua produzione influiscono anche le regole e le peculiarità del genere letterario e in ogni caso la situazione in cui viene prodotto.

I lessicografi hanno sempre parlato di un linguaggio tecnico e di uno poetico, di uno popolare e di uno dotto ecc.[6] È legittimo chiederci se attualmente possediamo, della struttura dell'universo linguistico, una conoscenza sufficiente a decidere le dimensioni e la composizione di un campione rappresentativo e, addirittura, a stabilire se e in quali limiti il rapporto tra testo e lingua sia legittimamente equiparabile, con rigore metodologico, al rapporto campione-universo.

2.2.1 Vi sono almeno due maniere diverse di guardare la lingua come un universo statistico.

Secondo una prima concezione, le unità del sistema linguistico sarebbero caratterizzate, oltre che dai tratti qualitativi emergenti dalle opposizioni e dalle relazioni che formano la struttura del sistema stesso, anche dalle loro rispettive probabilità di uso. Queste probabilità non sono direttamente osservabili, come non lo è, del resto, il sistema; esse però si tradurrebbero nel fatto (dato per certo) che le unità linguistiche ricorrerebbero nei testi, parlati e scritti, con frequenze relativamente stabili. In questa prospettiva le frequenze osservabili nei testi vengono assunte come *approssimazioni* delle probabilità non osservabili del sistema.

Tale concezione ha ricevuto una formulazione teorica esplicita da P. Gui-

5 È questo il caso, per esempio, dei sondaggi d'opinione prima delle elezioni. Per gli iscritti alle liste elettorali si conoscono con precisione la ripartizione geografica, la condizione socio-professionale, l'età, il sesso ecc.

6 Si veda per esempio B. Migliorini, *Che cos'è un vocabolario?*, Firenze, 1951, 2ª ed., p. 10 e segg. e p. 42 e segg.

raud,[7] anche se essa già traspariva da studi precedenti,[8] e fu accettata e sviluppata in primo luogo da G. Herdan[9] e poi da altri cultori della materia, i quali in essa riconoscono il fondamento teorico sul quale la statistica linguistica pone la propria autonomia come scienza.[10]

Queste premesse non hanno però trovato, almeno fino ad oggi, un consenso generale.[11]

Già gli autori del *Français fondamental* introducendo la nozione di disponibilità scossero per primi questa concezione e R. Moreau, al convegno di Strasburgo del 1964 sul tema *Statistique et analyse linguistique*, delineava esplicitamente questa situazione affermando: «Les premiers pas de la statistique appliquée à la linguistique ont précisément consistés à admettre des règles du jeu qui soient simples. C'est ainsi qu'on a énoncé (voir par exemple Guiraud) que la fréquence des mots était constante dans la langue, ce qui supposait donc que l'on pouvait assimiler le choix d'un mot au tirage d'une boule dans une urne dont la composition reste inchangée au cours du temps» (R. Moreau, *Intervention...*, 1966, p. 130). Anche M.me Hirscheberg e Ch. Muller hanno dimostrato che questa regolarità non si verifica: le frequenze delle parole, salvo casi eccezionali, non sono stabili, ma variano, in dipendenza dallo stile e dal tema,[12] da testo a testo e ciò sarebbe verificabile per-

7 P. Guiraud dice espressamente che la lingua potrebbe essere intesa come un ordine sistematico di engrammi, ciascuno presente con la sua probabilità, la quale costituisce un attributo oggettivo come la forma o il significato (P. Guiraud, *Problèmes et méthodes*, 1960, p. 16 e p. 25). Egli ha anche formulato le basi teoriche della possibilità di analisi della « langue » e della « parole » equiparate a « codice » e « messaggio », con i concetti e le tecniche della teoria dell'informazione. Molti aspetti del funzionamento della « langue » posti dal Saussure a fondamento della sua dottrina, quali « la différentialité, la segmentation, la linéarité, l'arbitraire du signe linguistique... sont... des caractères généraux et pré-linguistiques communs à toute catégorie de signes informationnels» (P. Guiraud, *Langage et communication*, 1954, p. 119).

L. Heilmann riassume così: «Se, dunque, oggi vogliamo tener presenti tutti i fattori che si riconoscono nel funzionamento della lingua, noi dobbiamo considerare quest'ultima come un sistema qualitativo e quantitativo di segni arbitrari doppiamente articolati (sul piano del significante e su quello del significato), caratterizzati da forma, contenuto e frequenza, attuantisi in un processo individuale fisiopsichico sottoposto al determinismo statistico» (L. Heilmann, *Considerazioni statistico-matematiche...*, 1963, p. 37).

8 D.W. Reed, in un articolo apparso in « Word » nel 1949, anticipa chiaramente il concetto di frequenza come carattere costitutivo di una forma linguistica e insiste sul rapporto tra la frequenza in testi particolari e la *probabilità* in lingua (D.W. Reed, *A statistical approach...*, 1949, pp. 235-236).

9 Herdan propone l'equiparazione tra l'antinomia saussuriana « langue-parole » da una parte e il rapporto « popolazione statistica-campione » dall'altra. La « langue » equivale alla « totality of engrams in the brains of members of the speech community together with their probabilities of occurence » e la « parole » « for random samples from it » (G. Herdan, *Type-Token...*, 1960, p. 34).

10 « Ca si alte ramuri ale lingvisticii moderne, lingvistica statistică - în calitate de disciplină constituită - se reclamă ca avînd bazele teoretice în conceptia despre limbă a lui F. de Saussure» (A. Roceric, *Fonostatistica...*, 1968, p. 18).

11 Alcuni autori, e non solo linguisti ma anche specialisti di statistica, hanno rilevato una eccessiva semplificazione nelle affermazioni di Herdan e del primo Guiraud e hanno proposto una maggiore cautela, consigliando di utilizzare e, in caso, di attendere il risultato degli spogli di testi che lo sviluppo della « linguistica computazionale » rende sempre più numerosi e attendibili.

12 Sono interessanti a questo proposito le considerazioni di Ch. Muller sui rapporti tra lessico *(lexique)* e vocabolario *(vocabulaire)*: egli riserva il primo di questi

fino « pour les *mots de relation*, qui sont les éléments les moins thématiques du lexique, car beaucoup d'entre eux, la plupart même, subissent visiblement l'effet des situations stylistiques » (Muller, *Initiation...*, 1968, p. 141).

Detto per inciso, i dati da noi ottenuti negli spogli per il lessico di frequenza dell'italiano confermano l'esistenza di variazioni, nella frequenza delle *parole di relazione*, tra i cinque sottoinsiemi esaminati, variazioni che appaiono stret-

termini alla « langue, » il secondo al « discours ». In questa terminologia, le unità che compongono il lessico sono « lessemi » *(lexèmes)*; quando queste unità virtuali sono attualizzate nel discorso, ciascuna di esse è un vocabolo *(vocable)*, cui corrisponde un certo numero di occorrenze nel testo. Si definisce dunque come vocabolario di un testo l'insieme di vocaboli presenti, ciascuno con una propria frequenza, nel testo in esame. Il vocabolario di un testo presuppone l'esistenza di un lessico, che trascende il testo, del quale il vocabolario non è che un campione. Si può concepire una serie di insiemi lessicali, ciascuno dei quali contiene il successivo: il lessico di un idioma, nel senso estensivo, entro certi limiti cronologici; il lessico di una lingua, considerato sincronicamente; il lessico di un gruppo umano limitato appartenente a una comunità linguistica; il lessico di un individuo di questo gruppo; il lessico dell'individuo in una situazione definita stilisticamente o tematicamente.

La nozione di lessico di situazione *(lexique de situation)* ingloba due tipi di elementi: i primi sono d'ordine stilistico legati all'interlocutore e all'effetto che il parlante vuol produrre; i secondi sono di ordine tematico, legati a ciò che il parlante vuol comunicare, al contenuto del messaggio. La situazione provoca, all'interno del lessico individuale, « un déplacement des probabilités d'emploi »; certe unità sono escluse dal discorso, e cadono a probabilità zero. Si può ammettere cioè che, al momento della redazione di un testo o della concezione di un enunciato, l'autore abbia un certo numero di unità del suo lessico « in gioco », mentre le altre sono escluse per motivi stilistici e tematici. Tra le unità lessicali di una lingua, o lessemi, e gli insiemi citati si potrebbero stabilire delle relazioni di inclusione o di esclusione. Se queste relazioni potessero essere stabilite con rigore, sarebbe possibile anche la quantificazione del lessico, mentre è possibile quantificare con rigore solo il vocabolario. La relazione di inclusione può essere stabilita di solito attraverso l'inventario del vocabolario; ma la relazione di esclusione può essere affermata solo per ragioni diacroniche, dunque in un piccolo numero di casi (Ch. Muller, *Initiation...*, 1968, pp. 136-140).

Si è potuto sperare, continua Muller, che la frequenza delle unità lessicali fosse stabile presso l'individuo, variabile nel gruppo: perciò si attribuivano alla statistica lessicale straordinarie possibilità di risolvere problemi filologici di attribuzione. Egli però teme che questa stabilità sia eccezionale. Ne deriva che, se gli elementi quantitativi semplici che compongono la struttura del vocabolario di un testo sono facilmente riconoscibili per mezzo dello spoglio (numero dei vocaboli del testo, frequenza di ciascuno, classi di frequenza e rapporto tra loro), non è invece possibile trasportare questa nozione di struttura nell'ambito del lessico, poiché la probabilità di un lessema nel lessico di un individuo può essere conosciuta o stimata solo se la frequenza è stabile (tenuto conto delle variazioni aleatorie di cui conosciamo la natura) in tutti i discorsi di questo individuo indipendentemente dalla situazione che è, invece, determinante sia sul piano stilistico sia su quello tematico. Quanto alla probabilità di un lessema nel lessico di un gruppo umano, si dovrebbe supporre che la sua frequenza sia costante presso tutti gli individui del gruppo, e Muller giudica questa possibilità ancora meno verisimile.

Ciononostante egli persiste ad ammettere una struttura quantitativa nel lessico, che si definisce con delle probabilità associate ai lessemi, ma non le riconosce « quelque réalité que dans un lexique de situation » (Muller, *Initiation...*, 1968, p. 141). Per il lessico di un individuo, o di un gruppo, bisognerebbe fare una media di tutte le situazioni di discorso nel quale l'individuo o gli individui del gruppo potrebbero trovarsi, ponderando ciascuna con la propria probabilità. Questo compito gli sembra meno utopistico nel caso di un gruppo, nel quale il grande numero dei membri diminuisce la difficoltà di ponderare la situazione.

2.2.3 La determinazione dei centri d'interesse avrà evidentemente qualcosa di arbitrario e la varietà di strutture e partizioni che abbiamo riscontrato nella campionatura dei Dizionari di frequenza lo conferma.[15] Gli studiosi del problema consigliano di aumentare quanto più è possibile il numero degli strati: « Stratifiez à outrance, » raccomanda Moreau (*Au sujet de...*, 1962, p. 157).

ha senso, dice Moreau: infatti il CREDIF ha preferito non indicarne la frequenza nel francese fondamentale.

15 Gli autori dei dizionari di frequenza hanno quasi sempre composto il corpus da spogliare con brani scelti da testi diversi, definiti da una differenza di origine (autore o soggetto parlante) o altro (genere letterario, ambiente sociale, situazione concreta, ecc.). Non solo le dimensioni complessive del campione scelto a rappresentare una determinata lingua sono, come abbiamo visto, notevolmente diverse, ma la scelta stessa dei testi risponde a criteri differenti. Molti sforzi sono volti a riprodurre nella composizione del corpus campione la complessità di strati e di stili della lingua nel suo insieme. Si hanno classificazioni e suddivisioni particolarmente complesse: ci si chiede quali siano i motivi di queste a volte incredibili diversità di composizione e di dimensioni del corpus assunto come campione. Per esempio Thorndike ha usato le fonti più svariate, dalla Bibbia e dalle opere drammatiche di Shakespeare, fino alla corrispondenza privata e ai libri di scuola, senza spiegare le ragioni della scelta (E.L. Thorndike, I. Lorge, *The teacher's word book...*, 1944, p. 224). Più spesso, invece, ci si preoccupa di definire con maggior precisione almeno i limiti cronologici: G.E. Vander Beke (*French Word Book*, 1929) ha spogliato 33 romanzi, 13 drammi, 14 tra giornali e riviste, 13 opere scientifiche e filosofiche, 16 opere storiche e critiche, che vanno da Balzac e Musset fino a Proust e Bergson; I. Rodriguez Bou (*Recuento ...*, 1952) ha esaminato più di 7 milioni di parole; esse si articolano in 3 aree principali, ciascuna suddivisa in gruppi minori:

1) *Vocabulario de expresión* (circa 3.400.000 parole di testo):
   a) vocabolario parlato;
   b) vocabolario di associazione libera e controllata;
   c) vocabolario di compiti di scuole medie superiori.
2) *Vocabulario de riconocimiento* (circa 3.011.000 parole):
   a) periodici;
   b) programmi radio;
   c) letteratura religiosa;
   d) i materiali dello spoglio di Buchanan.
3) *Vocabulario de autores* (circa 666.000 parole):
   a) libri di testo;
   b) letture supplementari.

Si è già detto della composizione del campione scelto dal CREDIF: conversazioni registrate in diversi ambienti sociali per il vocabolario delle frequenze; inchieste sui sostantivi e sui verbi disponibili per diversi centri di interesse; spoglio di giornali per il vocabolario letterario e di opere specializzate per il vocabolo scientifico.

M. A. Buchanan (*A Graded Spanish...*, 1927) divide il suo campione di circa 1.200.000 parole in 7 parti, come segue:

1) Drammi (7 unità)
2) Novelle (8 unità)
3) Poesie (3 unità)
4) Folklore (3 unità)
5) Prose varie (8 unità)
6) Prosa tecnica (7 unità)
7) Periodici (4 unità)

L'estesa spiegazione di V. García Hoz per la ripartizione delle sue fonti in 4 categorie (corrispondenza privata; periodici; documenti ufficiali: politici, religiosi, sindacali; libri vari), ci sembra un buon esempio per illustrare la soggettività che informa generalmente (e, come vedremo, non potrebbe per ora essere altrimenti) le scelte dei testi. Egli dice di voler tener conto delle manifestazioni fondamentali della vita umana, o, più precisamente, delle relazioni umane, dal momento che il linguaggio « es principalmente un medio de relación » (v. García Hoz, *Vocabulario...*, 1953, p. 18). Quattro aspetti della vita umana gli paiono chiaramente differenziabili, perché sono diversi in se stessi e perché possiedono un proprio mezzo di espressione (v. García Hoz, pp. 18-19):

a) *Vida familiar*: è rappresentata nel suo campione dalla corrispondenza privata: 620 lettere, raccolte in 15 diverse province spagnole tra gente di non elevata condizione.

b) *Vida social indiferenciada*, dell'uomo della strada, che vive le quotidiane preoccupazioni e conquiste sociali. La rap-

D'altra parte la linguistica quantitativa ha già ottenuto alcuni importanti risultati, soprattutto per quanto riguarda l'applicazione e l'adattamento dei metodi classici della statistica ai problemi della campionatura lessicale.

Riteniamo che gli studi tradizionali di stilistica e di lessicografia possano fornire i dati di partenza per una primissima stratificazione a priori della lingua, o meglio i dati per identificare, per lo meno come ipotesi di lavoro, alcuni sottoinsiemi sufficientemente definiti ed egualmente caratterizzati rispetto ad un parametro comune. Ai risultati degli spogli si applicheranno poi le tecniche statistiche da un lato per verificare l'omogeneità delle strutture quantitative dei sottogruppi, dall'altro per identificare le eventuali parole tematiche.

L'effettivo riconoscimento degli strati di lingua dovrà derivare da un lavoro accurato e documentato di induzione da spogli sempre più numerosi, naturalmente ammesso che questo stesso lavoro provi l'esistenza, sul piano delle strutture quantitative, di sottoinsiemi linguistici sicuramente definibili. Questo atteggiamento, che va sempre più diffondendosi, è, del resto, favorito dallo sviluppo della lessicografia automatica in molti paesi ove il numero degli spogli disponibili si accresce giorno per giorno grazie all'attività di centri specializzati.

In questa stessa direzione gli spogli e le elaborazioni per il nostro lessico di frequenza dell'italiano contemporaneo costituiscono un primo contributo[16] sistematico allo studio statistico della lingua italiana come universo lessicale, attorno al quale potrà organizzarsi l'elaborazione, per lo stesso fine, dei dati provenienti dagli ormai numerosissimi spogli che arricchiscono progressivamente la biblioteca elettronica del CNUCE (Centro Nazionale Universitario di Calcolo Elettronico) di Pisa.

presentano articoli di 25 periodici editi in diverse regioni. Gli articoli, scelti in modo da coprire l'intero arco dei mesi dell'anno, sono ripartiti così da rispecchiare quantitativamente la complessa composizione dei periodici. Delle 4000 parole prese per ogni giornale, 500 provengono da articoli dottrinali, 2000 dalla cronaca, 1000 dagli spettacoli, 500 dagli annunci.

c) *Vida social regulada,* dell'uomo in quanto membro della società organizzata. La prima società che sembra regolare la vita dell'uomo è lo stato. « Pero no es sólo el Estado la sociedad que protege o apresa al individuo; la vida de éste resulta incompleta y sin sentido quando no tiene trascendencia sobrenatural. En España, tal vez como en ningún otro país, nos encontramos al hombre desenvolviendo su vida religiosa en el marco de una nueva sociedad: la Iglesia. Por último, ha de subrayarse el hecho de que, en la actualidad, la vida profesional se desenvuelve regulada por los sindicatos: y no hay por qué considerar que los sindicatos, en su forma actual, tengan o no garantías de continuidad; sean de una o otra forma las organizaciones, el hecho sindical tiene alcance universal. De aquí el considerar que la vida social regulada tiene su expresión en el *Boletín Oficial del Estado,* en los *Boletines Eclesiásticos* y en las publicaciones sindicales. » Queste tre categorie di documenti sono ugualmente rappresentate: 33.333 parole per ciascuna.

d) *Vida cultural,* patrimonio culturale cui partecipa l'uomo comune. Il criterio per la scelta dei 22 libri non è stato il valore intrinseco, ma la diffusione, appurata dalle statistiche editoriali. Le parole di un libro sono prese non da un unico blocco di pagine consecutive, ma da pagine singole distribuite a intervalli regolari nel testo.

16 Per i motivi sopra esposti, altri progetti di dizionari di frequenza annunciati per l'italiano saranno indubbiamente utili perché basati su « corpora » diversi, ed anzi sembra più che mai doveroso assicurare la comparabilità dei dati, attraverso la compatibilità delle metodologie e delle modalità di registrazione.

3 IL CORPUS CAMPIONE DEL LIF

3.1 La lingua è, per la sua stessa natura, un mezzo di comunicazione prevalentemente orale: la lingua scritta è sempre venuta in un secondo tempo e, specialmente presso i popoli di antiche e nobili tradizioni culturali, ha mantenuto una maggiore fissità, venendo così a differenziarsi sempre di più dalla lingua parlata, che subiva più rapide trasformazioni.

Per ciò che si riferisce alla lingua italiana, per parecchi secoli della sua lunga storia, la lingua scritta, fortemente ancorata a modelli trecenteschi e cinquecenteschi e frenata nel suo evolversi da grammatici e da puristi, si è sempre maggiormente staccata dalla lingua parlata che, per naturale evoluzione, seguiva un continuo sviluppo. È però vero che, verso la metà dell'Ottocento, con la riforma manzoniana, la lingua letteraria italiana subì un notevole mutamento che, dal suo immobilismo, la ravvicinò alla lingua parlata dalle persone colte in Toscana e specialmente a Firenze.

Ma dal 1840 al 1970 sono passati centotrent'anni densissimi di avvenimenti importanti per l'evoluzione della vita e della cultura del nostro paese, talché molte innovazioni della riforma manzoniana, accettate o non accettate dai grammatici e dai puristi, sono ormai superate o considerate addirittura arcaiche, mentre molte altre si sono venute affermando. La prosa neorealistica che domina quasi incontrastata nella letteratura contemporanea italiana di quest'ultimo dopoguerra, offre, è vero, soluzioni stilistiche assai varie, ma quasi tutte o per la maggior parte con tendenze fortemente popolari. Le opere di scrittori come Moravia, Vittorini, Pavese e Pratolini contribuiscono a far ormai considerare regolari, o per lo meno accettabili nella lingua scritta, parecchie forme regionali, dialettali o comunque di carattere piuttosto volgare, prima raramente accettate o per lo meno non in tale proporzione, nella lingua scritta. I successi più notevoli, anche dal punto di vista editoriale, sono stati in questi ultimi anni, quelli di autori, come Cassola e Calvino, che tendono sempre più verso una « lingua comune ».

D'altra parte, la crescente diffusione dei grandi mezzi di comunicazione (stampa, radio, televisione) e del cinema e la diffusione ed estensione fino a 14 anni della scuola dell'obbligo hanno contribuito decisamente a colmare la distanza che separava la lingua scritta (che andava perdendo, a sua volta, sempre più il suo tradizionale carattere aulico) dalla lingua parlata, e d'altra parte riducevano progressivamente l'uso dei dialetti anche in quelle regioni in cui l'espressione dialettale mostrava ancora una grande vitalità. Così, attraverso un duplice movimento convergente, non solo aumenta ogni giorno il numero di coloro che intendono e parlano la lingua nazionale, ma gli stessi dialetti, anche se tuttora usati, si annacquano e si italianizzano (v. T. De Mauro, *Storia...*, 1970, 2ª ed.). Ci è sembrato quindi opportuno, per questo nostro lessico di frequenza dell'italiano contemporaneo, prenderé come base delle fonti scritte che si avvicinassero il più possibile alla realtà dell'italiano contemporaneo, non solo per la prevalenza del dialogo, ma anche per gli argomenti trattati.

3.2 In questi ultimi anni sono apparsi due importanti dizionari di frequenza (A. Juilland, E. Chang Rodriguez, *Frequency dictionary of spanish words*,

The Hague, 1964 e A. Juilland, P.M.H. Edwards, I. Juilland, *Frequency dictionary of rumanian words*, The Hague, 1965), primi di una collana, *The romance languages and their structures* diretta da A. Juilland, che si propone lo studio quantitativo del lessico, della grammatica e della fonematica delle principali lingue romanze (spagnolo, rumeno, francese, italiano, portoghese).

Questi due volumi, sebbene siano stati soggetti a molte critiche, cosa d'altra parte inevitabile, poiché sono due lavori metodologicamente nuovi, rappresentano quanto di meglio si è fatto finora in questo campo.

Basati su larghi spogli per un complesso di 500.000 parole circa per ogni lingua, suddivise in cinque diversi gruppi, sono dedicati in modo speciale alla lingua moderna e contemporanea. Questo numero è stato scelto anche da noi per il presente saggio, e ciò soprattutto allo scopo di avere più agevoli comparazioni non solo con quelli redatti per due lingue geneticamente affini all'italiano, come lo spagnolo e il rumeno, ma soprattutto con quello italiano che, come abbiamo visto, rientra nel programma di Juilland e che, data la diversità del corpus campione, non sarà certo una ripetizione del nostro lessico.

La principale differenza fra il nostro corpus e quelli dei dizionari già pubblicati dallo Juilland consiste probabilmente nella maggiore sincronicità e attualità dei testi da noi spogliati (1947-1968) e in una diversa suddivisione del campione: alla categoria *saggistica* di Juilland, che comprende saggi di storia e di critica letteraria, abbiamo sostituito la categoria *cinema*, cioè testi dialogati di film. È da notare, inoltre, che per *testi tecnici* abbiamo preso i *sussidiari* delle scuole elementari.

3.3 Per ciò che riguarda i limiti cronologici dei nostri testi, per cercare di rendere la nostra rappresentazione più aderente alla lingua contemporanea, come *terminus a quo* abbiamo preso il 1945, cioè la fine del secondo conflitto mondiale (in realtà il meno recente fra i testi spogliati è stato edito nel 1947, il più recente nel 1968): questa data, come del resto le altre fissate per i periodi storici o letterari, pur non rappresentando un confine netto, è significativa soprattutto perché rappresenta la chiusura di un periodo storico che ha avuto notevoli ripercussioni sulla lingua e un diverso orientamento nelle fonti d'informazione. Prima di tutto avviene un notevole rinnovamento del lessico con l'uscita dall'uso di molte voci legate a particolari istituzioni storico-politiche del passato regime, in secondo luogo viene a cessare o a diminuire moltissimo l'influsso di modelli francesi, lasciando il posto a modelli prevalentemente anglosassoni (inglesi e soprattutto americani).

Le 500.000 occorrenze del nostro campione sono state tratte in parti uguali (100.000) dai 5 gruppi, in cui abbiamo diviso il nostro corpus: *Teatro*, *Romanzi*, *Cinema*, *Periodici*, *Sussidiari*. I brani sono stati scelti dai seguenti testi:

Teatro

V. BOMPIANI, *Teresa Angelica*, in *Il teatro italiano del dopoguerra*, a cura di V. Pandolfi, Bologna, 1956

L. SQUARZINA, *L'esposizione universale*, in *Il teatro italiano del dopoguerra*, a cura di V. Pandolfi, Bologna, 1956

C. TERRON, *Avevo più stima dell'idrogeno*, in *Il teatro italiano del dopoguerra*, a cura di V. Pandolfi, Bologna, 1956

G. CASSIERI, *Il salto mortale*, in « Sipario », marzo, 1962

C. FRUTTERO, *Una donna uccise per deduzione*, in « Sipario », agosto-settembre, 1963
G. PATRONI GRIFFI, *D'amore si muore*, in *Teatro*, Milano, 1965
F. BRUSATI, *La pietà di Novembre*, in « Sipario », maggio, 1966
D. FO, *Gli arcangeli non giocano a flipper*, Torino, 1966
M.S. CODECASA, *La gara*, Rai-Tv[17]
M. FIOCCO e M. VERGOZ, *Una mattina d'estate*, Rai-Tv[17]

Romanzi

V. PRATOLINI, *Cronache di poveri amanti*, Firenze, 1947
C. CASSOLA, *La ragazza di Bube*, Torino, 1960
A. MORAVIA, *La noia*, Milano, 1960
E. VITTORINI, *Il Sempione strizza l'occhio al Fréjus*, Milano, 1962
D. BUZZATI, *Un amore*, Milano, 1963
E. ROVERSI, *Registrazione di eventi*, Milano, 1964
G. BASSANI, *Il giardino dei Finzi-Contini*, Torino, 1966
I. CALVINO, *Le Cosmicomiche*, Torino, 1966
C. PAVESE, *La Luna e i falò*, Torino, 1967
A. BEVILACQUA, *L'occhio del gatto*, Milano, 1968

Cinema

P. GERMI, *L'uomo di paglia*, Rocca San Casciano, 1958
R. ROSSELLINI, *Era notte a Roma*, Rocca San Casciano, 1960
L. VISCONTI, *Rocco e i suoi fratelli*, Rocca San Casciano, 1960
DE SICA - FELLINI - MONICELLI - VISCONTI, *Boccaccio 70*, Rocca San Casciano, 1962
M. ANTONIONI, *Deserto rosso*, Rocca San Casciano, 1964
F. FELLINI, $8^1/_2$, Rocca San Casciano, 1965
V. DE SETA, *Un uomo a metà*, Rocca San Casciano, 1966
M. BELLOCCHIO, *La Cina è vicina*, Rocca San Casciano, 1967

Periodici

« Quattrosoldi », agosto, 1967
« Corriere della Sera », 15 maggio 1968
« Corriere della Sera », 26 maggio 1968
« Oggi », 6 giugno 1968
« Corriere della Sera », 28 giugno 1968
« Corriere della Sera », 9 luglio 1968

Sussidiari

M.G. GABRIELI, *Vele*, sussidiario per la 3ª classe elementare, Bergamo, 1967
M.G. GABRIELI, *Vele*, sussidiario per la 4ª classe elementare, Bergamo, 1967
M. MALFATTI, C. NOTA, U. PETRINI, *Primato*, sussidiario per la 5ª classe elementare, Bergamo, 1967.

17 Dattiloscritti gentilmente messici a disposizione dalla direzione Rai-Tv di Firenze.

4.1 A conclusione dello spoglio lessicale del nostro campione, abbiamo ottenuto per ciascun lemma non solo la sua frequenza complessiva nel corpus, considerato come un unico blocco, ma anche le sue 5 frequenze parziali, una per ciascuno dei 5 sottoinsiemi descritti al capitolo precedente (teatro, romanzi, cinema, periodici, sussidiari).

In concreto, si è ottenuto un elenco nel quale ciascun lemma è accompagnato da 6 distinte frequenze (la frequenza complessiva [F] e le frequenze parziali f_1, f_2, f_3, f_4, f_5) e da un indice R che è il numero di sottoinsiemi nei quali il lemma appare. Riportiamo qui, a titolo di esempio, alcune righe di tale elenco, nel quale compaiono alcuni lemmi (15), tutti con $F = 15$, scelti tra i 117 che hanno tale frequenza nel nostro corpus in modo da dare un'idea delle varietà delle ripartizioni incontrate.

Lemma	F	f_1 (teatro)	f_2 (romanzi)	f_3 (cinema)	f_4 (periodici)	f_5 (sussidiari)	R
accertare	15	0	1	0	14	0	2
colloquio	15	0	5	0	10	0	2
deposito	15	0	5	0	6	4	3
divinità	15	0	0	0	0	15	1
finanziario	15	1	0	0	13	1	3
giacimento	15	0	0	0	2	13	2
gonfio	15	3	3	2	2	5	5
minerario	15	0	0	0	0	15	1
parcheggio	15	0	0	0	15	0	1
poliziotto	15	3	3	3	3	3	5
proprietario	15	1	2	1	10	1	5
sesso	15	5	3	1	6	0	4
uniforme	15	4	7	0	4	0	3
valido	15	0	1	1	12	1	4
vizio	15	3	4	3	5	0	4

L'esame dei dati ottenuti per i 15.750 lemmi che figurano nel nostro corpus ha mostrato chiaramente che le frequenze della maggior parte dei lemmi non sono ripartite uniformemente nei 5 sottoinsiemi.

Ad un estremo troviamo il caso di lemmi che occorrono in uno solo dei 5 sottoinsiemi e, all'estremo opposto, il caso di lemmi che ricorrono con eguale frequenza in tutti e 5 i sottoinsiemi del corpus.

Nel primo caso una delle frequenze parziali è uguale alla frequenza complessiva, mentre le altre 4 frequenze sono uguali a zero: ad esempio, per il lemma *minerario*, $f_5 = F = 15$, mentre $f_1 = f_2 = f_3 = f_4 = 0$.

Nel secondo caso, invece, le 5 frequenze parziali sono eguali tra loro e, dunque, uguali alla frequenza media del lemma; ad esempio per il lemma *poliziotto*, $f_1 = f_2 = f_3 = f_4 = f_5 = F/5 = 15/5 = 3$.

Tra questi due casi limite esiste, naturalmente, una grande varietà di modi di ripartizione, come hanno constatato, per lingue diverse, tutti gli autori di lessici di frequenza che hanno suddiviso in sottoinsiemi il corpus assunto come campione.

Da questa constatazione nasce « l'idée de compléter la notion de fréquence par celle de *stabilité de la fréquence*, ou, si l'on préfère, de corriger la fréquence par la façon dont celle-ci se distribue dans le corpus » (Ch. Muller, *Fréquence, Dispersion et Usage...*, 1965, p. 34).

Dopo che lo spoglio lessicale ha fornito l'elenco dei lemmi presenti nei testi del corpus assunto come campione, gli autori, normalmente, operano una selezione e includono nel lessico di frequenza solo una parte dei lemmi trovati. Nell'eseguire tale scelta, essi prendono in esame non solo la frequenza di ciascun lemma, ma anche il modo più o meno uniforme nel quale essa è ripartita tra i diversi sottoinsiemi del corpus.

Evidentemente si suppone che se la frequenza di un lemma è accumulata in uno solo dei sottoinsiemi, essa sia legata a situazioni particolari, quale il tema del sottoinsieme o di uno dei testi che lo compongono, le quali potrebbero non verificarsi in una diversa campionatura.

Al contrario, se la frequenza del lemma è stabile nei diversi sottoinsiemi o almeno in buona parte di essi si suppone che essa rimarrà stabile anche in altri corpus-campione, composti diversamente, della stessa lingua e pertanto si preferisce includere nel lessico di frequenza lemmi di questo tipo piuttosto che del tipo precedente.

Allo stesso modo, nell'ordinare i lemmi all'interno del lessico di frequenza, si dà maggior rilievo, e cioè la precedenza nell'ordinamento, ai lemmi del secondo tipo.

4.2 Per tener conto della ripartizione, alcuni hanno proposto di assumere semplicemente l'indice *R*, che rappresenta, come abbiamo visto, il numero dei sottoinsiemi nei quali il lemma appare:[18] questo indice (che è stato chia-

18 H.A. Kenniston *(A basic list...*, 1933) divide il suo vocabolario dello spagnolo in sei classi: le parole della lista I sono reperite almeno nell'80 % dei testi studiati, quelle della lista II in almeno il 67,7 %, ecc. Il principio sottinteso è che la distribuzione sia meno soggetta della frequenza alle variazioni della selezione casuale e che essa fornisca un indice migliore dell'uso dell'unità lessicale che non il numero delle occorrenze. Un altro esempio è quello di G.E. Vander Beke, *French word book*, 1929. Egli chiama *range* il numero *R* dei sottoinsiemi nei quali l'unità lessicale è rappresentata: poiché egli operava su 88 testi distinti, *R* era dunque un numero intero che variava da 1 a 88. Il suo vocabolario di base elimina le unità lessi-

mato di volta in volta con nomi diversi: rango, ripartizione, classe di dispersione, ecc.) può variare nel nostro corpus, evidentemente, da 1 a 5. Per restare agli esempi precedenti, *R* sarà 5 per il lemma *gonfio*, 4 per *valido*, 3 per *deposito*, 2 per *colloquio*, 1 per *parcheggio*.

In generale, se il corpus è stato suddiviso in *n* sottoinsiemi, *R* può variare da 1 a *n*.

Non è difficile osservare che lemmi la cui frequenza è ripartita in modo assai diverso possono avere lo stesso indice *R*: per esempio la frequenza complessiva 15 è ripartita in modo molto più uniforme per *poliziotto* (3, 3, 3, 3, 3), che per *proprietario* (1, 2, 1, 10, 1).

cali con *R* inferiore a 5, e ordina le restanti secondo il decrescere di *R*, indipendentemente dalla frequenza.

Il vocabolario russo di H.H. Josselson (*Russian...*, 1953), ad eccezione delle prime 204 parole scelte tenendo conto anche delle frequenze, ordina tutte le altre, estratte da un universo di 1.000.000 di occorrenze e diviso in 100 sottoinsiemi, esclusivamente in base a *R*, senza che la frequenza abbia alcun influsso sull'ordinamento.

Nel *Français fondamental du premier degré* del CREDIF, gli autori chiamano *répartition* l'indice *R* sopra definito (G. Gougenheim, R. Michéa, R. Rivenc, A. Sauvageot, *L'élaboration...*, 1964, p. 69). Lo spoglio dei testi ha fornito solo una parte del lessico accettato nel francese fondamentale. Qui ci interessa riassumere il procedimento con il quale sono stati utilizzati i dati provenienti dai 163 testi disponibili, lunghi, in media, duemila occorrenze ciascuno. La priorità è data alla frequenza, che fornisce la soglia di eliminazione (fissata a venti occorrenze) e determina l'ordinamento dei lemmi; la ripartizione interviene solo in un secondo tempo per eliminare i lemmi che, con una frequenza uguale o superiore a venti, appaiono in meno di cinque dei 163 testi spogliati e poi, per ordinare tra loro i lemmi che hanno uguale frequenza (per una valutazione del metodo si veda Ch. Muller, *Fréquence...*, 1965, p. 35).

V. García Hoz distingue il vocabolario risultante dai suoi spogli in tre zone, per così dire, concentriche:

a) *Vocabulario usual*

Opposto a « vocabulario total », comprende sia il « léxico activo » costituito dall'insieme delle parole che l'uomo comune impiega correntemente nella conversazione e nella scrittura spontanea, sia il « léxico latente», formato dalle parole che, pur non essendo usate in modo spontaneo, sono comprese senza difficoltà, quando si leggono e si odono. Questa dicotomia si trova già in J. Casares (*Nuevo concepto...*, 1941), al quale García Hoz si rifà esplicitamente. Egli si chiede come stabilire il limite preciso tra le parole usuali e le non usuali. Si potrebbe cercare di stabilire il numero di parole conosciute dall'uomo medio: ma le ricerche in questo senso hanno condotto a risultati contrastanti (v. García Hoz, *Vocabulario...*, 1953, pp. 23-24).

A.F. Watts (*The language...*, 1946, pp. 57-58) suggerisce una media di diecimila parole all'età di 14 anni, e, ammettendo un accrescimento di circa 400 parole all'anno colloca a 15.000 il vocabolario della maggioranza degli adulti. García Hoz accetta questa cifra e decide di accumulare spogli successivi fino al raggiungimento di un tal numero di lemmi diversi che egli accoglie in ordine alfabetico nel *vocabulario usual* (in realtà egli si ferma prima, dopo aver spogliato 400.000 parole e ottenuto circa 12.900 lemmi, supponendo che i rimanenti 2.000 possano appartenere alla porzione specialistica del lessico individuale). È da notare che il nostro spoglio ci ha fornito circa 15.700 lemmi, cifra curiosamente vicina a quella proposta da Watts. Questa coincidenza è probabilmente attribuibile solo alle dimensioni del nostro campione.

b) *Vocabulario común*

Una volta registrati tutti i lemmi con le frequenze rispettive in ciascun sottoinsieme, si separano dagli altri quelli che hanno almeno una occorrenza in ciascun sottoinsieme (complessivamente 1971). Si tratta, evidentemente, di un procedimento nel quale interviene solo *R*. Ma García Hoz si rende conto dello scarso potere discriminativo di questa formula. Egli confronta, come esempio, *proba-*

Questo limite della rappresentatività di *R* è tanto più evidente quanto minore è *n*; infatti, se il numero di sottoinsiemi è piccolo (per esempio 5 come nel nostro caso), *R* può dare delle indicazioni utilizzabili solo per frequenze molto basse, ma mette sullo stesso piano tutti i lemmi di frequenza elevata, qualunque sia la loro ripartizione, purché abbiano almeno un'occorrenza in ciascun sottoinsieme. Inoltre, *R* dà maggior rilievo a lemmi che appaiono in tutti i sottoinsiemi con frequenza molto bassa (per esempio: 1, 1, 1, 1, 1) rispetto a lemmi che appaiono con frequenza elevata in tutti i sottoinsiemi meno uno (per esempio: 50, 50, 50, 50, 50).

Per questo motivo, la maggior parte degli autori che ha adoperato l'indice

bilidad che ha 4 occorrenze complessive, ripartite una in ciascuno dei quattro sottoinsiemi, e *elemento* che, con 104 occorrenze complessive, figura in tre sottoinsiemi: manca solo nella corrispondenza privata. Egli sente l'esigenza di combinare assieme frequenza e ripartizione: « Esta última clase de palabras no puede figurar en el vocabulario común; mas ¿ puede decirse que sean de menos interés que algunas de los que figuran en él, teniendo menos frecuencias en total? » Egli aggiunge perciò al vocabolario comune una « relación adicional » (v. García Hoz, *Vocabulario...*, 1953, p. 386), cioè altri 212 lemmi che, pur non apparendo in tutti e quattro i sottoinsiemi superano la frequenza 40, da soli, o unendo insieme *(sic!)* lemmi diversi, ma strettamente affini tra loro. Il numero di 40 è stato fissato in relazione al calcolo di correlazione menzionato al punto *c*). Nell'insieme, al "vocabulario común" corrispondono 339.110 occorrenze.

c) *Vocabulario fundamental*

È costituito dai lemmi che hanno la frequenza distribuita praticamente in parti uguali nei sottoinsiemi. Per misurare la uniformità della distribuzione García Hoz ha applicato il calcolo delle correlazioni, in base al quale ha scelto 208 parole (che rappresentano 245.384 occorrenze). La frequenza media dei lemmi del vocabolario usuale è di 31, nel comune di 172, nel fondamentale di 1.324. Ma se si considerano i lemmi che appaiono rispettivamente: *a*) solo nel vocabolario usuale, *b*) solo nel vocabolario comune e *c*) solo nel vocabolario fondamentale, le frequenze medie rispettive sono: *a*) 4,2; *b*) 52; *c*) 1.324. La ricerca, egli conclude, ha reso evidente che il vocabolario usuale, per lo meno quello scritto, non è tutto omogeneo nei diversi sottoinsiemi (en las distintas manifestaciones de la vida humana), ma che vi è un piccolo numero di lemmi (circa 200) che è presente in tutti gli aspetti. Altre 2.000 parole sono presenti in tutte le manifestazioni, ma non in modo indifferenziato, bensì cariche di un valore specifico che le rende rappresentative di un aspetto sociale piuttosto che di un altro. Se ammettiamo che in una conversazione si pronunciano 160-170 parole al minuto, cioè 10.000 l'ora, le parole del vocabolario fondamentale avrebbero in media 33 ripetizioni ciascuna, quelle del vocabolario comune una o due, quelle dell'usuale non avrebbero praticamente probabilità di apparire (0,105): in altri termini le parole del vocabolario fondamentale apparirebbero una volta ogni due minuti, quelle del comune una volta l'ora, quelle dell'usuale una ogni 10 ore. Il metodo di García Hoz è discusso ampiamente da A. Juilland, *Spanish...*, 1964, p. XLVII e segg.

Anche Buchanan, *A graded spanish...*, 1927, propone una formula che combina assieme la frequenza e *R*: essa ha un certo potere discriminativo, specie in relazione alle parole di frequenza uguale distribuita in un numero diverso di sottoinsiemi:

$$U = \frac{F}{10} + R$$

dove *U* sta per valore di uso, coefficiente, a un tempo, per l'ordinamento e per la soglia di accettazione delle parole; *F* è la frequenza complessiva nel corpus; *R* è il numero dei sottoinsiemi nei quali appare. Evidentemente questa formula non ha potere discriminativo tra parole con la stessa frequenza non ripartita allo stesso modo in uno stesso numero di sottoinsiemi e dà un peso troppo grande alla ripartizione nei confronti della frequenza nel caso di parole di bassa frequenza e viceversa.

R ha suddiviso il corpus in un numero n di sottoinsiemi molto elevato, talora più di un centinaio, nel tentativo di rendere R più significativo.

4.3 Per queste e altre ragioni alcuni autori hanno cercato un indice più rappresentativo di R.

Tra i vari indici suggeriti abbiamo scelto e adottato nel LIF l'indice D (coefficiente di *Dispersione*) proposto da Juilland e Chang Rodriguez nel già citato *Frequency dictionary of spanish words* del 1964.

Seguendo da vicino l'esempio dei due autori, prima di descrivere tale indice descriveremo uno dopo l'altro tre altri possibili indici (che chiameremo d_1, d_2, e d_3), non perché essi siano stati di fatto adoperati in qualche precedente lessico di frequenza, ma perché riteniamo che l'esame delle loro caratteristiche possa rendere più agevole seguire, al § 4.5, la descrizione della formula per il calcolo dell'indice D a quanti non hanno familiarità con i metodi della statistica.

Dovremo usare in questo paragrafo espressioni come « uniformità della ripartizione della frequenza », « maggiore (o minore) uniformità della ripartizione della frequenza », « uniformità massima », « uniformità nulla », senza avere definito le nozioni corrispondenti, che assumiamo, invece, come immediatamente evidenti alla intuizione del lettore.

Se consideriamo lemmi con frequenza complessiva (F) uguale, una possibile valutazione dell'uniformità della ripartizione si può ottenere dalla differenza tra la sua frequenza parziale massima (f_{max}) e la sua frequenza parziale minima (f_{min}):

$$d_1 = f_{max} - f_{min}$$

Prendiamo in esame, a titolo di esempio, 4 lemmi (L_1, L_2, L_3, L_4) con identica frequenza complessiva ($F = 10$) ma ripartita diversamente nei 5 sottoinsiemi come segue:[19]

Lemma	F	f_1	f_2	f_3	f_4	f_5
L_1	10	2	2	2	2	2
L_2	10	3	2	2	2	1
L_3	10	4	3	1	1	1
L_4	10	10	0	0	0	0

19 Per comodità di esposizione e per rendere gli esempi più completi, semplici ed evidenti, preferiamo, anziché citare ripartizioni di lemmi che appaiono di fatto nel nostro campione, riferirci a possibili ripartizioni di ipotetici lemmi L_i.

Ciò non significa che gli esempi siano tutti fittizi e che non trovino riscontro nel nostro campione. Per esempio alle ripartizioni dei lemmi L_9, L_{10}, L_{11}, L_{13}, L_{14}, L_{15}, che adoperiamo più avanti, corrispondono rispettivamente, nel nostro

Le differenze tra la frequenza massima e la frequenza minima sono:

Lemma	f_{max}	f_{min}	d_1
L_1	2	2	0
L_2	3	1	2
L_3	4	1	3
L_4	10	0	10

Queste differenze indicano in modo inverso l'uniformità della ripartizione: minore è la differenza, maggiore è l'uniformità della ripartizione e viceversa. Però d_1 non può servire per confrontare l'uniformità di ripartizione di lemmi aventi frequenza complessiva diversa, come mostra l'esempio seguente nel quale la stessa differenza 100 si ottiene per entrambi i lemmi, mentre la frequenza di L_5 appare ripartita più uniformemente di quella di L_6.

Lemma	*F*	f_1	f_2	f_3	f_4	f_5	d_1
L_5	4600	900	900	900	900	1000	100
L_6	105	1	1	1	1	101	100

Un mezzo semplice di correzione è dividere la differenza in questione per la frequenza complessiva. Chiameremo questo indice d_2:

$$d_2 = \frac{d_1}{F}$$

Verifichiamo nell'esempio precedente:

Lemma	*F*	d_1	d_2
L_5	4600	100	100 : 4600 = 0,02
L_6	105	100	100 : 105 = 0,95

campione, le ripartizioni dei lemmi *rapidità, strazio, paesano, piazzale, capriola, alfabetico*. Come si vede, solo la ripartizione di L_{12} non è stata da noi riscontrata, ma la inseriamo egualmente tra gli esempi, appunto per rendere l'esemplificazione più completa ed efficace.

È da notare che questo indice d_2, come d_1, assume il valore zero quando la frequenza complessiva è ripartita in modo del tutto uniforme, cioè quando

$$f_1 = f_2 = f_3 = f_4 = f_5, \text{ per cui } (f_{max} - f_{min}) = 0$$

e dunque anche $d_2 = 0$.

A differenza di d_1, d_2 non può oltrepassare il valore 1, che raggiunge solo quando il lemma occorre in un unico sottoinsieme, caso limite nel quale possiamo dire che l'uniformità della ripartizione è nulla.

Infatti in tal caso $f_{max} = F$, $f_{min} = 0$,

e dunque
$$d_2 = \frac{F - 0}{F} = 1.$$

Anche d_2, come d_1, indica in modo inverso l'uniformità della ripartizione, ma è largamente più soddisfacente perché il suo valore è indipendente dalla frequenza complessiva e, variando tra 0 e 1, fornisce un metro comune di valutazione dell'uniformità della ripartizione che può essere applicato a lemmi con frequenze complessive diverse.

Però anche d_2 può perdere di efficacia, per esempio i due lemmi

Lemma	F	f_1	f_2	f_3	f_4	f_5	d_1	d_2
L_7	100	10	20	20	20	30	20	0.2
L_8	1000	100	150	200	250	300	200	0.2

hanno lo stesso valore di d_2, ma è manifesto che L_7 è ripartito diversamente da L_8: infatti le frequenze f_2, f_3 e f_4 di L_7, escluse cioè la minima e la massima, sono ripartite più uniformemente delle corrispondenti frequenze f_2, f_3 e f_4 di L_8.

Un indice capace di discriminare tra L_7 e L_8 deve dunque tener conto di tutte le n frequenze parziali; se con $\bar{f}$ indichiamo la frequenza media, cioè F/n, esso può essere:

$$d_3 = \frac{1}{F}(|f_1 - \bar{f}| + \ldots + |f_n - \bar{f}|)$$

cioè più brevemente

$$d_3 = \frac{1}{F} \sum_{i=1}^{n} |f_i - \bar{f}|$$

Infatti, affinché la ripartizione sia uniforme, occorre, come abbiamo visto, che ogni sottoinsieme abbia frequenze $f_i = \bar{f}$ quindi $|f_i - \bar{f}|$ dice quanto la frequenza dell'i^{mo} sottoinsieme si discosti dalla ripartizione uniforme; la som-

ma degli scarti delle singole frequenze dalla frequenza media, che appare in d_3, può servire perciò per indicare di quanto la ripartizione $f_1, f_2, \ldots f_n$ si discosti dalla ripartizione uniforme.

Il moltiplicatore $1/F$ serve per rendere d_3 indipendente dalla frequenza complessiva.

È facile verificare che d_3 riesce a discriminare L_7 e L_8.

Lemma	F	$\bar{f}$	$\lvert f_1-\bar{f}\rvert$	$\lvert f_2-\bar{f}\rvert$	$\lvert f_3-\bar{f}\rvert$	$\lvert f_4-\bar{f}\rvert$	$\lvert f_5-\bar{f}\rvert$	d_3
L_7	100	20	10	0	0	0	10	$\frac{20}{100} = 0.2$
L_8	1000	200	100	50	0	50	100	$\frac{300}{1000} = 0.3$

Ma anche d_3, a sua volta, non riesce sempre a discriminare, come mostra l'esempio seguente, nel quale, per maggiore evidenza e semplicità, facciamo apparire lemmi con uguale frequenza complessiva ($F = 5$) e dunque con uguale frequenza media ($\bar{f} = 1$).

Lemma	F	f_1	f_2	f_3	f_4	f_5	$\lvert f_1-\bar{f}\rvert$	$\lvert f_2-\bar{f}\rvert$	$\lvert f_3-\bar{f}\rvert$	$\lvert f_4-\bar{f}\rvert$	$\lvert f_5-\bar{f}\rvert$	d_3
L_9	5	1	1	1	1	1	0	0	0	0	0	0.0
L_{10}	5	2	1	1	1	0	1	0	0	0	1	0.4
L_{11}	5	2	2	1	0	0	1	1	0	1	1	0.8
L_{12}	5	3	1	1	0	0	2	0	0	1	1	0.8
L_{13}	5	3	2	0	0	0	2	1	1	1	1	1.2
L_{14}	5	4	1	0	0	0	3	0	1	1	1	1.2
L_{15}	5	5	0	0	0	0	4	1	1	1	1	1.6

Infatti d_3 ha lo stesso valore 0,8 per L_{11} e L_{12} e lo stesso valore 1,2 per L_{13} e L_{14} mentre è evidente che la ripartizione di L_{11} è diversa da quella di L_{12} e quella di L_{13} è diversa da quella di L_{14}.

4.4 L'indice proposto da Juilland e Chang Rodriguez nel loro *Frequency dictionary of spanish words* è capace di ovviare a questi inconvenienti. Tale indice, che i due autori chiamano, in inglese, *Dispersion*, e che viene indicato con D, è stato adottato in lavori successivi, a cominciare dal *Frequency dictionary of rumanian words* dello stesso Juilland, ed è oggi largamente adoperato,

anche se ha dato luogo a non poche discussioni, che ne hanno messo in evidenza pregi e difetti.[20]

Sappiamo che alcune modifiche alla formula di Juilland e di Chang Rodriguez sono state accolte in progetti di dizionari di frequenza ancora in corso, non tutti di prossima pubblicazione. Noi abbiamo preferito adottare senza modifiche la formula di Juilland e Chang Rodriguez perché, come abbiamo più volte ripetuto, ci sembra importante nello stato attuale delle ricerche di statistica linguistica assicurare la comparabilità tra i risultati di ricerche diverse.

Tuttavia, nel descrivere il procedimento per il calcolo di *D*, preferiamo adottare un sistema di simboli un po' diverso da quello degli autori; esso segue da vicino il sistema proposto da Ch. Muller nella recensione al *Frequency dictionary of spanish words* (Ch. Muller, *Fréquence, dispersion et usage*), che ci sembra più agevole e piano ai fini dell'esposizione.

4.5 Il procedimento richiede che si calcoli, innanzitutto:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f})^2$$

Si sommano qui non i valori assoluti delle differenze tra le frequenze parziali e la frequenza media, come per il calcolo di d_3, ma i quadrati di tali differenze. Questo accorgimento ovvia alla inefficienza di d_3, che abbiamo constatato per esempio nel confronto tra le ripartizioni di L_{11} e L_{12} e di L_{13} e L_{14}.

Non è difficile vedere che pur essendo i primi due scarti di L_{11} diversi dai primi due scarti di L_{12}, la somma dei valori assoluti dei primi due scarti di L_{11} è uguale alla somma dei valori assoluti dei primi due scarti di L_{12}. Invece la somma dei quadrati degli scarti di L_{11} è minore della somma corrispondente per L_{12}. Ciò è vero anche per la coppia L_{13} e L_{14}:

Lemma	$\lvert f_1 - \bar{f} \rvert$	$\lvert f_2 - \bar{f} \rvert$	$\lvert f_1 - \bar{f} \rvert$ + $\lvert f_2 - \bar{f} \rvert$	$(f_1 - \bar{f})^2$ + $(f_2 - \bar{f})^2$
L_{11}	1	1	2	2
L_{12}	2	0	2	4
L_{13}	2	1	3	5
L_{14}	3	0	3	9

20 Non li analizziamo qui, ma rinviamo alla già citata recensione di Ch. Muller, *Fréquence, dispersion et usage...*, 1965.

Per alcune modifiche proposte si vedano gli scritti di S. Allen (*Vocabulary data processing*, 1969) e di J.B. Carroll (*An alternative to Juilland's usage coefficient...*, 1970), e il Lessico di frequenza dello slovacco di J. Mistrík (*Frekvencia slov v slovenčine...*, 1969).

In generale, l'operazione di elevare al quadrato gli scarti prima di sommarli ha come conseguenza di aumentare, nella somma, l'incidenza di uno scarto proporzionalmente alla sua ampiezza.

Si calcola poi

$$S = \sqrt{S^2}$$

che, per ripartizioni uguali, cresce con la frequenza, e dunque con la frequenza media $\bar{f}$.

Definiamo coefficiente di variazione il rapporto

$$V = \frac{S}{\bar{f}}$$

che, a motivo del denominatore, è manifestamente indipendente dalla frequenza.

Questo coefficiente di variazione è nullo quando

$$f_1 = f_2 = f_3 = f_4 = f_5 = \bar{f};$$

infatti in tal caso sono nulli anche S^2 e S.

Il suo valore è invece massimo quando il lemma appare in uno solo dei sottoinsiemi; in questo caso si ha:[21]

$$V = \sqrt{n - 1}$$

21 Per dimostrare che V è massimo quando il lemma appare in uno solo degli n sottoinsiemi, dimostriamo innanzitutto che:

(1) $\sum_{i=1}^{n} (f_i - \bar{f})^2$ è massima se

$\sum_{i=1}^{n} f_i^2$ è massima.

Infatti

$$\sum_{i=1}^{n} (f_i - \bar{f})^2 = f_1^2 + f_2^2 + f_3^2 + \ldots$$

$$\ldots + f_n^2 + n\bar{f}^2 - 2f_1\bar{f} - 2f_2\bar{f} - 2f_3\bar{f} \ldots$$

$$\ldots - 2f_n\bar{f} = n\bar{f}^2 - 2\bar{f}\ (f_1 + f_2 + f_3 + \ldots$$

$$\ldots + f_n) + f_1^2 + f_2^2 + f_3^2 \ldots + f_n^2 =$$

$$= n\frac{F^2}{n^2} - 2\frac{F^2}{n} + f_1^2 + f_2^2 + f_3^2 + \ldots$$

$$+ f_n^2 = -\frac{F^2}{n} + \sum_{i=1}^{n} f_i^2$$

dove F e n restano costanti per qualsiasi diversa ripartizione di F.

Dimostriamo ora che il valore di

$$\sum_{i=1}^{n} f_i^2$$

è massimo quando le frequenze parziali di $(n - 1)$ sottoinsiemi sono eguali a 0.

Poiché

$$f_1 + f_2 + f_3 \ldots + f_n = F$$

Nel nostro corpus, suddiviso in 5 sottoinsiemi, il valore massimo di V è 2. Dividendo V, il coefficiente di variazione, per $\sqrt{n-1}$ si ottiene un indice che varia da 0, ripartizione perfettamente uniforme, a 1, quando il lemma appare in un solo sottoinsieme.

Il rapporto

$$\frac{V}{\sqrt{n-1}}$$

elevando al quadrato

$$f_1^2 + f_2^2 + f_3^2 \ldots + f_n^2 + 2f_1f_2 + \ldots$$

$$\ldots + 2f_1f_n \ldots + 2f_nf_{n-1} = F^2$$

cioè

$$f_1^2 + f_2^2 + f_3^2 \ldots + f_n^2 \leq F^2$$

(perché tutte le f_i sono positive o nulle)

ovvero ogni ripartizione di F in più di un sottoinsieme dà una somma di quadrati minori di F^2, mentre se F si trova in un solo sottoinsieme la frequenza al quadrato f_i^2 di questo sottoinsieme è uguale a F^2, perché tutti gli altri addendi sono nulli (si tratta, infatti, di doppi prodotti che contengono almeno un fattore = 0).

In altre parole

$$\sum_{i=1}^{n} f_i^2$$

è massima quando $(n-1)$ frequenze sono nulle e per la (1) in tal caso sarà massima anche

$$\sum_{i=1}^{n} (f_i - \bar{f})^2 .$$

Dimostriamo ora che

$$V = \sqrt{n-1}$$

quando la frequenza è accumulata in un unico sottoinsieme e supponiamo, a titolo di esempio, che

$$f_1 = F \quad \text{e} \quad f_2 = f_3 \ldots = f_n = 0$$

in tal caso

$$(f_1 - \bar{f})^2 = \left(F - \frac{F}{n}\right)^2$$

$$(f_2 - \bar{f})^2 = (f_3 - \bar{f})^2 \ldots = (f_n - \bar{f})^2 =$$

$$= \left(-\frac{F}{n}\right)^2$$

perciò

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f})^2 =$$

$$= \frac{1}{n}\left[\left(F - \frac{F}{n}\right)^2 + (n-1)\left(-\frac{F}{n}\right)^2\right] =$$

$$= \frac{1}{n}\left[F^2 - 2\frac{F^2}{n} + \frac{F^2}{n^2} + (n-1)\frac{F^2}{n^2}\right] =$$

$$= \frac{1}{n}\left[F^2 - 2\frac{F^2}{n} + \frac{F^2}{n}\right] =$$

$$= \frac{1}{n}\left(F^2 - \frac{F^2}{n}\right) = \frac{1}{n}\frac{n-1}{n}F^2 =$$

$$= \frac{F^2}{n^2}(n-1) .$$

Dunque

$$S = \frac{F}{n}\sqrt{n-1}$$

$$V = \sqrt{n-1} .$$

varia dunque tra 0 e 1, crescendo col diminuire dell'uniformità della ripartizione.

Se preferiamo un indice che, invece, cresca linearmente al crescere del grado di uniformità, basterà, per invertire il senso della variazione, introdurre il complemento all'unità.

Questa formula

$$D = 1 - \frac{V}{\sqrt{n-1}}$$

(che è da noi adottata e corrisponde a quella di A. Juilland ed E. Chang Rodriguez) qualifica la ripartizione della frequenza di ogni lemma del corpus per mezzo di un *coefficiente di dispersione D* che varia tra zero (frequenza accumulata in un solo sottoinsieme) e 1 (ripartizione perfettamente uniforme della frequenza del lemma negli n sottoinsiemi).

Verifichiamo per i lemmi L_9 e L_{15} del precedente esempio di 7 lemmi con $F = 5,\ n = 5,\ \bar{f} = 1$.

Per L_9

$$f_1 = f_2 = f_3 = f_4 = f_5 = \bar{f} = 1$$

e dunque

$$(f_1 - \bar{f})^2 = (f_2 - \bar{f})^2 = (f_3 - \bar{f})^2 = (f_4 - \bar{f})^2 = (f_5 - \bar{f})^2 = 0$$

e quindi

$$S^2 = \frac{0}{5} = 0$$

per cui

$$S = 0 \quad \text{e} \quad V = 0$$

e pertanto

$$D = 1 - \frac{V}{\sqrt{n-1}} = 1 - \frac{0}{2} = 1$$

Invece per L_{15}

$$f_1 = F = 5 \quad \text{e} \quad f_2 = f_3 = f_4 = f_5 = 0$$

e dunque

$$(f_1 - \bar{f})^2 = 16 \text{ e } (f_2 - \bar{f})^2 = (f_3 - \bar{f})^2 = (f_4 - \bar{f})^2 = (f_5 - \bar{f})^2 = 1$$

e quindi

$$S^2 = \frac{20}{5} = 4$$

per cui

$$S = 2 \quad \text{e} \quad V = \frac{S}{\bar{f}} = \frac{2}{1} = 2$$

cosicché il *coefficiente di dispersione* è

$$D = 1 - \frac{V}{\sqrt{n-1}} = 1 - \frac{2}{2} = 0$$

Se calcoliamo D per tutti e 7 i lemmi dell'esempio in questione, verifichiamo il suo potere di discriminazione, in particolare tra i lemmi L_{11} e L_{12}, e tra i lemmi L_{13} e L_{14}, che avevano d_3 uguale; infatti

per L_9 , $D = 1{,}000$
per L_{10}, $D = 0{,}685$
per L_{11}, $D = 0{,}577$
per L_{12}, $D = 0{,}450$
per L_{13}, $D = 0{,}368$
per L_{14}, $D = 0{,}225$
per L_{15}, $D = 0{,}000$

4.6 Abbiamo visto, al § 4.5, che nello scegliere i lemmi da includere nel lessico di frequenza e nell'ordinarli tra loro si pone il problema di tenere conto sia della ripartizione, che misuriamo con l'indice D, sia della frequenza F.

A. Juilland ed E. Chang Rodriguez esaminano dettagliatamente le diverse possibili formule per collegare F e D nelle sezioni 1. 2 3 1 (pagg. LXII - LXIV) e 1. 2 3 2 (pagg. LXIV - LXIX) del loro *Frequency dictionary of spanish words*, alle quali rinviamo per la discussione delle possibili alternative. I due autori concludono scegliendo la formula

$$U = F\,D\,.$$

U è l'indice (che chiamano *usage* e noi chiameremo *uso*) in base al quale i lemmi sono ordinati e accettati, oppure esclusi, dal loro vocabolario di frequenza.

Poiché la dispersione varia da 0 a 1, U è uguale a F quando la parola è ripartita uniformemente nel corpus, poiché in tal caso $D = 1$. Invece $U = 0$ quando le occorrenze sono concentrate tutte in un solo sottoinsieme, perché

in tal caso $D = 0$. Per lemmi di frequenza complessiva uguale, U è tanto più vicino a F quanto più la ripartizione è uniforme e decresce proporzionalmente al decrescere di D.

Per altri commenti rinviamo alla citata recensione di Ch. Muller. Egli avanza qualche critica: per esempio, nel caso particolare del campione a 5 sottoinsiemi, sono avvantaggiate le parole con frequenza 5 o multipla di 5. Così una parola con frequenza 10 può avere teoricamente un indice $U = 10$. Ma una parola di frequenza 9 non può raggiungere che $U = 8$ e per $F = 11$ non si può sorpassare $U = 10$. Questo difetto si riflette nell'ordinamento dei lemmi e potrebbe essere ridotto aumentando il numero dei sottoinsiemi.

La difficoltà di tale aumento risiede nel dover specializzare maggiormente il campione, cioè nell'immaginare un numero maggiore di strati o sottoinsiemi, e nel reperire i testi relativi. L'elaborazione e il calcolo non ne sarebbero però minimamente appesantiti o complicati, almeno se si usa il calcolatore; infatti, né la fase di lemmatizzazione né la fase di applicazione della formula comportano maggiori difficoltà di programmazione o di esecuzione.

4.7 Nell'elaborazione del LIF abbiamo preferito seguire per U la formula di Juilland e Chang Rodriguez, per garantire la confrontabilità tra i nostri dati e i loro. Ci siamo però discostati per quanto riguarda la scelta del limite inferiore di accettazione dei lemmi nel nostro lessico di frequenza. Essi, infatti, hanno scelto il materiale da pubblicare in un elenco di circa 9000 lemmi rimasti dopo aver eliminato, in un primo tempo, tutti i lemmi con $F < 4$ (circa 6000) e poi quelli con $R < 3$ (cioè presenti solo in 1 o 2 sottoinsiemi: circa 5000). Essi spiegano queste decisioni con motivi di convenienza pratica, per semplificare le operazioni di calcolo. Noi abbiamo invece calcolato D e U per tutti i lemmi prima di scegliere il limite inferiore. Dal punto di vista dell'elaborazione con il calcolatore, il loro procedimento ci è sembrato più complesso. Infatti il calcolo di U e D per tutti i lemmi richiede una piccola percentuale di tempo di elaborazione in più, offrendo in cambio il vantaggio di una scelta più oculata della soglia inferiore. In considerazione di questo abbiamo proceduto come segue:

a) abbiamo elencato i lemmi in ordine di F, D e U decrescenti, numerandoli progressivamente nei tre ordini;

b) sull'esempio di Juilland, abbiamo dapprima considerato la possibilità di accogliere nel LIF i primi 5000 lemmi in ordine di *Uso*. Il lemma 5000 ha $U = 2{,}00$; hanno questo stesso valore di U i lemmi dal 4945 al 5001;

c) a questo punto abbiamo osservato che, giungendo fino al lemma 5356, l'ultimo dei 260 che hanno $U = 1{,}78$, si includono tutti i lemmi con $R \geqslant 3$. Abbiamo pertanto fissato ad $U = 1{,}78$ la soglia inferiore entro la quale accogliere i lemmi nel LIF. (Si noti che 1,78 è il valore di U per $F = 3$ e $R = 3$.) Così facendo abbiamo aggiunto ai 5001 lemmi solo 95 lemmi con $R < 3$ e 260 lemmi con $R = 3$.
Poiché tra il lemma 1 e il lemma 5001 i lemmi con $R < 3$ sono 331, il LIF risulta costituito da 426 lemmi con $R < 3$ e da 4930 con $R \geqslant 3$.

5 L'ELABORAZIONE ELETTRONICA

5.1 L'uso dei calcolatori trova ormai così numerose e svariate applicazioni in diverse branche della linguistica e di scienze affini, che la linguistica computazionale [22] (d'ora in poi scriveremo LC) si prospetta secondo alcuni « non come una sottospecialità professionale, ma piuttosto come un insieme di tecniche utilizzabili da tutti i linguisti » (D.G. Hays, *Applied...*, 1969).[23]

La LC si distingue oggi chiaramente dalla linguistica applicata e dalla linguistica matematica, mentre in passato i tre termini venivano spesso usati indifferentemente.

È un dato di fatto innegabile che l'uso del calcolatore ha costituito e costituisce uno strumento di ricerca importante della linguistica matematica e una caratteristica essenziale della linguistica applicata. Ma è vero anche che sono venuti moltiplicandosi studi e ricerche appartenenti ai settori tradizionali della linguistica e delle discipline filologiche e letterarie in genere, nei quali l'impiego del calcolatore ha una funzione importante.[24]

22 La congerie di attività che sono oggi riunite sotto il nome di LC, negli Stati Uniti si è sviluppata attorno alla traduzione automatica come primo nucleo e spesso nel passato è stata identificata con questa. Se la traduzione automatica fosse tutta la LC, si potrebbe parlare di crisi di quest'ultima. La traduzione automatica è stata giudicata troppo costosa per essere utilizzabile, almeno nella situazione attuale della linguistica e dell'elaborazione automatica dei dati: per esempio nella riunione del 1966 dell'*Automatic language processing advisory committee*.

Tuttavia le relazioni tra LC, linguistica generale e traduzione automatica sono state al centro dell'interesse di diversi studiosi; si vedano per esempio le pagine di G. Lepschy (*La linguistica strutturale*, 1966, appendice).

23 Se l'affermazione di D.G. Hays, che del resto è stata in un certo senso anticipata da S. Lamb (*The digital computer...*, 1961), è in linea di principio esatta, in pratica la situazione della LC oggi è tale da richiedere, sia nel settore della ricerca sia in quello dell'insegnamento, la presenza di linguisti che possiedano una conoscenza approfondita e costantemente aggiornata di quell'insieme di tecniche in continuo sviluppo che costituiscono il patrimonio comune della LC.

24 Linguistica applicata (LA) è stato per lungo tempo, soprattutto negli USA, « un termine elegante per l'applicazione di idee e di scoperte della linguistica nell'insegnamento delle lingue » (D.G. Hays, *Applied...*, 1969). Anche in Italia è spesso usato come sinonimo di glottodidattica. In seguito ci si è accorti, osserva J.P. Vinay, che si delineavano due grandi tendenze, le quali si sono sviluppate parallelamente ma che si devono distinguere accuratamente: « Per alcuni la LA dovrebbe occuparsi esclusivamente dei problemi dell'insegnamento, per altri, invece, di applicare le teorie attuali alle macchine elettroniche che permettono il trattamento automatico dell'informazione, la traduzione automatica... » (J.P. Vinay, *Enseignement...*, 1968, pp. 699-700).

L'Associazione internazionale di linguistica applicata si è costituita in occasione del congresso di Nancy dell'ottobre 1964, organizzato appunto attorno a due nuclei di interesse:

1) l'automazione in linguistica,
2) la pedagogia delle lingue moderne.

L'impostazione del secondo congresso internazionale di LA, svoltosi nel settembre 1969 a Cambridge, ha accentuato la caratterizzazione plurivalente della LA e ha mostrato, anzi, che l'uso dell'elaboratore si va diffondendo anche nel settore della glottodidattica. Molte relazioni riguardavano l'utilizzazione, nell'insegnamento, delle tecniche della *computer assisted instruction* e si è affermata una nuova concezione del « laboratorio linguistico » nel quale il calcolatore ha un ruolo di coordinatore.

La caratteristica essenziale della LA è quella di applicare ritrovati e teorie della linguistica al perseguimento di scopi pratici diversi, esterni, per così dire, alla linguistica. Esistono, invece, numerosi progetti ed elaborazioni nei quali il calcolatore è usato come sussidio per ricerche « interne » alla linguistica; concordanze a livello fonematico e lessicale, archivi di dati

Questo orientamento è chiaramente illustrato in campo internazionale dalle comunicazioni presentate al secondo convegno di linguistica computazionale svoltosi a Stoccolma nel settembre 1969[25] e, sul piano nazionale, dal moltiplicarsi, negli ultimi anni, di progetti e ricerche a carattere linguistico presso il CNUCF di Pisa, ove si è costituita una sezione di LC[26] che ha messo a punto un insieme di procedure di programmi e di metodologie validi per alcuni dei principali settori della LC. Tra di essi abbiamo utilizzato per il Lessico di frequenza dell'italiano quelli approntati per gli spogli lessicografici di testi in lingua naturale e per le applicazioni di tecniche e formule statistiche ai dati quantitativi risultanti dagli spogli.

per lo studio e la descrizione sistematica di una lingua, ricerche nel settore della linguistica storica e comparata, elaborazioni di grandi « corpora » di dati dialettali, statistiche fonematiche e lessicali, verifica di modelli di grammatica, ecc. È chiaro quindi che l'area della LC è distinta da quella della LA, e la interseca solo nella misura in cui quest'ultima usa il calcolatore come strumento operativo, il che avviene soprattutto ad opera di quanti sono interessati al linguaggio non per sé ma alla sua funzione di veicolo primario dell'informazione nella società umana: in particolare gli « scienziati dell'informazione », i cultori di « information retrieval », di « content-analysis », di « traduzione automatica ».

Numerose trattazioni che si propongono una descrizione sintetica della linguistica matematica (LM) concordano nel riconoscervi alcuni settori nettamente distinti. Si parla per lo più di:

a) modelli quantitativi o statistici, *b*) modelli di teoria dell'informazione, *c*) modelli strutturali, *d*) traduzione automatica o, più in generale, linguistica matematica applicata.

Non sono mancate critiche a questo schema di classificazione (F. Kiefer, *Some aspects...,* 1964) anche se poi esso è pressoché costantemente seguito. Il denominatore comune alla base dell'accostamento dei diversi settori, che in un certo senso sembra giustificare la loro riunione sotto un'unica « etichetta » (cfr. G. Lepschy, *La linguistica strutturale,* 1966, p. 190), è l'introduzione nelle ricerche linguistiche di « metodi esatti » che, nelle scienze, sono inseparabilmente uniti alle matematiche (Akhmanova et alii, *Exact methods,* 1963, p. VII). Un elemento ritenuto comune e fondamentale è l'impiego del concetto di modello (cfr. J.J. Revzin, *Les modèles...* 1968) alla cui costruzione è indispensabile un elevato grado di formalizzazione; questa dovrebbe presentare numerosi vantaggi: assicurare la univocità terminologica, rendere possibili rappresentazioni esatte e più economiche, rendere esplicite le descrizioni.

Ma mentre nella LM la formalizzazione e l'adozione di tecniche matematiche sono richieste per conferire certe caratteristiche al modello (struttura assiomatica, esplicitazione nelle definizioni, verificabilità delle regole), nella LC la formalizzazione è richiesta per la eseguibilità meccanica delle elaborazioni, e non sempre e non necessariamente come caratteristica intrinseca della ricerca. A ciò si aggiunge che si può meccanizzare più di quanto si possa formalizzare. È vero che la natura algoritmica della programmazione stimola il ricercatore verso un grado di formalizzazione che, naturalmente, è prerogativa della LM. Ma spesso la formalizzazione dell'algoritmo non riguarda teorie linguistiche nel senso tradizionale della LM e il carattere linguistico è dato, più che dalla natura delle elaborazioni, dai materiali elaborati e dai risultati conseguiti. È però un dato di fatto che il calcolatore si è imposto come strumento di indiscussa utilità in tutti i settori della linguistica matematica: nel settore dei modelli quantitativi per la raccolta di dati esatti sulla base di una sufficiente quantità di spogli, possibile solo, di fatto, con il calcolatore; nel settore dei modelli strutturali per verificare i modelli con i metodi della simulazione e controllando l'esattezza formale della notazione e il corretto funzionamento delle regole relative, spesso assai complesse.

25 Si vedano a questo proposito gli Atti dell'*International conference on computational linguistics,* svoltasi a Stoccolma dal 31 agosto al 6 settembre 1969, attualmente in corso di pubblicazione. Per un rendiconto di questo convegno si veda A. Zampolli, *Cronaca...,* 1970.

26 In Italia il primo progetto di lessicografia automatizzata è stato quello dell'Index Thomisticus, iniziato da R. Busa S.J. a Gallarate nel 1949 (cfr. R. Busa,

5.2 Ci sembra ora opportuno precisare alcuni termini usati comunemente nella descrizione degli spogli lessicografici automatici: non ci proponiamo di darne delle definizioni scientifiche, ma solo di indicarne intuitivamente il contenuto così come si è venuto affermando nell'uso, al fine di agevolare e semplificare la descrizione delle diverse fasi delle nostre elaborazioni.

Chiamiamo *parole* (o *occorrenze*) le successive unità grafiche di cui è costituito un testo: per il programma l'unità grafica corrisponde a una o più lettere (o caratteri equivalenti) tra spazi o segni di interpunzione.[27]

L'insieme delle indicazioni che individuano la posizione della parola nel testo costituiscono il suo *riferimento*: per esempio i numeri di volume, pagina e riga, oppure i numeri di canto, strofa e verso.

In un testo sufficientemente lungo non tutte le parole sono diverse: alcune sono ripetute una o più volte. Chiamiamo *forme grafiche* le parole « diverse » presenti nel testo: l'elenco delle forme grafiche ci darà quindi tutte le parole diverse del testo esaminato: accanto a ciascuna forma grafica possiamo far scrivere il numero delle sue apparizioni nel testo, che chiamiamo *frequenza assoluta* della forma. Per esempio nella frase *chi non è con me è contro di me* le parole sono 9, ma le forme grafiche sono 7: *è* e *me* hanno frequenza 2, *chi*, *non*, *con*, *contro*, *di* hanno frequenza 1.

Come vedremo, a una stessa forma grafica possono corrispondere unità linguistiche diverse: per esempio la forma grafica *danno* può essere sia la prima persona singolare presente indicativo di *dannare* sia la terza persona plurale presente indicativo di *dare* sia il sostantivo sinonimo di *guasto, lesione*. Diciamo

A. Zampolli, *Centre...*, 1968). Nell'anno accademico 1959-1960 fu discussa all'Università di Padova la tesi di laurea di A. Zampolli, *Studi di statistica linguistica eseguiti con impianti IBM* (cfr. A. Zampolli, *Recherche...*, 1968), sotto la direzione di C. Tagliavini, che costituisce il primo saggio di applicazione dei calcolatori allo spoglio e all'analisi di un testo italiano a livello fonematico, lessicale e morfologico (cfr. C. Tagliavini, *Applicazione...*, 1968). Nel 1964 furono eseguiti sotto la direzione di A. Duro i primi esperimenti di spoglio per il grande *Vocabolario storico della lingua italiana dell'Accademia della Crusca* (cfr. A. Duro, A. Zampolli, *Analisi...*, 1968). Nel 1965, in occasione dell'inaugurazione del CNUCE a Pisa, venne offerta al capo dello stato una copia delle concordanze della *Divina Commedia,* elaborate elettronicamente per iniziativa della IBM Italia a cura di C. Tagliavini. Nel 1966 l'Accademia della Crusca affidò l'elaborazione elettronica dei progetti di spoglio al CNUCE; il loro esempio e il successo delle applicazioni suscitarono presto altri progetti, cosicché oggi oltre 50 Istituti italiani di Università e del CNR e alcuni Istituti stranieri si avvalgono degli impianti, della collaborazione tecnica e scientifica e dei programmi di utilità della Sezione Linguistica del CNUCE, costituita nel 1968. La direzione di questa sezione è affidata ad A. Zampolli, coordinatore scientifico delle attività linguistiche e letterarie della IBM Italia, e conta oggi 15 collaboratori più alcuni borsisti.

Il macchinario utilizzato è costituito dagli elaboratori elettronici IBM installati presso il CNUCE, i quali sono stati anche dotati di dispositivi costruiti appositamente per elaborazioni e ricerche linguistiche (cfr. A. Zampolli, *La section...*, 1969).

27 I *caratteri codificati* che costituiscono il testo registrato su schede si dividono, per il programma, in due grandi categorie funzionali:

a) *separatori*, cioè codici che separano, interrompono le parole (per esempio lo spazio, i segni d'interpunzione, ecc.);
b) *lettere*, cioè codici che compongono le parole (per esempio: lettere, segni diacritici, apostrofo, ecc.).

Una parola è definita come una serie ininterrotta di *lettere* tra due *separatori*.

che la forma grafica *danno* è *omografa,* e corrisponde a tre *forme lessicali* distinte.

Chiamiamo *lemma* quella voce di base che rappresenta, in un dizionario, le varie forme di un testo. Per esempio, alle forme *dite* e *diremo* corrisponde il lemma *dire*; alle forme *va* e *andarono* il lemma *andare*, alle forme *bello, belle, belli* il lemma *bello*.

Chiamiamo *lemmatizzazione* o (come preferirebbe dire G. Devoto) *lemmazione*, l'operazione di ricondurre ciascuna forma e le relative occorrenze al rispettivo lemma.

La *frequenza assoluta del lemma* è il numero delle sue occorrenze nel testo, e corrisponde alla somma delle frequenze delle sue forme.

Nell'elenco dei *lemmi* (o delle *forme*) *per frequenza decrescente*, i lemmi (o le forme) sono elencati dal più frequente al meno frequente: lemmi (o forme) di frequenza uguale sono ordinati alfabeticamente entro la stessa *classe* di frequenza.

Chiamiamo *contesto* di una parola un insieme variamente delimitato di parole che le sono contigue, o comunque associate, nel testo, scelte di solito fra quelle che la precedono e la seguono immediatamente.

I contesti possono essere raggruppati sotto le rispettive forme ordinate alfabeticamente e allora parliamo di *concordanze delle forme*; oppure possono essere elencati sotto i relativi lemmi ordinati alfabeticamente e allora parliamo di *concordanze dei lemmi.*

5.3 Nelle figure da pag. XXXVIII a pag. XLI è riprodotto un diagramma operativo (inglese: *flowchart*) delle operazioni compiute per produrre il LIF. Le prime due figure si riferiscono alle operazioni di spoglio; le altre alle elaborazioni statistiche per la scelta, l'ordinamento e la stampa dei lemmi e delle forme che costituiscono il LIF. Ogni singola operazione è contrassegnata con un numero progressivo.

Operazione 1

Per ottenere che il testo sia leggibile dal calcolatore è necessario ricopiarlo integralmente sulla tastiera di una macchina perforatrice, la quale traduce in fori su schede meccanografiche i caratteri del testo.[28]

Per i testi campioni del LIF abbiamo riportato su schede le parole, la pun-

28 Esistono altri mezzi per introdurre un testo nel calcolatore: lo si può ricopiare su un nastro di carta perforato, oppure direttamente su nastro magnetico, ecc. Ma soprattutto la crescente diffusione e disponibilità di macchine che permettono di comunicare direttamente e a distanza con il calcolatore offrono interessanti prospettive per la registrazione dei testi, soprattutto per i casi nei quali la trascrizione non può essere affidata a comuni operatrici, ma deve, invece, essere eseguita da specialisti, addirittura dallo stesso ricercatore: testi di difficile decifrazione, manoscritti, ecc. Questa possibilità è da tenere presente anche per quanto riguarda le edizioni critiche: lo studioso potrebbe, anziché dattiloscrivere, comporre il testo su un terminale. In seguito il testo potrebbe essere pubblicato senza altre ricopiature (per esempio in fotocomposizione), con il considerevole vantaggio, per il curatore, di poter verificare e controllare il testo critico servendosi di indici, di concordanze e di altri tipi di documentazione automatica possibili immediatamente a partire dall'unica trascrizione iniziale.

teggiatura, le diverse modalità grafiche (corsivo, maiuscolo, ecc.) e la divisione in pagine e righe.

Ciascuna di queste schede, che siamo soliti chiamare *schede-testo*, contiene in media una riga e mezzo di testo.

Operazione 2

Le schede-testo perforate vengono immesse in una macchina detta verificatrice, sulla cui tastiera un'operatrice, diversa da quella che ha eseguito la perforazione, ribatte una seconda volta il testo. La macchina verifica che i tasti battuti corrispondano alle perforazioni già esistenti nella scheda-testo. Se v'è una discordanza, la macchina si blocca e l'operatrice controlla se la discordanza è dovuta a un proprio errore o a un errore dell'operatrice che ha perforato le schede.

Operazione 3

Il calcolatore « legge » le schede, scompone il testo nelle diverse unità elementari di elaborazione (nel nostro caso, le parole definite come sequenze di lettere tra due spazi o segni d'interpunzione; in altre ricerche di LC, i grafemi, i fonemi, le sillabe, i sintagmi, ecc.) e lo registra parola per parola su un nastro magnetico, detto *nastro-parola*. Ogni parola costituisce un'*unità indipendente* di registrazione e riceve un numero progressivo che la individua univocamente.

Contemporaneamente il calcolatore stampa il testo perforato, riproducendo il più fedelmente ed esplicitamente possibile tutte le informazioni registrate (*lista-testo*).

Operazione 4

Le parole del nastro-testo vengono ricopiate in ordine alfabetico su un altro nastro.

Operazione 5

Dal nastro delle parole ordinate alfabeticamente si ottiene un nuovo nastro con l'elenco alfabetico delle forme grafiche e delle rispettive frequenze. Questo elenco viene anche stampato.

Operazione 6

Le forme vengono disposte secondo l'ordine decrescente delle rispettive frequenze.

Operazione 7

Si stampa l'elenco delle forme in ordine di frequenza decrescente.

Operazione 8

Facendo la media su grandi quantità di righe perforate (qualche milione) si è constatato che, dopo la prima perforazione, le schede-testo contengono circa il 4% di schede errate, le quali con l'operazione di *verifica* vengono ridotte di solito allo 0,4%. Per scoprire questi ultimi errori, si esaminano gli elenchi alfabetici delle forme[29] e si collaziona

29 Essi, tra l'altro, permettono di identificare rapidamente forme « impossibili » nella lingua considerata (che per lo più, essendo dovute ad errori di perforazione, hanno frequenza uguale a 1) e di rilevare incoerenze e discordanze nella grafia di forme particolari che presentano la possibilità di varianti grafiche libere.

il testo stampato dal calcolatore con il testo originale.[30]

Gli errori trovati vengono elencati in moduli appositi, nei quali si riportano il numero progressivo che individua le registrazioni errate e le eventuali modifiche da introdurre.

Operazione 9

Dal modulo, le correzioni vengono perforate su apposite *schede-correzione*.

Operazione 10

Il calcolatore ricopia il nastro-parola correggendovi gli errori sulla base delle schede-correzione e contemporaneamente stampa una nuova *lista-testo*.

Operazione 11

Il calcolatore ricopia le parole del testo corretto aggiungendo a ciascuna il relativo contesto.[31]

Operazione 12

Le parole, con i relativi contesti, vengono ordinate alfabeticamente.

Operazione 13

Il calcolatore stampa la lista delle concordanze per forma; contemporaneamente ricopia su un nastro forme e contesti numerati progressivamente e produce per ogni forma una scheda che contiene la forma e il numero progressivo corrispondente.

30 In media, per perforare un testo di 100.000 parole (pressappoco 300-400 pagine) occorrono circa 80 ore di lavoro di una perforatrice, e altrettante sono necessarie per verificare il testo perforato e correggerne gli errori. La lettura della lista-testo, che equivale approssimativamente a correggere delle bozze di stampa, richiede altre 50-60 ore. A confronto con questi tempi « lunghi » stanno le poche ore necessarie per le fasi automatiche dello spoglio: le diverse liste di frequenza e le concordanze per forma si ottengono, con un elaboratore anche di non grandi dimensioni, in non più di 6 ore, 3 delle quali occupate dalla semplice stampa dei risultati. È logico quindi che molti sforzi siano riservati a semplificare e ad alleggerire la fase di preparazione del testo da immettere nel calcolatore (un esempio di questo tipo di studi è rappresentato dai lavori di A.J. Szanser), ma è importante soprattutto registrare i testi con criteri scientifici e tecnici uniformi, così che possano essere utilizzati per elaborazioni e ricerche successive da ricercatori diversi. Questa esigenza appare nelle sue reali dimensioni se si considera che oggi, grazie all'attività dei numerosi centri di LC operanti in molte nazioni, diviene sempre più probabile che un ricercatore possa trovare già perforato da altri il testo che desidera studiare (cfr. M. Kay, *Standards...*, 1967). Per esempio presso il CNUCE di Pisa sono stati registrati su nastro magnetico quasi 50 milioni di parole in 15 lingue con uniformità di codici, di tracciati e soprattutto di criteri. Queste « biblioteche elettroniche », nelle quali le unità linguistiche sono classificate e registrate secondo criteri linguistici e tecnici uniformi, permettono rapidi conteggi interamente automatici in qualsiasi settore del corpus, con l'immediata raccolta dei dati quantitativi richiesti dalle elaborazioni statistiche: come abbiamo detto l'applicazione e la verifica delle teorie e delle formule della statistica linguistica a un gran numero di testi, nei quali le unità linguistiche siano uniformemente e rigorosamente definite, sono condizioni essenziali per lo sviluppo della statistica linguistica come scienza autonoma.

31 Il programma di elaborazione tiene conto di diversi fattori (per esempio: la punteggiatura, la fine o l'inizio di un capitolo, ecc.) nel delimitare il contesto. Per questo motivo la parola di cui si dà il contesto non è sempre al centro del contesto, ma la sua posizione è condizionata dalla presenza, alla sua destra o alla sua sinistra, di tali fattori.

Operazione 14

In assenza di un *dizionario di macchina*[32] italiano adeguato ai nostri scopi, la lemmatizzazione è stata eseguita da una équipe di 10 ricercatori, che hanno apposto opportune indicazioni nella lista delle concordanze per forma. Quest'operazione è descritta dettagliatamente più avanti a § 6 e segg.

Operazione 15

Un'operatrice aggiunge (perfora e verifica) le indicazioni dei lemmatizzatori alle schede-forma perforate dal calcolatore.

Operazione 16

Per mezzo delle schede lemmatizzate, il calcolatore produce un nuovo nastro nel quale ogni parola del testo è accompagnata dal rispettivo lemma e dal proprio contesto.

Operazione 17

Si ordinano i contesti per lemma, forma e riferimento.

Operazione 18

Si stampano le concordanze per lemma.

Operazione 19

Sulle concordanze lemmatizzate vengono eseguiti alcuni controlli per accertare l'esattezza della lemmatizzazione.

Operazione 20

Le parole vengono rimesse nella stessa sequenza nella quale appaiono nei testi.

Operazione 21

Da ciascuna opera vengono eliminati i nomi propri, personali o geografici, e le voci dialettali e straniere sulla base, naturalmente, di specifici contrassegni apposti dai lemmatizzatori; si tolgono anche alcuni brani, in modo da ridurre esattamente a 500.000 le 650.000 parole sottoposte a spoglio.[33]

Operazione 22

Le 500.000 parole vengono ordinate alfabeticamente per lemma e forma.

Operazione 23

Un programma di riepilogo riassume le 500.000 occorrenze in un nastro contenente lemmi (15.750) e forme (30.616). Per ogni lemma sono

32 Il dizionario di macchina è costituito da una serie di voci registrate nella memoria centrale o nelle memorie ausiliarie (nastri o dischi) del calcolatore. Ciascuna voce si compone di due parti: un termine di ingresso, che serve per la ricerca, e una serie di informazioni su questo termine. Nell'ipotesi di una lemmatizzazione automatica o semiautomatica (cfr. A. Duro, A. Zampolli, *Analisi...*, 1968) ogni parola, estratta dal suo contesto, deve essere « ricercata » automaticamente nel vocabolario; le informazioni in esso reperite devono essere associate alla parola quando viene reinserita nel suo contesto. Nelle applicazioni lessicografiche, le informazioni fornite dal dizionario di macchina sono costituite principalmente dal lemma e da una serie di codici che classificano morfologicamente la parola, oppure la qualificano per un particolare trattamento nelle elaborazioni successive (cfr. A. Zampolli, *Intervento...*, 1968).

33 Quando scegliemmo i testi campione da sottoporre a spoglio, fissammo per ognuno il numero di parole da estrarre.

registrati: il lemma, il codice di categoria grammaticale, la frequenza in ciascuno dei 5 sottoinsiemi e la frequenza complessiva nell'intero campione. Per ogni forma sono registrati: la forma, il suo lemma e la rispettiva categoria grammaticale e 6 frequenze diverse, una per ciascuno dei sottoinsiemi più la frequenza complessiva.

Operazione 24

Un programma statistico calcola e aggiunge a ogni lemma e a ogni forma la Dispersione *(D)* che, moltiplicata per la Frequenza complessiva *(F)* dà l'Uso *(U)*.

Operazione 25

Si stampano gli elenchi dei lemmi presenti in un solo sottoinsieme.

Operazione 26

Si stampano le frequenze, assolute e percentuali, ripartite per categorie grammaticali, di lemmi, forme e occorrenze.

Operazione 27

I lemmi vengono ordinati per *F* decrescente.

Operazione 28

I lemmi vengono stampati per *F* decrescente.

Operazione 29

I lemmi vengono ordinati per *D* decrescente.

Operazione 30

I lemmi vengono stampati per *D* decrescente.

Operazione 31

I lemmi vengono ordinati per *U* decrescente.

Operazione 32

I lemmi vengono stampati per *U* decrescente.

Operazione 33

Le forme vengono ordinate per *F* decrescente.

Operazione 34

Le forme vengono stampate per *F* decrescente.

Operazione 35

Le forme vengono ordinate per *D* decrescente.

Operazione 36

Le forme vengono stampate per *D* decrescente.

Non conveniva certo contarne a mano le parole da consegnare alla perforazione. Perciò calcolando nei diversi testi la media delle parole per riga, ci accertammo che i brani prescelti superassero largamente la cifra desiderata. A spoglio avvenuto, risultò che le parole erano complessivamente 650.000: perciò oltre ai nomi propri, personali e geografici, e alle voci straniere e dialettali, decidemmo di eliminare con un procedimento di selezione casuale anche alcuni brani, in modo da ridurre esattamente al numero prefissato la quantità delle parole prese da ogni singolo testo.

Operazione 37

Le forme vengono ordinate per *U* decrescente.

Operazione 38

Le forme vengono stampate per *U* decrescente.

Operazione 39

L'esame di tutti questi elenchi fornisce i criteri per scegliere il valore di *U* al di sopra del quale accogliere nel LIF i lemmi e le forme relative.

Operazione 40

Si separano e si ricopiano su un nastro tutti e solo i lemmi con $U \geqslant 1{,}78$, valore prescelto, e le forme relative, i quali costituiscono il LIF.

Operazione 41

Si ordinano alfabeticamente i lemmi e le forme.

Operazione 42

Si stampa l'elenco alfabetico dei lemmi e delle forme.

Operazione 43

Si ordinano i lemmi per *F* decrescente e si numerano per ordine di posizione.

Operazione 44

Si ordinano i lemmi per *D* decrescente e si numerano per ordine di posizione.

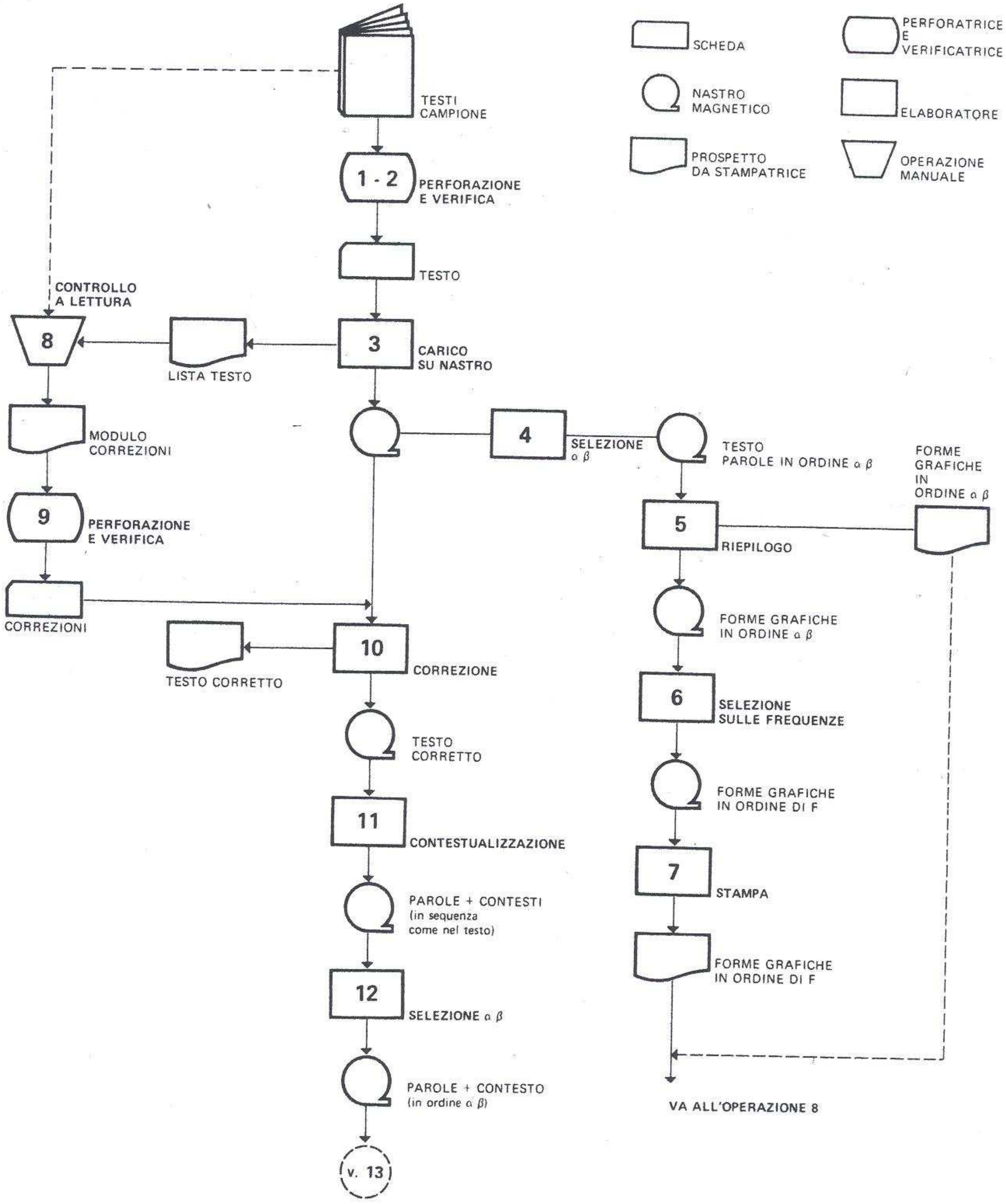
Operazione 45

Si ordinano i lemmi per *U* decrescente e si numerano per ordine di posizione.

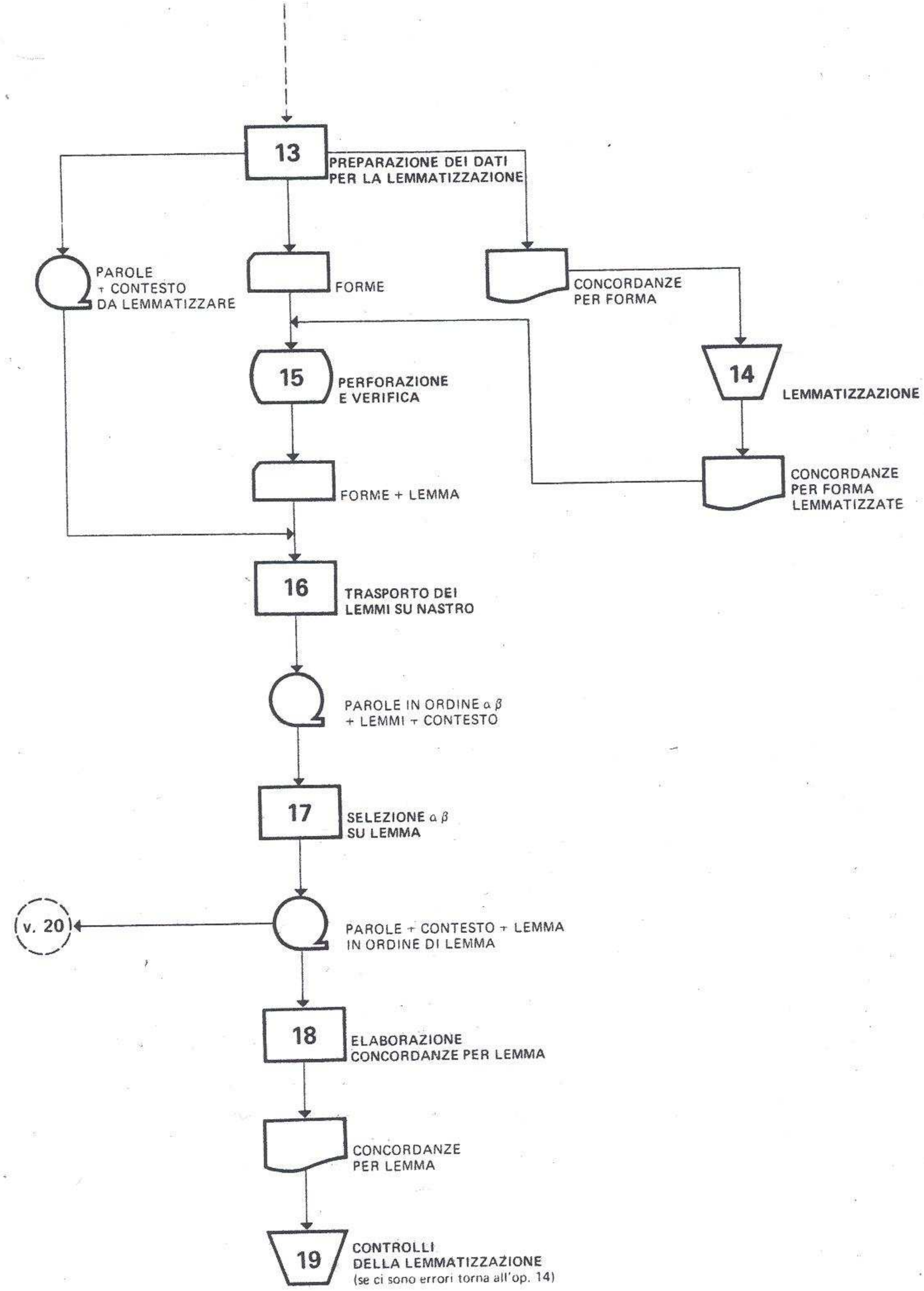
Operazione 46

Si leggono i lemmi nei vari ordinamenti stampando: dal nastro in ordine di *U*, il lemma, l'uso e il numero di posizione; dai rimanenti due nastri il numero di ordine.

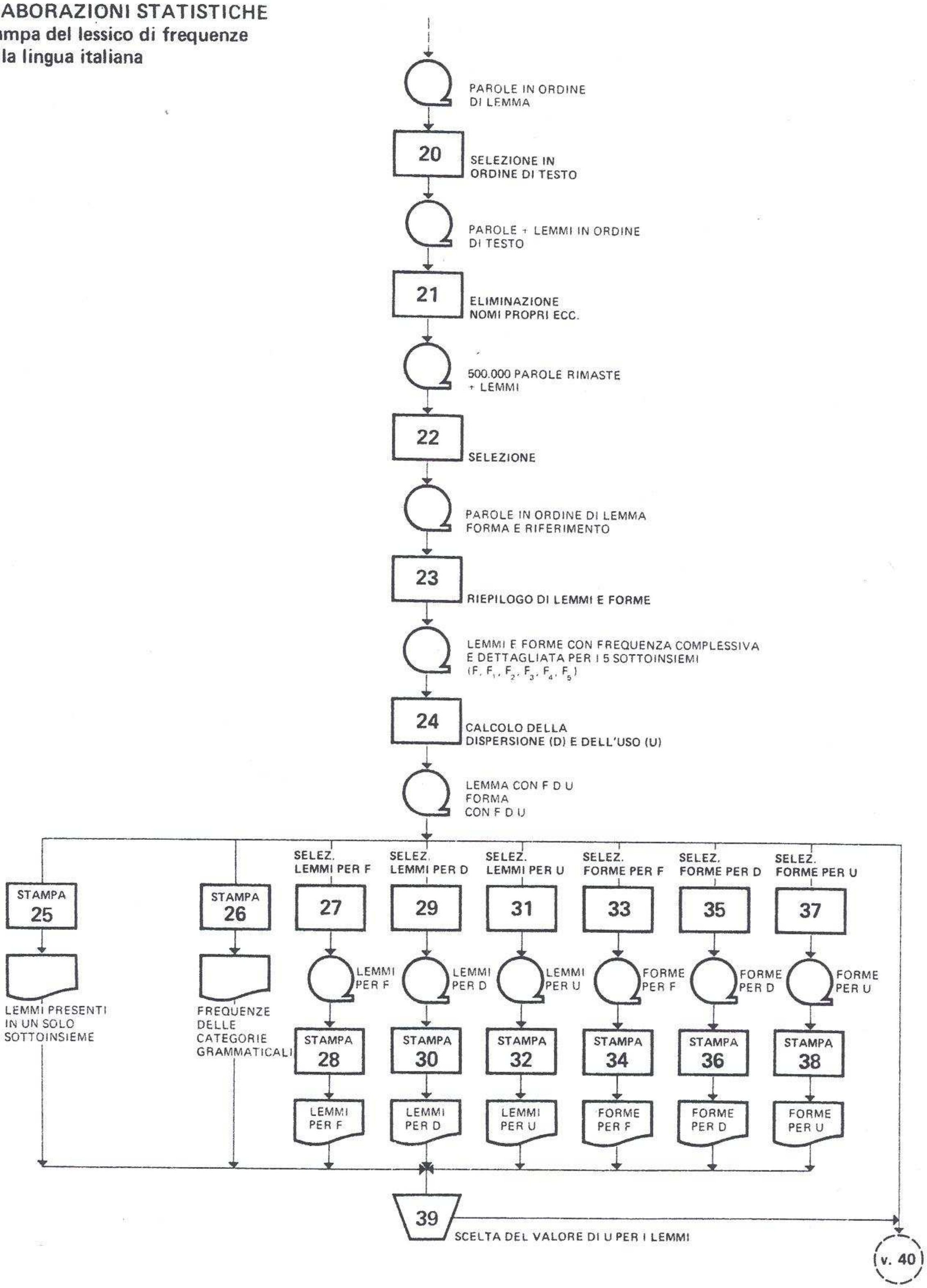
SPOGLIO
a) Fino alle Concordanze da lemmatizzare



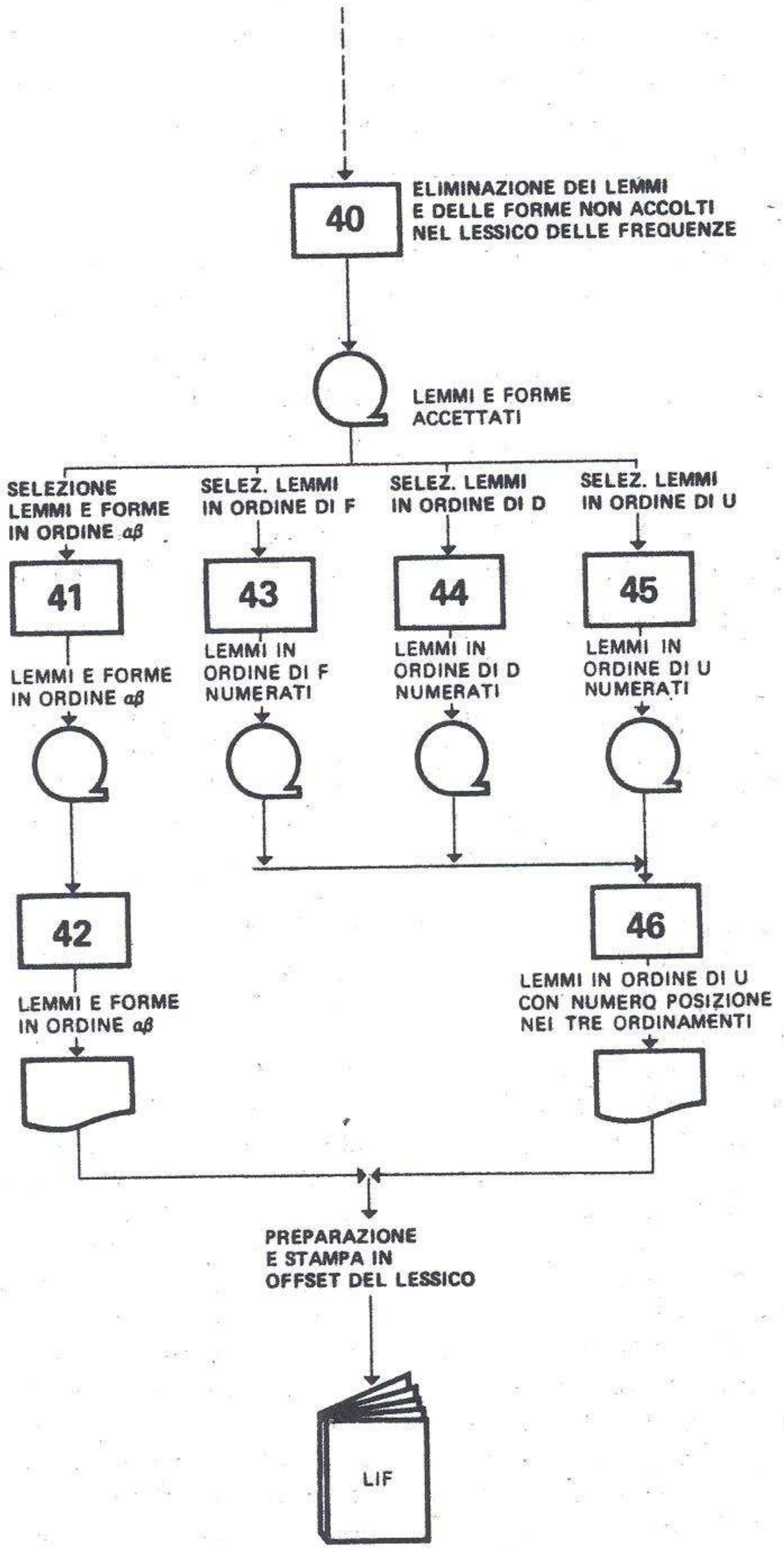
SPOGLIO
b) Lemmatizzazione



ELABORAZIONI STATISTICHE
Stampa del lessico di frequenze della lingua italiana



v. 40


40
ELIMINAZIONE DEI LEMMI E DELLE FORME NON ACCOLTI NEL LESSICO DELLE FREQUENZE
LEMMI E FORME ACCETTATI
SELEZIONE LEMMI E FORME IN ORDINE αβ
SELEZ. LEMMI IN ORDINE DI F
SELEZ. LEMMI IN ORDINE DI D
SELEZ. LEMMI IN ORDINE DI U
41
43
44
45
LEMMI E FORME IN ORDINE αβ
LEMMI IN ORDINE DI F NUMERATI
LEMMI IN ORDINE DI D NUMERATI
LEMMI IN ORDINE DI U NUMERATI
42
46
LEMMI E FORME IN ORDINE αβ
LEMMI IN ORDINE DI U CON NUMERO POSIZIONE NEI TRE ORDINAMENTI
PREPARAZIONE E STAMPA IN OFFSET DEL LESSICO
LIF

6 LEMMATIZZAZIONE: PROBLEMI E METODI

6.1 L'elaboratore ha effettuato lo spoglio dei testi prendendo come base le parole (= occorrenze, v. § 5.2) intese come unità grafiche e ci ha quindi fornito una lista di forme basata sull'identità grafica delle singole occorrenze.

Non sempre, però, l'unità grafica coincide con l'unità linguistica e anche in una lingua come l'italiano, dove la delimitazione della parola non è così problematica come in altre lingue (inglese, tedesco) e le tradizioni ortografica e tipografica ci danno una soluzione che, per lo più, è linguisticamente accettabile, vi sono alcuni casi in cui un'unità grafica rappresenta più unità lessicali come per esempio le preposizioni articolate, le forme pronominali congiunte ai verbi, le forme omografe (v. § 6.2).

D'altra parte più unità grafiche possono essere considerate una sola parola e per ragioni morfosintattiche come i tempi composti dei verbi e per ragioni lessicali come le locuzioni verbali, avverbiali, ecc., che potrebbero non essere formate occasionalmente dal parlante ma preesistere al «discours» nella «langue» (Ch. Muller, *Le mot...*, 1963); esempio: *far paura*, *aver sete*, *d'accordo*, *a malapena*, *a galla*.

Una corretta soluzione a problemi di questo genere presupporrebbe una teoria generale della parola o quanto meno una definizione soddisfacente che a tutt'oggi manca, nonostante i numerosi studi in proposito.[34]

34 La nozione di parola, pur passando generalmente per chiara, è in realtà una delle più discusse che s'incontrino nella linguistica e non esiste a tutt'oggi una definizione generalmente accettata e soddisfacente.

Alcuni linguisti escludono dalla linguistica il concetto di parola. Per esempio Ch. Bally (*Ling. gén. et ling. franç.*, 1963, pp. 338 e segg.) rifiuta il concetto di parola perché la sua posizione nel vocabolario, nella fonologia e nell'ortografia è ambigua e confusa. Invece di parola egli introduce il concetto di « semema » (es. *lup*-) e di « molecola sintattica » che include il semema e la « marca grammaticale » (es. *lupo, lupi*).

Altri linguisti non respingono il concetto di parola ma ne restringono l'uso. Essi concedono, per esempio, l'indipendenza di tale concetto nelle lingue indoeuropee, ma non in altre lingue (semitiche, bantu, amerindie), v. J. Krámský (*The word...*, 1969, p. 10). Secondo H. Seiler (*On defining...*, 1964) tutte le difficoltà della definizione di una parola hanno origine nel fatto che si vuol vedere nella parola un'unità nello stesso modo dei morfemi e dei fonemi, mentre la parola non è un'unità ma un costituente della frase o del periodo che è in relazione con la struttura a livello di frase, livello questo di relazioni e non di unità.

A. Sechehaye (*Essais sur...*, 1926) e A. Gardiner (*The theory of speech and language*, 1932) considerano la parola come unità di « langue » e la frase come unità di « parole ».

P.L. Garvin (*On linguistic method...*, 1964, p. 32) considera la parola come « properly defined distributional framework necessary for a precise determination of the distribution of the morphemes ». L'applicazione, però, di questa definizione al francese e all'inglese ha dato risultati negativi dovuti, secondo Garvin, all'assenza in queste lingue di catene di morfemi disposti uniformemente. Osservando poi che in lingue come il francese, il cèco e il polacco il limite di parola è individuabile in base all'accento, Garvin ha cercato di definire la parola basandosi su questo criterio. Comunque di nuovo senza successo, poiché restano escluse forme proclitiche ed enclitiche non accentate, ma tradizionalmente considerate parole separate. Ugualmente senza successo anche i tentativi di definire la parola in francese, inglese, vietnamita sulla base del criterio fonologico dell'accento o dell'ordine fisso dei morfemi o della combinazione di entrambi: Garvin conclude (op. cit., p. 33) « I have come to the conclusion that a proper definition of the word, that is, a consistent one without unaccountable residue, will emerge only

D'altra parte uno dei princìpi basilari della statistica è di operare su dati ben definiti: ed è ovvio, infatti, che da definizioni o da convenzioni diverse si otterranno risultati numericamente ben differenti. Noi, per esempio, abbiamo riunito (dando però sempre la possibilità di conteggi separati) sotto il lemma dell'articolo anche tutte le occorrenze di questo nelle preposizioni articolate e abbiamo ottenuto un totale di 54.752 occorrenze, di cui quasi la metà (22.657) sono le occorrenze dell'articolo in unione alle preposizioni. Abbiamo anche riunito sotto un unico esponente, come vedremo meglio in seguito, le forme del pronome personale soggetto e complemento e anche le forme enclitiche che abbiamo estratto dai verbi.

Così sotto l'esponente *io* su un totale di 7824 occorrenze, 2369 sono le occorrenze di *io*, 1150 le occorrenze enclitiche separate dai verbi, le altre sono le occorrenze del pronome personale di prima persona in funzione di complemento nelle forme toniche e atone. È chiaro che sia le preposizioni articolate sia tutte le forme dei pronomi personali avrebbero potuto essere considerate ciascuna un'unità di lessico a sé, dando luogo a valutazioni numeriche diverse. Naturalmente, anche secondo le distinzioni semantiche o funzionali operate sia nelle parole «forti» (v. nota 40) sia nelle parole grammaticali, si otterranno risultati numericamente differenti.

from a complete distributional reanalysis of the morphemic structure, using a properly defined temporary framework such as the utterance».

Altre definizioni si basano sulla distribuzione di unità significative e possono essere qualificate come morfologiche o grammaticali (per un panorama generale della questione cfr. A. Rosetti: *Le mot esquisse d'une théorie générale,* in *Linguistica,* The Hague, 1965, e K. Togeby: *Qu'est-ce qu'un mot?*, TCLC, v (1949), pp. 97-111). A queste appartiene la celebre definizione di parola di L. Bloomfield (*Language,* 1950, p. 178) come «the minimal free form». Questa definizione, sebbene concepita perché corrispondesse almeno approssimativamente al concetto ordinario di parola, presenta alcune conseguenze che sono in netto contrasto con la nozione tradizionale. Così, per esempio, in italiano le forme enclitiche e proclitiche dei pronomi personali non sarebbero delle parole indipendenti ma degli affissi uniti al verbo.

Recentemente J. Krámský (*The word...*, 1969) ha trattato ampiamente questo annoso problema. Punto di partenza della sua analisi è scoprire o, meglio, definire la posizione che la parola occupa nel sistema del linguaggio: a differenza del fonema, del morfema e della frase che hanno un posto ben determinato nel sistema linguistico e appartengono solo ed esclusivamente al piano rispettivamente fonetico, morfologico e sintattico, la parola appartiene a tre diversi livelli e cioè: morfologico, sintattico e lessicale e solo in quest'ultimo può essere considerata l'unità-base (poiché nel piano morfologico l'unità-base è il morfema, in quello sintattico la frase). Si dovrà quindi distinguere la funzione della parola nel piano morfologico e sintattico; nel primo, saranno le relazioni all'interno della parola tra i particolari morfemi che la compongono, nel secondo, le relazioni tra parole come realizzatrici della frase. Le parole come unità sintattiche sono in mutua relazione sintagmatica, mentre possono essere considerate morfologicamente solo in termini di relazioni paradigmatiche.

Solo questo duplice aspetto della parola, egli dice, potrà far luce sulla questione di quali forme di una stessa parola rappresentino differenti parole o solo varianti di una stessa parola. Krámský, dopo aver esaminato i criteri di indipendenza della parola che sono comunemente usati per la identificazione, conclude affermando che il concetto linguistico di parola come unità è un «universale» linguistico, ma che la sua posizione all'interno dei vari sistemi linguistici differisce nelle varie lingue. In alcune la parola è nel centro del sistema, in altre è più o meno distante da questo, in altre ancora è alla periferia; similmente l'indeterminatezza della parola come unità cresce dal centro alla periferia del sistema del linguaggio.

Ci si è quindi prospettata la necessità di stabilire una norma che sottraesse tali decisioni all'arbitrarietà del lemmatizzatore e all'ispirazione del momento e cercasse di ridurre al minimo i casi in cui è necessario ricorrere al contesto. L'analisi linguistica, però, non dà mai delle classificazioni nette, lascia sempre delle zone di indeterminatezza; nel continuum della lingua è, infatti, ben difficile poter tracciare dei limiti netti e dare delle classificazioni precise. Così se non vi saranno esitazioni nel distinguere *le* articolo da *le* pronome plurale femminile e da *le* pronome singolare dativo femminile, al contrario, chi volesse distinguere l'uso sostantivato degli aggettivi o l'uso aggettivale dei participi, troverebbe un gran numero di casi, come vedremo meglio in seguito, di fronte ai quali, anche in un'analisi approfondita, il linguista più esperto potrebbe restare in dubbio.

Tutti questi motivi ci hanno spinto verso la soluzione, del resto suggerita da precedenti esperienze, che è quella di seguire, o per lo meno di allontanarsene nel minor numero possibile di casi, la norma tradizionale dei dizionari, malgrado alcune fondate obiezioni di linguistica teorica. I lessicografi, infatti, pur conoscendo bene le incertezze della definizione di parola, non dovendo valutare quantitativamente gli elementi del lessico, hanno potuto lasciare molti problemi in sospeso, poiché il fare due articoli separati o uno solo o due paragrafi distinti sotto un unico lemma per una sola voce non è per loro un problema essenziale, poiché possono sempre fare la storia della parola o completarne le descrizioni, presupporre delle soluzioni diverse, esprimere dubbi, ecc.

Inoltre le loro decisioni si fondano piuttosto sull'evoluzione storica della lingua che su una descrizione sincronica che è invece l'unica possibile in un lavoro come il nostro. D'altra parte, però, abbiamo pensato che le lacune e le inesattezze della norma lessicografica sarebbero state compensate dalla sua comodità e relativa stabilità e che la stessa definizione di parola, che sembra di poter ricavare dalla pratica dei dizionari, pur confusa e istintiva, ha largamente contribuito a fissare i limiti dell'unità di parola nello spirito dei non specialisti.

Tra i molti dizionari della lingua italiana abbiamo dato la preferenza a quello del Migliorini[35] che, per l'autorità di linguista e di storico della lingua dell'Autore, ci è sembrato offrire le maggiori garanzie.

In questa fase, che abbiamo chiamato di lemmatizzazione, le operazioni da svolgere si possono ricondurre sostanzialmente a due:

1) riunire sotto uno stesso esponente tutte le occorrenze di uno stesso lemma;

2) separare le forme omografe risalenti a vocaboli diversi (tratteremo a parte i molti problemi postici dalla classificazione delle parole grammaticali).

La prima non comporta particolari difficoltà: si tratta di raggruppare sotto un unico esponente tutte le forme flesse e derivate con suffissi diminutivi, accrescitivi, vezzeggiativi, ecc.

35 B. Migliorini: *Vocabolario della lingua italiana*. Edizione rinnovata del vocabolario della lingua italiana di G. Cappuccini e B. Migliorini, Torino, 1965.

Verbi: l'unità semantica delle forme verbali flesse è evidente e non ha mai posto dubbi anche se alcune etimologicamente risalgono a lemmi diversi (*andiamo: andare; vanno: *vadere*). La sola incertezza riguarda le forme nominali che possono avere funzioni grammaticali diverse, ma di queste ci occuperemo in seguito. L'esponente del lemma è l'infinito attivo, la forma riflessiva dell'infinito compare come esponente solo quando *tutte* le occorrenze nei testi da noi spogliati sono riflessive.

Aggettivi: abbiamo raccolto sotto l'unico lemma del maschile singolare le forme flesse dell'aggettivo sia per gli aggettivi «biformi» cioè con terminazione diversa per il maschile e il femminile (per esempio: *bello*, *bella*, *belli*, *belle*), sia per gli aggettivi con terminazione unica per ambo i generi (*felice*, *felici*); per questi non abbiamo ritenuto opportuno dare anche le occorrenze effettive per genere. Anche per le forme alterate e derivate dell'aggettivo, cioè il superlativo e le forme derivate con suffissi diminutivi, peggiorativi, ecc. ci siamo attenuti alla prassi dei dizionari che le raggruppano sotto il lemma del positivo salvo i casi eccezionali del tipo: *migliore*, *ottimo*, *peggiore*, *pessimo*.

Sostantivi: come non abbiamo esitato a classificare insieme le diverse forme dell'aggettivo, così possiamo riunire plurale e singolare dei sostantivi. Ma non possiamo fare altrettanto per ciò che riguarda il genere, poiché mentre le quattro forme dell'aggettivo hanno fra di loro una precisa relazione semantica e morfologica come le due forme (singolare e plurale) dei sostantivi, due sostantivi opposti per il loro genere possono avere una genesi e un'evoluzione separata sia dal punto di vista semantico sia da quello formale.

Per la categoria dell'animato, dove si classificano nomi di esseri umani e un buon numero di nomi di animali, il genere è significativo e vi è un rapporto fra i due nomi che designano l'essere maschile e quello femminile corrispondente.

Dal punto di vista morfologico, tuttavia, questa corrispondenza può esistere in diversi gradi che vanno dall'identità completa in forme come *il (la) cantante* alla più completa diversità in forme anche di radice diversa come: *uomo ~ donna*, *marito ~ moglie*, *fratello ~ sorella*, ecc. tramite gradi intermedi come *amico ~ amica*, *pittore ~ pittrice*, *poeta ~ poetessa*, ecc.

D'altra parte anche la corrispondenza semantica tra i due elementi della coppia è variabile e si può facilmente constatare che il genere dei sostantivi non è una flessione nel senso proprio della parola, ma una corrispondenza più o meno forte per forma e senso, fra due sostantivi designanti esseri di sesso diverso.

Sarebbe quindi imprudente riunire in una sola unità di lessico le coppie di sostantivi di questa categoria; abbiamo perciò riunito solo quelle coppie la cui corrispondenza formale è completa, come ad esempio: *il (la) collega*, *il (la) cantante*.

Per la categoria dell'inanimato il genere non è più distintivo di sessi diversi e le coppie sono nettamente separate da accezioni semantiche diverse, come ad esempio: *il fine*, *la fine; il fronte*, *la fronte; il pianeta*, *la pianeta; il cassetto*, *la cassetta; il melo*, *la mela;* ecc. Soltanto in pochi casi vi è non solo una perfetta corrispondenza formale e semantica, ma anche un uso indiscriminato

di entrambi indipendentemente dal genere, e solo questi ci è sembrato perciò opportuno riunire sotto un unico lemma, il più frequente nel nostro corpus, come per esempio: *cioccolato ~ cioccolata.*

Queste soluzioni abbastanza semplici e accettabili pongono una sola condizione: che la parola si classifichi senza incertezze come verbo, come sostantivo, come aggettivo; i casi dubbi e cioè le forme nominali del verbo, gli aggettivi sostantivati e i sostantivi aggettivati, come anche le parole grammaticali e la flessione dei pronomi, pongono dei problemi particolari che affronteremo in seguito nella distinzione degli omografi.

Anche la classificazione delle forme derivate del sostantivo pone non pochi problemi, poiché se alcuni diminutivi o accrescitivi hanno assunto accezioni semantiche nuove che li separano nettamente dal sostantivo da cui derivano, come per esempio *cappa, cappella,* per molti altri il limite non è né così netto né facilmente definibile; ogni caso, inoltre, presupporrebbe un'analisi approfondita della coscienza linguistica del parlante. Abbiamo perciò preferito attenerci alle divisioni date dal vocabolario della lingua italiana del Migliorini, anche se non sempre ci sono sembrate tutte ugualmente soddisfacenti e adeguate.

Locuzioni: Abbiamo preferito non considerare lemmi indipendenti le locuzioni significative e le connettive,[36] riservando la possibilità di raggruppamenti a uno stadio ulteriore della ricerca; abbiamo riunito solo alcune locuzioni che presentavano nei nostri testi la grafia unita e separata come: *fin che (finché), dopo tutto (dopotutto), dopo che (dopoché)*, ecc. e quelle come: *a galla, a malapena,* dove l'elemento semanticamente più importante non esiste più indipendentemente o quanto meno non compare in tale forma nei nostri testi-campione.

Varianti: Se la parola ha più varianti è necessario sceglierne una come lemma; c'è, infatti, chi pronuncia e scrive *fisionomia* e chi *fisonomia,* chi preferisce *obbiettivo* e chi *obiettivo*, chi *pronunzia* e chi *pronuncia,* chi *diritto* e chi *dritto* aggettivo (il sostantivo sempre *diritto*), ecc.

In questi casi, in cui i due usi pressappoco si bilanciano, il lessicografo per dare la preferenza all'una o all'altra delle forme deve esercitare un certo arbitrio poiché è difficile stabilire una qualche distinzione (v. Migliorini, *Che cos'è un vocabolario?,* 1951, pp. 24-25). In altri casi, invece, tra le varie forme tende a stabilirsi una certa differenza semantica, come fra *coltura e cultura, focolaio* e *focolare,* oppure di uso, così le forme con dittongo *uo* dopo palatale, come per esempio: *spagnolo ~ spagnuolo, fagiolo ~ fagiuolo, figliolo ~ figliuolo,* che perdono sempre più terreno anche nella lingua scritta e cominciano a sembrare pedantesche (le abbiamo trovate, infatti, solo nei sussidiari).

Noi abbiamo preferito assegnare come lemma la forma più frequente nel nostro corpus anche se in alcuni casi il nostro lemma non coincide con quello dato dal vocabolario del Migliorini o dal DOP come il più comune o il più corretto.

36 Prendiamo queste denominazioni da J. Casares, *Introducción a la lexicografía moderna,* Madrid, 1950, pp. 167-184, cui rimandiamo per uno studio approfondito.

6.2 *Separazione degli omografi*

La separazione delle forme omografe risalenti a lemmi diversi,[37] che lingua italiana possiede in gran numero, pone senz'altro i maggiori probler al lemmatizzatore, poiché, oltre a difficoltà puramente pratiche, esige del decisioni di principio che spesso non sono né semplici né facili. Possiamo i nanzitutto decidere che due forme omografe rappresentano lemmi distinti s quando vi riconosciamo un significato diverso sia quando, pur avendo lo stes: significato, le due forme esplicano funzioni morfosintattiche diverse.

Il primo è il caso dell'omonimia: ricordiamo che generalmente due seg linguistici si dicono omonimi quando hanno significati identici e significa diversi, come per esempio: *bugia* (menzogna), *bugia* (candeliere); *collo* (par del corpo), *collo* (bagaglio). Per quanto si possa ammettere con E. Buysse (*Les langages et le discours*, 1943, § 60-61) che l'omonimia « est un défaut perspective qui ne se produit que lorsqu'on isole artificiellement le signe discours», tuttavia, nella delimitazione e classificazione delle unità sincronich è uno degli aspetti più importanti e difficili. La causa più comune dell'omor mia è l'evoluzione fonetica convergente: per influsso di regolari cambiamer fonetici, due o più parole che avevano un tempo forme diverse, vengono coincidere nella lingua parlata e in quella scritta, per esempio: *fiera* (bestia *fiera* (mercato) (omofoni e omografi), e talvolta solo in quella scritta (om grafi ma non omofoni), esempio: *pésca, pèsca; àncora, ancóra; balìa, bàli*

Questa forma di omonimia è comunissima nelle lingue con molti termi monosillabici, e perciò particolarmente frequente in inglese e in frances mentre è meno diffusa, per esempio, in italiano e in tedesco (v. O. Jesperse *Monosyllabism in English*, nel volume *Linguistica*, 1933, pp. 384-408).[38]

L'omonimia può determinarsi anche in seguito ad uno sviluppo divergen del senso: infatti quando due o più significati della stessa parola si distanziar a tal punto che non esiste più nella coscienza del parlante il senso dell'uni origine, la polisemia lascia il posto all'omonimia e l'unità della parola distrutta.

Nella lingua comune vi sono molti omonimi secondari di questo tipo e so un esperto in etimologia sarebbe in grado di connettere *penna* per scrive con *penna* di uccello, *collo* bagaglio con *collo* parte del corpo.

Non ci soffermeremo qui su casi di omografi ma non omofoni (con *pésca, pèsca; bàlia, balìa; tócco, tòcco*) poiché l'aggiunta di un accento grafi risolve facilmente tutti questi casi in cui l'omografia è causata solo da u deficienza del nostro sistema grafico, né sugli omonimi parziali, cioè appart

37 Non abbiamo ritenuto necessario distinguere forme omografe risalenti allo stesso lemma ed esercitanti funzioni morfologiche diverse, molto frequenti in italiano, soprattutto nella flessione verbale, come per esempio: *che io, tu, egli, faccia; amate,* ind. pres. 2 p. pl. o part. pass. f. pl.; *mangiate,* ind. pres. 2 p. pl. o cong. pres. 2 p. pl. o part. pass. f. pl. Abbiamo, invece, distinto forme omografe, ma non omofone come per esempio, *làvati* e *lavàti,* metten l'accento sulla forma sdrucciola e lascian inaccentate le forme piane.

38 Interessanti sono a questo proposi i dati statistici sulla frequenza degli om nimi in una data lingua e sul rapporto f omonimia e lunghezza delle parole e lo struttura per cui rimandiamo a B. Trnk *A phonological analysis of presentday sta dard english,* 2ª ed., Tokyo, 1966.

nenti a classi diverse di parole, come per esempio *(la) faccia* sostantivo femminile (latino classico *facies*, volgare *facia*) e *faccia* congiuntivo presente del verbo *fare* (latino *faciam, —as, —at)* o *piatto* aggettivo e *piatto* sostantivo che non presentano difficoltà di separazione né ambiguità nella classificazione, poiché il criterio morfosintattico è evidente e i contesti ambigui sono poco probabili.

In questi casi è stato quindi sufficiente indicare la categoria grammaticale per evitare l'ambiguità.

Restano i cosiddetti omonimi effettivi in cui una stessa forma esprime due o più significati senza che la funzione morfosintattica ne sia modificata, per esempio *cappa* (copertura), *cappa* (mollusco), *cappa* (lettera dell'alfabeto). Per alcuni di questi casi è veramente difficile dire dove finisce la polisemia e dove comincia l'omonimia poiché, come ha detto Bloomfield, non possiamo misurare « il grado di prossimità dei significati »; solo in una prospettiva diacronica il contrasto fra omonimi e unità a significato variabile (polisemia) è netto.

È, infatti, sull'etimologia che per lo più si basa il lessicografo nel decidere se registrare certi omonimi dubbi come una parola sola o come due. In linguistica sincronica, invece, il vero criterio sarebbe il legame semantico che vi è fra le due accezioni, ma poiché questo legame non può essere valutato che nella coscienza dei parlanti ed è ovviamente diverso da parlante a parlante; è quindi impossibile farne una norma generale. Ci è sembrato che l'unica soluzione fosse di seguire le suddivisioni date dal vocabolario.

Per individuare questi lemmi, nelle nostre liste, abbiamo aggiunto accanto, fra parentesi, una sommaria definizione sia nel caso in cui due o più omografi sono *attuali*, cioè compaiono nelle nostre liste, come per esempio: *ripartire* (dividere) e *ripartire* (tornar via), sia anche nel caso di omografi *possibili*, cioè quando un solo omografo compare nelle nostre liste.[39] In questo caso, però, non abbiamo indicato possibili omografie con termini scientifici (per esempio: *seno* « angolo »), storico-letterari (*dieta* « assemblea », *marca* « paese »), rari o limitati all'uso regionale (*testo* « coccio », *chiasso* « vicolo »).

Se il numero dei casi imbarazzanti di omonimia è fortunatamente limitato, il numero, invece, delle parole che possono assumere un altro valore sintattico per passaggio di categoria grammaticale si può dire infinito e mentre in alcuni casi è abbastanza semplice stabilire il limite del passaggio poiché è chiaramente definibile, in molti altri è praticamente impossibile determinarlo.

Così mentre *l'uso sostantivato* di *infiniti, participi passati* e *presenti* è facilmente delimitabile, per *l'uso aggettivale* del *participio presente* e soprattutto del *participio passato* è estremamente difficile fissare i limiti e talvolta quasi impossibile assegnare loro una funzione puramente verbale o aggettivale. Esempi: *impiegato, laureato, riservato* ecc.

39 È da notare che alcuni casi di omografia sono tali solo per la forma-base, che abbiamo chiamato « lemma », ma non per altre forme della stessa voce; così per esempio: *tema* « paura » o « argomento », ma *temi* è univocamente il plurale di *tema* « argomento ».

In altri casi, d'altra parte, l'omografia è solo a livello di forma e non di lemma, così per esempio: *cameriere* femm. pl. e m. sing., *vite* femm. sing. e femm. pl., *tratti* voce verbale di *trattare* e di *trarre* ecc. Questo tipo di distinzione comparirà perciò solo nelle liste delle forme.

Anche gli *aggettivi sostantivati* e i *sostantivi* in funzione *aggettivale* presen tano dei casi in cui è veramente impossibile determinare il limite del passaggi di categoria poiché sono passaggi che avvengono spesso e con la più grand facilità e di solito senza alcun cambiamento di significato, per esempio: *cr minale, esemplare, rivale.*

In questi casi la separazione è difficile e chi opera queste distinzioni potrebb essere portato ad adottare delle decisioni differenti per i diversi casi sopra c tati; abbiamo perciò preferito, anche in questo caso, mantenere le distinzio date dal dizionario.

Aggettivi usati avverbialmente: in quali condizioni un aggettivo possa fu gere da avverbio è un problema da lungo tempo dibattuto fra i grammatici le soluzioni sono varie e spesso contrastanti. Sono indicative al proposito variazioni che noi troviamo nei dizionari e che rivelano la difficoltà che vi nel decidere se si tratta di un uso aggettivale o piuttosto di una vera lessicali zazione come avverbio, per esempio: *molto, tutto, vicino, alquanto.*

In questi e in analoghi casi di ambiguità funzionale, anche se nell'individu zione delle unità lessicali abbiamo tenuto conto della loro specifica funzio grammaticale (i cui dati saranno oggetto di un prossimo studio statistico sul parti del discorso), abbiamo lasciato unite sotto lo stesso lemma sia le occo renze dell'uso aggettivale sia quelle dell'uso avverbiale poiché riteniamo ch ai fini del calcolo della Frequenza, Dispersione e Uso debbano essere tratta assieme. Accanto al lemma abbiamo dunque segnato la sigla di entrambe categorie grammaticali sottointendendo, naturalmente, che l'uso avverbia è attribuibile solo alla forma maschile singolare.

6.3 *Parole grammaticali*

Buona parte delle cosiddette parole (o strumenti) grammaticali sono vo prive di un valore semantico proprio, determinanti per lo più relazioni si tattiche fra le diverse parti del discorso[40]. Molti autori hanno preferito e

40 Una distinzione importante connessa con la funzione grammaticale della parola è quella, adottando la terminologia introdotta da Marty (*Satz und Wort,* Reichenberg, 1925, p. 208 e segg.) e O. Funke (*Innere Sprachform.* Reichenberg, 1924, passim), fra « parole autosemantiche » e « parole sinsemantiche ».

In altri termini si vuol intendere parole che hanno un significato autonomo dal contesto (sostantivi, aggettivi, verbi, avverbi) e parole invece che essendo portatrici solo di una funzione semantico-sintattica, acquistano il loro significato dal contesto in cui si trovano (articoli, preposizioni, pronomi, congiunzioni e interiezioni). Questa dicotomia risale ad Aristotele (v. R.H. Robins, *Ancient and mediaeval grammatical theory...,* 1951, pp. 19-2 e si è poi ripresentata in varie forme e c diverse denominazioni in molte opere filosofia e di linguistica. Nei diversi mome storici e dai vari studiosi si sono date verse divisioni e denominazioni, ne ric diamo alcune: Aristotele: φωναὶ σημαντι e φωναὶ ἄσημου; Harris: parole *princip* e *accessorie*; Husserl: termini *categoren tici* e *sincategorematici*; Marty e Funk parole *autosemantiche* e *sinsemantic* grammatici cinesi: parole *piene* e *vuo* Sweet: parole *piene* e parole *forma*; Rus e Carnap: *operatori* e *parole proposizion* Guiraud: *mots forts* e *mots outils*; par di *significazione* e parole di *struttu* Ullmann: parole *piene* e *particelle* o par *autonome* e parole *accessorie.*

minarle dalle liste di frequenza: tuttavia, questa soluzione radicale, pur sopprimendo l'estrema difficoltà di una classificazione di questi elementi, ne pone un'altra che consiste nel tracciare un limite netto fra elementi significativi ed elementi di relazione, cosa che spesso è estremamente ardua e difficile. La linea di confine, infatti, come la maggior parte dei confini linguistici, non è affatto definita e può presentare dei punti di fluidità; basti pensare a una tale distinzione fra avverbi, locuzioni prepositive, congiuntive, ecc.

D'altra parte l'eliminazione delle parole grammaticali potrebbe falsare fatti linguistici e stilistici interessanti proprio perché poco conosciuti.[41] Abbiamo quindi preferito fornire le liste di frequenza anche delle parole grammaticali, lasciando piuttosto uniti certi casi di ambiguità funzionale, come l'uso preposizionale e avverbiale di: *contro, dentro, dietro, dopo, fuori,* ecc. o l'uso come avverbio e congiunzione di: *allora, come, quindi, tuttavia.*

Per le forme variabili, cioè: articoli, preposizioni articolate, pronomi, abbiamo già visto come cambino i risultati numerici secondo che si considerino unità di lessico le singole forme o si facciano dei raggruppamenti. Abbiamo raggruppato tutte le forme e le occorrenze dell'*articolo* sotto il lemma dell'articolo maschile singolare *(il)* e abbiamo anche unito le occorrenze di questo nelle *preposizioni articolate* (queste sono facilmente identificabili nei nostri elenchi perché sono precedute da un trattino indicante appunto la separazione dalla preposizione).

Pronomi personali: abbiamo raggruppato sotto il lemma del pronome personale soggetto anche le forme del pronome personale complemento in posizione tonica e atona, mentre sono state tenute separate le forme del singolare da quelle del plurale, poiché anche se, per una tradizione grammaticale, come *voi* è considerato plurale di *tu*, *noi* è considerato il plurale di *io*; in realtà, sia nelle lingue indoeuropee sia in altri gruppi linguistici, i pronomi significanti *noi* appartengono a radici diverse e, tranne pochissime eccezioni, non sono mai come valore semantico un *io*+*io*, ma un *io*+*tu*, *io*+*egli*, in quanto un *io*+*io* sarebbe irrazionale.[42]

Abbiamo quindi preferito separare *io* e le forme complemento del singolare da un lato, da *noi* e le forme complemento del plurale dall'altro e per analogia abbiamo anche tenuto separate le forme della seconda e terza persona singolare da quelle del plurale, dove in realtà la divisione non era necessaria, poiché *voi* è *tu*+*tu* o *tu*+*egli*. Sono state anche separate dai verbi le forme enclitiche dei pronomi personali, che compaiono nelle nostre liste sotto il pronome soggetto corrispondente, precedute da un trattino.

Pronomi e Aggettivi possessivi, dimostrativi e indefiniti: una tradizione grammaticale largamente diffusa distingue i cosiddetti aggettivi possessivi dai pronomi, obbligando anche nelle scuole medie a distinguere nell'analisi gram-

41 Il significato stilistico di queste parole è stato studiato da Ch. Muller, *Les «pronoms de dialogue»: interprétation stylistique d'une statistique de mots grammaticaux en français*, in *Actes du Xème Congrès international de linguistique et philologie romanes*, Strasbourg, 1962, pp. 605-612.

42 v. Brugmann-Delbrück, *Grundriss...*, Strassburg, 1897 e sgg., II, 2, p. 378.

maticale: *il mio libro* (aggettivo) da *questo libro è mio* (pronome). Se una tale divisione è lecita anzi utile, se non necessaria, in lingue in cui le forme attributive e predicative sono formalmente distinte (come in francese: *mon livre* ma *ce livre est le mien*), non ci è sembrato utile distinguere solo in base alla funzione ora di attributo ora di predicato, quindi esclusivamente sintattica, voci che appartengono dal punto di vista storico a una medesima radice evidentemente pronominale e che non si differenziano né sul piano semantico né su quello morfologico. Lo stesso vale anche per i pronomi dimostrativi e per alcuni pronomi indefiniti, che abbiamo perciò riunito sotto un unico lemma, indicandone accanto la doppia funzione.

Forme invariabili: *Avverbi, congiunzioni, preposizioni*: mentre alcune non pongono problemi per una classificazione essendo la loro funzione morfologica univoca, come per esempio le preposizioni semplici o gli avverbi di modo, altre invece, come abbiamo detto sopra, possono esplicare più funzioni. Non abbiamo perciò distinto l'uso preposizionale da quello avverbiale di: *appresso, assieme, attorno, avanti, circa, contro, davanti, dentro, dietro, dinanzi, dopo, fino, fuori, innanzi, meno, oltre, presso*; né l'uso come avverbio da quello come congiunzione di: *allora, come, comunque, pure, quindi, tuttavia*, né l'uso aggettivale da quello preposizionale di *extra* e *super*.

7.1 I dati lessicali e quantitativi che abbiamo ottenuto appaiono, a un primo esame, molto interessanti. In particolare ci pare che essi permetteranno di attribuire sul piano quantitativo caratteristiche diverse ai sottoinsiemi del corpus da noi spogliato e ci ripromettiamo di approfondire in seguito l'analisi in questa direzione.

Qui, dato lo scopo del presente volume, riportiamo alcuni dati dell'analisi, esposti anche in forma di tabelle, e le relative rappresentazioni grafiche, che ci sembrano riassumere alcune delle principali caratteristiche quantitative del nostro lessico.

7.2 Classi di frequenza

Riportiamo qui di seguito sia per il LIF nel suo complesso (tabella 1) sia per ciascuno dei sottoinsiemi (tabelle 2÷6) alcuni dati relativi a diverse classi di frequenza.

Le classi sono state delimitate elencando i lemmi in ordine decrescente di frequenza, e poi dividendoli a gruppi di 500.

Colonna a numero d'ordine della classe.

Colonna b numero di lemmi contenuti nella classe.

Colonna c numero progressivo del primo e dell'ultimo lemma della classe nell'ordine di frequenza.

Colonna d somma delle frequenze di tutti i lemmi della classe.

Colonna e somma delle frequenze di tutti i lemmi da quello avente numero progressivo 1 all'ultimo lemma della classe in esame.

Colonna f frequenza media dei lemmi della classe (col. *d* diviso col. *b*).

Colonna g frequenza percentuale delle occorrenze dei lemmi della classe rispetto al totale delle occorrenze nel LIF (tabella 1) o nel rispettivo sottoinsieme (tabelle 2÷6).

Tabella 1 - LIF

a Numero ordine	b Numero lemmi	c Classe	d Frequenza complessiva della classe	e Sommatoria delle frequenze	f Frequenza media	g Frequenza percentuale sul LIF
1	500	1 - 500	390.358	390.358	780,716	80,652
2	500	501 - 1000	32.317	422.675	64,634	6,677
3	500	1001 - 1500	18.387	441.062	36,774	3,799
4	500	1501 - 2000	12.411	453.473	24,822	2,564
5	500	2001 - 2500	8.792	462.265	17,584	1,816
6	500	2501 - 3000	6.573	468.838	13,146	1,358
7	500	3001 - 3500	4.965	473.776	9,930	1,025
8	500	3501 - 4000	3.811	477.614	7,622	0,787
9	500	4001 - 4500	2.922	480.536	5,844	0,604
10	500	4501 - 5000	2.302	482.838	4,604	0,476
11	356	5001 - 5356	1.164	484.002	3,269	0,240

Tabella 2 - TEATRO

1	500	1 - 500	83.994	83.994	167,988	85,183
2	500	501 - 1000	6.055	90.049	12,110	6,140
3	500	1001 - 1500	3.149	93.198	6,298	3,193
4	500	1501 - 2000	1.939	95.137	3,878	1,966
5	500	2001 - 2500	1.288	96.425	2,576	1,306
6	500	2501 - 3000	1.000	97.425	2,000	1,014
7	500	3001 - 3500	501	97.926	1,002	0,508
8	500	3501 - 4000	500	98.426	1,000	0,507
9	178	4001 - 4178	178	98.604	1,000	0,182

Tabella 3- ROMANZI

a Numero ordine	b Numero lemmi	c Classe	d Frequenza complessiva della classe	e Sommatoria delle frequenze	f Frequenza media	g Frequenza percentuale sul LIF
1	500	1 - 500	79.776	79.776	159,552	82,194
2	500	501 - 1000	6.551	86.327	13,102	6,749
3	500	1001 - 1500	3.610	89.937	7,220	3,719
4	500	1501 - 2000	2.392	92.329	4,784	2,464
5	500	2001 - 2500	1.663	93.992	3,326	1,713
6	500	2501 - 3000	1.167	95.159	2,334	1,202
7	500	3001 - 3500	914	96.073	1,828	0,941
8	500	3501 - 4000	500	96.573	1,000	0,515
9	484	4001 - 4484	484	97.057	1,000	0,498

Tabella 4 - CINEMA

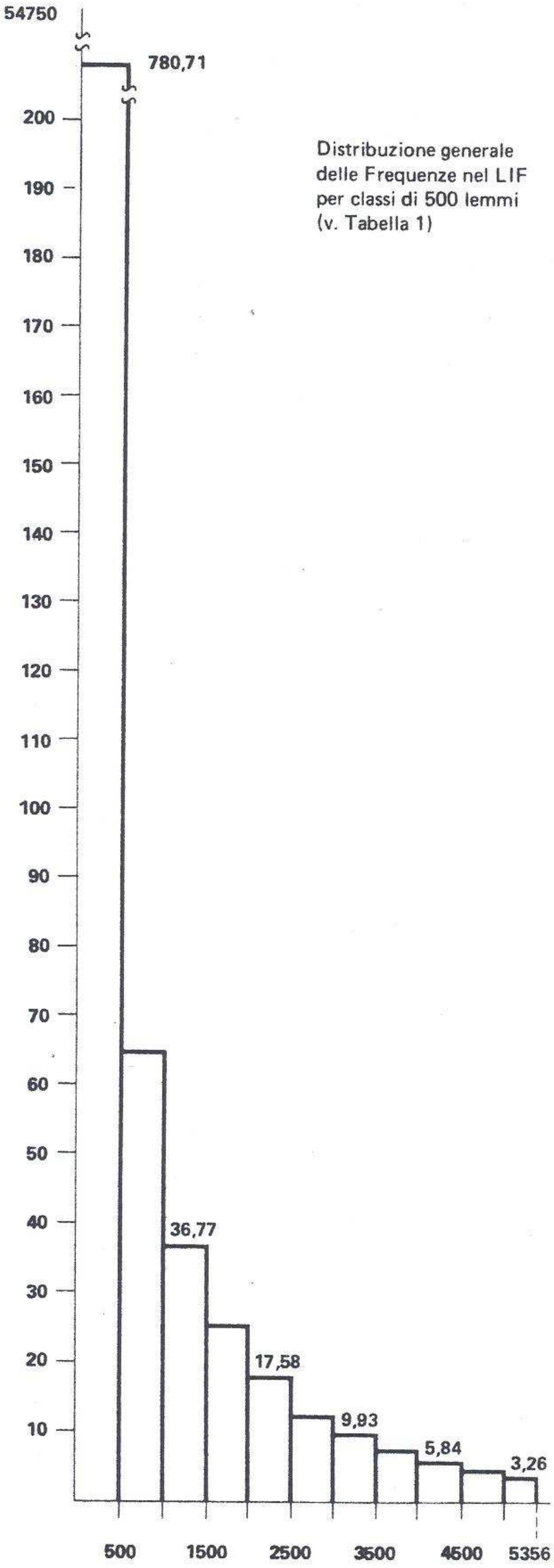
1	500	1 - 500	87.404	87.404	174,800	88,501
2	500	501 - 1000	5.241	92.645	10,482	5,306
3	500	1001 - 1500	2.519	95.185	5,038	2,550
4	500	1501 - 2000	1.490	96.648	2,980	1,508
5	500	2001 - 2500	996	97.644	1,992	1,008
6	500	2501 - 3000	500	98.144	1,000	0,506
7	500	3001 - 3500	500	98.644	1,000	0,506
8	116	3501 - 3616	116	98.760	1,000	0,117

Tabella 5 - PERIODICI

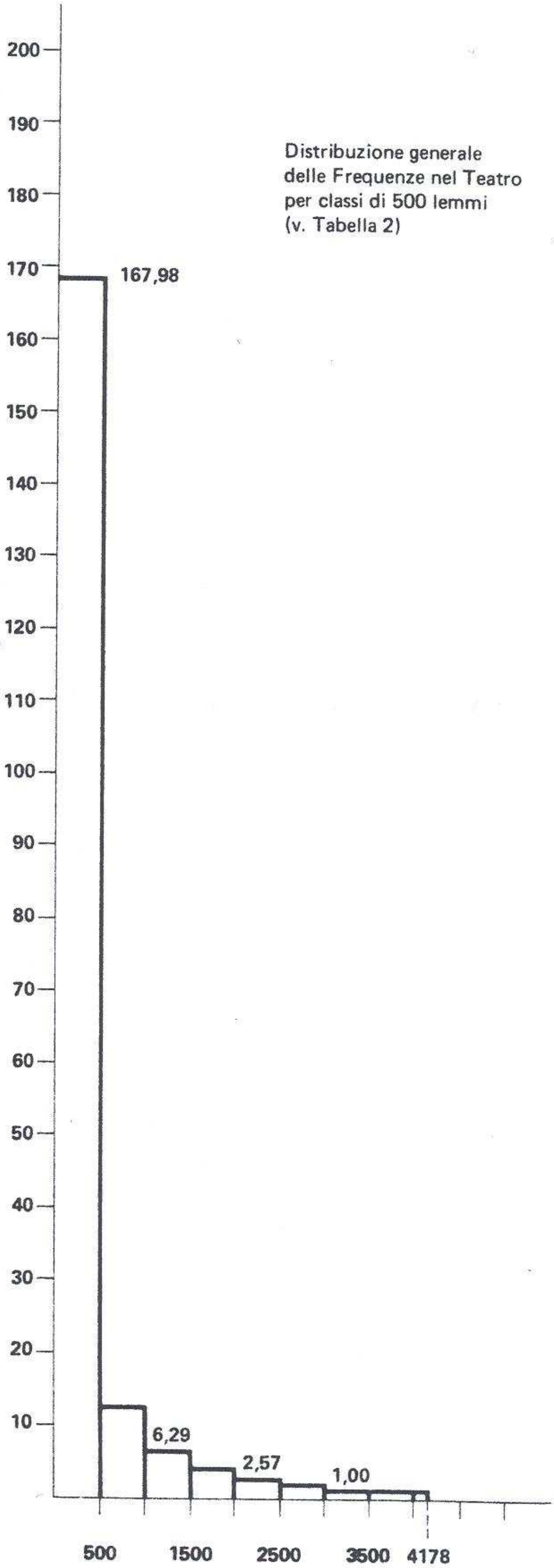
a Numero ordine	b Numero lemmi	c Classe	d Frequenza complessiva della classe	e Sommatoria delle frequenze	f Frequenza media	g Frequenza percentuale sul LIF
1	500	1 - 500	73.962	73.962	147,920	78,455
2	500	501 - 1000	7.726	81.688	15,452	8,195
3	500	1001 - 1500	4.562	86.250	9,124	4,839
4	500	1501 - 2000	2.899	89.149	5,798	3,075
5	500	2001 - 2500	1.901	91.050	3,802	2,016
6	500	2501 - 3000	1.311	92.361	2,622	1,390
7	500	3001 - 3500	952	93.313	1,904	1,009
8	500	3501 - 4000	500	93.813	1,000	0,530
9	460	4001 - 4460	460	94.273	1,000	0,487

Tabella 6 - SUSSIDIARI

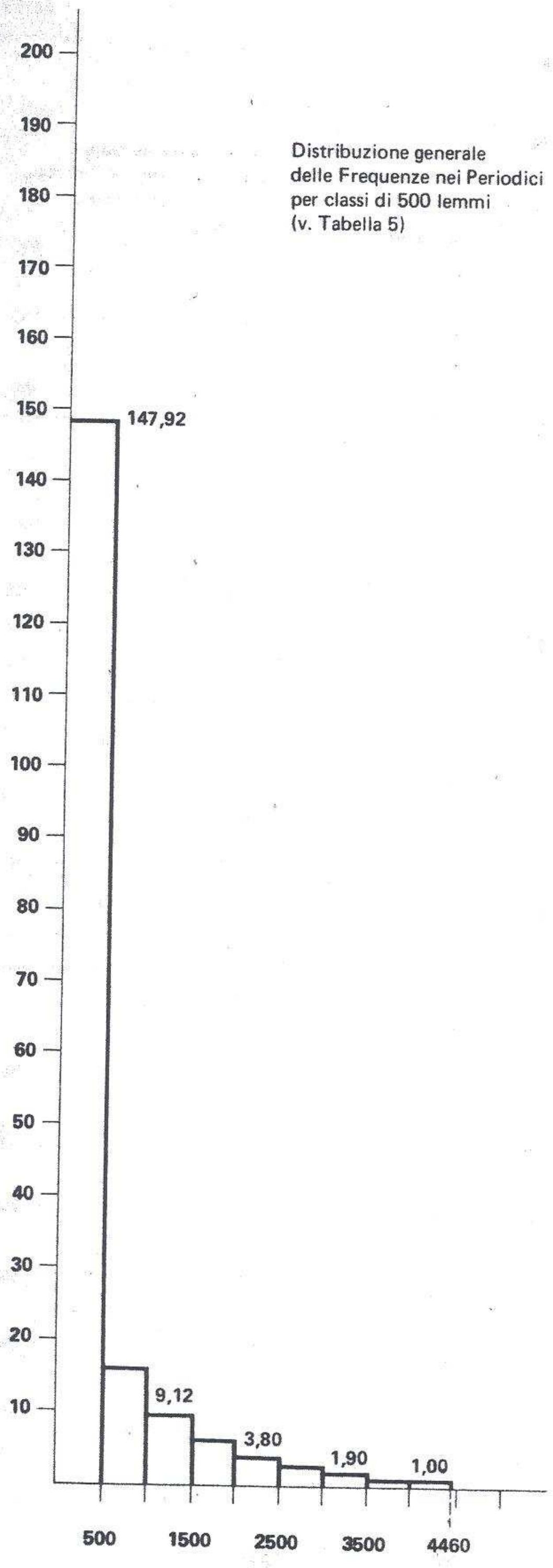
1	500	1 - 500	77.338	77.338	154,676	81,145
2	500	501 - 1000	7.847	85.185	15,694	8,233
3	500	1001 - 1500	4.130	89.315	8,260	4,333
4	500	1501 - 2000	2.423	91.738	4,846	2,542
5	500	2001 - 2500	1.553	93.291	3,106	1,629
6	500	2501 - 3000	1.000	94.291	2,000	1,049
7	500	3001 - 3500	525	94.816	1,050	0,550
8	492	3501 - 3992	492	95.308	1,000	0,516



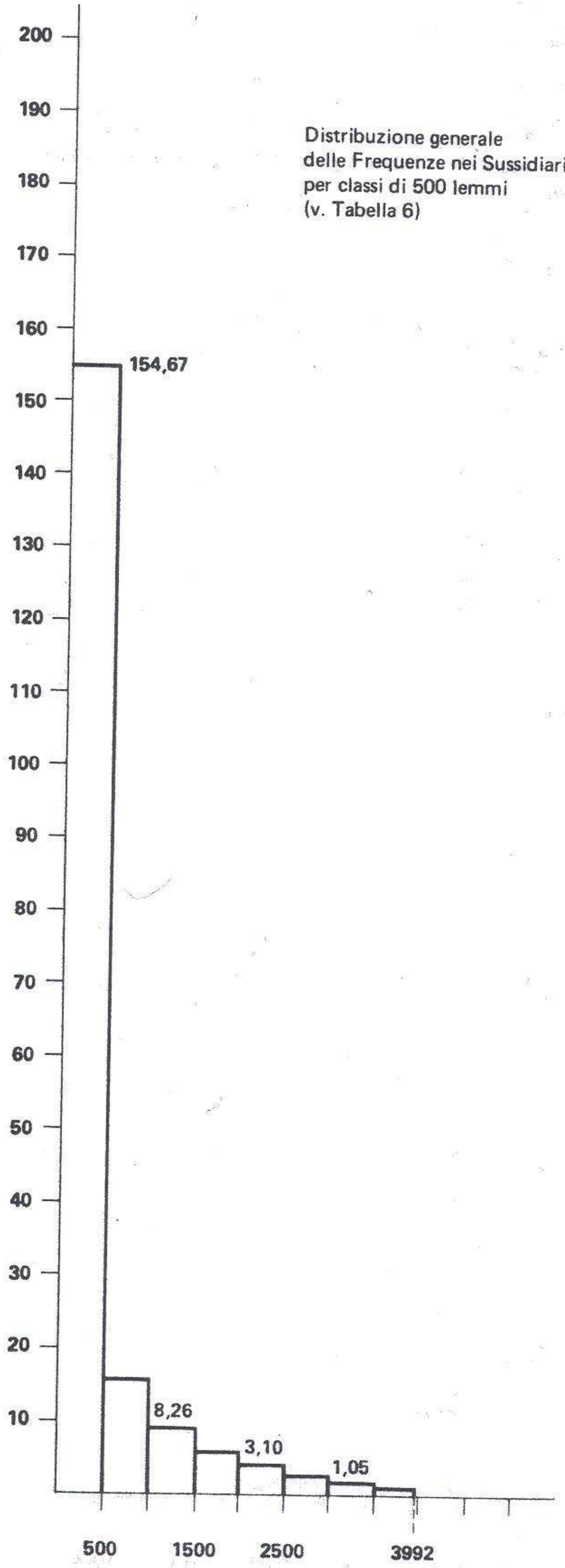
Distribuzione generale delle Frequenze nel LIF per classi di 500 lemmi (v. Tabella 1)



Distribuzione generale delle Frequenze nel Teatro per classi di 500 lemmi (v. Tabella 2)



Distribuzione generale delle Frequenze nei Periodici per classi di 500 lemmi (v. Tabella 5)



Distribuzione generale delle Frequenze nei Sussidiari per classi di 500 lemmi (v. Tabella 6)

7.3 *Valori di frequenza*

7.3.1 Metà (halves)

Se i 5356 lemmi del LIF sono divisi in 2 parti uguali, i primi 2678 assommano 464.833 occorrenze, pari al 96,039% del totale delle occorrenze, con una frequenza media di 173,574 occorrenze per lemma (tabella 7 a).

I secondi 2678 lemmi assommano 19.169 occorrenze, pari al 3,961% del totale delle occorrenze, con una frequenza media di 7,158 occorrenze per lemma (tabella 7 b).

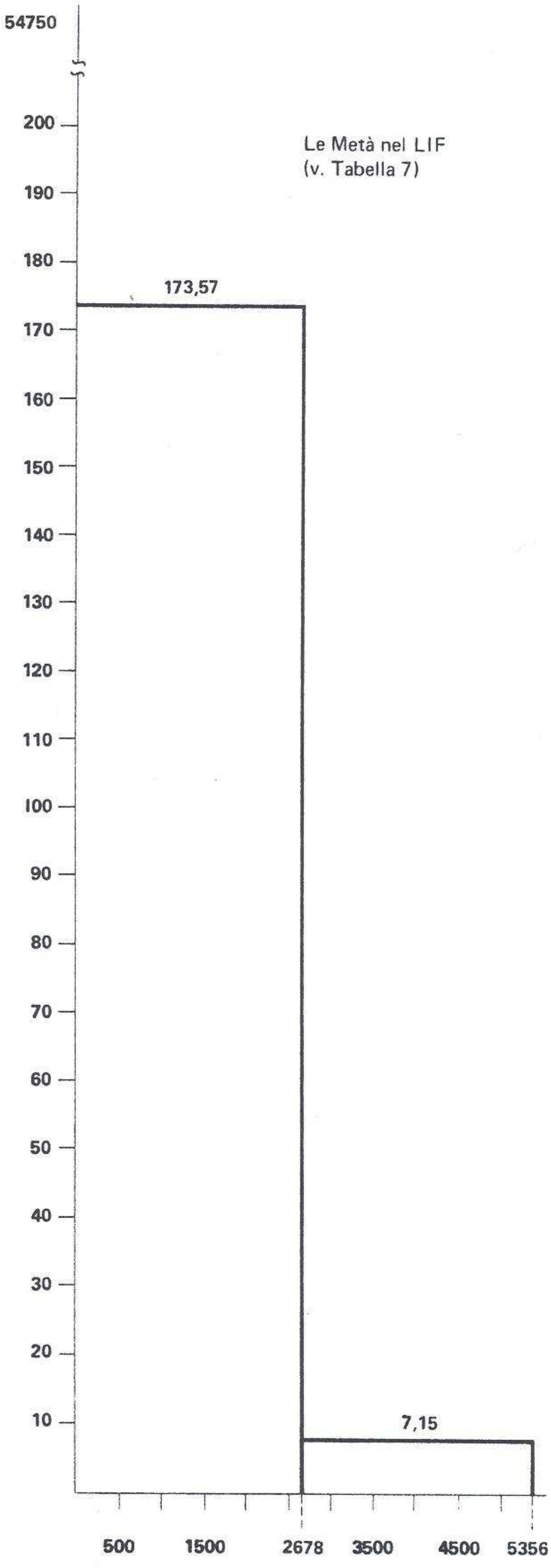
La prima riga della tabella riporta questi dati per il LIF nel suo complesso; le altre righe si riferiscono, invece, a ciascuno dei sottoinsiemi.

Tabella 7 a

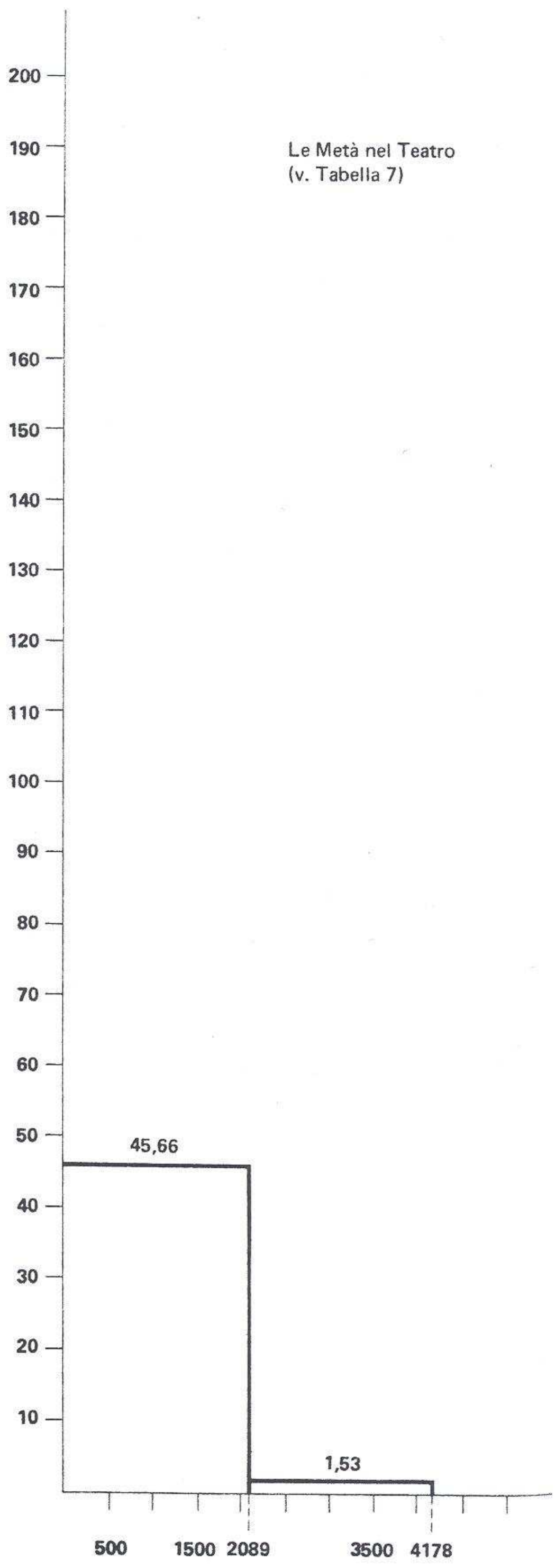
	Numero complessivo lemmi	Numero complessivo occorrenze	Numero lemmi	prima metà dei lemmi		
				Numero occorrenze	Percentuale sul totale occorrenze	Frequenza media per lemma
LIF	5.356	484.002	2.678	464.833	96,039	173,574
teatro	4.178	98.604	2.089	95.404	96,754	45,667
romanzi	4.484	97.057	2.242	93.218	96,044	41,578
cinema	3.616	98.760	1.808	96.160	97,367	53,185
periodici	4.460	94.273	2.230	90.105	95,578	40,405
sussidiari	3.992	95.308	1.996	91.722	96,237	45,952

Tabella 7 b

				seconda metà dei lemmi		
LIF	5.356	484.002	2.678	19.169	3,961	7,158
teatro	4.178	98.604	2.089	3.200	3,242	1,531
romanzi	4.484	97.057	2.242	3.839	3,955	1,712
cinema	3.616	98.760	1.808	2.600	2,632	1,438
periodici	4.460	94.273	2.230	4.168	4,421	1,869
sussidiari	3.992	95.308	1.996	3.586	3,762	1,796

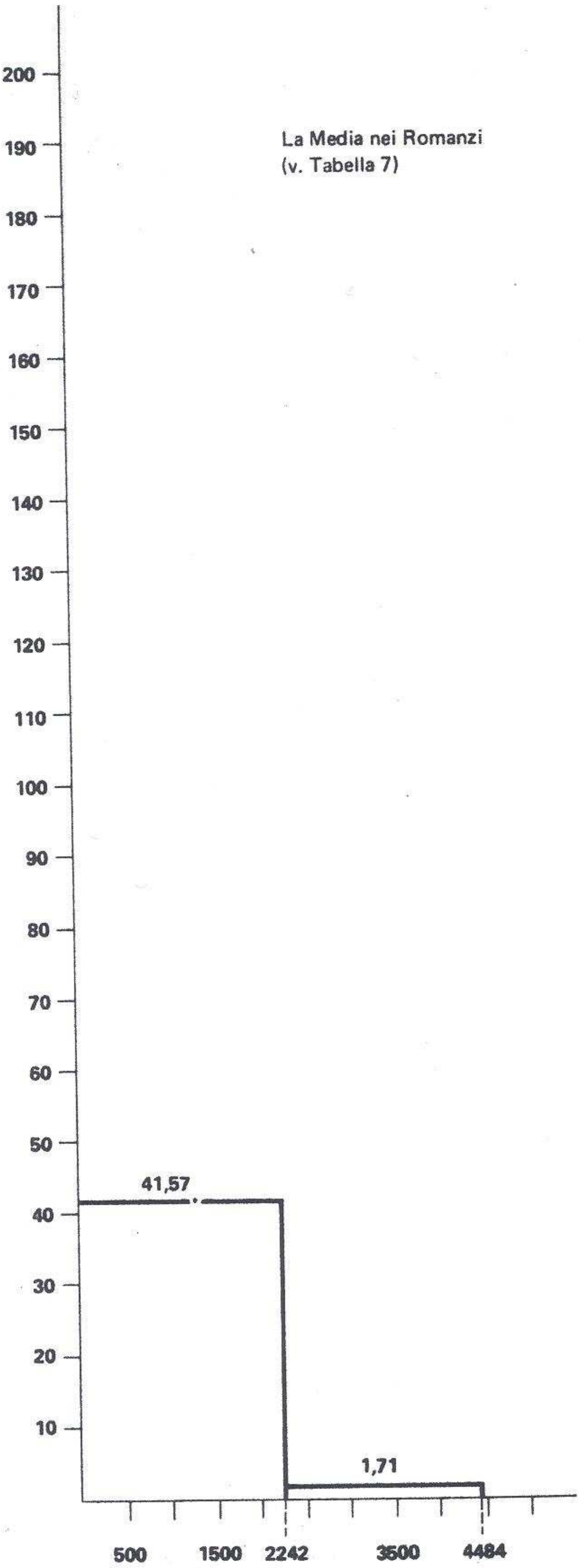


Le Metà nel LIF
(v. Tabella 7)

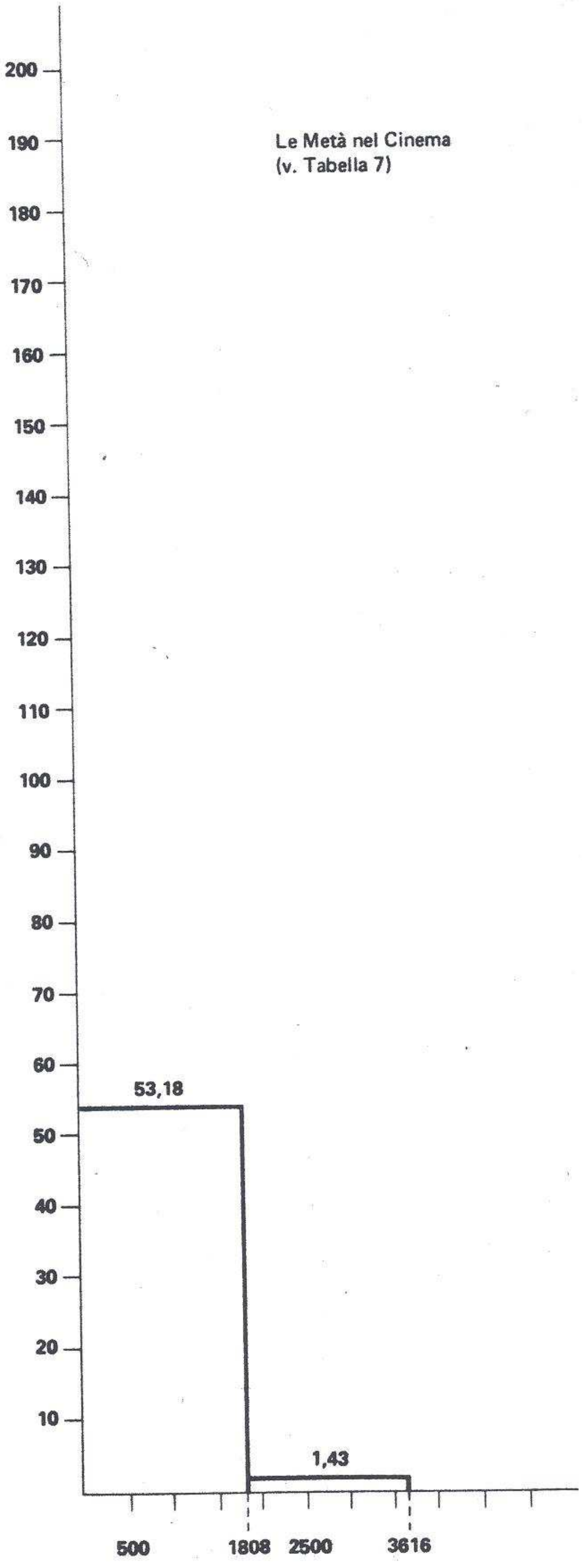


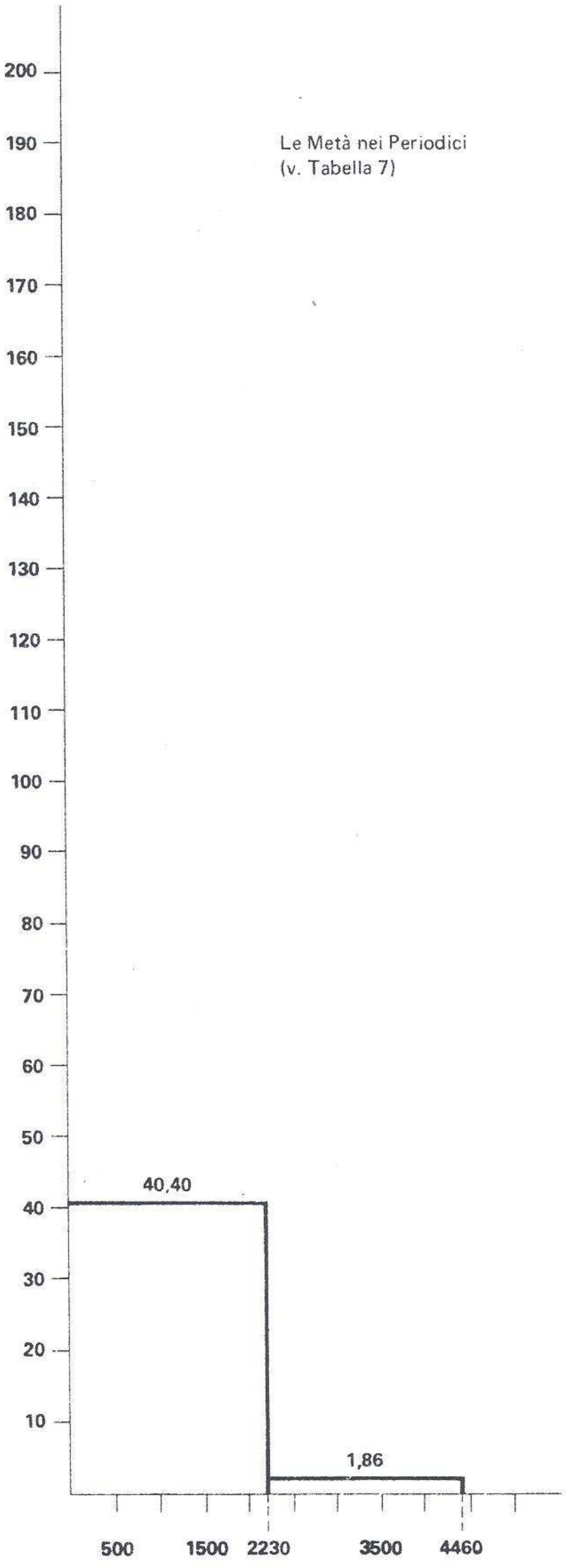
Le Metà nel Teatro
(v. Tabella 7)

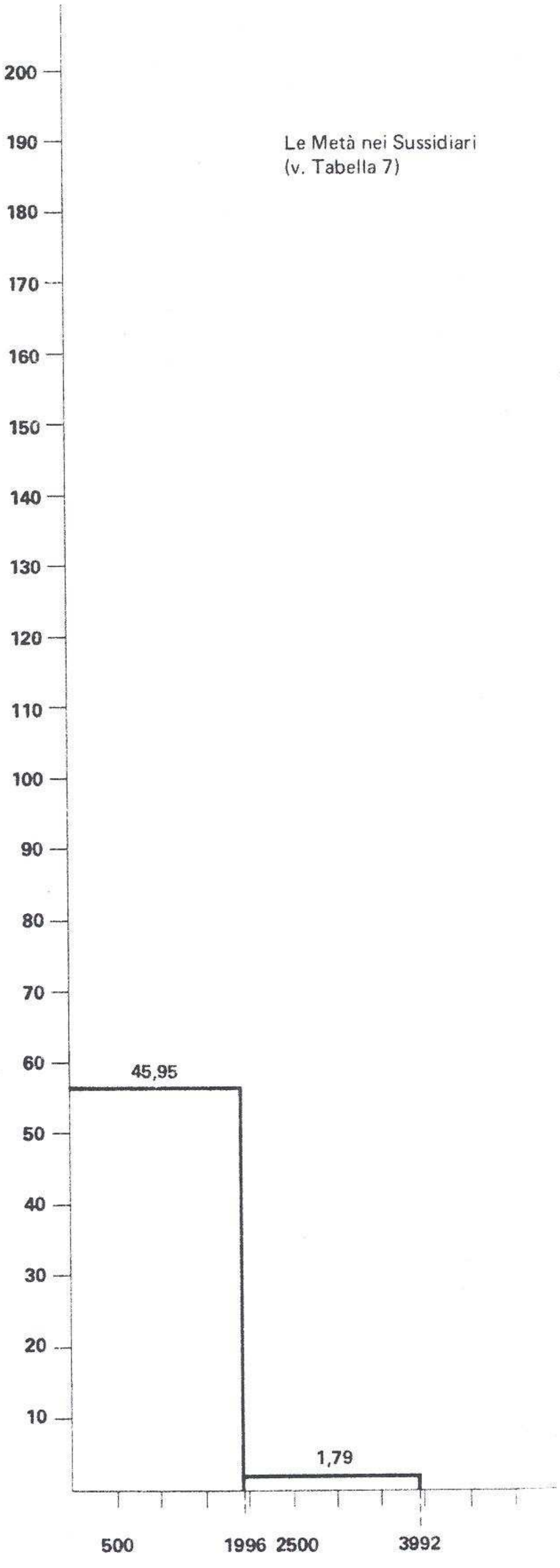
La Media nei Romanzi
(v. Tabella 7)



Le Metà nel Cinema
(v. Tabella 7)




Le Metà nei Periodici
(v. Tabella 7)
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
40,40
1,86
500
1500
2230
3500
4460


Le Metà nei Sussidiari
(v. Tabella 7)
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
45,95
1,79
500
1996
2500
3992

7.3.2 Media

Se il totale delle 484.002 occorrenze del LIF nel suo complesso è diviso per i 5356 lemmi corrispondenti, si ottiene la frequenza media di 1 lemma nel LIF, che è 90,366.

Ci sono 525 lemmi, e cioè il 9,802% dei 5356 lemmi, che hanno frequenza superiore a 90,366. Essi hanno frequenza media 747,945, e complessivamente assommano 392.671 occorrenze, pari all'81,130% del totale delle occorrenze (tabella 8 a).

I rimanenti 4831 lemmi, e cioè il 90,198% dei 5356 lemmi, hanno frequenza inferiore a 90,366. La loro frequenza media è 18,905 e assommano 91.331 occorrenze, pari al 18,870% del totale delle occorrenze (tabella 8 b).

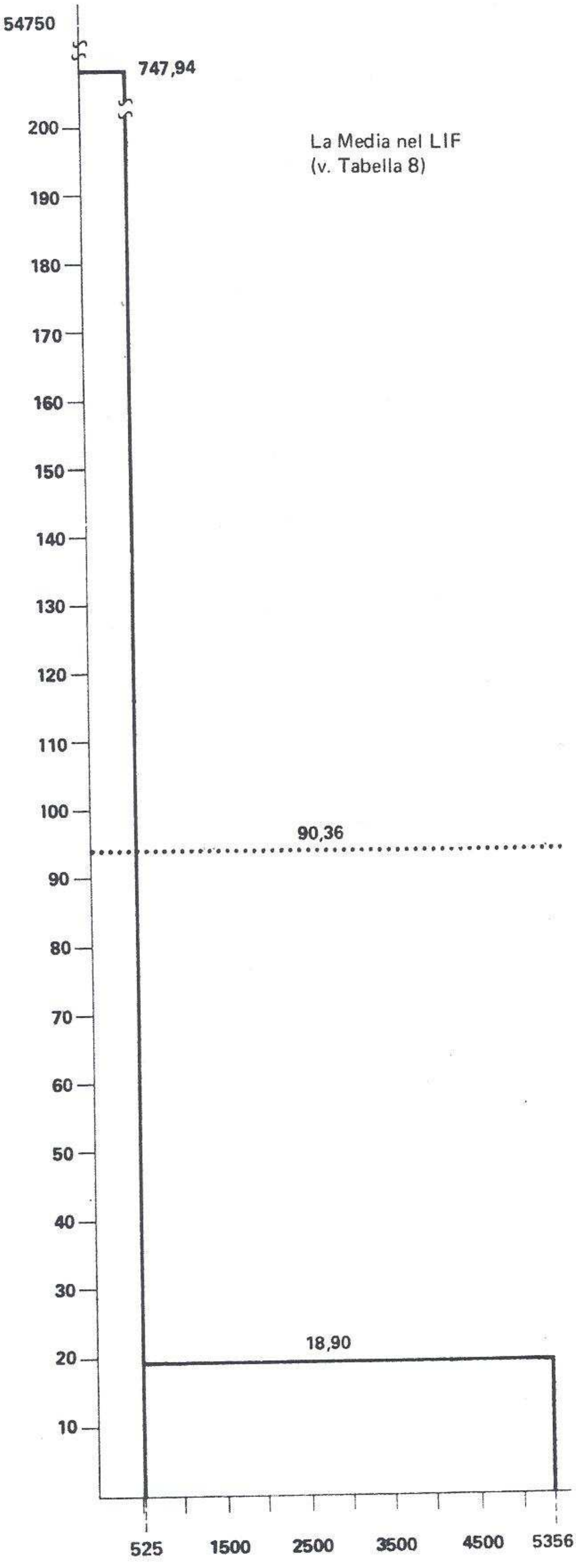
La prima riga della tabella riporta questi dati per il LIF nel suo complesso; le altre righe si riferiscono, invece, a ciascuno dei sottoinsiemi.

Tabella 8 a

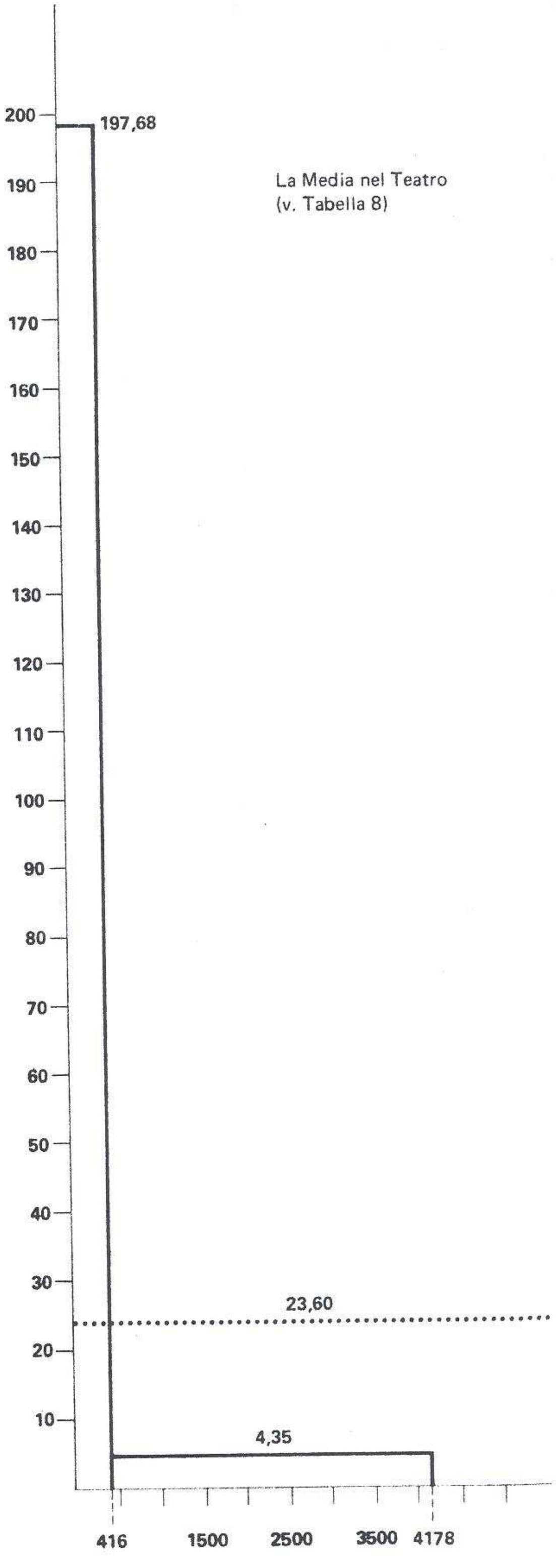
	Numero complessivo lemmi	Numero complessivo occorrenze	Frequenza media complessiva per lemma	Lemmi con frequenza maggiore della frequenza media: Numero lemmi	% sul N. complessivo lemmi	Numero occorrenze	% sul N. complessivo occorrenze	Frequenza media per lemma
LIF	5.356	484.002	90,366	525	9,802	392.671	81,130	747,945
teatro	4.178	98.604	23,600	416	9,956	82.239	82,348	197,689
romanzi	4.484	97.057	21,640	462	10,303	79.002	81,397	171,000
cinema	3.616	98.760	27,310	354	9,789	84.322	85,380	238,197
periodici	4.460	94.273	21,137	497	11,143	73.899	78,388	148,690
sussidiari	3.992	95.308	23,874	487	12,199	77.039	80,831	158,191

Tabella 8 b

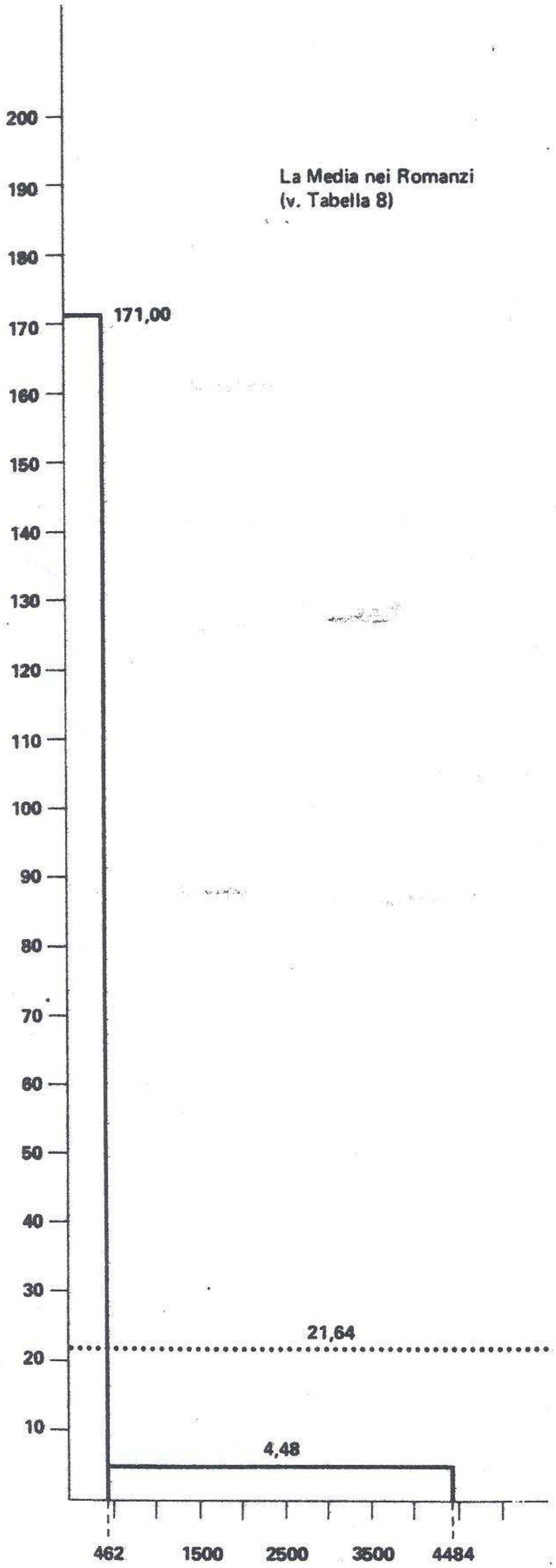
				Lemmi con frequenza minore della frequenza media				
LIF	5.356	484.002	90,366	4.831	90,198	91.331	18,870	18,905
teatro	4.178	98.604	23,600	3.762	90,043	16.365	16,583	4,350
romanzi	4.484	97.057	21,640	4.022	89,696	18.055	18,602	4,489
cinema	3.616	98.760	27,310	3.262	90,210	14.438	14,619	4,426
periodici	4.460	94.273	21,137	3.963	88,856	20.374	21,614	5,141
sussidiari	3.992	85.308	23,874	3.505	37,800	18.269	19,168	5,212

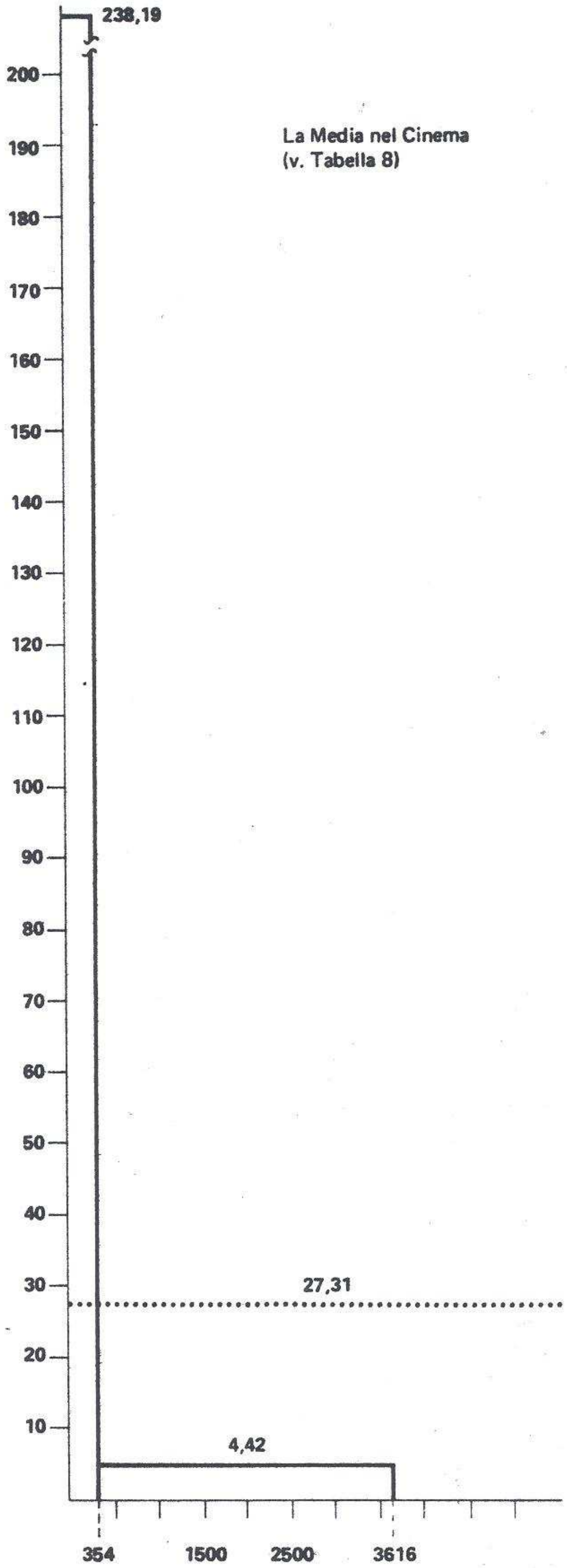


La Media nel LIF
(v. Tabella 8)

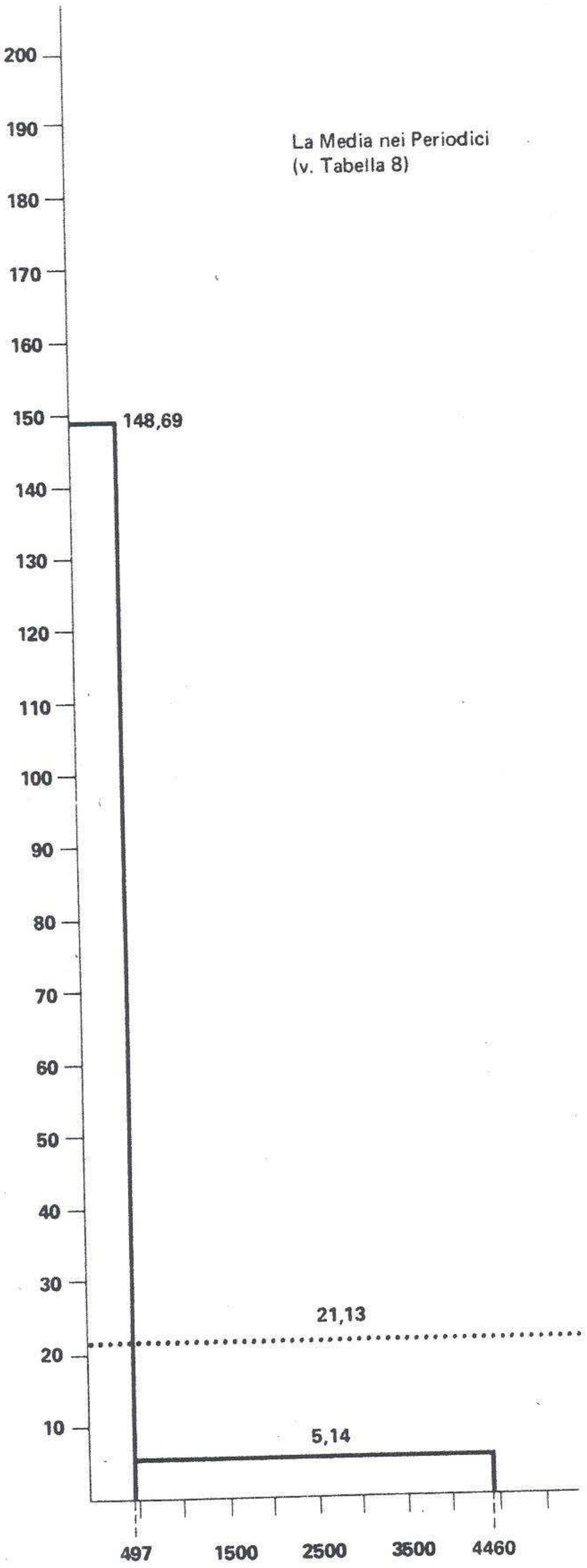


La Media nel Teatro
(v. Tabella 8)

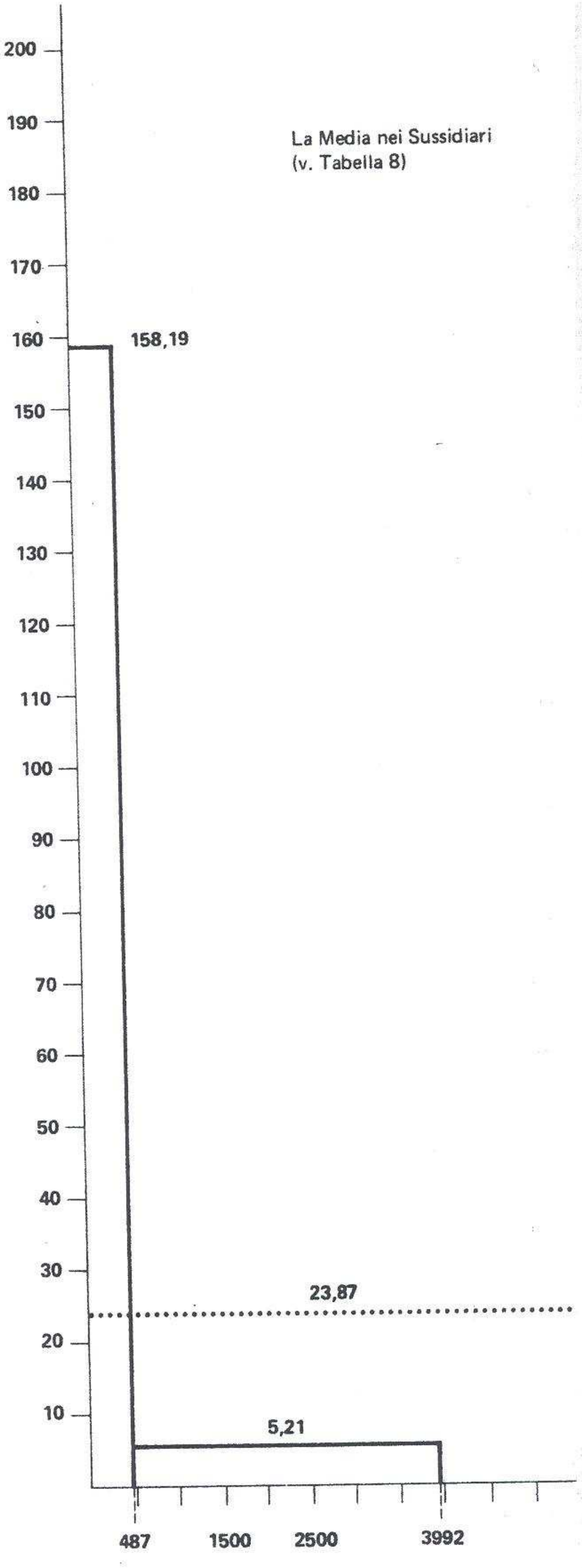

La Media nei Romanzi
(v. Tabella 8)
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
171,00
21,64
4,48
462
1500
2500
3500
4484


La Media nel Cinema
(v. Tabella 8)
238,19
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
27,31
4,42
354
1500
2500
3616

La Media nei Periodici
(v. Tabella 8)



La Media nei Sussidiari
(v. Tabella 8)



7.3.3 Mediana

Se dividiamo le 484.002 occorrenze dei 5356 lemmi del LIF in due parti di 242.001 ciascuna, i primi 29 lemmi, pari allo 0,541% del totale, corrispondono a un po' più della prima metà, cioè 242.715, con una frequenza media di 8369,483 (tabella 9 a).

I restanti 5327 lemmi, pari al 99,459% del totale, ricoprono un po' meno della seconda metà delle occorrenze: esattamente 241.287, con frequenza media 45,295 (tabella 9 b).

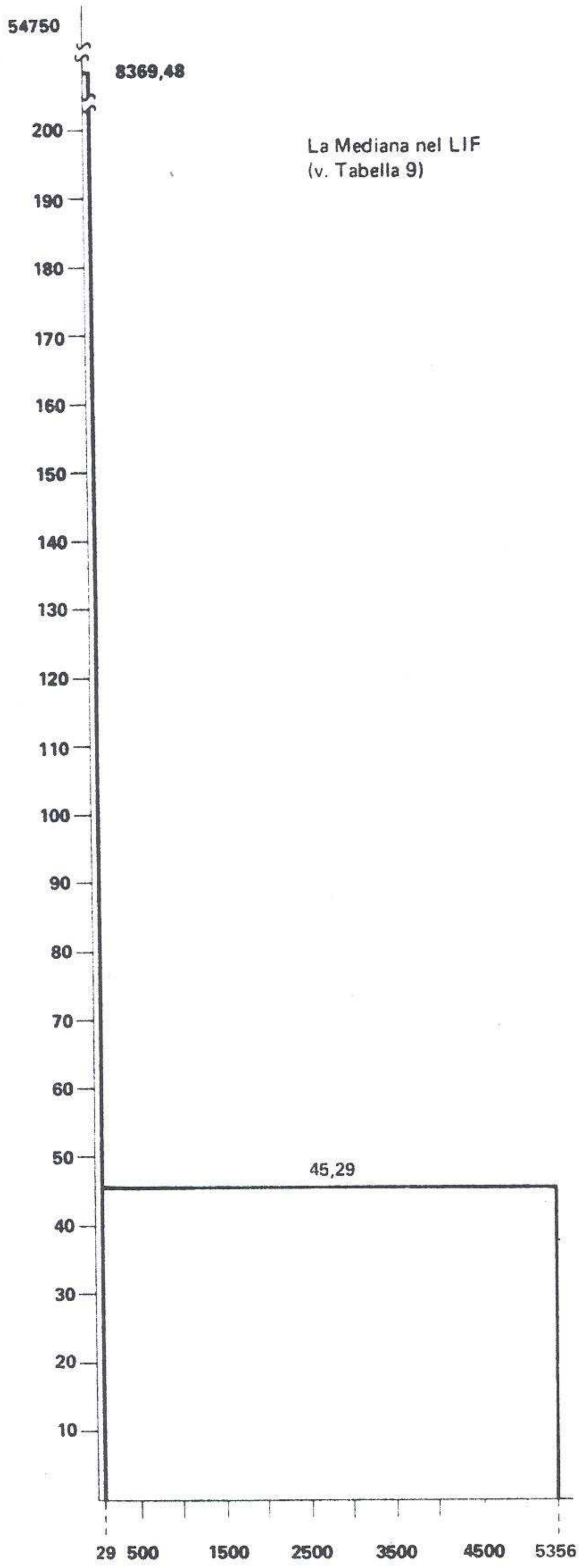
La prima riga della tabella riporta questi dati per il LIF nel suo insieme; le altre righe contengono, invece, gli stessi dati per ciascuno dei sottoinsiemi.

Tabella 9 a

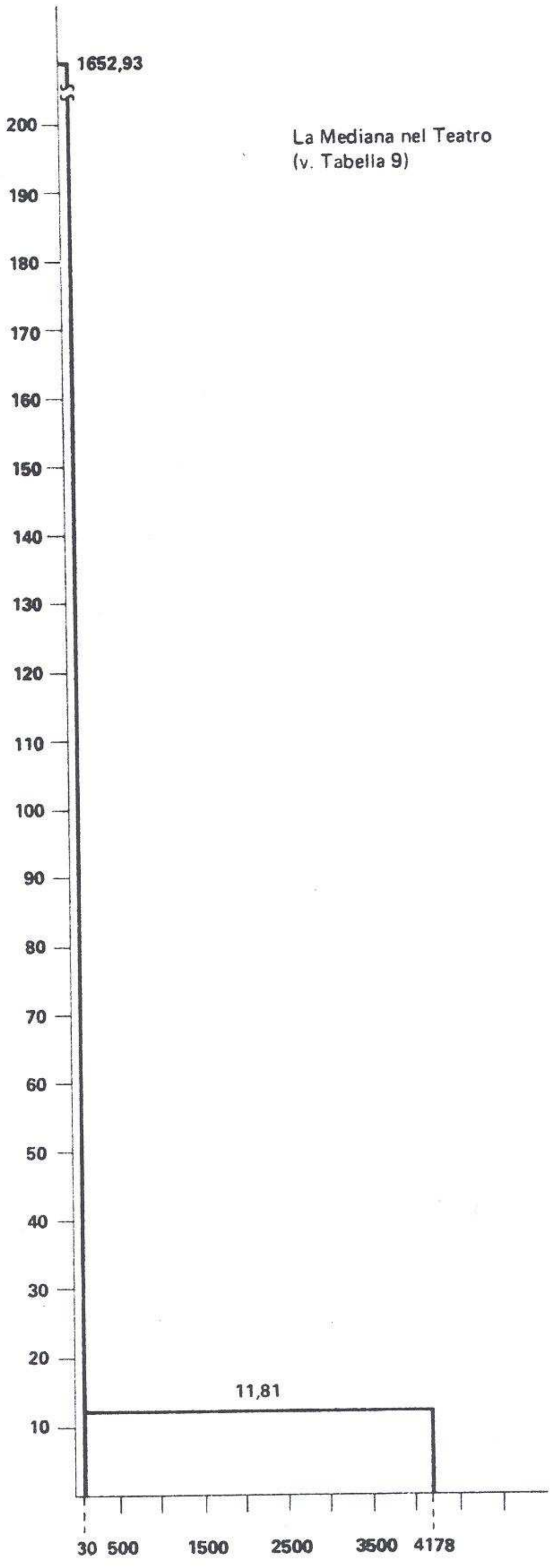
	Numero complessivo lemmi	Numero complessivo occorrenze	prima metà delle occorrenze			
			Numero occorrenze	Numero lemmi	% sul numero complessivo lemmi	Frequenza media per lemma
LIF	5.356	484.002	242.715	29	0,541	8.369,483
teatro	4.178	98.604	49.588	30	0,718	1652,930
romanzi	4.484	97.057	48.914	26	0,579	1881,307
cinema	3.616	98.760	49.754	29	0,801	1715,655
periodici	4.460	94.273	47.462	25	0,560	1898,480
sussidiari	3.992	95.308	47.871	25	0,626	1914,840

Tabella 9 b

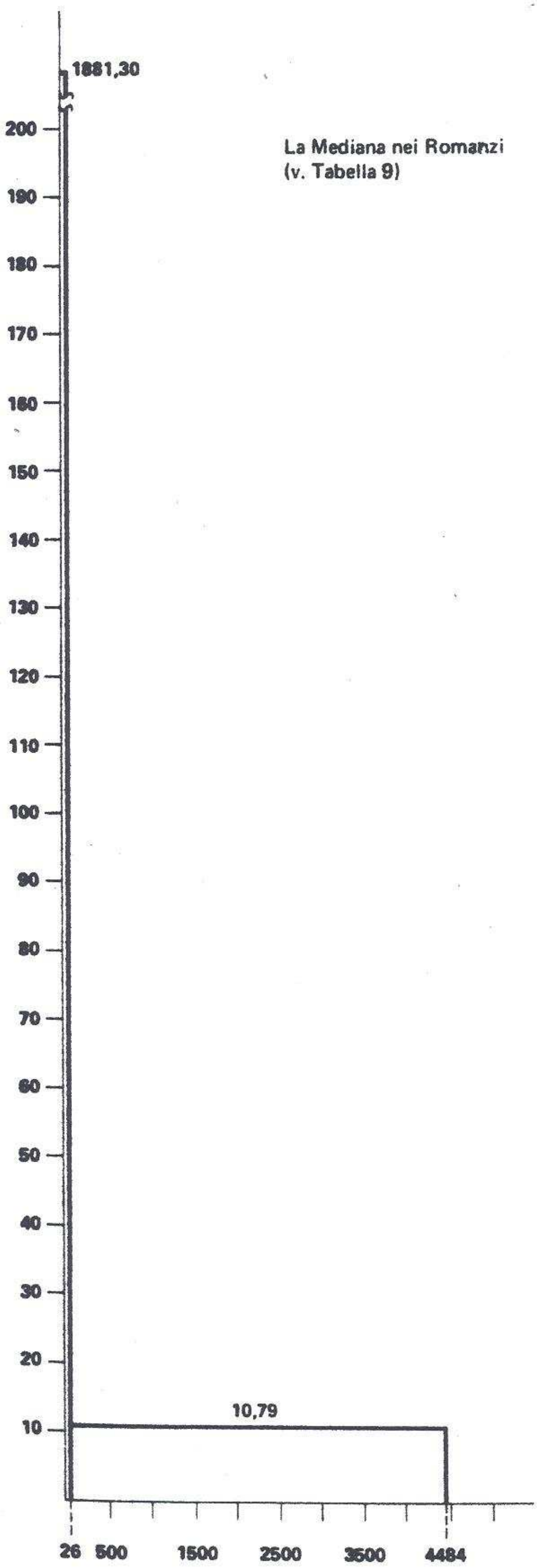
			seconda metà delle occorrenze			
LIF	5.356	484.002	241.287	5.327	99,459	45,295
teatro	4.178	98.604	49.016	4.148	99,281	11,816
romanzi	4.484	97.057	48.143	4.458	99,420	10,799
cinema	3.616	98.760	49.006	3.587	99,198	13,662
periodici	4.460	94.273	46.811	4.435	99,439	10,554
sussidiari	3.992	95.308	47.437	3.967	99,373	11,958



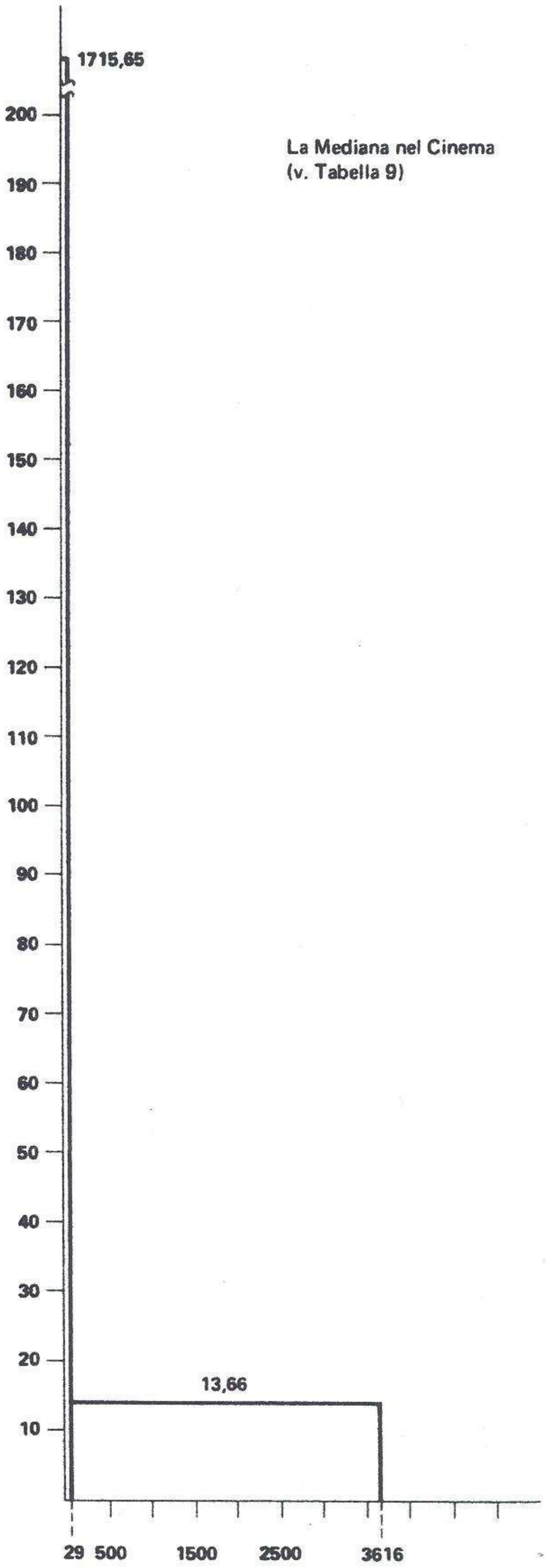
La Mediana nel LIF
(v. Tabella 9)



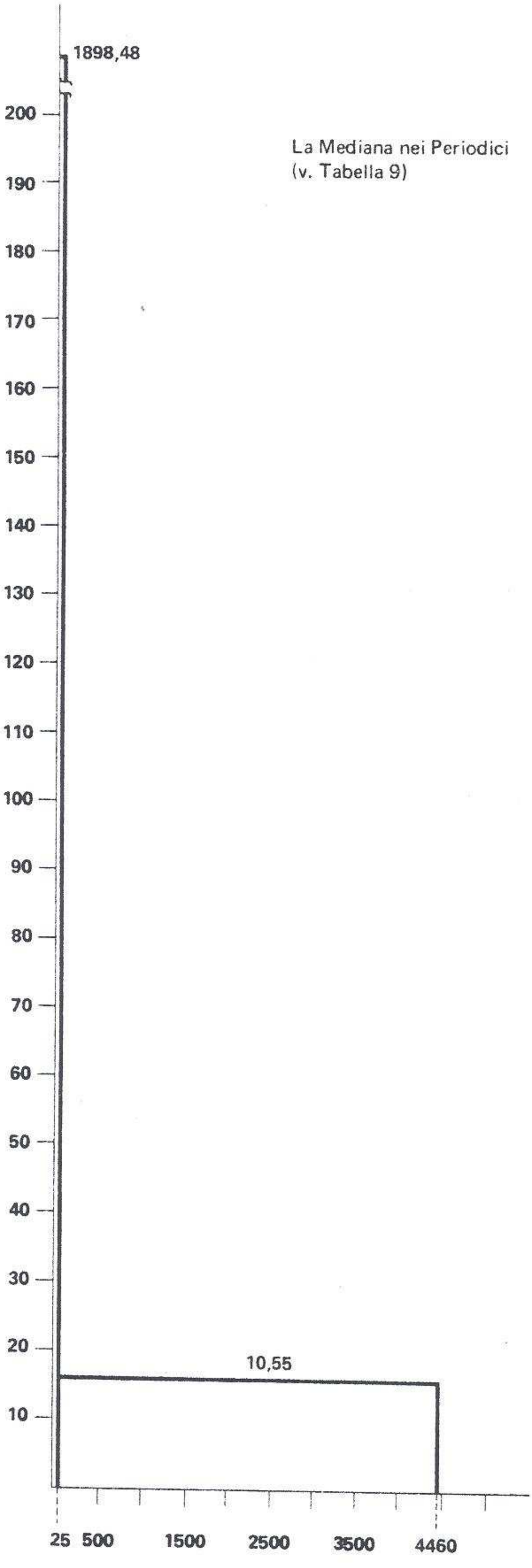
La Mediana nel Teatro
(v. Tabella 9)



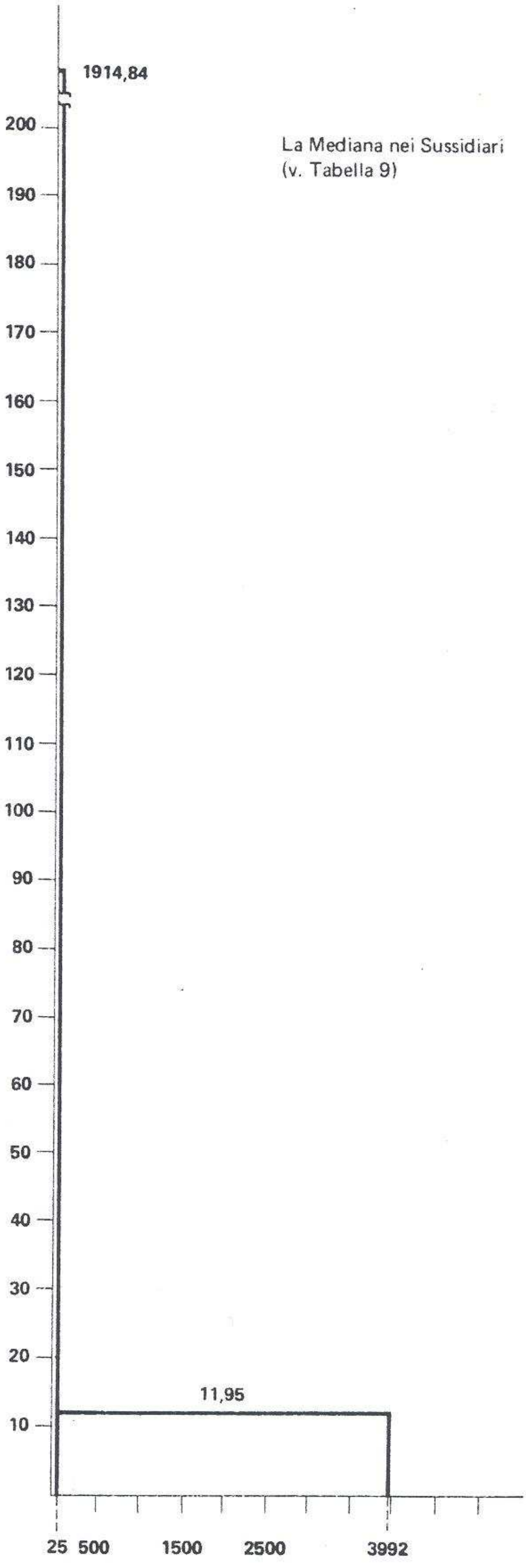
La Mediana nei Romanzi
(v. Tabella 9)



La Mediana nel Cinema
(v. Tabella 9)



La Mediana nei Periodici
(v. Tabella 9)



La Mediana nei Sussidiari
(v. Tabella 9)

7.4 *Classi di dispersione*

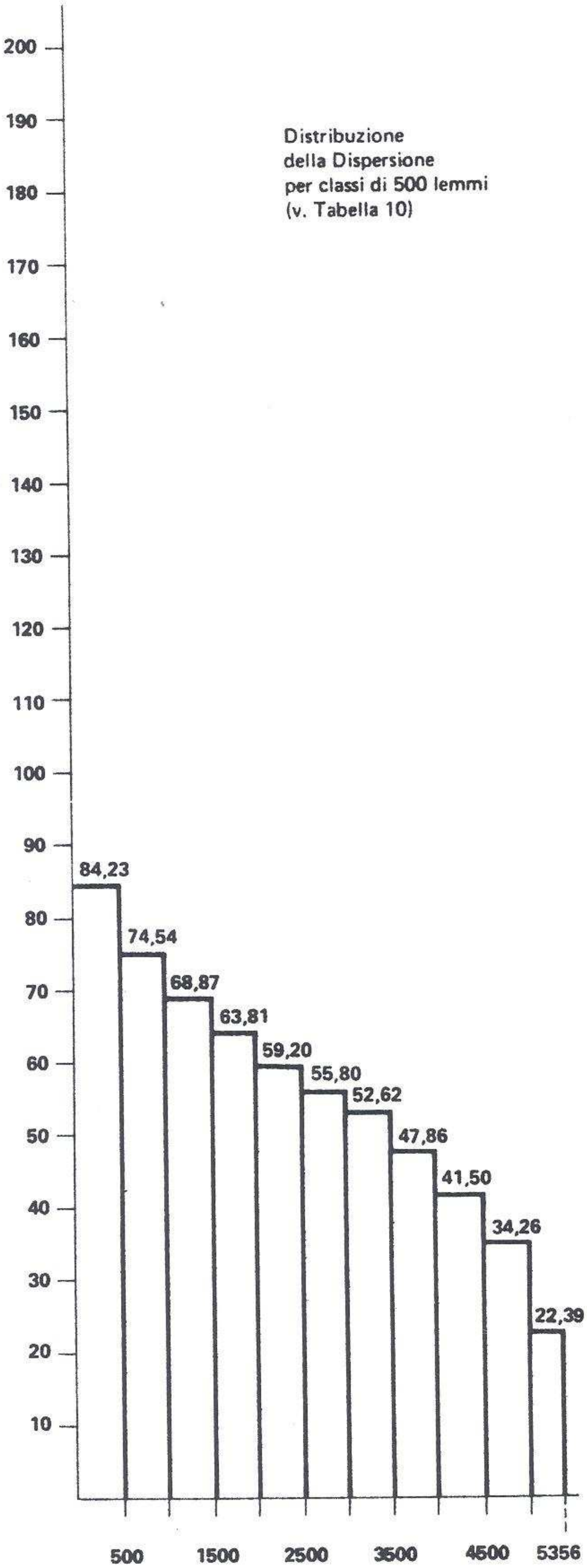
La tabella 10 riporta alcuni dati relativi a diverse classi di dispersione.

I lemmi sono elencati in ordine di dispersione decrescente e divisi in gruppi di 500.

Colonna a numero d'ordine della classe.
Colonna b numero di lemmi contenuti nella classe.
Colonna c numero progressivo del primo e dell'ultimo lemma della classe nell'ordine di dispersione.
Colonna d somma delle dispersioni di tutti i lemmi della classe.
Colonna e dispersione complessiva ottenuta sommando le dispersioni di tutti i lemmi, da quello avente numero progressivo 1 all'ultimo lemma della classe.
Colonna f dispersione media dei lemmi della classe (col. *d* diviso col. *b*).
Colonna g dispersione percentuale delle occorrenze dei lemmi della classe, rispetto alla dispersione complessiva.

Tabella 10

a Numero ordine	b Numero lemmi	c Classe	d Dispersione complessiva della classe	e Sommatoria delle dispersioni	f Dispersione media	g Dispersione percentuale
1	500	1 - 500	421,192	421,192	84,238	14,070
2	500	501 - 1000	372,731	793,923	74,546	12,451
3	500	1001 - 1500	344,391	1.138,314	68,878	11,505
4	500	1501 - 2000	319,086	1.457,401	63,817	10,659
5	500	2001 - 2500	296,012	1.753,413	59,202	9,889
6	500	2501 - 3000	279,017	2.032,431	55,803	9,321
7	500	3001 - 3500	263,132	2.295,563	52,626	8,790
8	500	3501 - 4000	239,322	2.534,886	47,864	7,995
9	500	4001 - 4500	207,503	2.742,390	41,500	6,932
10	500	4501 - 5000	171,342	2.913,733	34,268	5,724
11	356	5001 - 5356	79,713	2.993,446	22,391	2,663

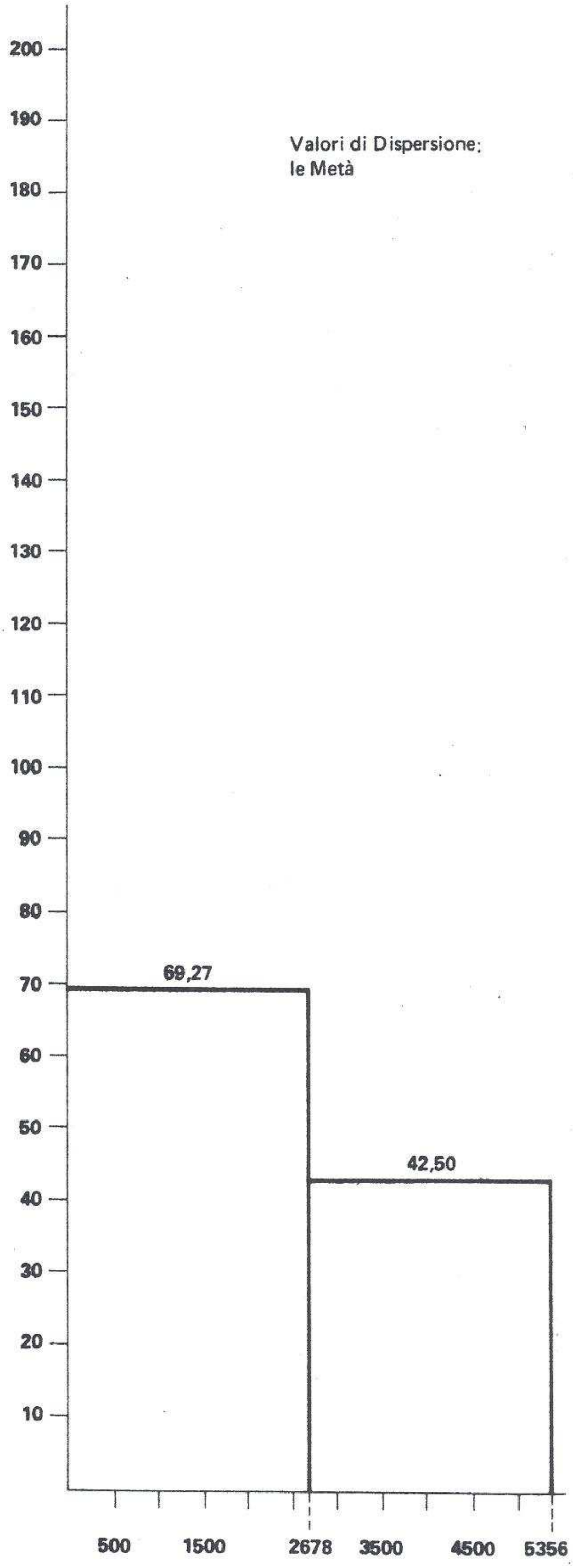

Distribuzione
della Dispersione
per classi di 500 lemmi
(v. Tabella 10)
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
84,23
74,54
68,87
63,81
59,20
55,80
52,62
47,86
41,50
34,26
22,39
500
1500
2500
3500
4500
5356

7.5 *Valori di dispersione*

7.5.1 Metà (halves)

Se dividiamo a metà i 5356 lemmi del LIF, la prima metà di 2678 lemmi ha dispersione complessiva 1855,287, pari al 61,978% del totale, e dispersione media 69,278.

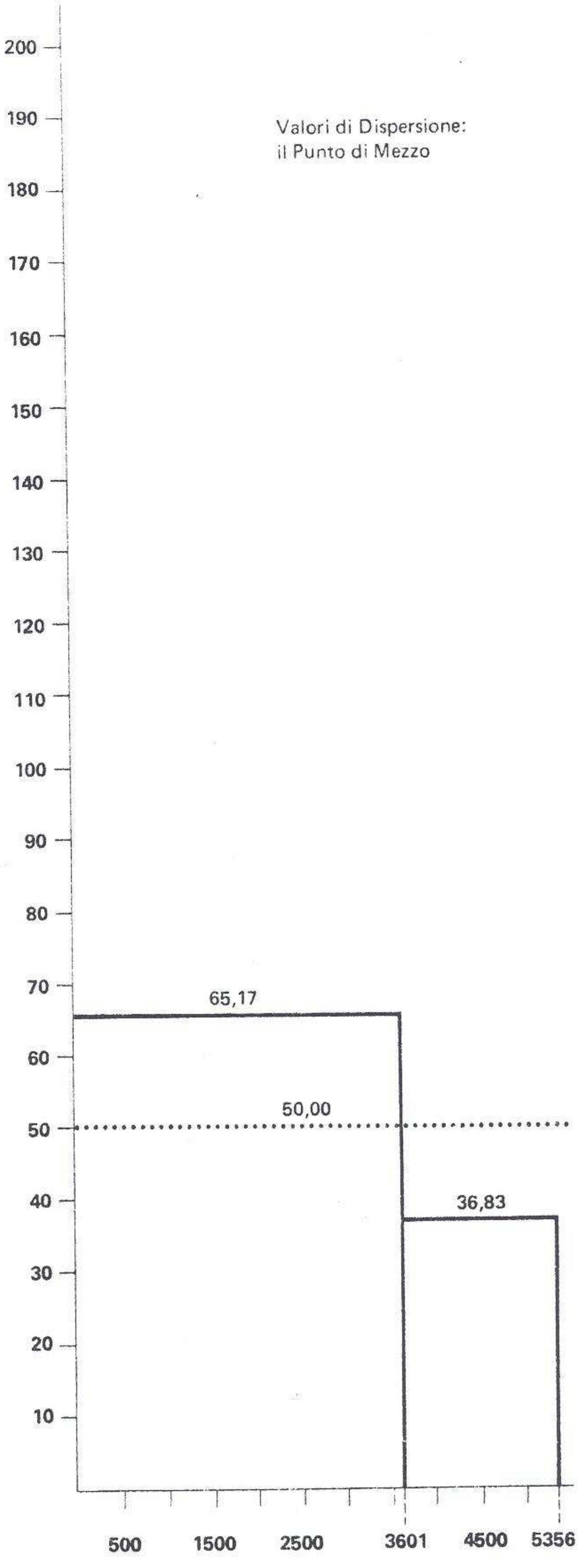
La seconda metà, anch'essa di 2678 lemmi, ha dispersione complessiva 1138,159, pari al 38,022% del totale, con dispersione media 42,50.



7.5.2 Punto di mezzo (mid-point)

Se dividiamo i 5356 lemmi in due gruppi relativi al punto di mezzo di dispersione, dal valore 1 allo 0,51 e dallo 0,50 allo 0,01, il primo gruppo conta 3601 lemmi, pari al 67,233%, con dispersione totale 2346,946 e dispersione media 65,174.

Il secondo gruppo conta invece 1755 lemmi, pari al 32,767%, con dispersione totale 646,50 e dispersione media 36,837.

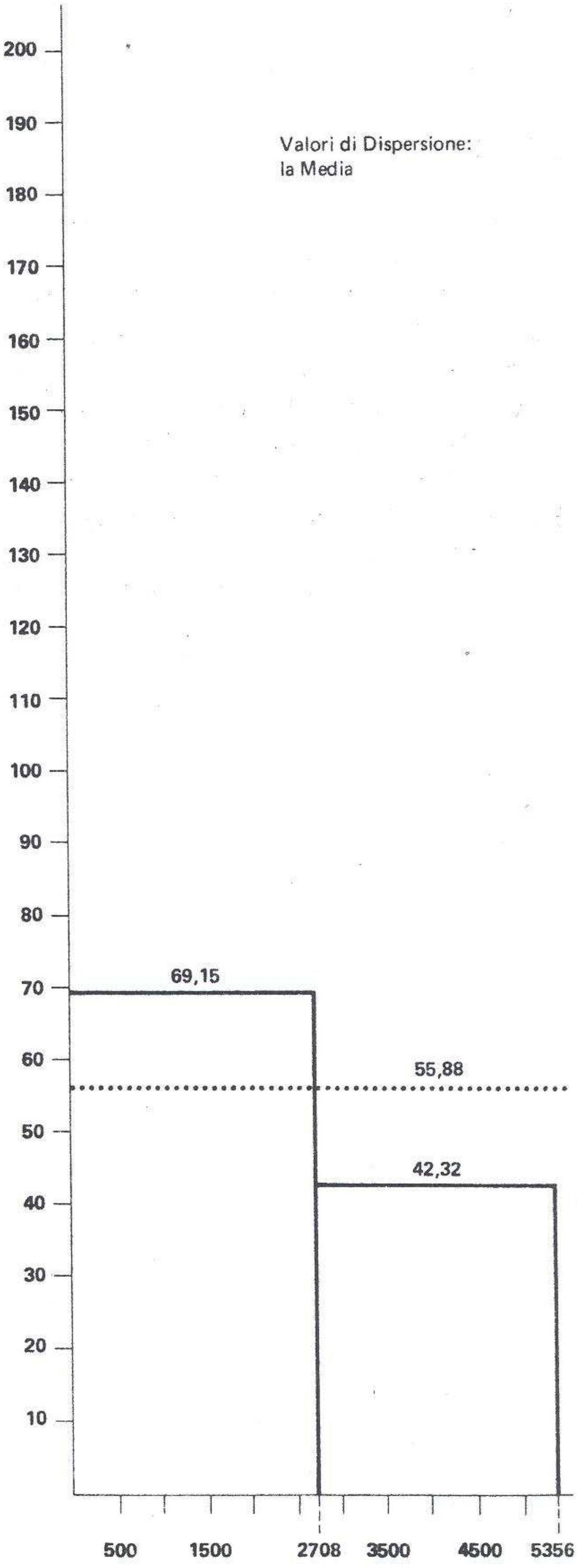


7.5.3 Media

Se la dispersione totale di 2993,446 è divisa per il numero dei lemmi (5356), la dispersione media è 55,889.

Ci sono 2708 lemmi, pari al 50,560%, con dispersione maggiore di 55,889; essi hanno dispersione totale 1872,676, pari al 62,559%, e dispersione media 69,153.

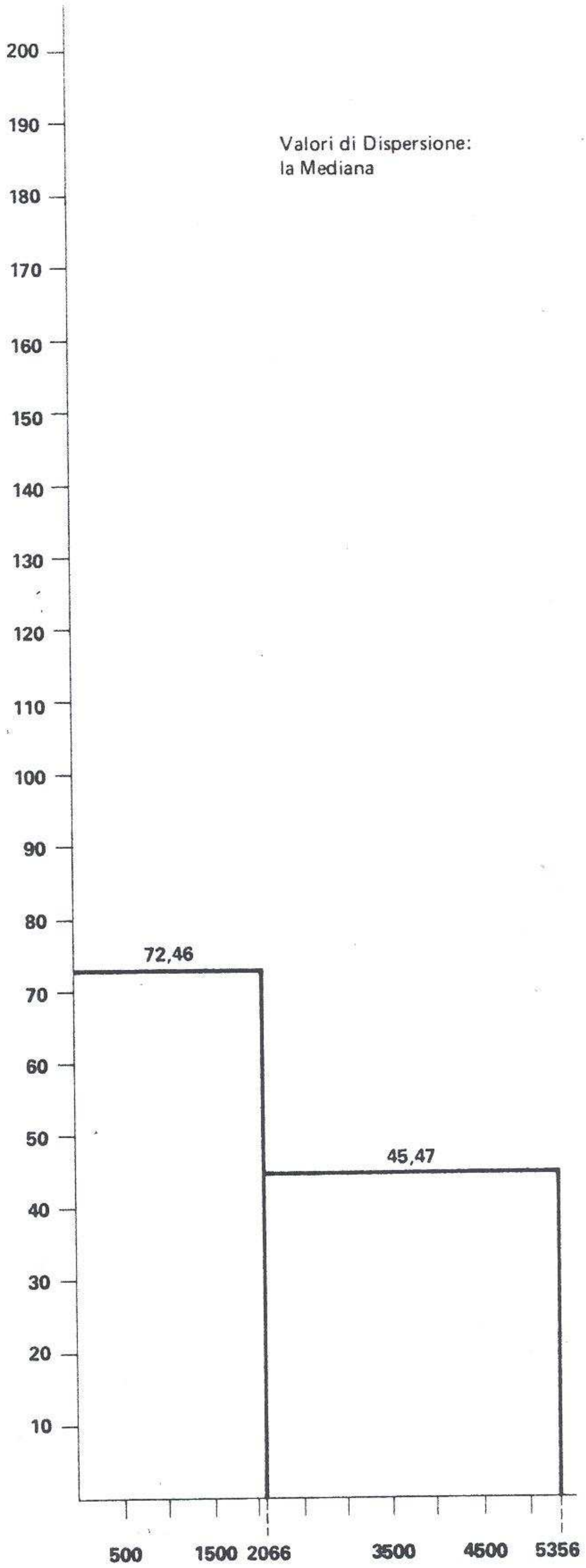
I lemmi con dispersione minore di 55,889 sono 2648, pari al 49,440%, e hanno dispersione totale 1120,770, pari al 37,441%, e dispersione media 42,325.



7.5.4 Mediana

Se la dispersione totale di 2993,446 è divisa in due metà di 1496,723 ciascuna, i primi 2066 lemmi, pari al 38,574% del totale, hanno dispersione complessiva 1497,198, pari al 50,016% della dispersione totale, con dispersione media 72,468.

I rimanenti 3290 lemmi, pari al 61,426% del totale, hanno dispersione complessiva 1496,248, pari al 49,984%, e dispersione media 45,478.



7.6 *Classi di uso*

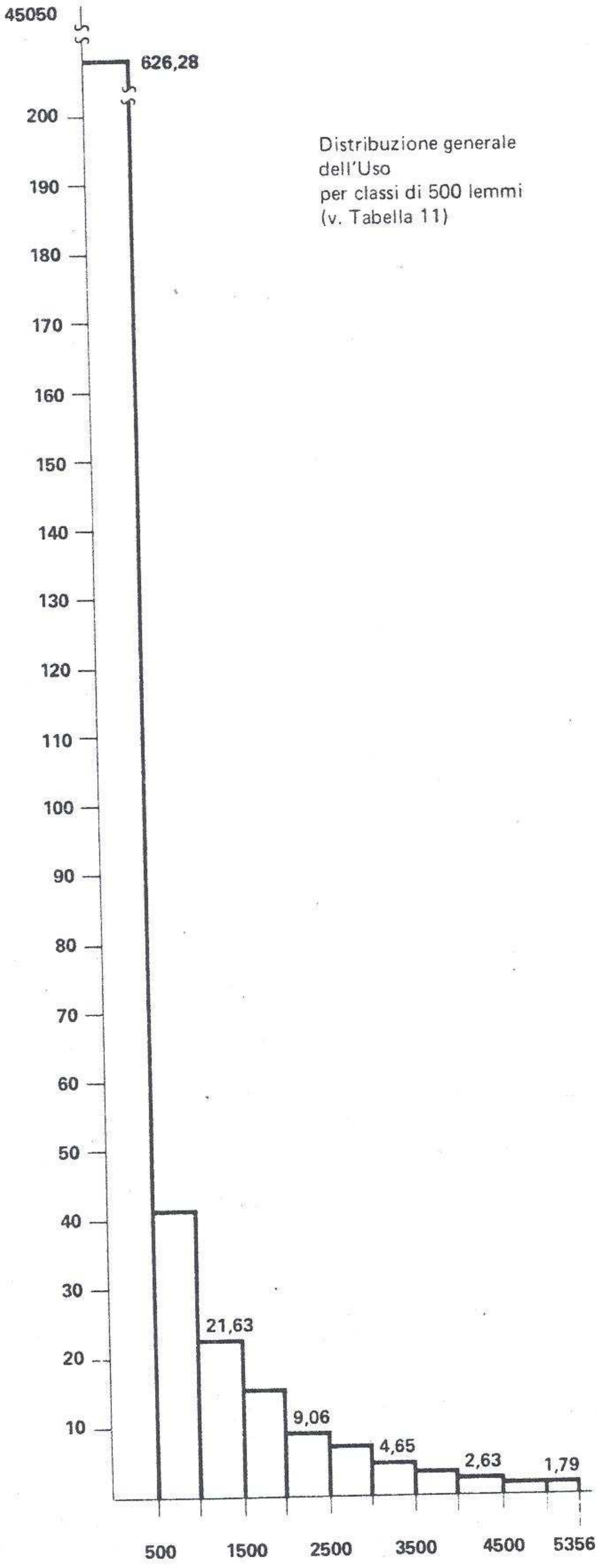
La tabella 11 riporta alcuni dati relativi a diverse classi di uso.

I lemmi sono stati elencati in ordine di uso decrescente e divisi in gruppi di 500.

Colonna a numero d'ordine della classe.
Colonna b numero di lemmi contenuti nella classe.
Colonna c numero progressivo del primo e dell'ultimo lemma della classe nell'ordine di uso.
Colonna d somma dell'uso di tutti i lemmi della classe.
Colonna e uso complessivo ottenuto sommando l'uso di tutti i lemmi da quello avente numero progressivo 1 all'ultimo lemma della classe.
Colonna f uso medio dei lemmi della classe (col. *d* diviso col. *b*).
Colonna g uso percentuale delle occorrenze dei lemmi della classe rispetto all'uso totale.

Tabella 11

a Numero ordine	b Numero lemmi	c Classe	d Uso complessivo della classe	e Sommatoria dell'uso	f Uso medio	g Uso percentuale sul LIF
1	500	1 - 500	313.143,37	313.143,37	626,287	85,529
2	500	501 - 1000	20.510,06	333.653,33	41,020	5,602
3	500	1001 - 1500	10.816,35	344.469,68	21,632	2,954
4	500	1501 - 2000	6.850,77	351.320,45	13,701	1,871
5	500	2001 - 2500	4.533,02	355.853,47	9,066	1,238
6	500	2501 - 3000	3.188,16	359.041,63	6,376	0,871
7	500	3001 - 3500	2.325,53	361.367,16	4,651	0,635
8	500	3501 - 4000	1.727,73	363.094,89	3,455	0,472
9	500	4001 - 4500	1.315,49	364.410,38	2,630	0,359
10	500	4501 - 5000	1.074,52	365.484,90	2,149	0,293
11	356	5001 - 5356	639,60	366.124,50	1,796	0,175



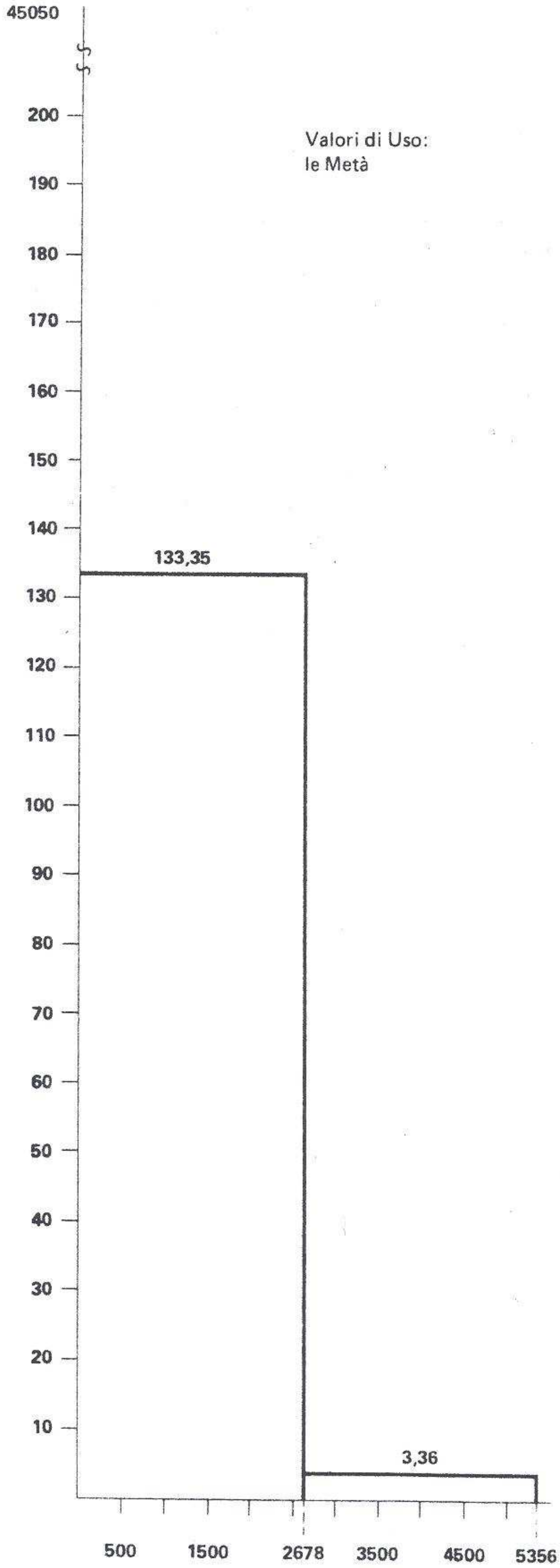
Distribuzione generale
dell'Uso
per classi di 500 lemmi
(v. Tabella 11)

7.7 *Valori di uso*

7.7.1 Metà (halves)

Se i 5356 lemmi sono divisi in 2 parti uguali, i primi 2678 lemmi hanno uso totale 357.114,35, pari al 97,539% del totale, e uso medio 133,351.

I secondi 2678 lemmi hanno uso totale 9010,15, pari al 2,461%, e uso medio 3,364.

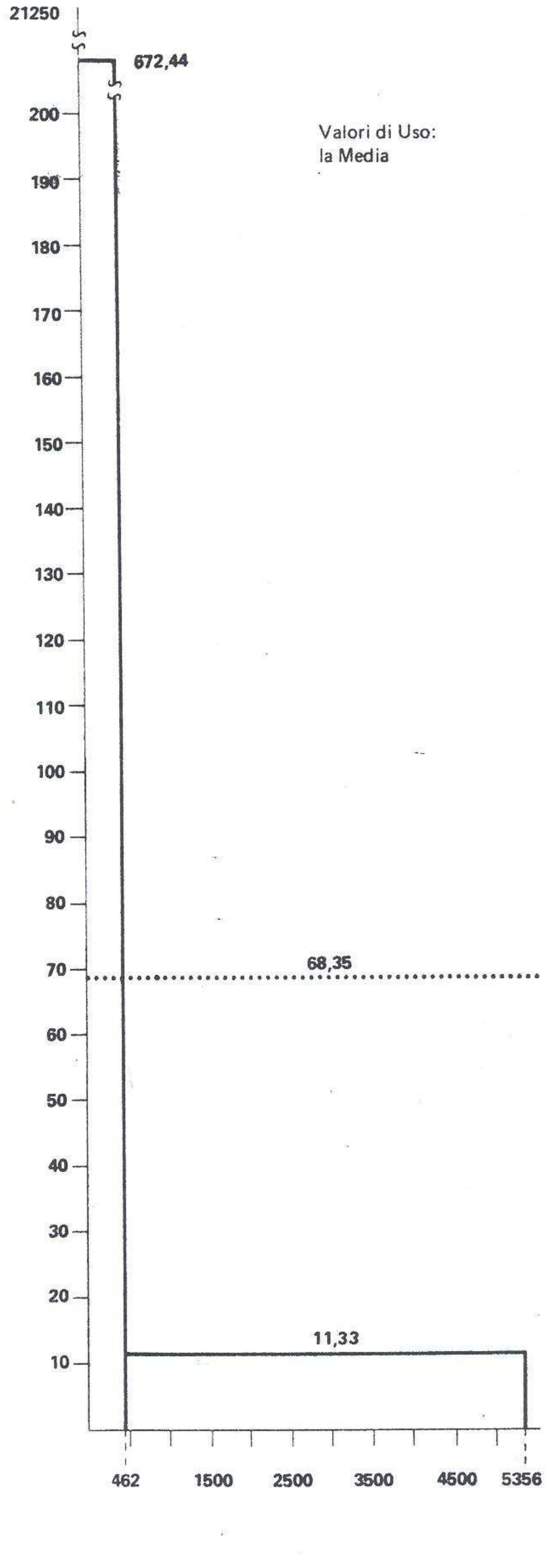


Valori di Uso:
le Metà

7.7.2 Media

Se l'uso totale 366.124,50 è diviso per 5356 lemmi, si ha uso medio 68,357. Ci sono 462 lemmi, pari all'8,626% del totale, con uso maggiore di 68,357; essi hanno uso complessivo 310.669,53, pari all'84,854% del totale, e uso medio 672,444.

I rimanenti 4894 lemmi, pari al 91,374% del totale, hanno uso complessivo 55.454,97, pari al 15,146% del totale, e uso medio 11,331.

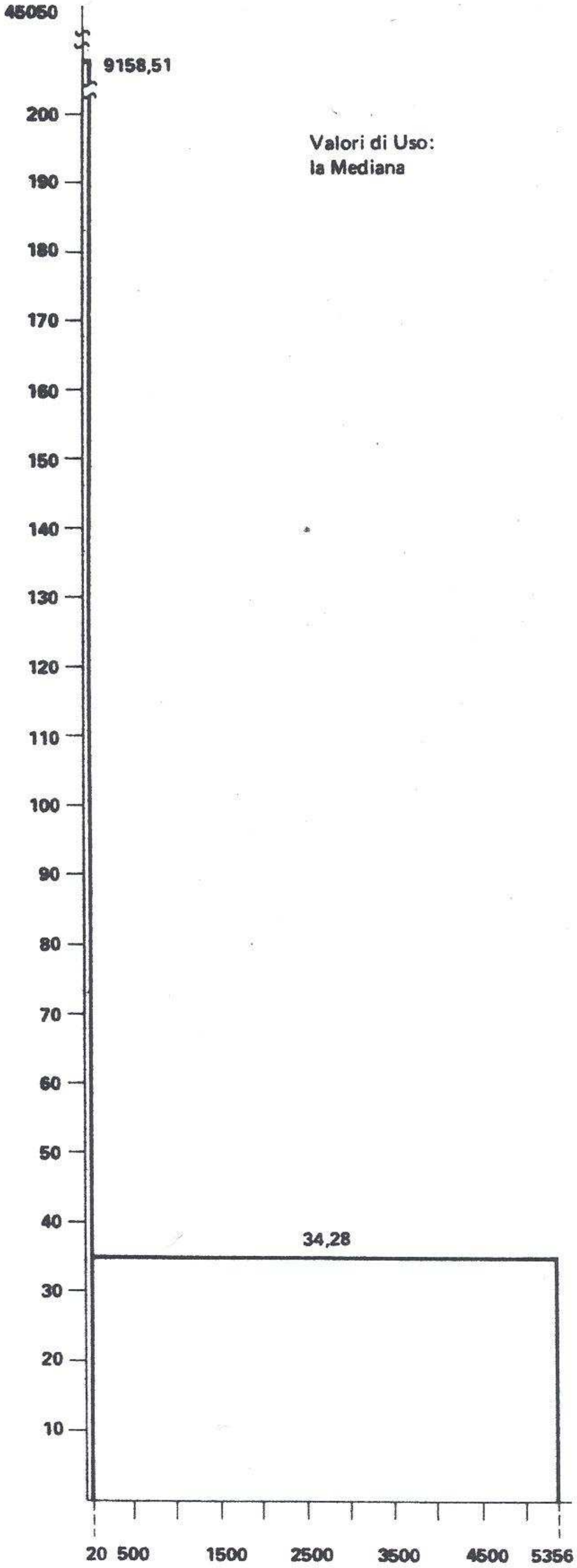


Valori di Uso: la Media

7.7.3 Mediana

Se l'uso totale 366.124,50 è diviso in due parti di 183.062,25 ciascuna, i primi 20 lemmi, pari allo 0,373%, hanno uso complessivo 183.170,37, pari al 50,030% del totale, e uso medio 9158,518.

I rimanenti 5336 lemmi, pari al 99,627%, hanno uso complessivo 182.954,13, pari al 49,970%, e uso medio 34,286.



Valori di Uso:
la Mediana

8.1 I nostri lessici, che riproducono fotograficamente gli elenchi stampati dall'elaboratore, constano di due parti e cioè:

1) *Lemmi e forme in ordine alfabetico*

In questo elenco sono riportati i lemmi in ordine alfabetico; sotto ciascun lemma sono elencate le forme relative, anch'esse in ordine alfabetico. Il lemma è contraddistinto da una sottolineatura tratteggiata ed è accompagnato da una sigla che ne specifica la categoria grammaticale come segue:

AG. aggettivo
AR. articolo
AU. ausiliare
AV. avverbio
C. congiunzione
E. esclamazione
N. numerale
PR. pronome
PZ. preposizione
S. sostantivo
V. verbo

Seguono, nell'ordine, i seguenti dati: Frequenza nel *Teatro*, Frequenza nei *Romanzi*, Frequenza nel *Cinema*, Frequenza nei *Periodici*, Frequenza nei *Sussidiari*, Frequenza *Totale*, coefficiente di *Dispersione*, indice di *Uso*.

Nello stesso ordine sono riportati poi i dati relativi a ogni singola forma, esclusa naturalmente l'indicazione della categoria grammaticale.

A titolo di esempio riportiamo qui sotto, in una tabella esplicativa, il lemma *altezza* e le sue due forme, singolare e plurale, con i dati relativi.

	Lemma e forme	Categoria grammaticale	Frequenza Teatro	Frequenza Romanzi	Frequenza Cinema	Frequenza Periodici	Frequenza Sussidiari	Frequenza Totale	Dispersione	Uso
Lemma	Altezza	S.	12	6	1	3	8	30	67.94	20.38
Forma 1	Altezza		11	6	1	3	5	26	67.59	17.57
Forma 2	Altezze		1	0	0	0	3	4	27.11	1.08

2) *Lemmi in ordine di Uso*

In questo elenco sono stampati tutti i lemmi accolti nel LIF in ordine decrescente di Uso. Accanto a ogni lemma è riportata la definizione grammaticale, poi l'indice di Uso e il rango, cioè il numero progressivo corrispondente alla posizione del lemma nell'ordinamento per Uso decrescente; nel caso che più

lemmi abbiano indice d'Uso identico, a tutti i lemmi del gruppo viene attribuito un rango convenzionale che risulta dalla media matematica tra il rango del lemma che precede il gruppo e quello del lemma che lo segue (quindi, se il numero dei lemmi nel gruppo è dispari, il rango sarà rappresentato da un numero intero corrispondente alla posizione centrale; se invece il numero è pari il rango sarà rappresentato da un numero seguito dal decimale 5), e l'ordinamento, all'interno del gruppo, è alfabetico. Seguono poi, rispettivamente nella quinta e nella sesta colonna, il rango che compete al lemma nell'ordinamento per Frequenza decrescente, e quello per coefficiente di Dispersione decrescente; anche qui, i lemmi appartenenti a gruppi con uguale Frequenza, o con uguale Dispersione, sono trattati con lo stesso criterio applicato per il rango in ordine di Uso.

Nell'edizione originale IBM Italia (fuori commercio) questa seconda parte è presentata in modo diverso: le tre liste (in ordine di Uso, di Frequenza, di Dispersione) sono stampate separatamente, su tre colonne, e per ogni lemma è riportato, dopo la sigla della categoria grammaticale, il rispettivo valore; i lemmi sono numerati in modo progressivo, indipendentemente dall'eventuale parità di due o più lemmi rispetto al valore considerato.

Nell'edizione originale vi è inoltre una terza parte « Forme in ordine di Uso, Frequenza, Dispersione » che ripete l'impostazione della seconda ma considerando le forme anziché i lemmi. Questa terza parte è omessa nella nostra edizione.