

L'esplorazione e l'analisi dei corpora

Metodi di indagine e di interrogazione

Liste di frequenza

Forma

- elenco di tutte le forme (*types*, tipi di parole)
- indici di frequenza (ossia il numero di occorrenze nel testo)
 - *frequenza relativa* (F_w/N)
 - rapporto tra le occorrenze della singola parola (F_w) e il numero di parole testuali del corpus (N)
 - *frequenza relativa normalizzata*

Presentazione

- per *frequenza decrescente*
 - al primo posto compare la parola testuale più frequente, all'ultimo la meno frequente.
- la forma che ha frequenza maggiore, e che si trova al primo posto, si dice di primo *rango*.

parole vuote

- *e, di, che, a, il, in*
- parole grammaticali

parole piene

- *Don, era*
- sostantivi, verbi, aggettivi, avverbi

Lista di frequenza del primo capitolo dei «Promessi Sposi»

255	4,1255%	e	41	0,6633%	come
195	3,1548%	di	39	0,6310%	una
162	2,6209%	che	38	0,6148%	ma
146	2,3621%	a	38	0,6148%	più
109	1,7635%	il	34	0,5501%	o
100	1,6179%	in	31	0,5015%	gli
100	1,6179%	un	28	0,4530%	don
97	1,5693%	non	28	0,4530%	da
80	1,2943%	la	26	0,4206%	due
78	1,2619%	per	25	0,4045%	se
55	0,8898%	le	24	0,3883%	poi
53	0,8575%	con	24	0,3883%	della
47	0,7604%	si	24	0,3883%	era
44	0,7119%	del	23	0,3721%	al
42	0,6795%	i	22	0,3559%	abbondio

I Frequenze assolute **II** frequenza relative **III** tipi di parole

Parole piene e vuote nei «Promessi sposi»

223.854 parole

Parole vuote 130.187 • 58%

Lemmi del vocabolario di base

- 6688 parole più frequenti, conosciute complessivamente da persone che abbiano un'istruzione pari alle medie inferiori
- 98,4% (6581 parole) è costituito da parole piene

Lemmi del vocabolario comune

- 39.700 parole note «a chiunque abbia un livello mediosuperiore di istruzione» (De Mauro 1999)
- 98,90% (39.265 parole) sono parole piene

Le fasce di frequenza

Fascia alta

- (poche parole): dalla frequenza massima

Fascia media

- (abbastanza poche parole): dalla prima coppia di parole della stessa frequenza

Fascia bassa

- (molte parole): occorrenze basse e *hapax legomena* (dalla prima parola dal basso che salta almeno un valore: 1, 2, 3, 3, 4, 4, 4, 5, 6, 8)

I lessici di frequenza

Lessici di frequenza

- liste lemmatizzate organizzate in ordine di frequenza decrescente
- permettono di osservare la distribuzione dei lessemi in relazione alle forme testuali che assumono nei testi
- forniscono un quadro delle principali fasce di uso dei lessemi e della loro copertura testuale

Impieghi

- sviluppo di risorse per la didattica delle lingue
- ricerca di lessicologia statistica
- produzione di dizionari-macchina per l'NLP (*Natural Language Processing*)
- integrazione di dati in applicazioni computazionali

Alcuni lessici di frequenza

Häufigkeitwörterbuch der deutschen Sprache

- curato da Friedrich W. Kaeding nel 1897
- frutto di uno spoglio manuale di testi

A Computational Analysis of Present-day American English

- di Kučera e Francis (1967)

Word Frequencies in Written and Spoken English

- cfr. Leech *et alii* 2001
- basato sull'analisi del *British National Corpus*

LIF - Lessico di frequenza della lingua italiana contemporanea

- Bortolini *et alii* 1971

COLFIS - Corpus e lessico di frequenza dell'italiano scritto

- Laudanna *et alii* 1995

LIP - Lessico di frequenza dell'italiano parlato

- De Mauro *et alii* 1993

Lemmi LIP in ordine alfabetico

A		0	0	3	1	3	7	4	abbraccio		0	10	0	0	0	10	0
E	3093								ABBRACCIO		0	2	0	0	13	15	2
a		0	0	3	1	3	7	4	S	3678							
A		2099	2500	2486	2320	2596	12001	11671	abbraccio		0	2	0	0	11	13	2
Pz	5								abbraccione		0	0	0	0	2	2	0
@a@		1	3	1	0	4	9	5	ABETE		0	0	5	0	0	5	0
@ar@		0	1	0	0	0	1	0	Cg	6505							
a		1320	1715	1379	1224	1643	7281	6860	abete		0	0	5	0	0	5	0
ad		34	32	137	141	113	457	337	ABILE		0	2	2	0	0	4	2
agli		22	10	27	28	14	101	84	Ag	4493							
ai		38	31	80	79	95	323	262	abile		0	0	1	0	0	1	0
al		296	235	329	323	302	1485	1413	abili		0	0	1	0	0	1	0
all'		123	122	188	167	116	716	649	abilissimo		0	2	0	0	0	2	0
alla		159	141	241	277	195	1013	893	ABILITA'		5	3	1	0	1	10	6
alle		89	200	82	51	103	525	401	S	2516							
allo		17	10	22	30	11	90	72	abilita'		5	3	1	0	1	10	6
A		5	9	7	12	16	49	40	ABILITARE		0	4	0	0	0	4	0
S	667								V	6505							
a		5	9	7	12	16	49	40	abilitarmi		0	1	0	0	0	1	0
ABBANDONARE		2	6	7	13	1	29	19	abilitata		0	1	0	0	0	1	0
V	1202								abiliti		0	2	0	0	0	2	0
abbandona		0	0	2	2	0	4	2	ABILITATO		0	4	0	0	0	4	0
abbandonando		0	0	1	2	0	3	1	S	6505							
abbandonano		0	0	0	2	0	2	0	abilitati		0	4	0	0	0	4	0
abbandonare		0	1	0	4	0	5	1	ABILITAZIONE		0	7	0	1	0	8	1
abbandonarsi		0	0	0	0	1	1	0									
abbandonata		2	0	2	2	0	6	4									
abbandonato		0	1	0	1	0	5	1									

I vocabolari fondamentali

Fascia

- massimo uso tra le fasce in cui si può suddividere il lessico di una lingua
- elaborazione metodologicamente più valida e precisa dei cosiddetti *word books*
- parole più usate in una lingua e costruiti come ausili nella didattica

An Italian Word List from Literary Sources

- Knease 1931-1933

Der grundlegende Wortschatz des Italienischen

- Migliorini 1943

Vocabolario di base

- De Mauro 1980

Il vocabolario di base (De Mauro 1980)

FO: fondamentale; tra i lemmi principali, sono così marcati 2049 vocaboli di altissima frequenza, le cui occorrenze costituiscono circa il 90% delle occorrenze lessicali nell'insieme di tutti i testi scritti o discorsi parlati

AU: di alto uso; sono così marcati 2576 vocaboli di alta frequenza, le cui occorrenze costituiscono un altro 6% circa delle occorrenze lessicali nell'insieme di tutti i testi scritti o discorsi parlati

AD: di alta disponibilità; sono così marcati 1897 vocaboli, relativamente rari nel parlare o scrivere, ma tutti ben noti perché legati ad atti e oggetti di grande rilevanza nella vita quotidiana (*alluce, batuffolo, carrozzeria, dogana, ecc.*)

I vocaboli fondamentali, di alto uso e di alta disponibilità (quest'ultimo è il gruppo più esposto al variare della cultura materiale e richiede aggiornamenti relativamente frequenti), costituiscono nell'insieme il «vocabolario di base» (De Mauro 1999a, XX).

**Lessico
fondamentale**

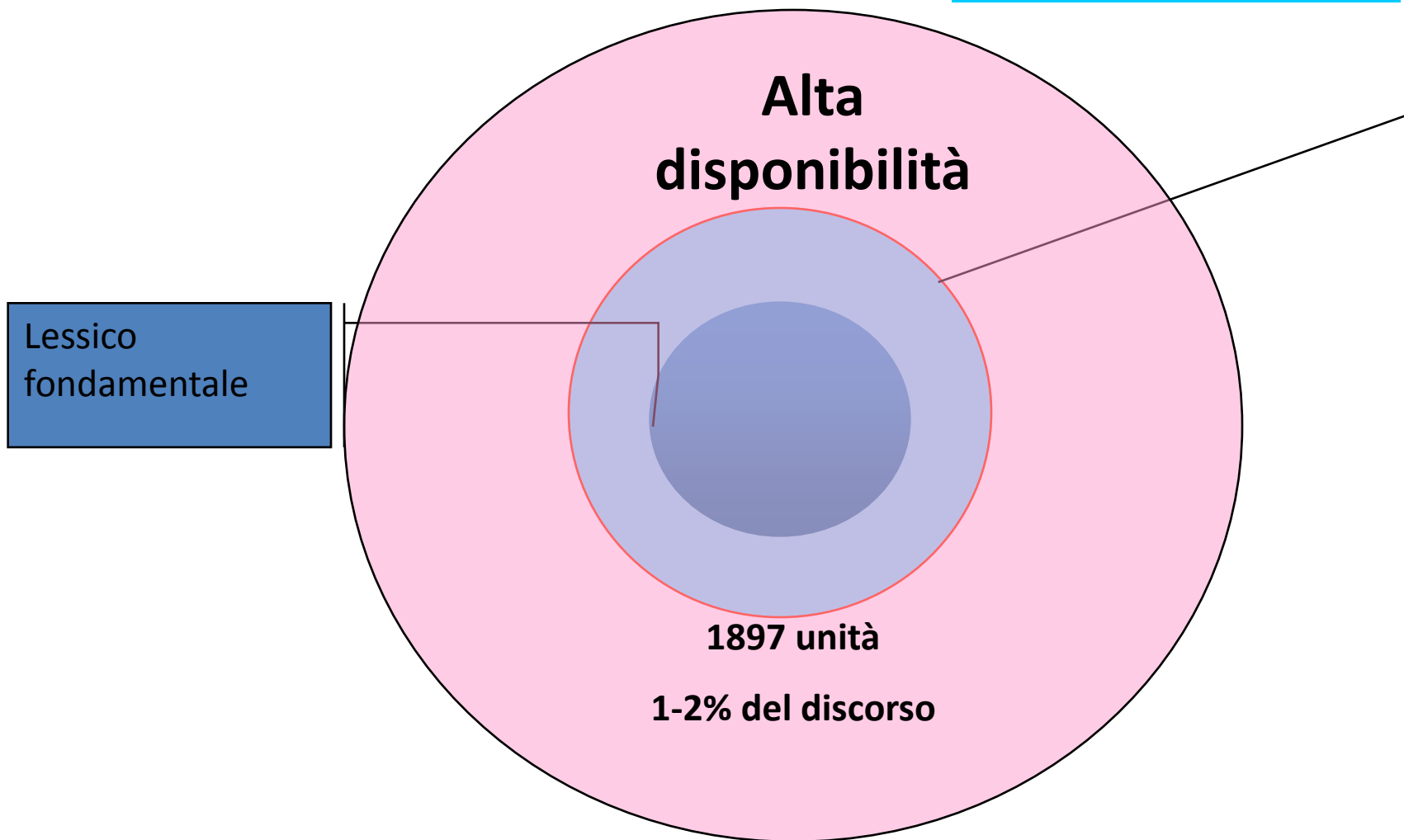
**2049 unità di
massima
frequenza**

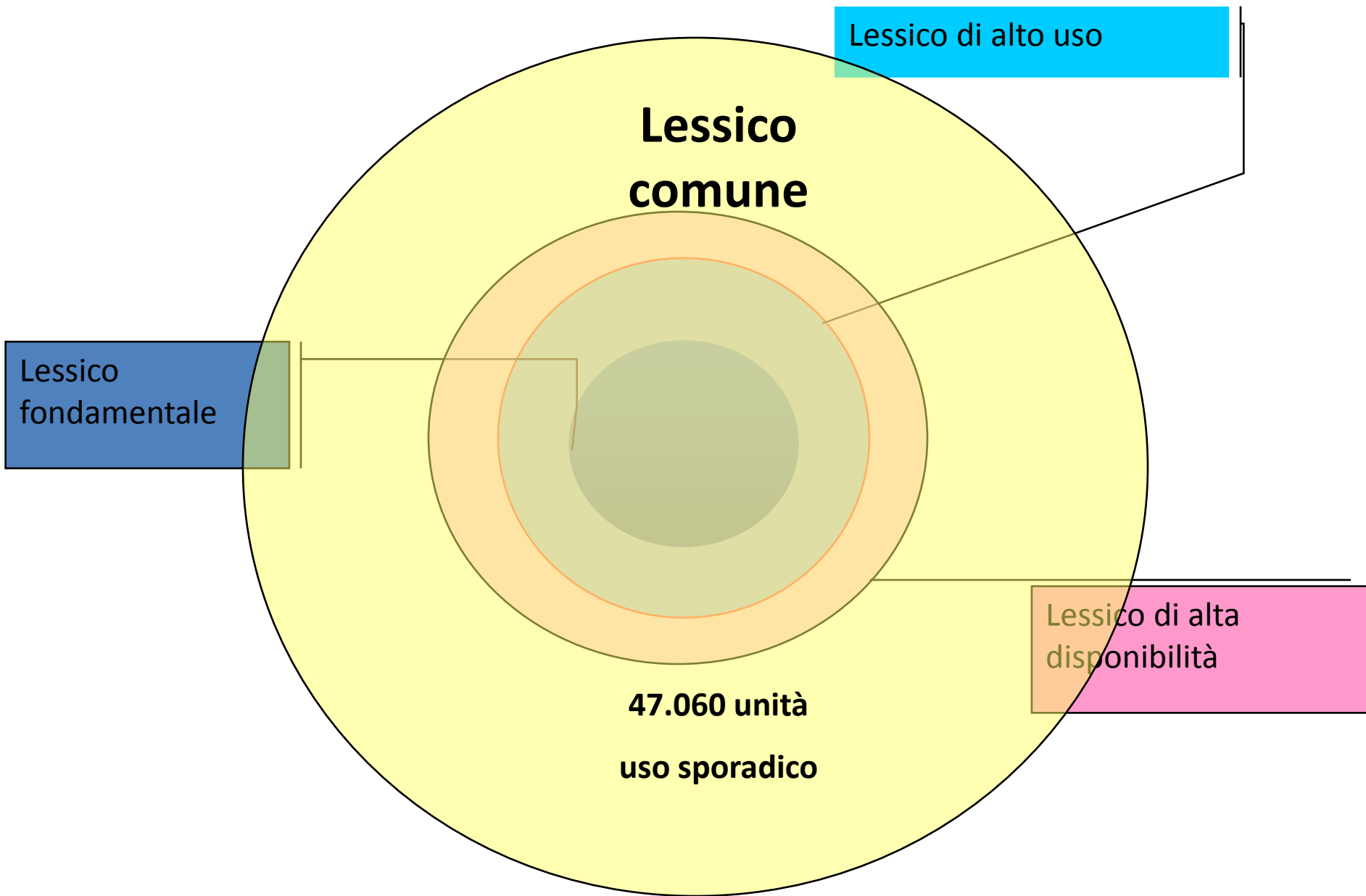
**90 % del
discorso**

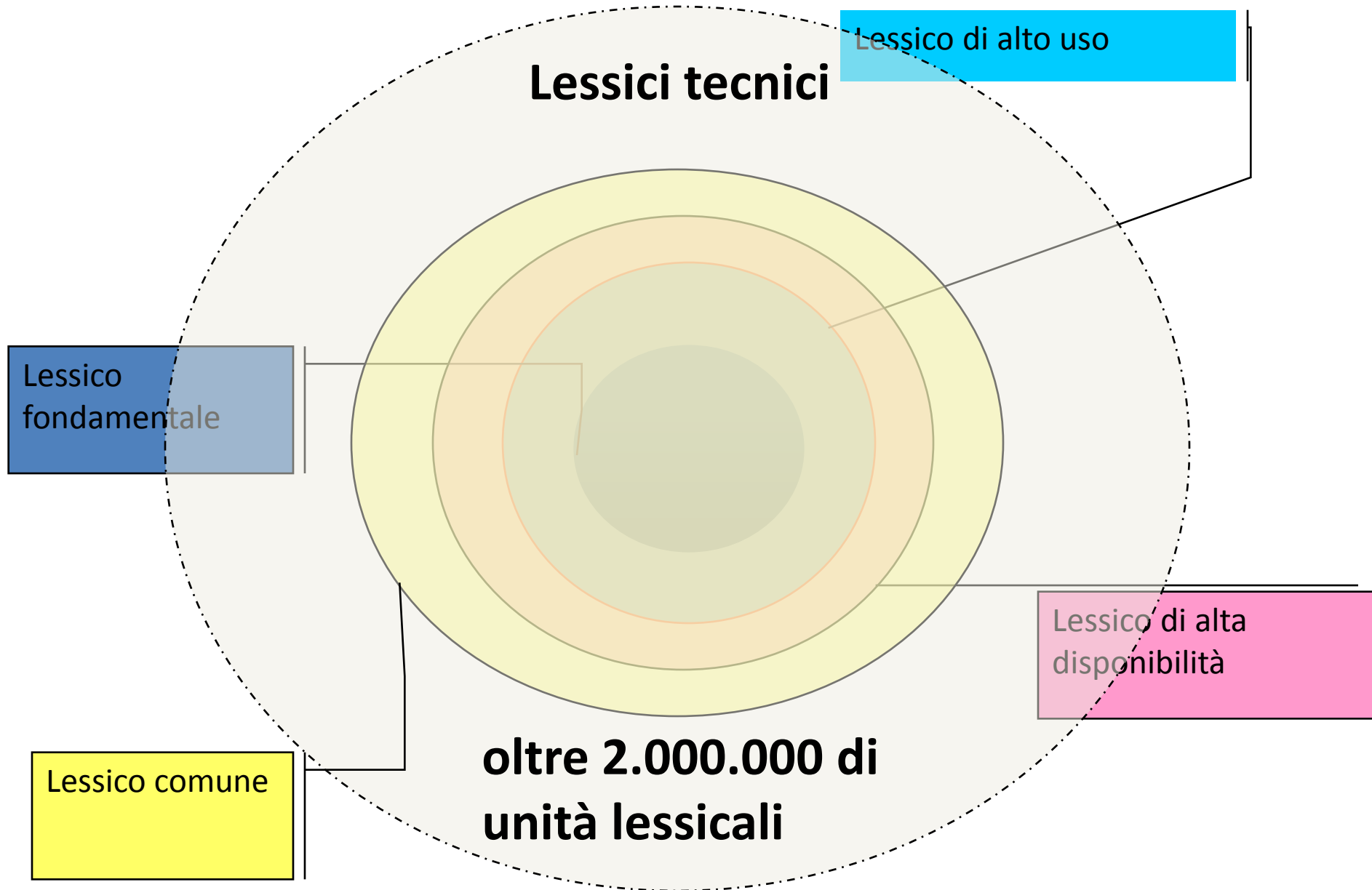


Da slides Massimo
Vedovelli

Lessico di alto uso







Lessico fondamentale +

Lessico di alto uso +

Lessico di alta disponibilità
=

Vocabolario di base

6522 parole, 98% del discorso

«Tagging»

Il *part-of-speech* (POS) *tagging*

- ossia l'etichettatura per categorie grammaticali
- un *tagger* è il dispositivo computazionale che opera un POS *tagging* su materiale testuale

Il *tagger* riceve in input una frase e restituisce in output le forme grafiche delle parole accompagnate da etichette che segnalano la categoria grammaticale di appartenenza

Le etichette applicabili sono definite da un insieme detto *tag-set*

«Tagging»

INPUT

- *la rapidità dello stile e del pensiero vuol dire soprattutto agilità*

Processing

- POS *TAGGER*

OUTPUT

- la: DET il
- rapidità: N rapidità
- dello: PRE del
- stile: N stile
- e: C e
- del: PRE del
- pensiero: N pensiero
- significa: V significare
- soprattutto: AVV soprattutto
- agilità: N agilità

Tagger «rule-based» e probabilistici

Un tagger basato su regole

- fonda la sua capacità di attribuzione della categoria grammaticale sull'accesso a una grammatica
- nella quale sono state formalizzate le regole di formazione dei diversi possibili sintagmi di una data lingua.
- i problemi principali di questo tipo di metodologia sono, da una parte, la complessità nella descrizione della grammatica necessaria, i tempi laboriosi, la necessità di avere in input solo frasi ben formate, e, dall'altra, l'impossibilità di risolvere ambiguità strutturali.
- TAGGIT, adoperato negli anni Settanta per etichettare il *Brown Corpus of Standard American English*
- 77% delle occorrenze

Un tagger di tipo probabilistico

- è basato su statistiche di frequenza delle parti del discorso e delle loro sequenze.
- *training corpora*
- CLAWS (Constituent Likelihood Automatic Word-tagging System) sul British National Corpus
- 96-97% delle occorrenze

La lemmatizzazione dei testi

Operazione

- ridurre le forme flesse di uno stesso lessema a una forma di citazione (lemma)
- la lista di frequenza conterrà solo le diverse forme di citazione come entrate: *essere, fare, libro, ecc.*
- *disambiguazione degli omografi*

Il	DET:def	Il
dottore	NOM	dottore
mi	PRO:pers	mi
raccomandò	VER:remo	raccomandare
di	PRE	di
non	ADV	non
ostinarmi	VER:infi	ostinarsi
a	PRE	a
guardare	VER:infi	guardare
tanto	ADV	tanto
lontano	ADJ	lontano

Lemmatizzazione

Funzione

- mettere in evidenza la relazione lessicale tra le parole
- osservare i lessemi e non le forme testuali

Ostacoli

- omografi
- polirematiche
- collocazioni
- ambiguità sintattiche

Lemmatizzatori automatici

- strumenti che usano diverse tecniche (disambiguazione sintattica per regole o statistico-probabilistica) per distinguere le forme omografe e riconoscere la struttura sintattica della frase

Interrogazione avanzata dei dati testuali

Ricerche avanzate

- sistema di estrazione di dati che sfrutta la capacità di combinare diversi criteri in modo da rispondere a interrogazioni che riguardano a un tempo diversi aspetti delle unità in analisi
- procedura di scoperta dei fatti linguistici

statistica testuale

- Frequenza / dispersione / Uso
- Indici di specificità
- Indici di associazione

La dispersione

Che cos'è?

- la dispersione ci indica se e dove vi sono **concentrazioni di occorrenze** nel corpus e/o in diverse tipologie testuali
- la dispersione dà un'immagine più precisa e corretta del modo con cui le parole compaiono nel corpus e serve per la determinazione dell'**uso** delle parole
- per valutare la dispersione degli elementi è necessario **suddividere il corpus** in parti (per lunghezza o per tipologia)

Metodi

- ci sono diverse formule per il calcolo della dispersione, una delle più note è il **coefficiente D**
 - linguista francese Alphonse Juilland

Il coefficiente D

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

v

- il *coefficiente di variazione* (rapporto tra la deviazione standard della frequenza – σ – e la frequenza media, dunque $v = \sigma/f_{media}$)

n

- il numero di testi diversi di cui è composto il corpus

La dispersione

- è sempre un numero **inferiore a 1** (maggiore quanto maggiore è il numero di testi in cui compare la parola)

L'uso

$$U = Df$$

Che cos'è l'uso?

- **l'uso** indica il modo con cui l'unità occorre nel corpus

Come si calcola?

- si calcola moltiplicando la **frequenza** per la **dispersione**
- maggiore è la dispersione (il valore di D si avvicinerà a 1), maggiore diventa la corrispondenza tra uso e frequenza
- il tasso di uso sarà sempre un valore inferiore o uguale al valore della frequenza, tanto inferiore quanto più la parola si trova in un numero basso di testi diversi

Liste di frequenza

	F (tot)	D	USO	STAMPA	NARR.	PR. ACC.	PR. GIUR.	MISC.	EPHEM
Internet	20.048	0,307	6159	1502	19	1354	1077	14.215	1880
Fax	2708	0,528	1428	618,0	54,6	180,0	105,0	890,0	860
E-mail	2127	0,229	487	152,5	12,3	70,0	67,5	1025,0	800
Flash	567	0,674	382	235,5	56,2	28,0	0,0	167,5	80
Jeans	371	0,533	198	100,5	188,5	20,0	5,0	42,5	15
Relax	290	0,541	157	64,0	20,0	16,0	2,5	87,5	100

La ricchezza del vocabolario

Type/token ratio

- V/N
 - V è il vocabolario del testo
 - N è la lunghezza in numero di parole

Hapax

- V_{hapax}/N
 - V_{hapax} è il numero delle parole che occorrono una sola volta nel testo

Coefficiente di Guiraud

$$G = \frac{V}{\sqrt{N}}$$

Esempio «Promessi sposi»

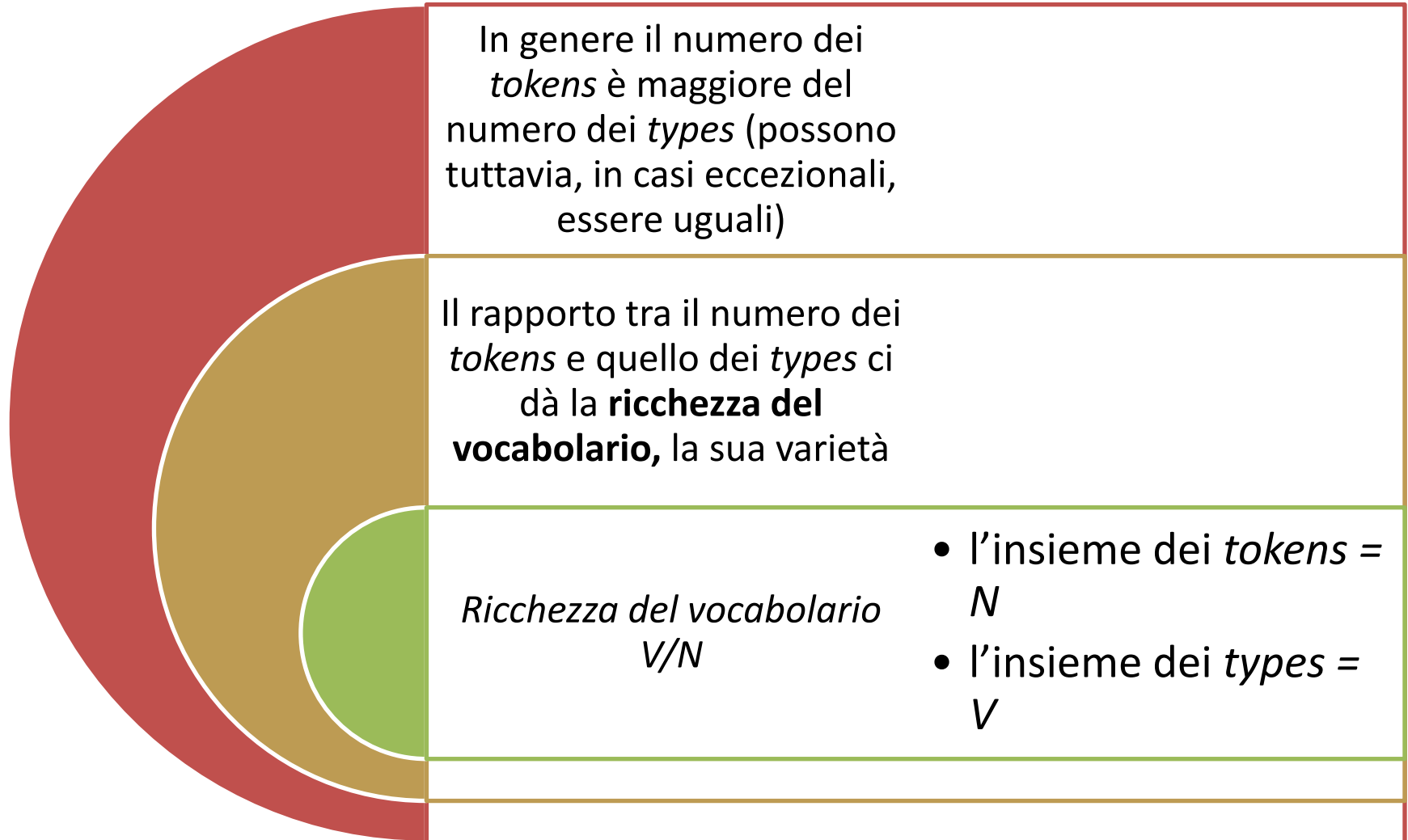
📖 Quel ramo del lago **di** Como, che volge a mezzogiorno, tra due catene non interrotte **di** monti, tutto a seni e a golfi, a seconda dello sporgere e del rientrare **di** quelli, vien, quasi a un tratto, a restringersi, e a prender corso e figura **di** fiume, tra un promontorio a destra, e un'ampia costiera dall'altra parte; e **il** ponte, che ivi congiunge le due rive, par che renda ancor più sensibile all'occhio questa trasformazione, e segni **il** punto in cui **il** lago cessa, e **l'**Adda ricomincia, per ripigliar poi nome **di** lago dove le rive, allontanandosi **di** nuovo, lascian **l'**acqua distendersi e rallentarsi in **nuovi** golfi e in **nuovi** seni.

Il testo contiene **116** parole testuali/grafiche (*tokens*)

- la congiunzione *e* occorre 10 volte
- le preposizioni *a* e *di*, rispettivamente 8 e 6 volte, ecc.

76 tipi di parole (*types*)

Il rapporto tra «types» e «tokens»



La ricchezza del vocabolario

Maggiore è il risultato di questo rapporto, maggiore è la ricchezza del vocabolario

- se $N=1000$ e $V=50$, il rapporto *types/tokens* è = 0,05, dunque poca varietà
- se $N=1000$ e $V=250$, il rapporto *types/tokens* è = 0,25, c'è dunque molta varietà

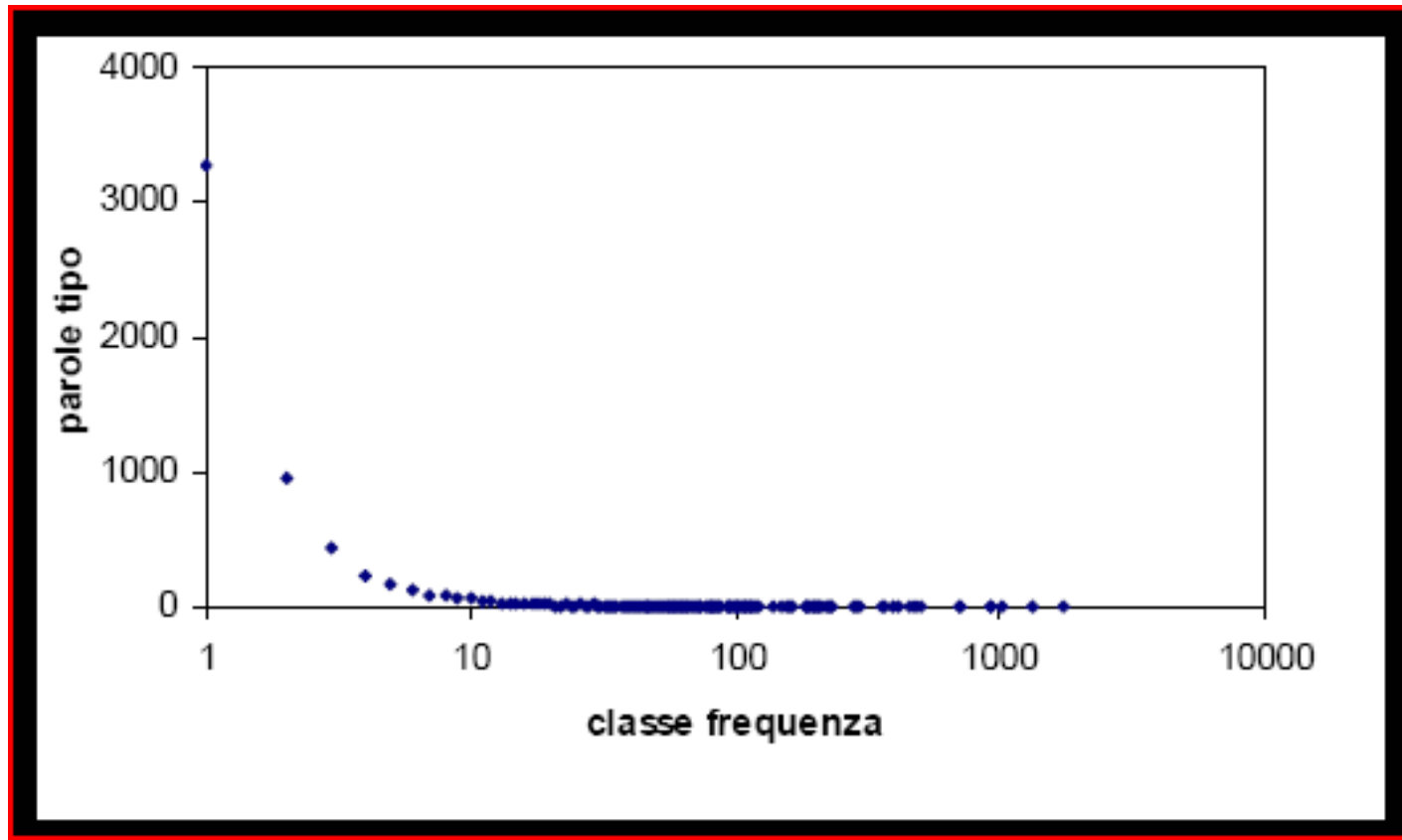
Il rapporto *token/types* del primo paragrafo dei *Promessi Sposi* è = 0,65

Ricchezza del vocabolario attraverso gli hapax legomena

$$V_{\text{hapax}}/N$$

- V_{hapax} è il numero delle parole che occorrono una sola volta nel testo
- Si osserva che in genere ci sono tante parole che occorrono una sola volta
- Questo è un indicatore della varietà lessicale
- Numero molto alto di eventi rari (Baayen)

Spettro delle frequenze lessicali di un testo “Pinocchio” (da Lenci et al 2005)

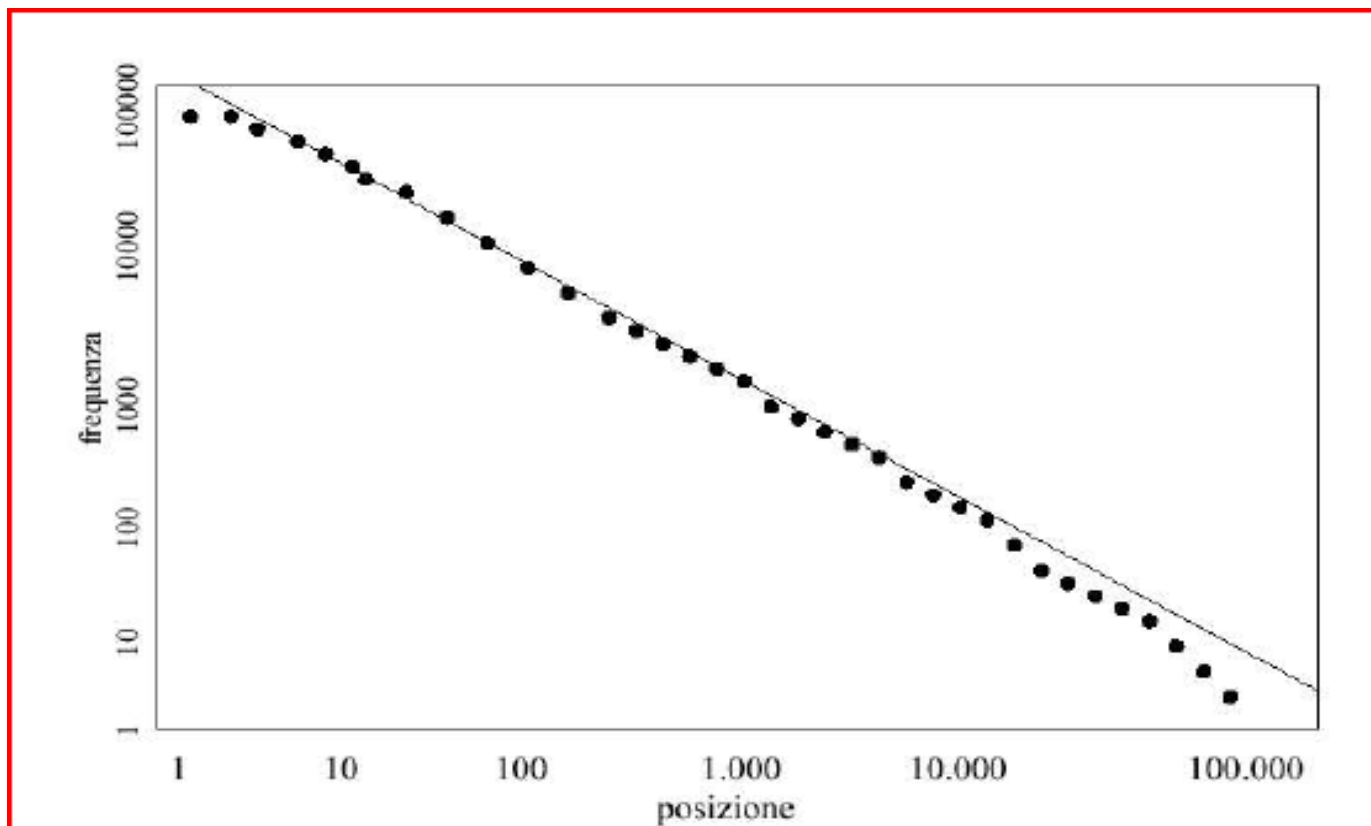


- sull'asse delle X
*le classi di freq
per valori
crescenti*

- sull'asse delle Y
*quante parole
tipo hanno
frequenza $i = |$
 $V_i |$*

La legge di Zipf

$$F \times \text{rango} = C$$



Conseguenze Zipf (da Lenci slides)

Le parole non si distribuiscono in maniera “normale” in un corpus

ci sono sempre poche parole molto frequenti

- corrispondono solitamente a parole appartenenti a “classi chiuse”
- (articoli, preposizioni, congiunzioni, ecc.)

Ci sono sempre moltissime parole a bassa frequenza e hapax

(LNRE, Large Number of Rare Events)

- sono parole “piene” (nomi, verbi, ecc.), solitamente estremamente
- informative sul contenuto di un documento

il vocabolario è aperto

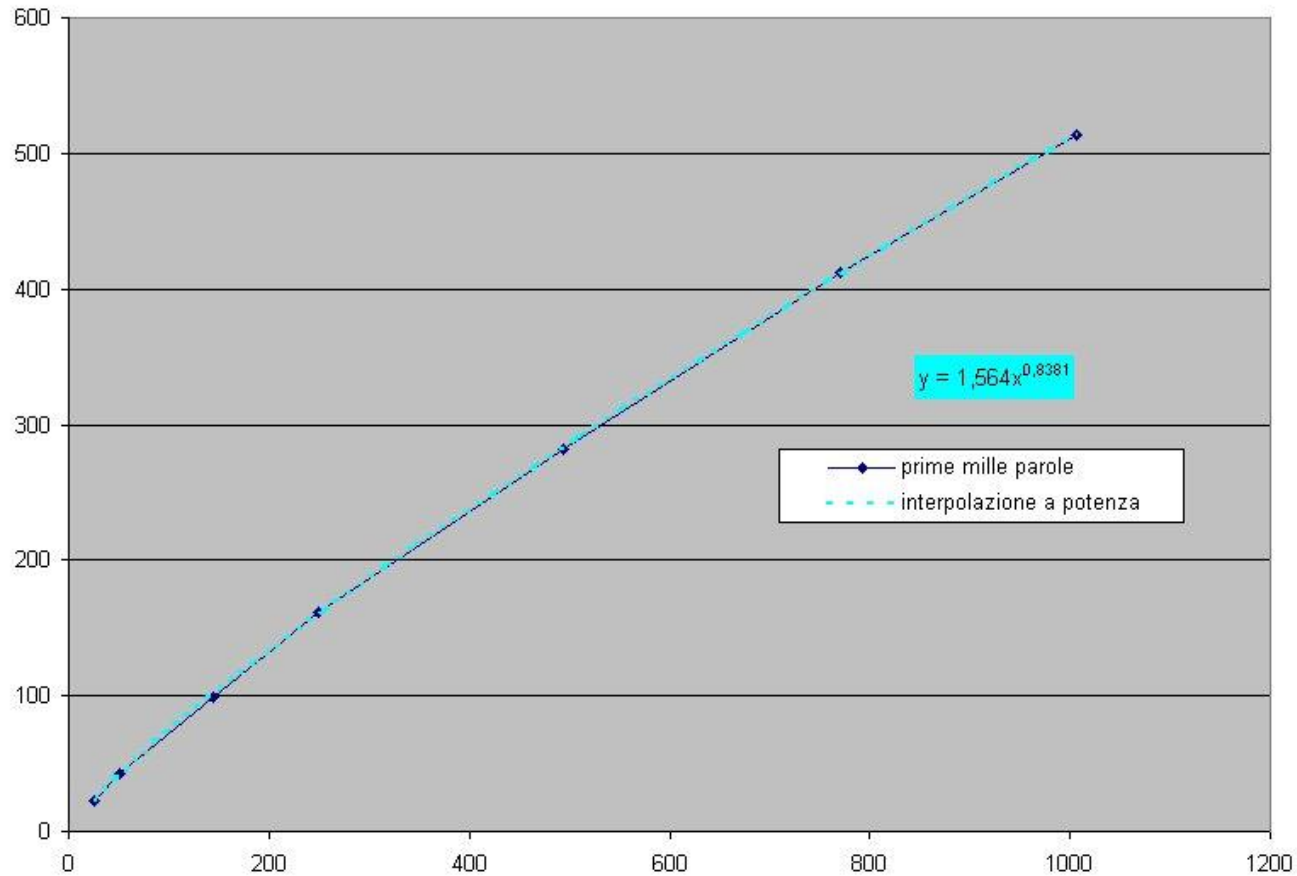
- nuovi temi e concetti portano a introdurre nuove parole
- produttività lessicale
- nuovi termini
- derivati morfologici, ecc.

Crescita del vocabolario (da Pirrelli slides)

- *il lessico di un testo* cresce quando introduciamo nel testo una parola mai usata prima
- intuitivamente la crescita di un lessico è rapida all'inizio, in quanto ogni parola che usiamo ha la tendenza ad essere nuova (raramente ci sono ripetizioni nella stessa frase)
- aumentando il numero di frasi, tuttavia, aumenta la probabilità di riusare parole già usate
- il ritmo di crescita del lessico di un testo tende quindi a diminuire all'aumentare del numero di frasi ...

da Pirrelli, continua...

crescita lessico



da Pirrelli, continua

- esistono classi di parole che è praticamente impossibile non ripetere all'interno di un testo anche molto breve
- queste classi sono formate dalle cosiddette parole “grammaticali” (articoli, preposizioni, ausiliari ecc.), che costituiscono l'impalcatura morfosintattica di una frase
- queste classi sono, tipicamente,
 - ◆ relativamente ristrette (contengono pochi elementi)
 - ◆ e “chiuse”, cioè non sono soggette ad espandersi attraverso processi produttivi del lessico come la [derivazione](#) o la [composizione](#)

Esempi letterari

Software Taltac2

Pinocchio

- Pretrattamento
- Parsing
- Fasce di frequenza
- Misure lessicometriche

www.alphabit.net

Liste di frequenza

- Wordsmith Tools
- Concordance
- AntConc
- ParaConc (corpora paralleli) MonoConc
- CONCAPP

POS Tagging

- TreeTagger