

Concordanze e collocazioni

Analisi del testo letterario 1

Isabella Chiari

Analisi degli usi con le concordanze

Il cotesto

- estrazione di informazioni linguistiche essenziali sugli usi della parola
- individuazione delle sequenze di parole che occorrono più abitualmente
 - *a guisa di, restare con un palmo di naso, giacenza di cassa*

Le concordanze

- è la presentazione delle parole di un testo, con l'indicazione della frequenza con la quale la parola occorre e il contesto linguistico precedente e successivo (cotesto).

Funzioni

- osservare i diversi usi di una parola
- esaminare i diversi contesti (semantici, sintattici o testuali) in cui occorre una parola
- analizzare la regolarità con la quale una parola è accompagnata ad altre nel suo cotesto

Concordanza di «anima» nella «Divina Commedia»

- [In.1.122] anima fia a ciò più di me degna
[In.2.45] l'anima tua è da viltade offesa
[In.2.58] O anima cortese mantoana
[In.3.88] E tu che se' costì, anima viva
[In.3.127] Quinci non passa mai anima buona
[In.5.7] Dico che quando l'anima mal nata
[In.6.55] E io anima trista non son sola
[In.10.15] che l'anima col corpo morta fanno
[In.12.74] saettando qual anima si svelle
[In.12.90] non è ladron, né io anima fuia
[Pu.4.3] l'anima bene ad essa si raccoglie
[Pu.18.44] e l'anima non va con altro piede

Presentazione delle concordanze

KWIC (*keyword in context*)

- la **parola chiave** (*keyword*) è la parola di cui si cerca l'uso, solitamente si trova nella **colonna centrale**.
- il cotesto (precedente e successivo) è stabilito dall'utente:
 - n° fisso di parole (3 – 3, ecc.)
 - frase o verso

Concordanze complete

- concordanze per tutte le parole del corpus
- voluminose e lente, più complesse

Concordanze specifiche

- concordanze per specifiche *keywords*
- veloci e facili

Esempio – software di concordanza

Concordance - LIPFOREIGNCONC.Concordance

File Text Search Edit Headwords Contexts View Tools Help

Headword	No.	Context...	Word	...Context	Line	Referen
chiacchierato	3	D: ah un hotel tre stelle ma non si	chiama	un hotel tre stelle sai	505	<F FA>
chiacchiere	8	A: signora non mi ricordo come si	chiama	riconosce che la signora	2272	<F FA>
chiacchieri	1	B: la sua mamma come si	chiama	?	2559	<F FA>
chiacchierona	1	ci no in cima insieme con lui c'era an...	chiama		3118	<F FA>
chiaia	1	/ mi ero rassegnata a non sapere n...	chiama	Paolo e m'ha	4771	<F FB>
chiam	4	come si	chiama	?	5330	<F >
chiama	187	settantacinque cinquantacinque ven...	chiama	da fuori	5722	<F >
chiamale	1	troppo cos' genera come si	chiama	troppo ampio il	5733	<F >
chiamala	2	mi	chiama	a i portatile? [sistenteunavocedallo...	5855	<F >
chiamalo	4	C: come si	chiama	lei?	6793	<F >
chiamami	11	B:	chiama	e anche al nero delle volte oltre tutt...	7188	<F >
chiamamo	1	A: si' come si	chiama	?	7509	<F FB>
chiamando	5	allora questa ragazza si	chiama	XYZ scritto XYZ	7901	<F FB>
chiamano	52	A: l'altro mi dimentico sempre come si	chiama	l'amico # va be' eh me	8154	<F FB>
chiamar	1	come si	chiama	la tua agenzia ?	9272	<F FC>
chiamarci	2		chiama		9713	<F FC>
chiamare	55	del meccanismo che si	chiama	scala mobile comunque eh eh c'era	10250	<F FC>
chiamarla	8	A: e la filos...	chiama		10583	<F FC>
chiamarle	2	A: eh come si	chiama	?	10805	<F FC>
chiamarlo	6	A: Forattini si	chiama	Forattini	10834	<F FC>
chiamarmi	3	storici e ha appunto promosso ques...	chiama	treno	10872	<F FC>
chiamarono	1	A: di? ahah...	chiama		10979	<F FC>
chiamarti	3	A: pero' si	chiama	reato d' opinione	11301	<F FC>
chiamarvi	1	riguarda la promulgazione e la pubbl...	chiama	anche	11681	<F FD>
chiamasse	2	questa economia che quindi fa capo...	chiama	economia	11968	<F FD>
chiamata	33	sostituzione # # si	chiama	cosi' perche' quando idealmente il n	12215	<F FD>

84% Resources C: 79%

Words	Tokens	At word	Word sort	Context sort
27930	522120	7094	Asc alpha (string)	Asc occurrence order

Tipologie di ricerca

Per forma specifica

- (parola testuale, forma flessa)
 - *psicologico*: si ottengono tutti i token di questo type

Per lemma

- *psicologico* (su corpus lemmatizzato): si ottengono tutte le occorrenze di tutte le forme flesse

Con l'uso di caratteri jolly (*wildcards*)

- *psicologic**: si ottengono tutte le forme flesse e tutti i derivati

Con le espressioni regolari

- *psicologic[aoih]?*: si ottengono tutte le forme che hanno uno dei caratteri tra parentesi [x] dopo la sequenza, più 0 o un carattere (esclude i derivati)

L'individuazione delle collocazioni e polirematiche

Collocazioni

- *prescrivere una ricetta, richiedere un ricovero ospedaliero*

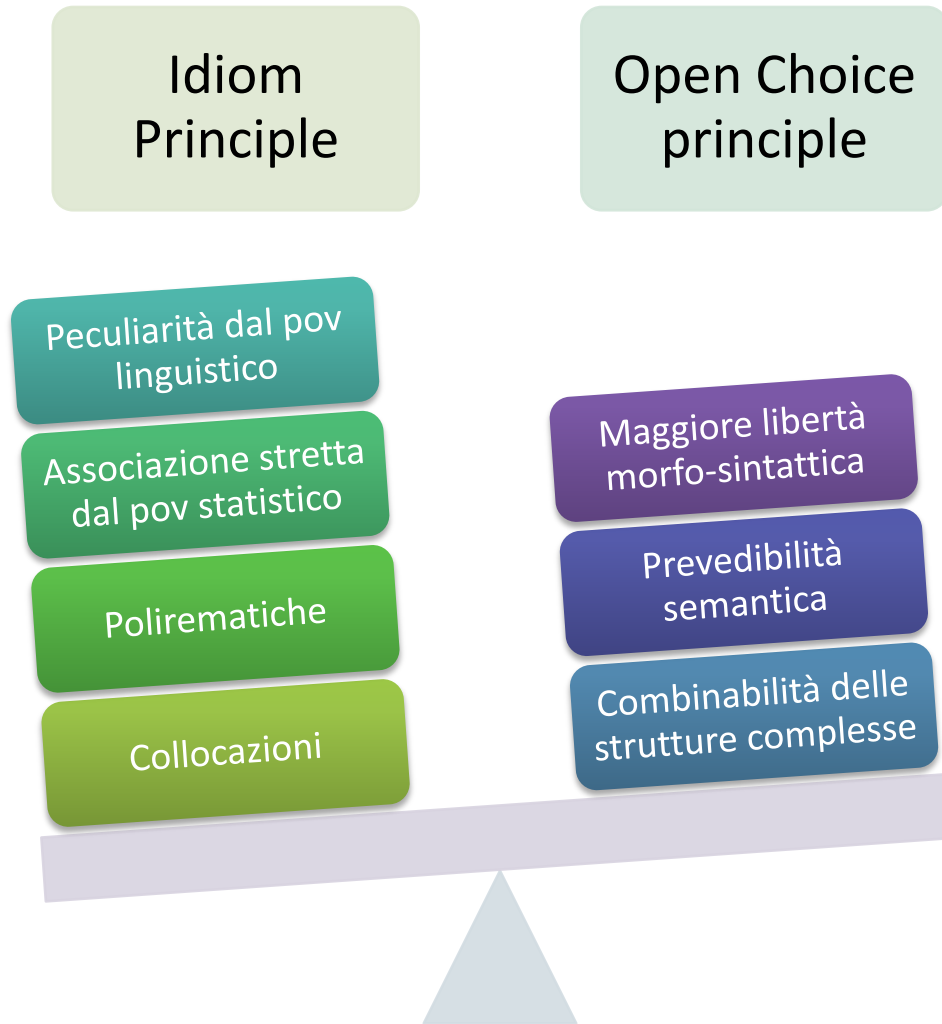
Polirematiche

- *damigella d'onore, navigazione a vista*

In una lista di frequenza vogliamo:

- contare separatamente le collocazioni dei lemmi semplici
 - il verbo *vedere* rispetto alla locuzione *vedere rosso*
- estrarre le caratteristiche statistiche delle parole che entrano in una locuzione

Bilanciamento (Sinclair)



Misure statistiche delle collocazioni

Intensità del legame

- diversi gradi di intensità nel legame tra due o più parole che co-occorrono in un testo
- procedure di estrazione automatica delle collocazioni

$$MI = \log \frac{O_{11}}{E_{11}}$$

Metodi

- la **mutual information** confronta la co-occorrenza effettiva di una coppia di parole con il valore di co-occorrenza che le due parole avrebbero casualmente
- lo **Z-score** e il **T-score** utilizzano il rapporto tra le co-occorrenze e la deviazione standard

Dove O sono le frequenze osservate ed E sono le frequenze attese di coppie di parole organizzate in tavole di contingenza

Limiti

- lingue a morfologia ricca (italiano, russo, ecc.)
- disambiguazione degli omografi

Esempi candidati collocazioni “rosso”

	<u>Freq</u>	<u>T-score</u>	<u>MI</u>	<u>MI3</u>	<u>log likelihood</u>	<u>min. sensitivity</u>	<u>salience</u>
<u>p/n</u> Brigate	5586	74.730	13.009	37.904	95134.098	0.035	112.242
<u>p/n</u> globulo	2315	48.108	12.918	35.271	38891.802	0.014	100.081
<u>p/n</u> giallo	4588	67.656	9.748	34.075	53306.711	0.028	82.188
<u>p/n</u> bianco	7969	89.084	8.909	34.829	83241.010	0.041	80.033
<u>p/n</u> semaforo	1572	39.626	10.811	32.047	20683.628	0.010	79.574
<u>p/n</u> blu	3709	60.822	9.579	33.293	42176.744	0.023	78.731
<u>p/n</u> colore	7741	87.777	8.742	34.578	79021.952	0.036	78.277
<u>p/n</u> vino	6042	77.564	8.870	33.991	62704.963	0.037	77.227
<u>p/n</u> scribanare	261	16.154	13.497	29.553	4838.353	0.002	75.155
<u>p/n</u> khmer	393	19.820	12.301	29.538	6115.273	0.002	73.515
<u>p/n</u> CORTONA	265	16.277	13.145	29.245	4594.235	0.002	73.394
<u>p/n</u> verde	4959	70.229	8.527	33.079	49039.637	0.031	72.560
<u>p/n</u> toga	591	24.299	11.011	29.425	7951.559	0.004	70.288
<u>p/n</u> bandiera	2378	48.672	9.032	31.463	25170.808	0.015	70.222
<u>p/n</u> cappuccetto	202	14.211	12.964	28.280	3412.168	0.001	68.880
<u>p/n</u> nero	4753	68.692	8.109	32.539	44212.337	0.023	68.659
<u>p/n</u> capello	2293	47.778	8.807	31.134	23541.408	0.014	68.153
<u>p/n</u> primula	334	18.270	11.716	28.484	4863.257	0.002	68.119
<u>p/n</u> rubino	656	25.594	10.474	29.189	8298.745	0.004	67.951
<u>p/n</u> croce	1679	40.901	9.095	30.521	17912.853	0.010	67.542
<u>p/n</u> cappuccettare	168	12.960	13.108	27.893	2896.522	0.001	67.244
<u>p/n</u> filo	2543	50.293	8.543	31.168	25165.675	0.016	66.991
<u>p/n</u> gambero	717	26.753	10.136	29.108	8716.243	0.004	66.662

Perché sono importanti le collocazioni?

In lessicografia computazionale servono, ovviamente, a estrarre espressioni da lemmatizzare come voce autonoma

Nella traduzione automatica servono per individuare traducenti cristallizzati e arricchire le banche dati terminologiche

Sono opportune nel *Natural Language Processing* per operare corrette analisi sintattiche e anche nella generazione linguistica

L'estrazione di collocazioni è utile inoltre nell'*information retrieval*

...nella disambiguazione dei sensi di una parola

...nel riconoscimento e nella sintesi del parlato

Siti di interesse

Collocations

- <http://www.collocations.de/>

UCS toolkit

- <http://www.collocations.de/software.html>

Collocate Finder

- <http://ell.phil.tu-chemnitz.de/collCollect/user/nph-index.cgi>

ConcGram

- <http://www.edict.com.hk/pub/concgram/>

Testo: “Codice penale”: fase 1

Testo grezzo

- 73.101 token
- 5.626 types
- Omografi 49,1%
- Sconosciuti: 5,83%

Forma grafica	Occorrenze totali	Lunghezza	CAT
amministrazione della giustizia	2	31	N
presidente della Repubblica	13	27	N
codice di procedura penale	6	26	N
Pubblica Amministrazione	10	24	N
ministro della Giustizia	5	24	N
comunicazione di massa	1	22	N
nell'adempimento degli	2	22	PREP
in conseguenza della	1	20	PREP
a disposizione della	1	20	PREP
nell'interesse della	4	20	PREP
Corte costituzionale	34	20	N
Camera dei Deputati	1	19	N
per quanto riguarda	4	19	PREP
per quanto concerne	1	19	PREP
polizia giudiziaria	4	19	N

Testo
normalizzato
da TALTA2

- 70.447 token
- **5.749** types
- Omografi 46,2%
- Sconosciuti: 1,42%

Tabelle e concordanze con più entrate

porre	2 05	V	VER:infi		inf_pres_inc	porre
porta	6 05	V	VER:fin		indic/impera	portare
porta	1 05	N	NOUN		s_f	porta
portante	1 08	V	VER:ppre		part_pres_s	portare
portare	1 07	V	VER:infi		inf_pres_inc	portare
portata	1 07	V	VER:ppast		part_pass_s	portare
portatore	2 09	N	NOUN		s_m	portatore
Portatori	1 09	N	NOUN		pl_m	portatore

Trovate 4 fg.

Forma grafica	Occ	ID Fr...	Intorno sinistro	Forma grafica	Intorno destro
porta	6	Inter...	sicurezza personali ; 4) quando la sentenza straniera	porta	condanna alle restituzioni o al risar
portare	1	Inter...	sostituita con l ' ergastolo . Art. 242 Cittadino che	porta	le armi contro lo Stato italiano Il cit
portante	1	Inter...	porta le armi contro lo Stato italiano Il cittadino che	porta	le armi contro lo Stato , o presta se
portata	1	Inter...	Usurpazione di titoli o di onori Chiunque abusivamente	porta	in pubblico la divisa o i segni distir
		Inter...	licenza dell ' Autorità , quando la licenza è richiesta ,	porta	un ' arma fuori della propria abitazi
		Inter...	fuori della propria abitazione o delle appartenenze di essa ,	porta	un ' arma per cui non è ammessa l

Interrogazione testuale

Concordanze semplici

- Tengono conto del lemma e del tag grammaticale

Espressioni regolari

- Permettono ricerche su sequenze di categorie grammaticali (abbinata a lemmi specifici, categorie personalizzate dall'utente, e LAG)

Permettono ricerche di:

- Strutture linguistiche grammaticali (sequenze NN, N Agg, ecc. V Prep)
- Comportamenti di lessemi (LEMMA + condizioni grammaticali)
- Analisi di specifiche classi (profilo grammaticale)

Profilo grammaticale del testo

Trovate 41 fg. Mostra tutte le fg *

Salva lista Concordanze del gruppo CONG

Forma grafica	Occ
o	2482
e	797
se	383
ovvero	347
Se	123
od	63
ma	47
ed	41
anche se	40
salvo che	39
senza	27
sia	19
prima che	15
nel momento in cui	10
nonché	9
sempre che	8
allorché	7
dopo che	6
dopo	5
in modo che	4
senza che	4
sul fatto che	4
mentre	3
un fatto che	3
a meno che	2
anziché	2
dal momento in cui	2
fino a quando	2
fra	2
il fatto che	2
oltre che	2

Trovate 2 fg. Mostra tutte le fg *

Salva lista Concordanze del gruppo FORM

Forma grafica	Occ
di cui all'articolo	6
per evitare che	1

salvo che i modi	1
salvo che il condannato	1
salvo che il decreto	2
salvo che il fatto	9
salvo che il giudice	3
salvo che la legge	7

La pena (Nome)

+ prep
articolata
+ NOUN

Espressione regolare da cercare (o incipit)

"LEMMA(pena) CATAAC(ARTPRE) CATAAC(NOUN)"

+
aggettivo

"LEMMA(pena) CATAAC(ADJ)"

pena dell' ammenda	4
pena dell' arresto	7
pena dell' ergastolo	20
pena della multa	4
pena della reclusione	32

pena accessoria	7
pena alternativa	1
pena detentiva	25
pena diversa	1
pena eccedente	1
pena medesima	1
pena ordinaria	1
pena pecuniaria	17
pena prevista	6
pena principale	2
pena privativa	3
pena restrittiva	12
pena scontata	1
pena stabilita	25
pena superiore	1
pena temporanea	1
pena unica	4

Pena (N)

Verbo + lag2 +
LEMMA(pena)

Espressione regolare da cercare (o incipit)

"CATGR(V) LAG2 LEMMA(pena)

	occorrenze		occorrenze	
aggiunte a pena	1	estingue la pena	2	
aggiunte a una pena	1	estinguere la pena	1	
aggiunto alla pena	1	ferma la pena	1	
aggravano la pena	1	importa la pena	2	
applica la pena	23	importa una pena	1	
applica soltanto la pena	2	infligga una pena	2	
applica una pena	1	infliggersi la pena	2	
commesso a pena	1	inflitta congiuntamente	1	
commesso, a pena	1	inflitta la pena	4	
comportano la pena	1	inflitta soltanto una pen	1	
condannato a pena	1	inflitta una pena	2	
condannato a una pena	1	iniziata della pena	1	
condannato ad una pena	1	opera sulla pena	1	
condannato alla pena	1	ordini che la pena	1	
consegue una pena	1	prevede la pena	1	
considerano come pena	1	punito con la pena	1	
considerate come pena	1	sconta la pena	1	
cumulata alla pena	1	scontare una pena	2	
determinare la pena	1	scontata la pena	1	
determinata, la pena	1	soggiacciono alla stessa	3	
diminuire la pena	2	soggiace alla pena	5	
è prevista la pena	1	soggiace soltanto alla pe	1	
escludono la pena	3	Sono applicabili alla pen	1	
eseguita la pena	1	stabilisca una pena	1	