Metodi linguistici di analisi di testi

Leggere i corpora
(2023/2024)

Giovanni Urraci giovanni.urraci@unipd.it

ALCUNI STRUMENTI

AntConc [https://www.laurenceanthony.net/software/antconc]

Voyant tools [https://voyant-tools.org]

NoSketch Engine e **Sketch Engine** [https://www.sketchengine.eu]

Tree Tagger [https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger]

READ-IT [https://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest]

ESPLORAZIONE DEI CORPORA

Linguistica computazionale, analisi dei dati testuali,

linguistica dei corpora, linguistica quantitativa...

↓

Etichette diverse, descrivono la stessa tipologia di approccio e

muovono dallo stesso oggetto di studio → il **corpus**.

Consentono di maneggiare efficacemente corpora di grandi dimensioni,

e di porre ai testi domande originali.

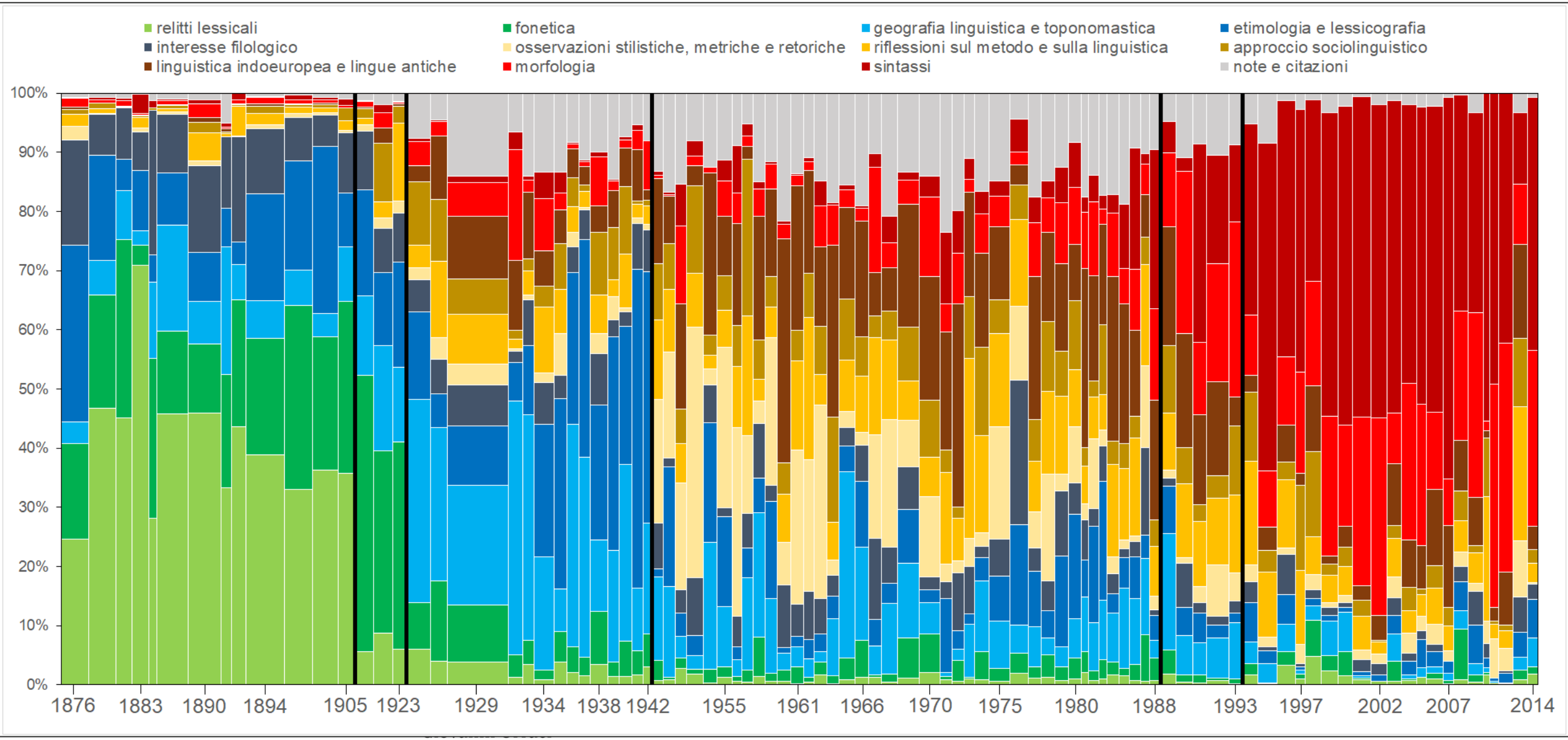
Tuttavia, non costituiscono un'alternativa ai tradizionali approcci qualitativi

↓

complementarietà

ANALISI STATISTICA DEI DATI TESTUALI

Profilo tematico dell'«Archivio Glottologico Italiano». Topic extraction, metodo di Reinert



NOZIONI FONDAMENTALI

Corpus raccolta omogenea di testi selezionati per la loro rappresentatività (di una lingua, una varietà, un dominio, una tipologia testuale). Oggi in formato digitale.

Possono essere di lingua scritta o trascritta, sincronici o diacronici, specialistici o 'di riferimento', di apprendenti, eccetera.

Es.: discorsi di insediamento dei presidenti della repubblica italiana; articoli pubblicati su la Repubblica tra il 1985 e il 2000.

Sub corpus porzione di un corpus: singolo testo o insieme dei testi che condividono una o più proprietà pertinenti ai criteri oggetto di analisi.

Indispensabili → l'analisi quantitativa si fonda sulla comparazione, necessaria per identificare le caratteristiche di uno specifico gruppo.

Es.: discorso di insediamento del presidente Sergio Mattarella; articoli di materia economica pubblicati su la Repubblica tra il 1985 e il 2000.

ALCUNI CORPORA

- KIParla – corpus di parlato spontaneo
- UniverS-ITA – testi formali scritti da studenti universitari
- la Repubblica (1985-2000)
- OpenSubtitles (2018) – raccolta di sottotitoli
- itTenTen20
- itWaC

Siti Internet e testi raccolti online

CREAZIONE CORPUS

La costruzione di un corpus varia a seconda delle variabili coinvolte e del software utilizzato.

Soluzione più frequente: testi in formato .txt (UTF-8), un documento per ogni sub corpus, raccolti in una cartella.

****001 *Rivista=AGI *Anno=1876

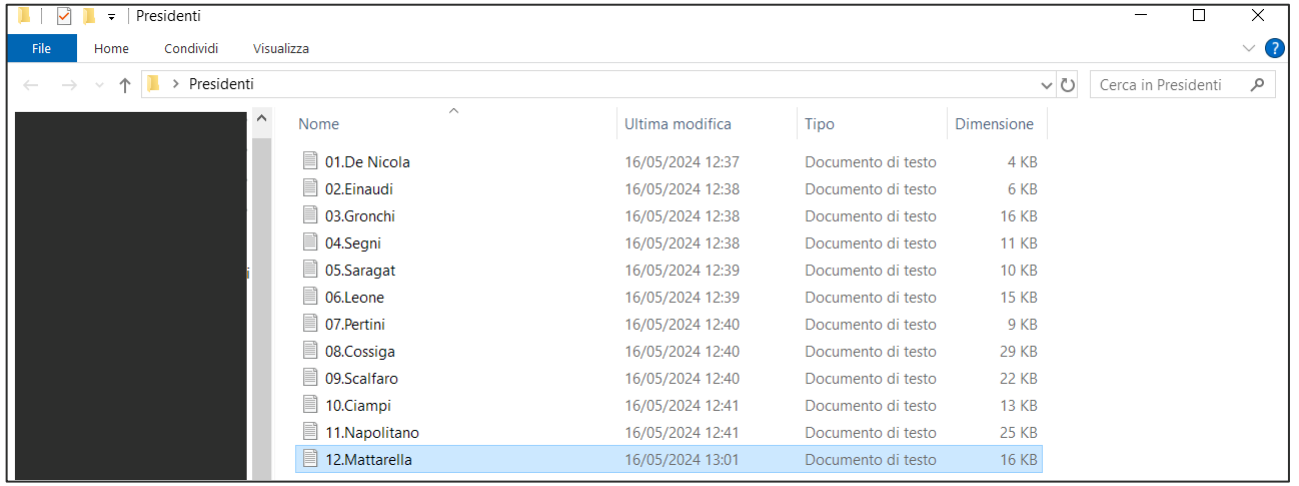
Postille etimologiche.
Saggio di un Glossario Modenese ossia studii del conte Giovanni Galvani intorno le probabili origini di alquanti idiotismi della città di Modena e del suo contado. Scrissi le seguenti postille etimologiche quattro e più anni sono; e le scrissi principalmente coll'intento di mettere per così dire a fronte due scuole, la vecchia e la nuova, la scuola senza metodo e quella del metodo. Attendendo per debito d'uffizio ad insegnar glottologia nell'Ateneo torinese, mi parve che dalla pubblicazione del Galvani venissemi non solo buona occasione, ma obbligo di dimostrare come nelle cose della linguistica più non valgano gran fatto di per sé soli né ingegno, né dottrina, né squisita coltura di lettere; pregi che ni uno avrebbe potuto negare al Galvani; ma si debba innanzi tutto chiedere a quella, che ora può dirsi ed è veramente scienza delle lingue, il metodo e i principj.
[...]

****002 *Rivista=AGI *Anno=1877

Fonetica del dialetto di Val Soana (canavese).
Il dialetto di Val Soana è parlato dalla popolazione dei quattro comuni della valle di questo nome, che sono Ingria, Ronco, Valprato e Campiglia. È inoltre parlato nei due comuni di Ribordone e Frassinetto, il primo de' quali sta a destra, l'altro a sinistra della valle. La popolazione di fatto ascendeva per cotesti comuni, il 31 dicembre del 1871, a 7582 anime, distribuite come segue: . Siccome però il censimento si faceva appunto in quel tempo dell'anno in cui la popolazione virile suole emigrare dalla valle, conviene aggiungere a questa cifra, per ottenere approssimativamente lo stato della popolazione di diritto, o meglio della vera popolazione effettiva, poco meno d'un altro migliajo d'anime; così che il numero delle persone, che ha per favella materna il dialetto valsoanino, riesce all'incirca di 8,500.
[...]

****096 *Rivista=LN *Anno=1939

Correnti dotte e correnti popolari nella lingua italiana.
In quell'operetta che si può considerare il più antico dei lessici italiani, il Vocabulario di cinque mila Vocabuli Toschi non men oscuri che utili e necessarii del Furioso, Bocaccio, Petrarcha e Dante, pubblicato a Napoli nel 1536, l'autore, Fabricio Luna, cita l'aneddoto di un gentiluomo che, volendo ordinare ai suoi staffieri di accorciargli la coreggia delle staffe, si esprimeva così: O famuli, o famuli, abreviatimi questi sustentacoli, che son troppo prolissi! E nello stesso anno un bolognese, Giovanni Filoteo Achillino, nelle Annotazioni della volgar lingua, citava così le parole del gentiluomo: Agricola, abbreviami esto sostentaculo ch'è nimio prolisso. Dalle origini fino ad oggi, nella pratica e nella teoria, in modi diversi eppure con grandissime conformità, affiorano le proteste contro la lingua troppo dotta; e ci permettono di cogliere al vivo, grazie appunto alle esagerazioni della polemica, uno degli aspetti perpetui dell'italiano (che del resto riappare, mutati i termini, per tutte le lingue colte d'Europa), il confluire nella lingua di correnti popolari e di correnti dotte.
[...]



ID	Day	Month	Year	Representative	Tweet
1	31	12	2013	Renzi	In Palazzo Vecchio, al lavoro per preparare la Giunta di fine anno. Si annuncia bella corposa
2	31	12	2013	Renzi	@Fiorello Nessuno è perfetto, Fiore! Auguri
3	31	12	2013	Renzi	Ho mantenuto l'impegno di andare nella terra dei fuochi. In silenzio e senza dichiarazioni. Graz
4	30	12	2013	Renzi	Oggi a #Firenze, Eataly ha aperto nella via Martelli pedonale. Sono122 posti di lavoro. E io son
5	28	12	2013	Renzi	Grazie a tutti, buona serata. Torneremo con #matteorisponde dopo Natale
6	27	12	2013	Renzi	Pronto per il #matteorisponde Tra cinque minuti si parte…
7	27	12	2013	Renzi	Colonna sonora di domani al @pdnetwork: "Resta ribelle" dei Negrita "la tua canzone". Piace?
8	27	12	2013	Renzi	Insieme a @bobogiac nel giorno in cui finisce lo sciopero della fame. pic.twitter.com/tsNq6hOz
9	20	12	2013	Renzi	Caro @beppe_grillo ti rispondo nei prossimi giorni con una #sorpresina che ti sto preparando.
10	17	12	2013	Renzi	Scusate il ritardo nelle risposte, sono stati giorni intensi. E bellissimi. Ma il meglio deve ancora
11	17	12	2013	Renzi	Grazie.
12	17	12	2013	Renzi	Giornata difficile da dimenticare… Ci vediamo alle 22 all'ObiHall (Firenze Sud) e in streaming s
13	14	12	2013	Renzi	Grazie a tutti i volontari che consentono le primarie e ai cittadini che stanno votando. Buon voto
14	14	12	2013	Renzi	La piazza di Empoli mi resterà nel cuore a lungo. Grazie ragazzi, è davvero #lavoltabuona
15	11	12	2013	Renzi	Grazie a tutte le volontarie e i volontari che stanno prendendo freddo ai tavolini nelle mille piazz
16	11	12	2013	Renzi	Mamma mia, quanto entusiasmo. Grazie a tutti. Adesso inizia la parte difficile: uno per uno, ca
17	11	12	2013	Renzi	Ultimo miglio e poi si #cambiaverso Arriviamo a Milano. Questa è la #voltabuona, ora o mai più
18	8	12	2013	Renzi	Tra qualche minuto a Tortona con Realacci, Bastioli, Angelantoni, Ghisolfi con le proposte su a
19	8	12	2013	Renzi	Un gigante, #Mandela. Ciao #Madiba

CONCORDANZE

Voyant Tools

Contesti

Documento	Sinistra	Parola	Destra
01.De Nicola	la mia breve ma intensa missione di Capo provvisorio dello	stato	inspirandomi ad un solo ideale: di servire con fedeltà e
01.De Nicola	nel quale i diritti dei cittadini e i poteri dello	stato	siano egualmente garantiti, trarrà dal passato salutari insegnamenti, consacrerà per
02.Einaudi	trapasso avvenuto il 2 giugno dall'una all'altra forma istituzionale dello	stato	fu non solo meraviglioso per la maniera legale, pacifica del
02.Einaudi	ceto politico migliore di quello che, all'alba del Risorgimento, era	stato	fornito dal suffragio ristretto. Or qui si palesa il grande
02.Einaudi	è garanzia della libertà della persona umana contro l'onnipotenza dello	stato	e la prepotenza privata; e garantire a tutti, qualunque siano
03.Gronchi	indicazione del Parlamento il mandato per la suprema magistratura dello	stato	. Quest'attesa e quest'ansia non mi sgomentano, ma fanno più presente
03.Gronchi	quali che siano le qualità benevolmente attribuitemi, l'elemento determinante dello	stato	d'animo comune a tanta parte del popolo italiano. Ma è
03.Gronchi	I problemi fondamentali del passato sono stati la ricostituzione dello	stato	nella sua organizzazione e nella sua autorità, la ricostruzione economica
03.Gronchi	sistema produttivo; ma essi hanno già nella organizzazione politica dello	stato	moderno un'influenza che è adeguata alla loro importanza economica. Io
03.Gronchi	il suffragio universale ha condotto sino alle soglie dell'edificio dello	stato	senza introdurle effettivamente dove si esercita la direzione politica di
03.Gronchi	insieme opera di progresso e di conservazione, di intervento dello	stato	e di rispetto dell'iniziativa privata. Nuove forme di organizzazione economica
03.Gronchi	e sud che travagliano penosamente l'efficienza operativa dell'economia nazionale. Lo	stato	può dare un valido concorso a nuove forme di rapporti
03.Gronchi	della nostra Costituzione. Distinguere in questo campo la responsabilità dello	stato	da quella dell'iniziativa privata non vuol dire contrapporre le due
03.Gronchi	vigile ardimento, la trasformazioni delle strutture economiche e sociali. Allo	stato	spetta in primissima istanza la responsabilità di mantenere le condizioni
03.Gronchi	le condizioni necessarie all'ordinato sviluppo democratico della comunità nazionale. Lo	stato	è imparziale tutore dei diritti di ciascuno, della libertà, dell'uguaglianza
03.Gronchi	di un'ordinata convivenza. Non è una definizione puramente giuridica lo	stato	di diritto: è l'espressione di un'esigenza politica e sociale alla
03.Gronchi	sospetta che, dietro la facciata di quel che si chiama	stato	, si cela il giuoco di potenti gruppi organizzati". Nessuna parola
03.Gronchi	definire il carattere e le responsabilità di un Capo di	stato	per quanto riguarda la piena osservanza della Costituzione, delle norme
03.Gronchi	sinceramente accettato di soggezione alla legge ed all'imparziale autorità dello	stato	. Onorevoli deputati, onorevoli senatori! La nuova fase della nostra vita
04.Segni	piena e di imparzialità assoluta nell'esercizio della suprema magistratura dello	stato	. A Luigi Einaudi, che gli successe, va il pensiero commosso
04.Segni	a me spetta determinare gli indirizzi politici nella vita dello	stato	, prerogativa questa del governo della Repubblica e massimamente di questo
04.Segni	appropriate al bene comune. Ma a me, quale Capo dello	stato	, incombe, nell'esercizio delle mie funzioni, il dovere di tutelare l'osservanza
04.Segni	affinché sia garantita, nella forma e nello spirito dell'attività dello	stato	, l'unità civile e morale della nazione italiana, una e indivisibile
04.Segni	continuità ed unità di questa nostra Repubblica che è uno	stato	di diritto, dotato di leggi giuste ed uguali per tutti
04.Segni	equanime giustizia. Considerando quelli che sono gli organi supremi dello	stato	ed i miei doveri verso di essi, mi sia consentito
05.Saragat	dei legislatori, degli uomini di governo e dei Capi di	stato	. Ma la pace si persegue creando con tenacia e con
05.Saragat	Costituzione a uno statuto di sovranità e indipendenza accanto allo	stato	sovrano e indipendente nella sfera propria. La Repubblica democratica difende
05.Saragat	il nostro secondo Risorgimento. Le relazioni tra il Capo dello	stato	ed il governo sono fissate dalla Costituzione; e sarà nella
05.Saragat	ordinata convivenza gioverà all'intero Paese. Dell'azione governativa i funzionari dello	stato	costituiscono lo strumento fondamentale. Comprendo tutto il travaglio cui questi
05.Saragat	la soluzione di essi consentirà di dare alla macchina dello	stato	quell'efficienza che i compiti odierni richiedono in particolare misura. Agli
05.Saragat	voi mi avete affidato con la più alta magistratura dello	stato	la custodia dei supremi valori della patria. Con animo commosso
06.Leone	potrà garantire la maggiore partecipazione del cittadino alla vita, dello	stato	, caratteristica essenziale della democrazia. Coerente con la linea politica di
06.Leone	nella società tecnologica contemporanea. Per quanto riguarda i rapporti tra	stato	e Chiesa, è nella Costituzione la direttrice di operare perché
06.Leone	fanno il giusto posto alla indipendenza ed alla sovranità dello	stato	e della Chiesa cattolica, ciascuno nel proprio ordine; si tratta
06.Leone	Alle forze armate, garanzia dell'indipendenza nazionale e della sovranità dello	stato	, nelle quali i nostri giovani trovano una grande scuola di
06.Leone	senso del dovere e nel sacrificio dei molti servitori dello	stato	un punto di fiducia. Va rinnovato in questa sede l'invito
06.Leone	sua alta coscienza giuridica e in una visione austera dello	stato	accompagnò l'opera di De Gasperi per la rinascita del Paese

105 contesto espandi Riduci

CONCORDANZE (2)

Keyword In Context

CONCORDANCE

Italian parliamentary debates (ParlaMint 2.1)

simple avversari • 258
8.4 per million tokens • 0.00084%

Get more space

KWIC

	Details	Left context	KWIC	Right context
1	CorsiniPaolo • ...	udico questo Governo? Un Governo di tregua; un Governo frutto di un armistizio. Del resto, la pace si stipula tra	avversari	. Tra loro lo scontro non è così cruento come tra consanguinei. Allora, come tradurre una costrizione, uno stato c
2	SustaGianluca •...	di costruire una democrazia dell'alternanza in cui reciprocamente ci si rispetti, ci si riconosca, anche come	avversari	. Questo non passa attraverso l'annullamento delle distinzioni o la rinuncia, anche in questo periodo di conviven:
3	SchifaniRenato ...	oma , possa avvalersi di una captatio benevolentiae, di una acquisizione di consenso maggiore rispetto agli altri	avversari	in quanto parlamentare in carica e vuole dunque sfrondare il campo da questo sospetto per presentarsi ai propri
4	ScilipotiIsgroD...	del premio di maggioranza alla Camera , poiché il partito o la coalizione che ha conseguito un voto in più degli	avversari	ottiene il 55 per cento dei seggi: tramite il premio di maggioranza, la coalizione guidata da Bersani ha conseguit
5	CentinaioGianMa...	na visione laica della vita. Pur non condividendo alcune sue esternazioni, spesso pungenti nei confronti dei suoi	avversari	, rispettiamo la forza d'animo e la convinzione con cui ha sempre cercato di far comprendere il suo pensiero. La
6	BondiSandro • 2...	. I partiti, attraverso questa forma di lotta politica, sono quasi sempre inflessibili nei confronti degli	avversari	politici e indulgenti verso la propria parte politica. Vero, senatore De Cristofaro ? Quando l'onorevole Vendola ,
7	BondiSandro • 2...	i limiti del rispetto per le persone, specialmente quelle come il ministro Alfano , conosciuto da tutti in quest'Aula,	avversari	e amici, per la sua integrità morale e il suo senso di rispetto delle istituzioni. Manteniamo dunque il confronto ent
8	CasiniPierFerdi...	sul rispetto che sempre si deve nutrire per i membri del Parlamento , in particolare - per chi tali li ritiene - per gli	avversari	politici, e che possa responsabilmente compromettere l'esito di questo Governo. Lo sforzo che gli italiani stanno
9	BondiSandro • 2...	capacità, a cui lei ha fatto riferimento nel suo discorso, di comprendere le ragioni degli altri, le ragioni degli	avversari	politici? Ricordo quell'epoca, onorevole Letta , l'epoca del compromesso storico (lo ricorda come me), della
10	CastaldiGianluc...	. Signor Presidente , colleghi, sento il dovere, in coerenza con l'educazione che uso nei confronti di tutti, amici e	avversari	, nonché per i rapporti cordiali che ho intrapreso con molti di voi, soprattutto i colleghi della 10a Commissione , d
11	ScilipotiIsgroD...	che allora avevano assunto una posizione completamente diversa abbiano capito oggi che il rispetto degli	avversari	è qualcosa di importante e che la democrazia si fonda essenzialmente sul dialogo costruttivo con l'avversario.
12	RomaniMaurizio ...	torale, eternamente rimandata per interesse momentaneo. L'interesse è farla nel momento giusto per fregare gli	avversari	e mantenere la poltrona, non farla in modo giusto per garantire al popolo un'equa rappresentanza. Come possia
13	QuagliarielloGa...	si affronta questa materia bisogna innanzitutto leggere e meditare sulle parole e sulle argomentazioni dei propri	avversari	ed evidentemente trarne insegnamento e, laddove possibile, conforto. Cito quanto ha scritto qualche mese fa un
14	ZizzaVittorio •...	che consiste non nella presenza di Berlusconi sulla scena istituzionale e politica (come invece tanti suoi	avversari	affermano), ma nel fatto che in vent'anni non si è riusciti ad abbattere Berlusconi per via elettorale e si tenta di
15	MinzoliniAugust...	? Innanzitutto il ricordo indelebile di un omicidio politico, di un ritorno a quel passato drammatico in cui gli	avversari	venivano liquidati con tutti i mezzi possibili, meno che con quelli leciti della competizione politica. Ma resterà
16	GasparriMaurizi...	anche oggi fare quello che fa sempre: ignorare il mandato improvvidamente ricevuto di senatore a vita. A tutti gli	avversari	politici credo che dobbiamo dire con serenità che Silvio Berlusconi continuerà a svolgere il suo ruolo politico. No
17	ToniniGiorgio •...	approvare il provvedimento e si trovano nell'impossibilità di farlo (mi riferisco, in particolare, ai colleghi e	avversari	di Forza Italia). Allo stesso modo, sento il bisogno di rivolgere delle scuse ai colleghi delle opposizioni che sono
18	LettaEnrico • 2...	ome il passaggio da una situazione di contrapposizione tossica tra nemici a un sistema di competizione sana tra	avversari	; un passaggio obbligato dall'esito del voto di febbraio, ma, soprattutto, dalla necessità - che io giudico, oggi più
19	SacconiMaurizio...	per essere rimessa al piccolo gioco tattico o, ancor peggio, per essere utilizzata come ennesima clava contro gli	avversari	in una dialettica politica già viziata da molte diffidenze. Il metodo ragionevole è quello che muove dalle forze
20	CandianiStefano...	qualcosa. Penso all'infelice battuta che l'altro giorno ha fatto la senatrice Pezzopane quando, rivolta ai suoi	avversari	politici, ha detto: "sterminiamoli"; poi, si è corretta dicendo che intendeva dire: "asfaltiamoli". Magari fa

Rows per page: 20 1–20 of 258 1 / 13

COSTRUZIONE E ANALISI DI UN CORPUS

PRE-PROCESSING

- Tokenizzazione
- Normalizzazione

CLASSIFICAZIONE DELLE ENTITÀ D'ANALISI

- Named-entity recognition
- Segmentazione in frasi e assegnazione POS
- Lemmatizzazione e stemming

MISURE LESSICOMETRICHE ESSENZIALI

- Frequenza, rango e hapax
- Media e mediana
- TTR

INDIVIDUAZIONE DEL LESSICO RILEVANTE

- Frequenza
- Keyword
- collocazioni

PRE-PROCESSING

Fase di preparazione del corpus.

Determina validità delle operazioni successive e qualità dei risultati.

Impone decisioni legate allo scopo dell'indagine.

1. **Tokenizzazione** Segmentazione in token, le più piccole unità di analisi: sequenze di caratteri isolate da separatori.

La macchina deve essere istruita,
impostazioni né banali né scontate.
Es.: in un corpus di tweet # e @ sono rilevanti.
Oppure: cosa fare con gli apostrofi? E - ?

Distinzione fondamentale:

- word-type: parole diverse attestate in un corpus = graficamente distinte
- word-token: occorrenze

Sopra	la	panca	la	capra	canta	sotto	la	panca	la	capra	crepa	
1	2	3	4	5	6	7	8	9	10	11	12	TOKEN → 12
1	2	3	2	4	5	6	2	3	2	4	7	TYPE → 7

PRE-PROCESSING (2)

Fase di preparazione del corpus.

Determina validità delle operazioni successive e qualità dei risultati.

Impone decisioni legate allo scopo dell'indagine.

2. Normalizzazione Uniformazione grafica e pulitura del testo. Passaggio fondamentale.

Es.: maiuscolo > minuscolo (presidente ≠ Presidente)

apostrofi e accenti (perchè > perché)

uniformazione dei composti

eventuali interventi su grafie estemporanee.

Quali variazioni sono rilevanti e quali un ostacolo all'analisi?

Cosa fare con <h> etimologica e oscillazioni ortografiche?

CLASSIFICAZIONE DELLE UNITÀ D'ANALISI

1. **Named-entity recognition** Riconoscimento e categorizzazione semantica di unità lessicali mono- e polirematiche: persone (Sergio_Mattarella), luoghi (Monaco_di_Baviera), eventi (Seconda_Guerra_Mondiale), organizzazioni (ONU), professioni, eccetera. Inoltre: specifiche categorie di dominio (es.: farmaci, molecole).

   ↓

   Eseguito attraverso il confronto con liste di rifermento.

2. **Segmentazione in frasi** shallow parsing: analisi dei costituenti e individuazione dei sintagmi

   full parsing: albero sintattico completo

   ↓

   assegnazione POS

3. **Lemmatizzazione** Raggruppamento delle forme flesse sotto i lemmi di riferimento.

   Operazione complessa per testi non canonici (es. reiterazione vocalica e refusi nei testi da Internet): necessario l'addestramento su corpora pertinenti annotati.

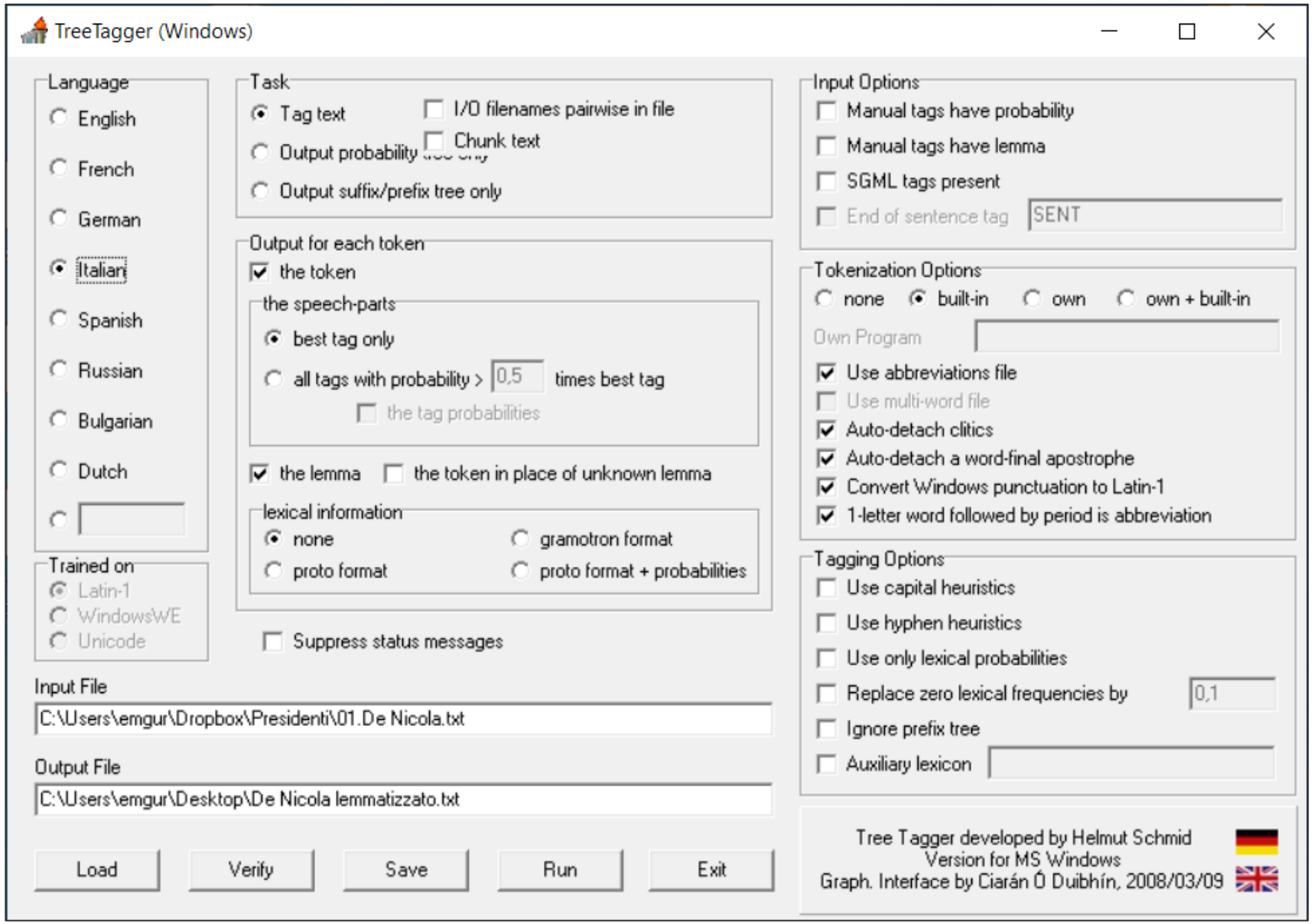
   Possibile anche lo stemming: riduzione alla radice – problematico per it.

Ambiguità risolte considerando il contesto.
Es. Marco porta una torta

TreeTagger

Demo:

https://cental.uclouvain.be/treetagger/



LEMMATIZZAZIONE (2)

	A	B	
1	FORMA	POS	LEMMA
2	Giuro	VER:pres	giurare
3	davanti	ADJ	davanti
4	al	PRE:det	al
5	popolo	NOM	popolo
6	italiano	ADJ	italiano
7	,	PON	,
8	per	PRE	per
9	mezzo	NOM	mezzo
10	della	PRE:det	del
11	Assemblea	NOM	assemblea
12	Costituente	ADJ	costituente
13	,	PON	,
14	che	PRO:rela	che
15	ne	ADV	ne
16	è	VER:pres	essere
17	la	DET:def	il
18	diretta	ADJ	diretto
19	e	CON	e
20	legittima	ADJ	legittimo
21	rappresentanza	NOM	rappresentanza
22	,	PON	,
23	di	PRE	di
24	compiere	VER:infi	compiere
25	la	DET:def	il
26	mia	PRO:poss	mio
27	breve	ADJ	breve
28	ma	CON	ma
29	intensa	ADJ	intenso
30	missione	NOM	missione
31	di	PRE	di
32	Capo	NOM	capo
33	provvisorio	ADJ	provvisorio
34	dello	PRE:det	del
35	Stato	NOM	stato
36	inspirandomi	VER:geru	inspirare
37	ad	PRE	ad
38	un	DET:indef	un
39	solo	ADV	solo
40	ideale	ADJ	ideale
41	:	PON	:
42	di	PRE	di
43	servire	VER:infi	servire
44	con	PRE	con
45	fedeltà	NOM	fedeltà

```
Italian tagset used in the TreeTagger parameter file

(Copyright Prof. Achim Stein, University of Stuttgart)

ABR     abbreviation
ADJ     adjective
ADV     adverb
CON     conjunction
DET:def definite article
DET:indef       indefinite article
FW      foreign word
INT     interjection
LS      list symbol
NOM     noun
NPR     name
NUM     numeral
PON     punctuation
PRE     preposition
PRE:det preposition+article
PRO     pronoun
PRO:demo        demonstrative pronoun
PRO:indef       indefinite pronoun
PRO:inter       interrogative pronoun
PRO:pers        personal pronoun
PRO:poss        possessive pronoun
PRO:refl        reflexive pronoun
PRO:rela        relative pronoun
SENT    sentence marker
SYM     symbol
VER:cimp        verb conjunctive imperfect
VER:cond        verb conditional
VER:cpre        verb conjunctive present
VER:futu        verb future tense
VER:geru        verb gerund
VER:impe        verb imperative
VER:impf        verb imperfect
VER:infi        verb infinitive
VER:pper        verb participle perfect
VER:ppre        verb participle present
VER:pres        verb present
VER:refl:infi   verb reflexive infinitive
VER:remo        verb simple past
```

Testo originale

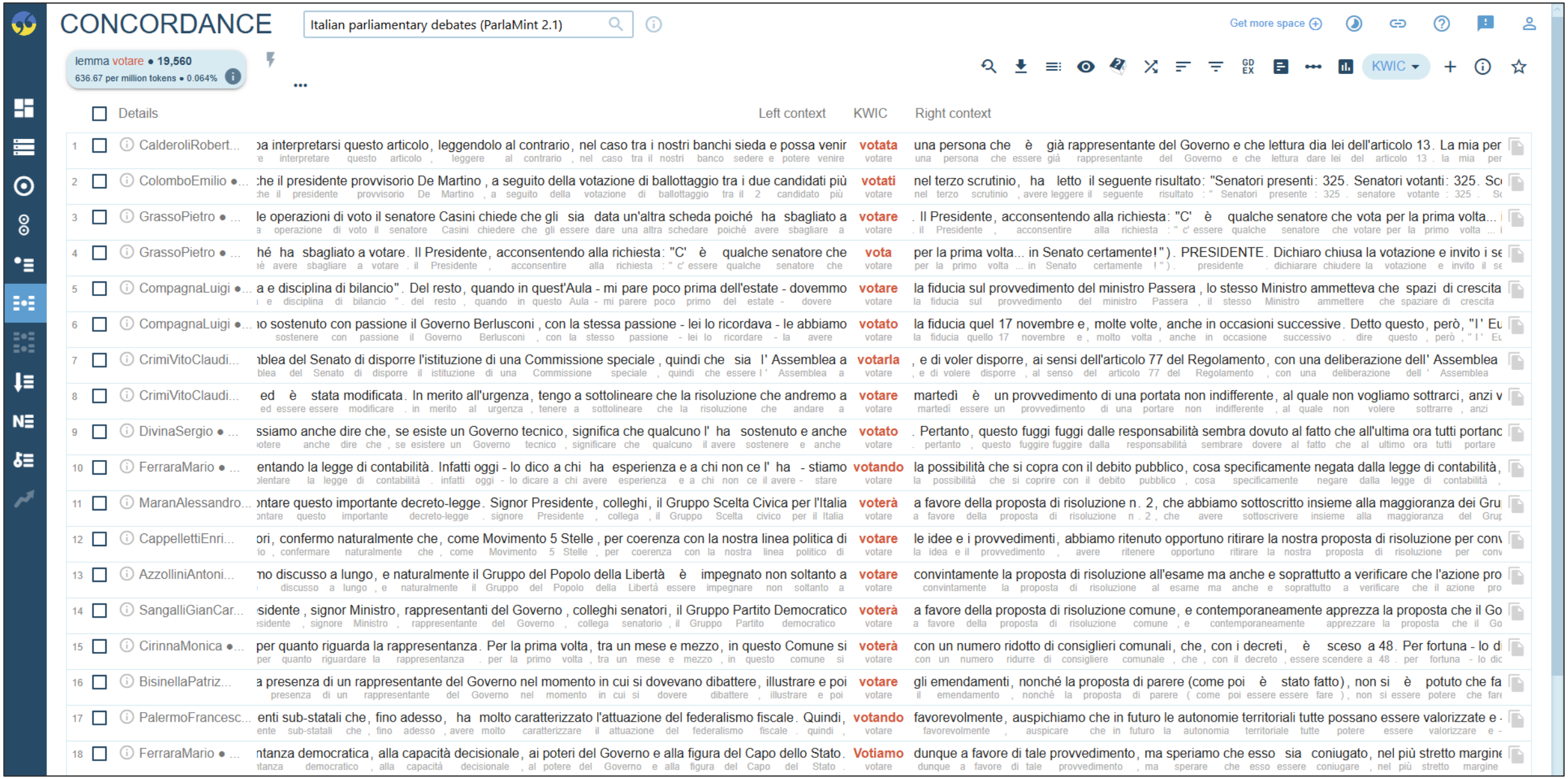
Giuro davanti al popolo italiano, per mezzo della Assemblea Costituente, che ne è la diretta e legittima rappresentanza, di compiere la mia breve ma intensa missione di Capo provvisorio dello Stato inspirandomi ad un solo ideale: di servire con fedeltà e con lealtà il mio Paese.

Per l'Italia si inizia un nuovo periodo storico di decisiva importanza. All'opera immane di ricostruzione politica e sociale dovranno concorrere, con spirito di disciplina e di abnegazione, tutte le energie vive della nazione, non esclusi coloro i quali si siano purificati da fatali errori e da antiche colpe.

Testo lemmatizzato

giurare davanti al popolo italiano, per mezzo del assemblea costituente, che ne essere il diretto e legittimo rappresentanza, di compiere il mio breve ma intenso missione di capo provvisorio del stato inspirare ad un solo ideale: di servire con fedeltà e con lealtà il mio paese. per il Italia si iniziare un nuovo periodo storico di decisivo importanza. al opera immane di ricostruzione politico e sociale dovere concorrere, con spirito di disciplina e di abnegazione, tutto il energia vivo del nazione, non escluso **colorare** il quale si essere purificare da fatale errore e da antico colpa

LEMMATIZZAZIONE (3)



MISURE LESSICOMETRICHE ESSENZIALI

Termini

	Parola	Conteggio	Relativa
1	stato	105	4,000
2	libertà	83	3,162
3	repubblica	79	3,009
4	ogni	76	2,895
5	popolo	75	2,857
6	costituzione	74	2,819
7	essere	71	2,705
8	paese	64	2,438
9	vita	61	2,324
10	politica	59	2,247
11	pace	59	2,247
12	l'italia	55	2,095
13	giustizia	55	2,095
14	sempre	52	1,981
15	parlamento	52	1,981
16	democrazia	52	1,981
17	sociale	51	1,943
18	presidente	50	1,905
19	senza	46	1,752
20	forze	46	1,752
21	nazionale	45	1,714
22	lavoro	45	1,714
23	solo	44	1,676
24	italiano	43	1,638
25	mondo	42	1,600
26	deve	41	1,562
27	responsabi…	40	1,524
28	può	40	1,524
29	valori	39	1,486
30	onorevoli	39	1,486
31	saluto	38	1,448
32	parte	38	1,448
33	comunità	38	1,448

Frequenza Assoluta: numero di occorrenze di una parola nel corpus

Relativa: rapporto tra la frequenza relativa di una parola e le dimensioni del corpus (token); può essere normalizzata moltiplicandola per una base prestabilita (standard: 1 milione)

↓

Perché è importante?

Rango Ordine nella lista delle parole ordinate secondo la loro frequenza assoluta

Distribuzione normale:

- le parole più frequenti sono parole grammaticali – è atteso, non è un risultato dell'analisi! Possono essere rimosse?
- il numero delle occorrenze declina rapidamente, già tra rango 1 e 2
- pochi elementi con frequenza altissima, molti con frequenza bassa – specialmente hapax.

↓

% hapax è un indicatore rilevante.

Dovrebbe essere <50%

MISURE LESSICOMETRICHE ESSENZIALI (2)

Media (*mean*) Somma di tutti i valori di una variabile divisa per il numero totale dei valori.

Mediana (*median*) È il valore che divide la distribuzione a metà dopo aver ordinato i valori dal più piccolo al più grande.

TTR Rapporto tra type e token, è un indice di ricchezza lessicale. Esprimibile come una percentuale moltiplicandolo per 100.

Risponde alla domanda: quanto è vario il lessico utilizzato in un testo? → Più è alto il valore, più il lessico è ricco.

Documents

	Title	Words	Types	Ratio	Words/Sentence
1	01.De Nicola	549	342	62%	34.3
2	02.Einaudi	825	464	56%	55.0
3	03.Gronchi	2,357	1,015	43%	37.4
4	04.Segni	1,684	670	40%	31.2
5	05.Saragat	1,530	662	43%	30.0
6	06.Leone	2,094	904	43%	30.3
7	07.Pertini	1,338	615	46%	20.6
8	08.Cossiga	4,425	1,516	34%	39.9
9	09.Scalfaro	3,456	1,270	37%	27.4
10	10.Ciampi	1,920	854	44%	27.8
11	11.Napolitano	3,738	1,442	39%	34.6
12	12.Mattarella	2,336	1,029	44%	17.4

La misura ha valore quando impiegata per confrontare corpora o sub corpora.

Poiché il numero dei token influenza quello dei type è possibile ricorrere al

TTR standardizzato [TTR medio calcolato su segmenti di lunghezza predefinita] – ha altre debolezze…

INDIVIDUAZIONE DEL LESSICO RILEVANTE

Come selezionare le forme su cui concentrare l'analisi?

Si possono considerare diversi soglie e pesi.

Quanti e quali dipende degli scopi della ricerca.

- Frequenza → soglia di frequenza, o un determinato tasso di copertura del corpus.

  Inoltre: includere o escludere le stop words?

- Estrazione delle keyword

- Collocazioni

ESTRAZIONE DELLE KEYWORD

Per evidenziare le peculiarità di un testo è necessario il confronto tra corpora o sub corpora.

Si possono impiegare diverse misure di keyness (= rilevanza)

1. Simple maths

Compara il focus corpus con un corpus di riferimento.

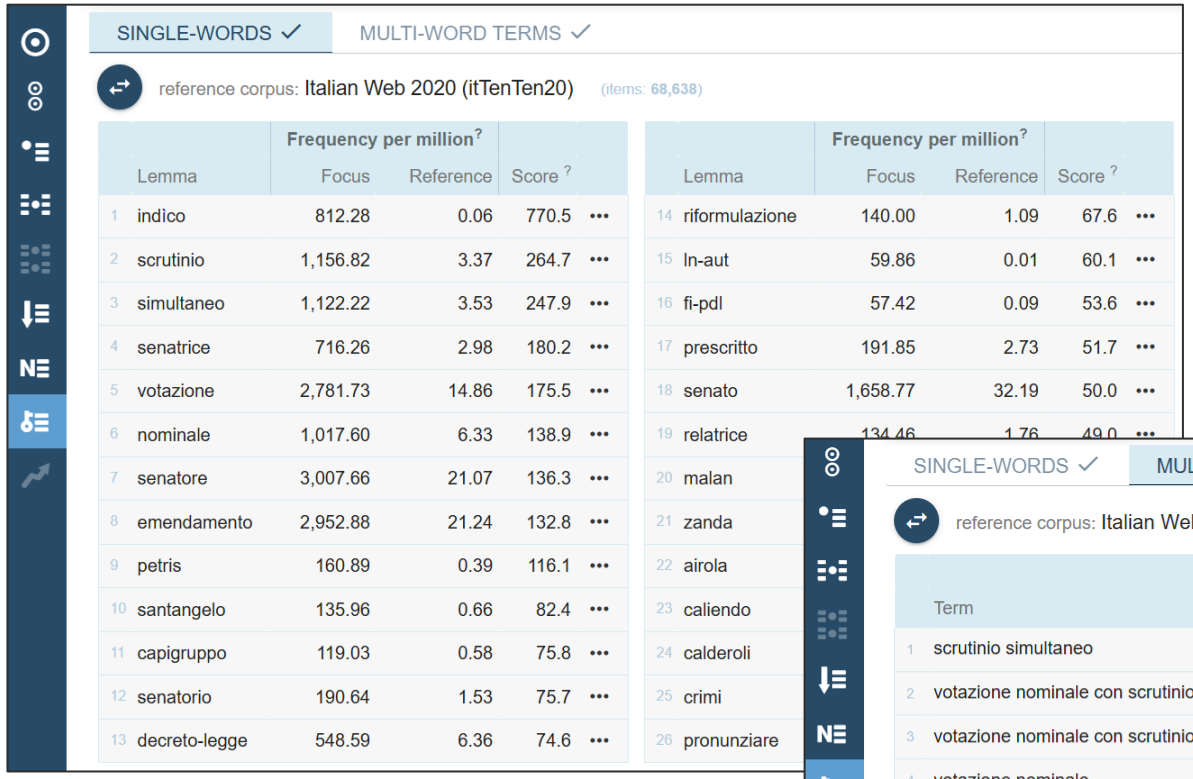
Considera il rapporto tra la frequenza normalizzata di una parola nel focus corpus e in quello di riferimento → il numero ottenuto ne rappresenta la keyness:
più il valore è alto, più la parola è distintiva.

Misura spesso impiegata per estrarre la terminologia.

Raffinabile con l'aggiunta di una costante: 1 [trova parole di bassa frequenza], 100, 1000 [trova parole di alta frequenza].

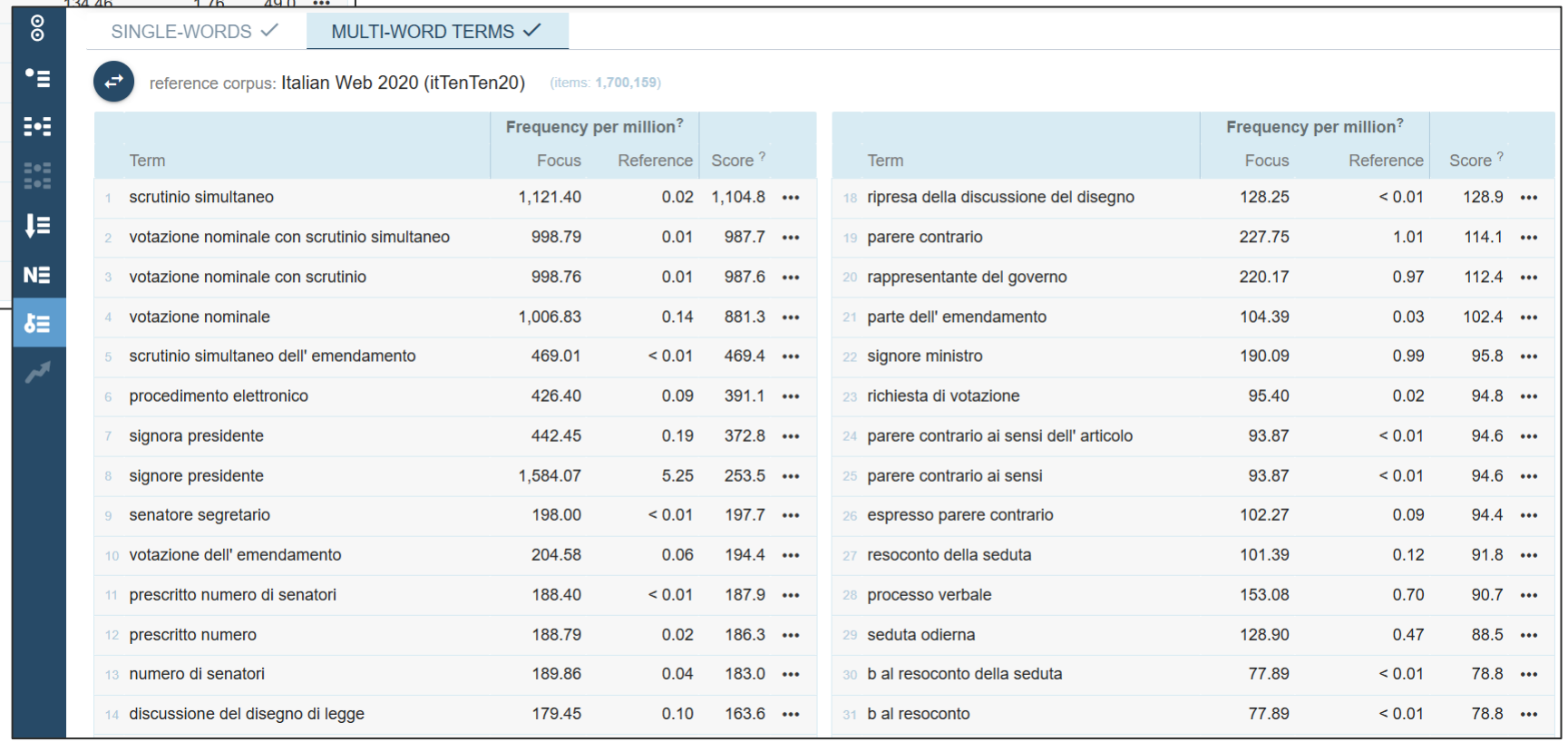
$$\frac{fpm_{rmfocus} + N}{fpm_{rmref} + N}$$

SIMPLE MATHS



	Lemma	Frequency per million? Focus	Reference	Score ?
1	indico	812.28	0.06	770.5
2	scrutinio	1,156.82	3.37	264.7
3	simultaneo	1,122.22	3.53	247.9
4	senatrice	716.26	2.98	180.2
5	votazione	2,781.73	14.86	175.5
6	nominale	1,017.60	6.33	138.9
7	senatore	3,007.66	21.07	136.3
8	emendamento	2,952.88	21.24	132.8
9	petris	160.89	0.39	116.1
10	santangelo	135.96	0.66	82.4
11	capigruppo	119.03	0.58	75.8
12	senatorio	190.64	1.53	75.7
13	decreto-legge	548.59	6.36	74.6

	Lemma	Frequency per million? Focus	Reference	Score ?
14	riformulazione	140.00	1.09	67.6
15	In-aut	59.86	0.01	60.1
16	fi-pdl	57.42	0.09	53.6
17	prescritto	191.85	2.73	51.7
18	senato	1,658.77	32.19	50.0
19	relatrice	134.46	1.76	49.0
20	malan			
21	zanda			
22	airola			
23	caliendo			
24	calderoli			
25	crimi			
26	pronunziare			



	Term	Frequency per million? Focus	Reference	Score ?
1	scrutinio simultaneo	1,121.40	0.02	1,104.8
2	votazione nominale con scrutinio simultaneo	998.79	0.01	987.7
3	votazione nominale con scrutinio	998.76	0.01	987.6
4	votazione nominale	1,006.83	0.14	881.3
5	scrutinio simultaneo dell' emendamento	469.01	< 0.01	469.4
6	procedimento elettronico	426.40	0.09	391.1
7	signora presidente	442.45	0.19	372.8
8	signore presidente	1,584.07	5.25	253.5
9	senatore segretario	198.00	< 0.01	197.7
10	votazione dell' emendamento	204.58	0.06	194.4
11	prescritto numero di senatori	188.40	< 0.01	187.9
12	prescritto numero	188.79	0.02	186.3
13	numero di senatori	189.86	0.04	183.0
14	discussione del disegno di legge	179.45	0.10	163.6

	Term	Frequency per million? Focus	Reference	Score ?
18	ripresa della discussione del disegno	128.25	< 0.01	128.9
19	parere contrario	227.75	1.01	114.1
20	rappresentante del governo	220.17	0.97	112.4
21	parte dell' emendamento	104.39	0.03	102.4
22	signore ministro	190.09	0.99	95.8
23	richiesta di votazione	95.40	0.02	94.8
24	parere contrario ai sensi dell' articolo	93.87	< 0.01	94.6
25	parere contrario ai sensi	93.87	< 0.01	94.6
26	espresso parere contrario	102.27	0.09	94.4
27	resoconto della seduta	101.39	0.12	91.8
28	processo verbale	153.08	0.70	90.7
29	seduta odierna	128.90	0.47	88.5
30	b al resoconto della seduta	77.89	< 0.01	78.8
31	b al resoconto	77.89	< 0.01	78.8

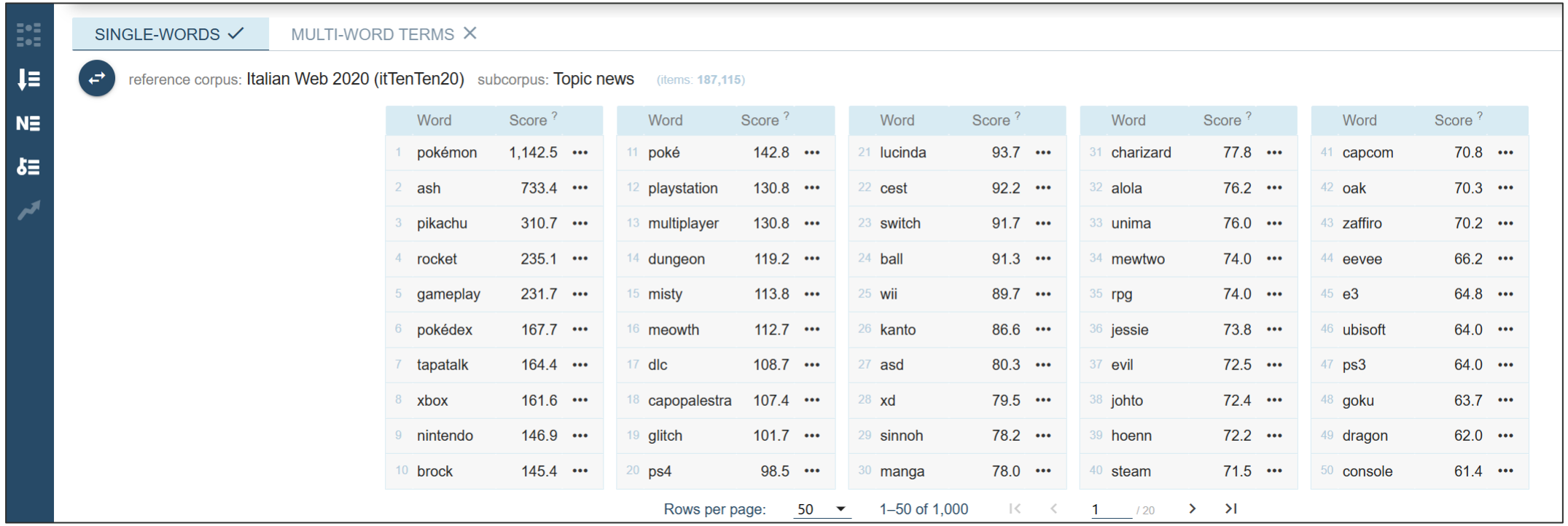
Funzione **keywords** in (No)Sketch Engine

Confronto corpus ParlaMint vs ItTenTen20

SIMPLE MATHS (2)

Funzione **keywords** in (No)Sketch Engine

Confronto sub corpora games vs news



ESTRAZIONE DELLE KEYWORD (2)

Per evidenziare le peculiarità di un testo è necessario il confronto tra corpora o sub corpora.

Si possono impiegare diverse misure di keyness (= rilevanza)

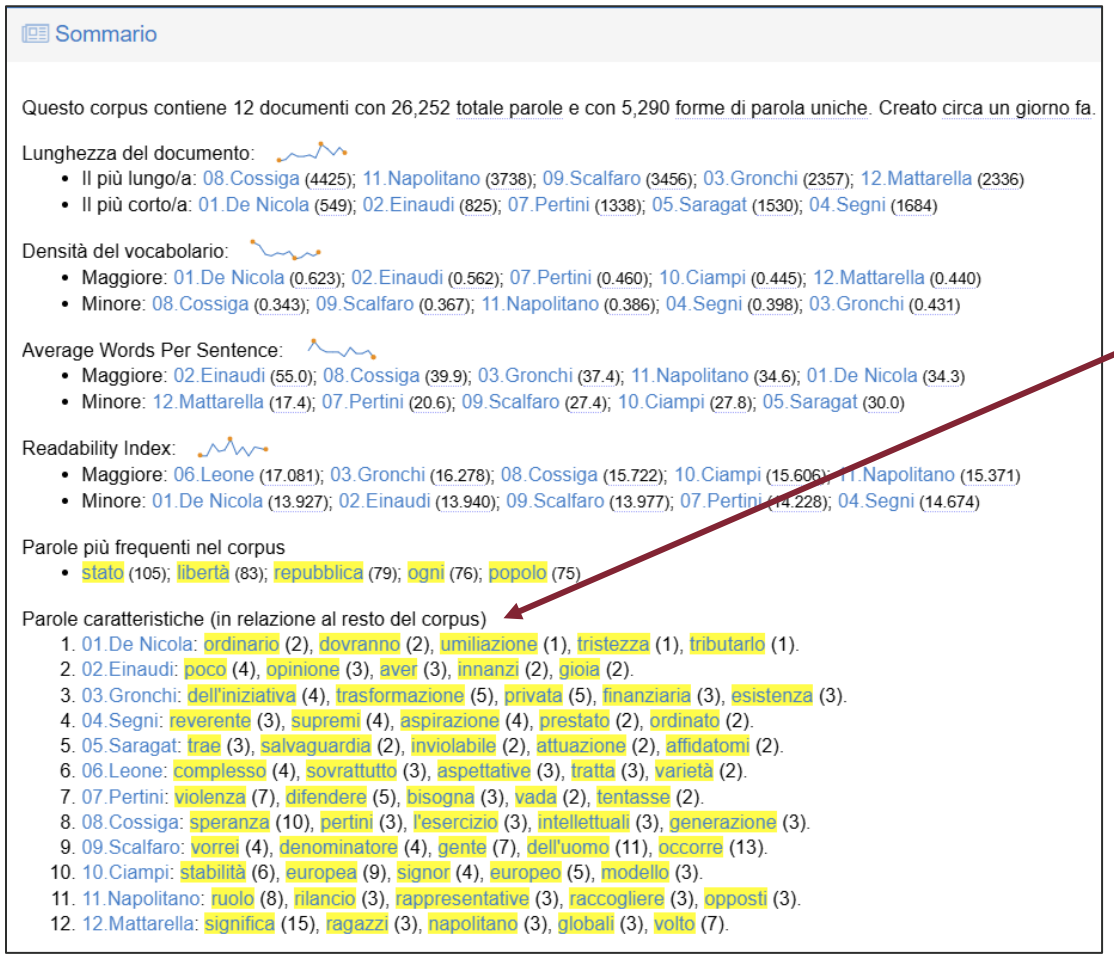
2. TF-IDF Mette in relazione la frequenza di una parola in un sub corpus (TF), con la frequenza inversa della parola in tutti i documenti (IDF). $TF \times \log 10 \frac{M}{m}$

M = numero di testi
m = testi in cui la parola è attestata

Se una parola è molto frequente in un testo, e assente in quasi tutti gli altri,

allora è fortemente distintiva → lessico peculiare di un sub corpus

TF-IDF



Sommario

Questo corpus contiene 12 documenti con 26,252 totale parole e con 5,290 forme di parola uniche. Creato circa un giorno fa.

Lunghezza del documento:

- Il più lungo/a: 08.Cossiga (4425); 11.Napolitano (3738); 09.Scalfaro (3456); 03.Gronchi (2357); 12.Mattarella (2336)
- Il più corto/a: 01.De Nicola (549); 02.Einaudi (825); 07.Pertini (1338); 05.Saragat (1530); 04.Segni (1684)

Densità del vocabolario:

- Maggiore: 01.De Nicola (0.623); 02.Einaudi (0.562); 07.Pertini (0.460); 10.Ciampi (0.445); 12.Mattarella (0.440)
- Minore: 08.Cossiga (0.343); 09.Scalfaro (0.367); 11.Napolitano (0.386); 04.Segni (0.398); 03.Gronchi (0.431)

Average Words Per Sentence:

- Maggiore: 02.Einaudi (55.0); 08.Cossiga (39.9); 03.Gronchi (37.4); 11.Napolitano (34.6); 01.De Nicola (34.3)
- Minore: 12.Mattarella (17.4); 07.Pertini (20.6); 09.Scalfaro (27.4); 10.Ciampi (27.8); 05.Saragat (30.0)

Readability Index:

- Maggiore: 06.Leone (17.081); 03.Gronchi (16.278); 08.Cossiga (15.722); 10.Ciampi (15.606); 11.Napolitano (15.371)
- Minore: 01.De Nicola (13.927); 02.Einaudi (13.940); 09.Scalfaro (13.977); 07.Pertini (14.228); 04.Segni (14.674)

Parole più frequenti nel corpus

- stato (105); libertà (83); repubblica (79); ogni (76); popolo (75)

Parole caratteristiche (in relazione al resto del corpus)

1. 01.De Nicola: ordinario (2), dovranno (2), umiliazione (1), tristezza (1), tributarlo (1).
2. 02.Einaudi: poco (4), opinione (3), aver (3), innanzi (2), gioia (2).
3. 03.Gronchi: dell'iniziativa (4), trasformazione (5), privata (5), finanziaria (3), esistenza (3).
4. 04.Segni: reverente (3), supremi (4), aspirazione (4), prestato (2), ordinato (2).
5. 05.Saragat: trae (3), salvaguardia (2), inviolabile (2), attuazione (2), affidatomi (2).
6. 06.Leone: complesso (4), sovrattutto (3), aspettative (3), tratta (3), varietà (2).
7. 07.Pertini: violenza (7), difendere (5), bisogna (3), vada (2), tentasse (2).
8. 08.Cossiga: speranza (10), pertini (3), l'esercizio (3), intellettuali (3), generazione (3).
9. 09.Scalfaro: vorrei (4), denominatore (4), gente (7), dell'uomo (11), occorre (13).
10. 10.Ciampi: stabilità (6), europea (9), signor (4), europeo (5), modello (3).
11. 11.Napolitano: ruolo (8), rilancio (3), rappresentative (3), raccogliere (3), opposti (3).
12. 12.Mattarella: significa (15), ragazzi (3), napolitano (3), globali (3), volto (7).

Voyant Tools

Le **parole caratteristiche** vengono individuate sulla base del TF-IDF.

MISURE DI COLLOCAZIONE

Individuano co-occorrenze, e ne segnalano il peso.

1. Frequenza osservata Numero di co-occorrenze tra due parole in contesti di lunghezza predefinita.

È necessario rimuovere le stopwords per ottenere risultati utili.

Es.: funzioni 'sintagmi' e 'collocazioni' di Voyant tools

Voyant Tools

Collocazioni

Parola	Collocazione	Conteggio (contesto)
popolo	italiano	36
repubblica	presidente	20
ogni	ogni	18
libertà	giustizia	15
repubblica	viva	10
libertà	democrazia	8
ogni	parte	8
stato	diritto	7
stato	capo	7
repubblica	italiana	7
repubblica	democratica	7
ogni	giorno	7
ogni	cittadino	7
essere	presidente	7
politica	sociale	7
stato	responsabilità	6
libertà	pace	6
repubblica	parlamento	6

MISURE DI COLLOCAZIONE (2)

Individuano co-occorrenze, e ne segnalano il peso.

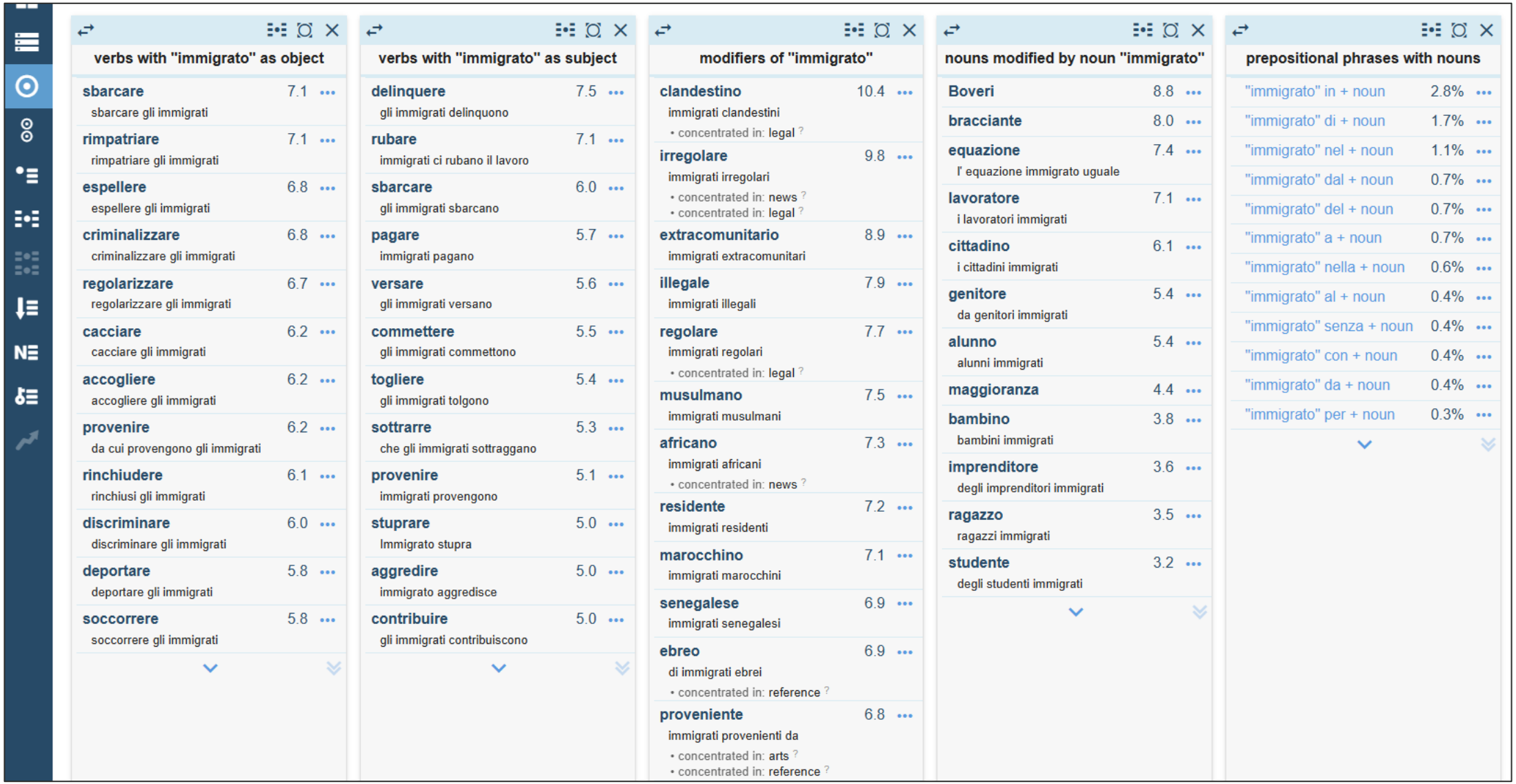
2. LogDice Mette in relazione la frequenza relativa delle parole con quella della collocazione.

↓

Assegnato un valore alto se le parole considerate occorrono (quasi) esclusivamente nella collocazione considerata → associazione forte.

Es.: funzione ‘word sketch’ di Sketch Engine.

MISURE DI COLLOCAZIONE (2)



verbs with "immigrato" as object

sbarcare	7.1
sbarcare gli immigrati	
rimpatriare	7.1
rimpatriare gli immigrati	
espellere	6.8
espellere gli immigrati	
criminalizzare	6.8
criminalizzare gli immigrati	
regolarizzare	6.7
regolarizzare gli immigrati	
cacciare	6.2
cacciare gli immigrati	
accogliere	6.2
accogliere gli immigrati	
provenire	6.2
da cui provengono gli immigrati	
rinchiudere	6.1
rinchiusi gli immigrati	
discriminare	6.0
discriminare gli immigrati	
deportare	5.8
deportare gli immigrati	
soccorrere	5.8
soccorrere gli immigrati	

verbs with "immigrato" as subject

delinquere	7.5
gli immigrati delinquono	
rubare	7.1
immigrati ci rubano il lavoro	
sbarcare	6.0
gli immigrati sbarcano	
pagare	5.7
immigrati pagano	
versare	5.6
gli immigrati versano	
commettere	5.5
gli immigrati commettono	
togliere	5.4
gli immigrati tolgono	
sottrarre	5.3
che gli immigrati sottraggano	
provenire	5.1
immigrati provengono	
stuprare	5.0
Immigrato stupra	
aggredire	5.0
immigrato aggredisce	
contribuire	5.0
gli immigrati contribuiscono	

modifiers of "immigrato"

clandestino	10.4
immigrati clandestini	
• concentrated in: legal ?	
irregolare	9.8
immigrati irregolari	
• concentrated in: news ?	
• concentrated in: legal ?	
extracomunitario	8.9
immigrati extracomunitari	
illegale	7.9
immigrati illegali	
regolare	7.7
immigrati regolari	
• concentrated in: legal ?	
musulmano	7.5
immigrati musulmani	
africano	7.3
immigrati africani	
• concentrated in: news ?	
residente	7.2
immigrati residenti	
marocchino	7.1
immigrati marocchini	
senegalese	6.9
immigrati senegalesi	
ebreo	6.9
di immigrati ebrei	
• concentrated in: reference ?	
proveniente	6.8
immigrati provenienti da	
• concentrated in: arts ?	
• concentrated in: reference ?	

nouns modified by noun "immigrato"

Boveri	8.8
bracciante	8.0
equazione	7.4
l' equazione immigrato uguale	
lavoratore	7.1
i lavoratori immigrati	
cittadino	6.1
i cittadini immigrati	
genitore	5.4
da genitori immigrati	
alunno	5.4
alunni immigrati	
maggioranza	4.4
bambino	3.8
bambini immigrati	
imprenditore	3.6
degli imprenditori immigrati	
ragazzo	3.5
ragazzi immigrati	
studente	3.2
degli studenti immigrati	

prepositional phrases with nouns

"immigrato" in + noun	2.8%
"immigrato" di + noun	1.7%
"immigrato" nel + noun	1.1%
"immigrato" dal + noun	0.7%
"immigrato" del + noun	0.7%
"immigrato" a + noun	0.7%
"immigrato" nella + noun	0.6%
"immigrato" al + noun	0.4%
"immigrato" senza + noun	0.4%
"immigrato" con + noun	0.4%
"immigrato" da + noun	0.4%
"immigrato" per + noun	0.3%

CONFRONTO TRA CORPORA

È possibile prendere in considerazione diverse misure di similarità e distanza. Le più 'semplici':

1. **Indice di Jaccard** Considera il numero di parole comuni in rapporto al numero di parole totali.
   Il valore è compreso tra 0 (niente in comune) e 1 (identici).

2. **Distanza di Jaccard** Esprime la diversità tra due corpora. Si calcola invertendo la misura precedente: 1 – Indice di Jaccard.

Analisi delle corrispondenze: periodizzazione dell'AGI basata sui profili lessicali delle annate