ParlaMint – IT – Il corpus del Senato Italiano. Una guida pratica per l'interrogazione del corpus ParlaMint-IT con *NoSketch Engine*, a supporto dell'analisi del discorso politico.

Dario Del Fante*
Istituto di Linguistica Computazionale "A. Zampolli" - CNR, Pisa

8 maggio 2022

Dissemination level	Internal
Actual submission date	08/05/2022
Funding	CLARIN FOE 2018 – CUP B54I18010400001
Task	Promozione del dataset Parlamint-IT
Type	Tutorial
Author(s)	Dario Del Fante
Supervisor(s)	Francesca Frontini Monica Monachini Valeria Quochi
Version	v1.1
Number of pages	p.1-p.23

*Lo sviluppo e la revisione di tale tutorial sono state svolte sotto la supervisione della Dott.ssa Francesca Frontini e della Dott.ssa Valeria Quochi, tuttavia l'autore rimane responsabile di quanto scritto.

1 Introduzione

ParlaMint è un progetto promosso da CLARIN ERIC, l'infrastruttura europea per le risorse e tecnologie del linguaggio[1]; tale progetto ha come obiettivo principale quello di creare una risorsa multilingue di dibattiti parlamentari di facile accesso per ricercatori e per i singoli cittadini, che permetta di svolgere analisi contrastive tra lingue diverse e tra contesti politici, culturali e socio-economici differenti. Attualmente sono presenti corpora di diversi parlamenti europei in 16 lingue (bulgaro, croato, ceco, danese, olandese, inglese, francese, ungherese, islandese, italiano, lettone, lituano, polacco, sloveno, spagnolo, turco). I dati sono accessibili secondo due modalità. La versione "plain-text" [2] contiene i corpora codificati secondo le linee guida Parla-CLARIN [3], un formato TEI P5 che annota i metadati [4] di tipo testuale in aggiunta ad informazioni extra-testuali relative al parlamento. Per esempio, sono indicati dettagli relativi a chi parla: nome, sesso, anno di nascita, ruolo all'interno del parlamento, affiliazione a un determinatopartito, se alla maggioranza o all'opposizione. I discorsi sono divisi in termini temporali, secondo la sessione parlamentare o la riunione di riferimento. Sempre all'interno della piattaforma *NoSketchEngine* e utilizzando la metadatazione, si possono effettuare ricerche restringendo il campo relativamente ad un solo oratore o a quelli relativi ad un partito. Sono stati anche inseriti elementi che indicano la presenza di accadimenti come interruzioni impreviste o applausi o fischi. La versione "marked-up"[5], aggiunge anche informazioni di tipo linguistico, come la tokenizzazione [6], la segmentazione in frasi, la lemmatizzazione[7], l'annotazione per Parti del Discorso (POS tag) o anche detta morfosintattica[8], secondo Universal Dependencies (UD)[9], per dipendenze sintattiche[10], e infine, le Entità Nominate [11]. Sono possibili poi ulteriori annotazioni linguistiche secondo schemi più specifici per ogni lingua. In generale, i corpora del progetto sono codificati secondo le linee guida Parla-CLARIN TEI. I file possono essere scaricati all'indirizzo di cui sopra e includono i seguenti: il corpus annotato linguisticamente in formato XML TEI; il corpus derivato

[1] https://www.clarin.eu/

[2] Reperibile a questo link ¡http://hdl.handle.net/11356/1432¿.

[3] Informazioni reperibili presso https://clarin-eric.github.io/parla-clarin/

[4] Il metadato è un'informazione che descrive un sistema di dati. Possiamo pensare ad un metadato come una scheda che ha lo scopo di descrivere il contenuto e/o gli attributi di uno degli elementi a cui è assegnato. La funzione principale di un sistema di metadati è quella di ricercare, localizzare, selezionare il dato in maniera efficente. In aggiunta, la presenza di metadati consente l'interoperabilità semantica, che permette di usare il dato in ambiti disciplinari diversi grazie a una serie di equivalenze fra descrittori.

[5] Reperibile a questo link ¡ http://hdl.handle.net/11356/1431¿

[6] Tokenizzare un testo significa dividerlo in sequenze di caratteri, cioè unità minime di analisi dette 'token'. Nel caso dell'italiano un token tipicamente corrisponde ad una parola e alla punteggiatura (punti, virgole etc.)

[7] La lemmatizzazione consiste nel processo di riduzione di una forma flessa alla sua forma canonica - non marcata - ovvero il lemma. Es: gatti; gatte; gatto; gatta → GATTO (lemma)

[8] Tale annotazione assegna ad ogni 'token' un'etichetta corrispondente alla parte del discorso a cui esso appartiene. Vi sono varie tipologie di categorizzazione. Comunemente il testo si annota per: Nomi; Verbi; Aggettivi; Articoli; Pronomi; Articoli; Avverbi; Congiunzioni.

[9] Nel caso specifico dei corpora ParlaMint l'annotazione morfosintattica è stata fatta secondo la filosofia e usando le etichette delle Universal Dependencies https://universaldependencies.org/guidelines.html

[10] L'analisi di una frase per dipendenze sintattiche consiste nell'annotazione delle relazioni sintattiche tra parole che la compongono, utilizzando etichette come soggetto, oggetto, modificatore etc.

[11] Un'entità nominata o Named Entity è un oggetto del mondo reale, come una persona, un luogo, un'organizzazione, un prodotto etc. che può essere indicata con un nome proprio. Può essere astratto o avere un'esistenza fisica. Il riconoscimento di Entità Nominate, o Named Entity Recognition (NER), è una tecnica di estrazione dell'informazioni per identificare e classificare entità denominate nel testo. Nel caso di ParlaMint, le entità identificate sono state categorizzate in base alle 4 classi definite dal progetto CoNLL-2003: Persone (PER); Organizzazioni (ORG); Località (LOC); Nomi di altre entità (MISC)

in CoNLL-U [12] con metadati in formato TSV [13]; i file verticali o concordanze (con file di registro), adatti all'uso con software per l'analisi del corpus come CWB[14], NoSketch Engine[15] o KonText[16].
In riferimento a questo tutorial, ci si concentrerà sul corpus italiano ParlaMint-IT, che contiene le trascrizioni di una selezione di sedute del Senato della Repubblica Italiana. Verrà utilizzata la piattaforma NoSketchEngine che, come anticipato, ospita tutti i corpora del progetto ParlaMint ed è ad accesso libero. In relazione all'attuale guida pratica, utilizzeremo la versione 2.1.

In questo tutorial, in particolare, verranno affrontati i seguenti argomenti:

- La costruzione e l'uso di sotto-corpora
- La creazione di liste di parole e la comparazione per parole chiave
- L'estrazione di collocazioni e concordanze e la loro distribuzione diacronica

2 ParlaMint -IT: I sotto-corpora - percorsi e potenzialità

La versione italiana di ParlaMint comprende tutti i discorsi parlamentari del Senato della Repubblica Italiana tenuti dal marzo 2013 all'ottobre 2020. Come da riferimento nella 1, il corpus è molto ampio e offre varie possibilità di ricerca.

ParlaMint-IT V.2.1	
Tokens	30,615,130
Words	26,571,966
Sentences	1,087,465
Paragraphs	214,300
Documents	79,283

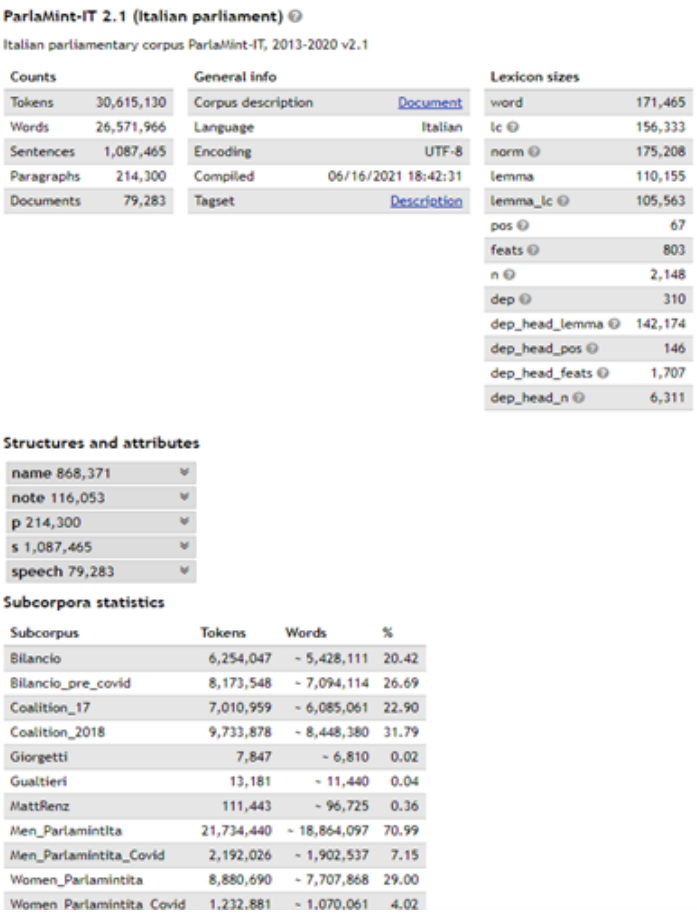
Tabella 1. Il corpus in cifre.

Sfruttando i metadati che sono stati prodotti in relazione al corpus, una delle funzionalità principali è la possibilità di creare sotto-corpora: ovvero di dividere il corpus in base ai propri interrogativi di ricerca o interessi.

Figura 1: Esempio d'uso delle opzioni

Dopo aver raggiunto la pagina iniziale del progetto [17] e aver selezionato la voce *ParlaMint - IT 2.1*, come mostrato nella 1, si presenterà davanti a voi una schermata in cui sono racchiuse tutte le informazioni più importanti del corpus ParlaMint – IT (2).

[12] https://universaldependencies.org/format.html
[13] I file TSV appartengono principalmente a Excel di Microsoft. TSV è l'estensione del nome dei file di testo ASCII che contengono valori separati da tabulazione. Questi file sono usati principalmente per scambiare dati tra database e applicazioni di fogli di calcolo. I file TSV contengono dati di testo strutturati in cui ogni valore di un record è delimitato da una scheda. Puoi aprire o creare tali file TSV con Excel, Calc, o un editor di testo come Sublime.
[14] Reperibile presso https://sourceforge.net/projects/cwb/
[15] Reperibile presso ¡http://clarin.si/noske/index-en.html¿
[16] Reperibile presso ¡ https://www.korpus.cz/kontext/corpora/corplist¿
[17] http://clarin.si/noske/index.html



ParlaMint-IT 2.1 (Italian parliament)

Italian parliamentary corpus ParlaMint-IT, 2013-2020 v2.1

Counts	
Tokens	30,615,130
Words	26,571,966
Sentences	1,087,465
Paragraphs	214,300
Documents	79,283

General info	
Corpus description	Document
Language	Italian
Encoding	UTF-8
Compiled	06/16/2021 18:42:31
Tagset	Description

Lexicon sizes	
word	171,465
lc	156,333
norm	175,208
lemma	110,155
lemma_lc	105,563
pos	67
feats	803
n	2,148
dep	310
dep_head_lemma	142,174
dep_head_pos	146
dep_head_feats	1,707
dep_head_n	6,311

Structures and attributes

name 868,371
note 116,053
p 214,300
s 1,087,465
speech 79,283

Subcorpora statistics

Subcorpus	Tokens	Words	%
Bilancio	6,254,047	~ 5,428,111	20.42
Bilancio_pre_covid	8,173,548	~ 7,094,114	26.69
Coalition_17	7,010,959	~ 6,085,061	22.90
Coalition_2018	9,733,878	~ 8,448,380	31.79
Giorgetti	7,847	~ 6,810	0.02
Gualtieri	13,181	~ 11,440	0.04
MattRenz	111,443	~ 96,725	0.36
Men_Parlamintlta	21,734,440	~ 18,864,097	70.99
Men_Parlamintita_Covid	2,192,026	~ 1,902,537	7.15
Women_Parlamintita	8,880,690	~ 7,707,868	29.00
Women_Parlamintita_Covid	1,232,881	~ 1,070,061	4.02

Figura 2: Dettagli del corpus ParlaMint-IT 2.1

Nella sezione *structures and attributes*, invece, come vedete nella 3, vengono indicate le varie tipologie di annotazioni che sono state applicate al corpus.

Per name, ad esempio, si indica la categorizzazione per entità nominate, secondo le tipologie indicate dal progetto ConLL-2003 (ORG, PER, LOC), come visto in precedenza. Mentre note si riferisce a due tipi di annotazioni, note-content e note-type. Nel primo caso, il corpus è annotato sulla base del contenuto specifico come, per esempio, "brusio" oppure "commenti del gruppo del PD". Nel secondo caso, il corpus è annotato sulla base di azioni generali che stanno avvenendo all'interno del parlamento come, per esempio, l'ingresso di parlamentari, l'uscita di parlamentari, le risate di parlamentari. La sezione p e s invece si riferiscono rispettivamente all'annotazione di ogni paragrafo o di ogni frase del corpus. Ogni frase o paragrafo sono annotati sulla base dei seggi attivi per ogni seduta. Ad esempio, nel caso di *ParlaMint-IT_2016-05-18-LEG17-Sed-628.seg243*, l'annotazione ci dice che il paragrafo appartiene al discorso pronunciato dal parlamentare del seggio 243 durante la seduta 628 della 17 legislatura il 18 maggio del 2016. La sezione speech si riferisce all'annotazione del corpus in base ai discorsi. Ad ogni discorso è assegnato un identificativo speech.id. All'interno di questa classificazione, sono presenti ulteriori annotazioni più specifiche. Per ogni seduta del parlamento, un speech.text.id o title o sitting sono stati definiti. Tali identificativi numerici contengono l'insieme dei discorsi per ogni seduta del parlamento. Inoltre, è possibile selezionare:

- from e to - una data specifica, attraverso l'annotazione;
- party_status - il ruolo del partito, se di opposizione, maggioranza o misto;

Figura 3: I metadati

- speaker_birth - l'anno di nascita del parlante;
- speaker_gender - il genere del parlante;
- speaker_id – l'identificativo numerico assegnato al parlante specifico;
- speaker_name - il nome del parlante, ad esempio Filippo Civati o Matteo Renzi;
- speaker_party - l'identificativo numerico assegnato al partito;
- speaker_party_name - il nome del partito;
- speaker_type - il tipo di parlante, cioè se membro del parlamento oppure ospite esterno;
- speaker_role - il ruolo, se presidente, membro regolare o ospite [18] ;
- term - la legislatura, in questo caso XVII o XVIII.

Si possono sfruttare tutti questi descrittori per creare sotto-corpora con caratteristiche specifiche. Per fare questo, bisogna selezionare la finestra WordList sulla sinistra e poi cliccare su create new nella schermata che appare sulla destra, come mostra la Fig. 4.

Una volta cliccato, vi troverete davanti una schermata che vi darà diverse possibilità di creare dei sotto-corpora (Fig. 5) in base ai metadati sopra presentati.

[18] La differenza tra un presidente e un membro regolare sta nello stile formulare del primo rispetto al secondo dovuto alle caratteristiche intrinseche delle sedute parlamentari.

Figura 4: Schermata Wordlist

Una delle prime possibilità che vi si presenta è selezionare i due sotto-corpora già presenti su NoSketchEngine. Il COVID corpus contiene tutti i dibattiti parlamentari relativi al periodo della pandemia (Gennaio 2020 – Ottobre 2020). Il REFERENCE corpus contiene tutti i discorsi del parlamento italiano dal 2013 al Dicembre 2019. La 2 fa un riassunto dei dettagli di entrambi i corpora. Come mostra la Fig. 5, è possibile però anche combinare le varie categorie e creare dei sotto-corpora specifici.

	ParlaMint-IT REFERENCEs	**ParlaMint-IT COVID**
Tokens	27,190,223	3,424,907
Words	23,574,948	2,997,018
Documents	6,238	73,045

Tabella 2. Dettagli del Reference Corpus e Covid Corpus

Per esempio, in vista di questo tutorial, si possono creare quattro sotto-corpora. Selezionando prima il sotto-corpus COVID e poi speaker_gender M oppure F, si creano due sotto-corpora che chiameremo Men_ParlamintIta_COVID e Women_ParlimintIta_COVID, rispettivamente contenenti i discorsi parlamentari esclusivamente maschili o femminili durante il COVID. Successivamente, selezionando il sotto-corpus REFERENCE e speaker_gender M e F, si creano Men_ParlamintIta e Women_ParlamintIta, contenenti rispettivamente i discorsi parlamentari esclusivamente maschili o femminili precedenti il COVID, come si vede nella tabella 3.

	Men_ParlamintIta_COVID	**Women_ParlimintIta_COVID**	**Men_ParlamintIta**	**Women_ParlamintIta**
Tokens	21,734,440	1,232,881	2,192,026	8,880,690
Words	18,864,097	1,070,061	1,902,537	7,707,868
% [19]	70.99	4.02	7.15	29.00

Tabella 3. Corpus "Men" e "Women" a confronto

Le possibilità di esplorazione sono molteplici. Ad esempio, incrociando COVID e REFERENCES e la categoria speaker_party, si possono creare sotto-corpora contenenti i discorsi di un singolo politico prima e durante la pandemia COVID, dando la possibilità di poter svolgere diverse tipologie di ricerche. La possibilità di creare sotto-corpora combinando le diverse tipologie di annotazione rappresenta un potentissimo strumento poiché permette di studiare dati da prospettive diverse e quindi di far emergere risultati in ottica contrastiva. Nelle prossime sezioni verranno mostrati alcuni dei principali percorsi di analisi con NoSketchEngine e ParlaMint-IT.

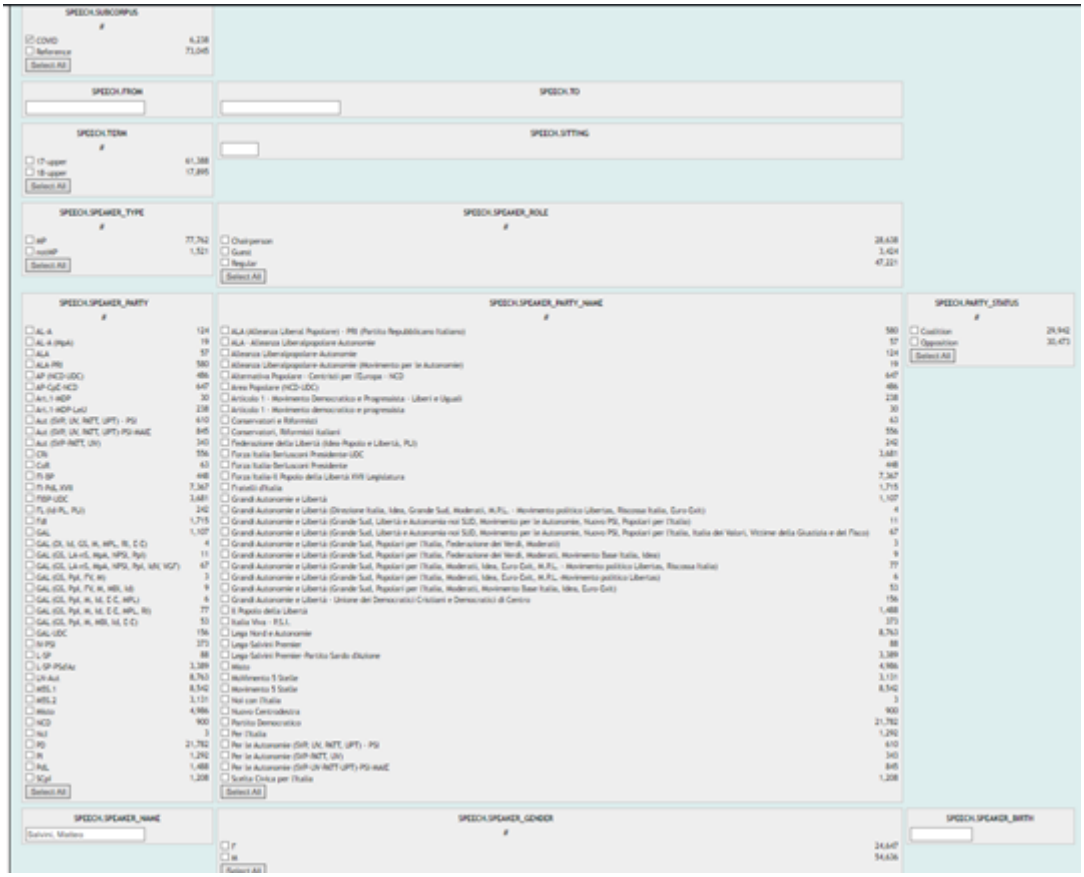


Figura 5: Schermata Sotto-corpora

3 Liste di parole e comparazione per parole chiave – come la frequenza aiuta a capire l'uso delle parole

Le liste di parole – Wordlist – corrispondono a liste di frequenza di token che vengono estratti dai corpora. Queste liste mostrano le parole all'interno del corpus ordinate secondo la loro frequenza o secondo altri indicatori statistici (Average Reduced Frequency – ARF[20]). Se consideriamo il corpus come un campione selezionato per essere rappresentativo di una varietà linguistica, le parole più frequenti di un dato corpus sono una informazione significativa in relazione alla rappresentatività del campione. Anche le differenze nella frequenza delle parole tra i vari sotto-corpora possono essere altrettanto significative. Per esempio, la lista di parole del Parla-Mint – COVID mostra le parole più frequenti utilizzate nel parlamento italiano durante il periodo pandemico e può mostrare delle parole come appunto mascherina, distanziamento o sanificazione che normalmente hanno una frequenza relativamente bassa all'interno dei discorsi politici e che mai ci aspetteremmo di trovare.

Come si crea una lista di parole? Si seleziona dalla colonna a sinistra Wordlist (Fig. 6) e poi successivamente nella sezione corpus ParlaMint-IT 2.1.

Una volta selezionato il nostro sub-corpus Men_ParlamintIta, si possono produrre liste di parole in base ai positional attributes (7), cioè utilizzando le informazioni aggiuntive

[20]Average Reduced Frequency è un dato di frequenza modificato che tiene in considerazione la distribuzione all'interno del corpus per evitare che il risultato sia eccessivamente influenzato dal fatto che la parola analizzata sia altamente concentrata in una parte del corpus. Se la parola è distribuita uniformemente nel corpus, ARF e frequenza assoluta saranno simili o identiche. Maggiori dettagli al sito ¡https://www.sketchengine.eu/documentation/average-reduced-frequency/¿

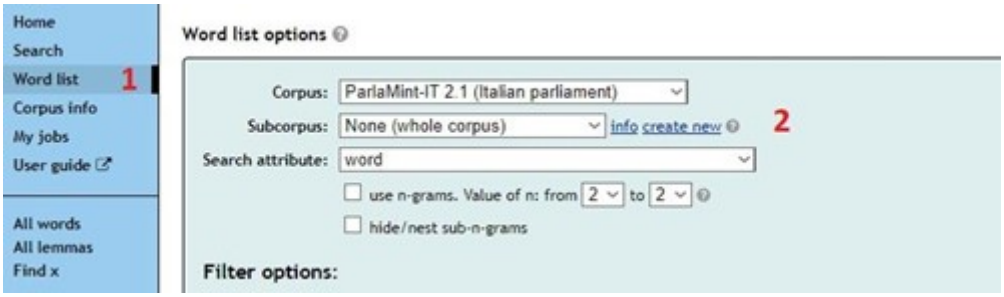


Figura 6: Graficoschermata per creare "Wordlist"

per ogni parola.

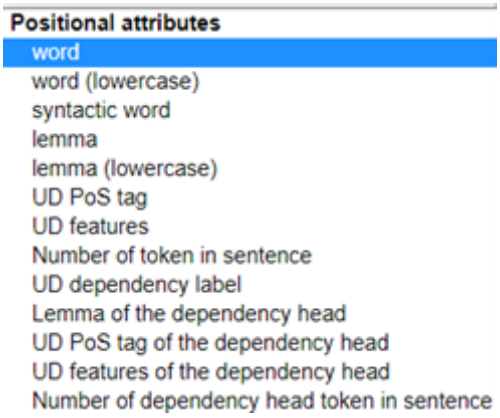


Figura 7: lista di positional attributes

Per esempio, si può produrre una lista di parole semplici, o token (8), di lemmi (CITTADINO: cittadino, cittadina, cittadini, cittadine) (9), di parti del discorso secondo le etichette del formato Universal Dependencies (ADJ, NOM, VERB, ...) (10).

Alternativamente, si possono creare delle liste sulla base dei tipi di testi (11). Questo può essere utile per vedere quale partito è più presente, oppure quale politico, oppure quale legislatura è caratterizzata dal maggior numero di parole. Inoltre, si può scegliere di filtrare i risultati (Fig. 12) con espressioni regolari [21], stabilendo una frequenza minima e massima, concentrandosi solo su una serie di parole definite dall'utente – whitelist – oppure escludendo una serie di parole sempre definite dall'utente – blacklist. Per quanto riguarda la modalità di visualizzazione dei risultati, si può scegliere tra la frequenza assoluta [22] , la frequenza relativa per milioni di parole[23], l'Average Reduced Frequency, il numero di

[21] Un'espressione regolare (regular expression o regexp, regex o RE) è una sequenza di simboli che identifica un insieme di stringhe. Esse vengono utilizzate per cercare dei pattern specifici come per esempio parole che iniziano allo stesso modo o che finiscono allo stesso modo.

[22] Per frequenza assoluta si intende il numero di volte che una parola appare all'interno del corpus. Essa è potenzialmente un dato compreso tra 0 e il totale delle parole (numero dei token) che compongono il corpus. Per esempio, il lemma PRESIDENTE ha una frequenza assoluta di 132.771 occorrenze su 30.615.130 milioni di occorrenze totale del corpus ParlaMint-IT 2.1

[23] La frequenza relativa è il rapporto tra la frequenza assoluta e il numero delle parole che compongono il corpus. La frequenza relativa può essere espressa in punti percentuali e in NoSketchEngine è espressa "per million words", cioè relativizzata ad un corpus di 1 milione di parole. Per esempio, la frequenza relativa del lemma PRESIDENTE sarà calcolata secondo la seguente formula: (numero di occorrenze : numero totale di occorrenze del corpus) * 1.000.000 In questo caso corrisponderà a (132.771 : 30.615.130) * 1.000.000 = 4.336,78 La decisione di utilizzare una misura come questa nasce che dal fatto che essa permette la comparazione di corpora con dimensioni diverse. Infatti, maggiore è la somma totale delle parole che compongono un corpus, maggiore sarà il numero di occorrenze di una singola parola o un lemma.

Word list

Corpus: **ParlaMint-IT 2.1 (Italian parliament)**
Subcorpus: **Men_ParlamintIta**
Total number of items: **51,430**
Total frequency: **18,732,971**

Page 1 Go Next >

word	frequency
di	707,959
che	525,266
e	474,148
la	344,138
il	320,738
in	278,562
è	255,489
del	249,420
a	235,947
per	224,784
non	223,970
un	201,326
della	163,062
l'	162,456
una	150,665
con	137,240
si	136,259
le	132,786
i	118,340
dell'	109,965
dei	101,962
al	97,069
da	96,783
questo	95,706
ha	90,270
anche	87,648
sono	86,035
delle	84,019
alla	74,972
come	71,881

Figura 8: Lista di lemmi

Word list

Corpus: **ParlaMint-IT 2.1 (Italian parliament)**
Subcorpus: **Men_ParlamintIta**
Total number of items: **29,686**
Total frequency: **18,791,154**

Page 1 Go Next >

lemma	frequency
il	1,223,317
di|il	755,294
di	728,257
essere	624,936
che	526,924
e	511,490
uno	401,527
a|il	330,700
in	296,508
a	289,267
avere	260,471
non	248,287
per	238,595
questo	207,634
in|il	168,607
si	149,068
con	141,576
da|il	131,776
da	100,709
presidente	99,031
fare	94,894
anche	93,406
quello	89,718
potere	87,405
su|il	85,889
ci	85,846
ma	81,805
come	79,135
dovere	74,354
senatore	71,741

Figura 9: Lista di parole

Word list

Corpus: **ParlaMint-IT 2.1 (Italian pa**
Subcorpus: **Men_ParlamintIta**
Total number of items: **38**
Total frequency: **21,734,391**

pos	frequency
NOUN	4,479,805
PUNCT	2,614,262
VERB	2,173,571
DET	2,110,403
ADP	1,993,339
ADP|DET	1,473,037
ADJ	1,424,736
ADV	1,217,881
PRON	1,154,624
AUX	1,001,060
CCONJ	672,365
PROPN	613,245
SCONJ	383,908
NUM	320,932
VERB|PRON	70,024
INTJ	10,504
X	9,838
AUX|PRON	5,580
VERB|PRON|PRON	1,768
PROPN|PRON	1,009
SYM	616
VERB|DET	419
PROPN|DET	396
NOUN|PRON	361

Figura 10: Lista di parole secondo parti del discorso/categorie lessicali

documenti/discorsi in cui appare una determinata forma oppure prodotti da un singolo parlante.

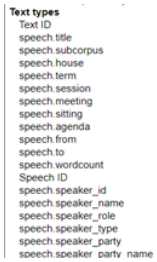


Figura 11: Diverse opzioni di annotazione dei testi.

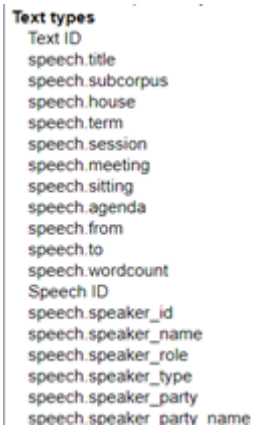


Figura 12: Diverse opzioni per il calcolo delle lista di parole.

Per esempio, la Fig. 13 e 14 mostrano rispettivamente i politici che hanno prodotto più parole tra il periodo precedente la pandemia (2013-2019) e il periodo pandemico (2020-2021). Questo non vuol dire necessariamente che siano intervenuti più spesso di altri (perché possono aver ipoteticamente fatto un discorso molto lungo diviso in poche occasioni) né che abbiano inciso maggiormente rispetto ad altri (la quantità non sempre corrisponde alla quantità). Allo stesso modo, le figure 15 e 16 mostrano le politiche donne che hanno parlato maggiormente dal periodo pre-pandemia a dopo la comparsa del Covid.

Infine, un'altra funzione molto utile quando si ha l'interesse della comparazione è la comparazione per parole chiavi o keywords comparison. L'estrazione di parole chiave tra due corpora consiste nel confrontare le liste di parole e far emergere le parole più sovra / sotto rappresentate in un corpus dato rispetto al corpus di riferimento. Utilizzando un'espressione metaforica per spiegare questo concetto, si potrebbe pensare alla comparazione per parole chiave come al passaggio di un liquido attraverso un filtro: il corpus di test è infatti comparabile ad un liquido che viene fatto passare attraverso un filtro, che è invece comparabile al corpus di riferimento scelto. Possiamo dire che "il materiale" che rimane sul filtro corrisponde a quelle parole che sono altamente frequenti/ oppure presenti soprattutto nel corpus di test. Mentre il liquido che passa attraverso il filtro corrisponde a tutte quelle parole che sono frequenti sia nel corpus di test che nel corpus di riferimento. In questo caso, bisogna ripetere la stessa procedura di prima. Sempre nella colonna a sinistra, selezionare wordlist, poi corpus e infine sub-corpus. In questo caso, facendo nuovamente riferimento alla Fig. 17, nella sezione Output type, selezionare Keyword. Sempre in questa sezione, selezionare il corpus e poi il sotto-corpus di riferimento. È possibile

speech.speaker_name	frequency
Calderoli, Roberto	1,001,185
Grasso, Pietro	942,550
Gasparri, Maurizio	732,267
Malan, Lucio	461,920
Giovanardi, Carlo	323,606
Candiani, Stefano	317,042
Barani, Lucio	259,244
Caliendo, Giacomo	256,182
Arrigoni, Paolo	241,418
De Cristofaro, Peppe	225,618
Crimi, Vito Claudio	214,065
Divina, Sergio	211,355
D'Ali', Antonio	207,332
Endrizzi, Giovanni	198,530
Uras, Luciano	190,579
Centinaio, Gian Marco	184,274
Consiglio, Nunziante	172,136

Figura 13: Periodo pre-pandemia

speech.speaker_name	frequency
Conte, Giuseppe	74,217
Calderoli, Roberto	64,348
Malan, Lucio	47,464
La Russa, Ignazio	46,431
Romeo, Massimiliano	45,652
Urso, Adolfo	42,282
Salvini, Matteo	39,424
Gasparri, Maurizio	38,640
Speranza, Roberto	38,008
Zaffini, Francesco	37,214
Bagnai, Alberto	33,806
Faraone, Davide	31,359
Errani, Vasco	30,476
Comincini, Eugenio	29,812
Renzi, Matteo	27,976
de Bertoldi, Andrea	26,967
Briziarelli, Luca	26,633

Figura 14: Periodo post-pandemia

speech.speaker_name	frequency
De Petris, Loredana	700,056
Lanzillotta, Linda	433,679
Alberti Casellati, Maria Elisabetta	299,084
Fedeli, Valeria	262,801
Stefani, Erika	214,581
Bonfrisco, Anna Cinzia	211,998
Petraglia, Alessia	155,564
Bernini, Anna Maria	150,504
Fucksia, Serenella	150,111
Mussini, Maria	144,757
Montevecchi, Michela	139,747
Taverna, Paola	138,725
Nugnes, Paola	133,107
Bencini, Alessandra	125,239
Blundo, Rosetta Enza	116,265
Rossomando, Anna	110,093
Lo Moro, Doris	106,493
Di Giorgi, Rosa Maria	101,500
Paglini, Sara	95,313
Bisinella, Patrizia	95,241

Figura 15: Periodo pre-pandemia

speech.speaker_name	frequency
Alberti Casellati, Maria Elisabetta	75,128
De Petris, Loredana	65,982
Rossomando, Anna	40,911
Binetti, Paola	35,336
Gallone, Maria Alessandra	30,984
Rauti, Isabella	27,302
Bernini, Anna Maria	26,815
Modena, Fiammetta	26,476
Boldrini, Paola	25,239
Taverna, Paola	23,589
Garavini, Laura	23,042
Conzatti, Donatella	22,960
Valente, Valeria	21,883
Sbrollini, Daniela	21,541
Rizzotti, Maria	21,540
Castellone, Maria Domenica	20,909
Parente, Annamaria	20,560
Rivolta, Erica	19,135
Bonino, Emma	17,773
Stefani, Erika	17,311

Figura 16: Periodo post-pandemia

scegliere cosa far emergere dal confronto: se concentrarsi sulle parole più frequenti oppure su quelle meno frequenti. Il riquadro bianco in cui inserire il valore, invece, si riferisce alla misura statistica utilizzata per calcolare le parole chiave, simple math[24] (Fig. 17).

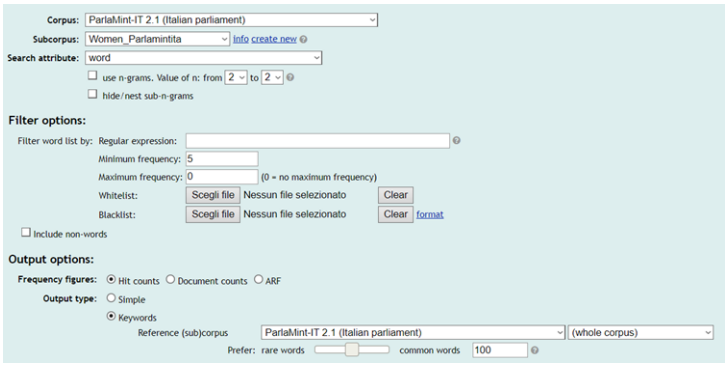


Figura 17: Analisi per Parole Chiave

In generale, più alto è il valore (100, 1000, ...) di simple math, maggiore è la possibilità che vengano estratte parole a più alta frequenza (parole più comuni). Mentre, più basso il valore (1, 0,1, ...), maggiore è la possibilità che vengano estratte parole a bassa frequenza (parole più rare). Una volta presa questa decisione, si può procedere. Nella Fig. 18, per esempio, la tabella mostra i dati relativi alla comparazione tra il corpus Men_ParlamintIta e Women_ParlamintIta, che avevamo precedentemente creato. In questo caso, si è scelto di mostrare quelle parole che avessero una frequenza maggiore rispetto a quelle meno frequenti. Provate a vedere cosa accade cambiando il valore di Simple Math nella schermata iniziale. Per esempio, la Fig. 18 mostra quelle che sono le parole più utilizzate dagli uomini, essendo il corpus relativo alle donne selezionato come corpus di riferimento. Se cliccate sul pulsante Switch focus and reference (sub)corpus, vi accorgerete che le parole cambieranno poiché in questo secondo caso il corpus da cui vengono estratte le parole chiave sarà quello relativo alle donne, come mostrato in Fig. 19.

Il numero di comparazioni e di analisi per parole chiave che possono essere prodotte è molto vario. Qui si è mostrato come, cambiando l'ordine del corpus principale e del corpus di riferimento, la risultante lista di parole estratte cambia. Questo è un aspetto fondamentale da comprendere per questo tipo di analisi. In questo senso, la fase preliminare di selezione dei corpora da comparare diviene fondamentale. Inoltre, considerando le varie possibilità di creazione offerte da NoSketchEngine in relazione al corpus ParlaMint-IT, questa tipologia di analisi può rivelarsi molto fruttuosa in termini di risultati. Sempre per dare un esempio, nelle successive figure 20 e 21 sono stati comparati il corpus dei discorsi maschili (fig.20) e femminili (fig. 21) durante il Covid e prima del Covid con quelli relativi al periodo pre-pandemico. Le figure riportano solo le prime 20 parole chiave. Come potrete notare, i lemmi più utilizzati nei discorsi prodotti dai parlamentari di sesso sia maschile che femminile durante il Covid, rispetto agli stessi discorsi nel periodo precedente la pandemia, sono molto simili tra i generi. Questo risultato è probabilmente viziato dall'occorrenza pandemica. È altamente probabile che il confronto tra un corpus di discorsi pronunciati durante una pandemia e un corpus di discorsi pronunciati in un

[24] *Simple Math* è la misura statistica usata per calcolare il punteggio di keyness in Sketch Engine. Questo punteggio è usato per identificare keywords o parole chiave. In concreto, identifica gli elementi che appaiono più frequentemente nel corpus 1 rispetto al corpus 2. Utilizza frequenze relative (per milione) e, pertanto, permette di comparare corpora di dimensioni disuguali. Per maggiori dettagli consultare la voce del sito a questo indirizzo: ¡https://www.sketchengine.eu/documentation/simple-maths/¿.

	ParlaMint-IT 2.1 (Italian parliament) : Men_ParlamintIta		ParlaMint-IT 2.1 (Italian parliament) : Women_Parlamintita		
word	frequency	frequency/mill	frequency	frequency/mill	Score
sostanzialmente	2,755	126.8	529	59.6	1.4
convinto	1,363	62.7	135	15.2	1.4
segreto	1,863	85.7	293	33.0	1.4
Senatore	4,774	219.7	1,159	130.5	1.4
Votazione	3,721	171.2	869	97.9	1.4
RAI	1,933	88.9	339	38.2	1.4
Se	10,297	473.8	2,870	323.2	1.4
signor	12,611	580.2	3,570	402.0	1.4
PdL	1,968	90.5	371	41.8	1.3
partito	3,585	164.9	865	97.4	1.3
credo	10,069	463.3	2,855	321.5	1.3
giudice	2,269	104.4	471	53.0	1.3
maggioranza	13,457	619.2	3,898	438.9	1.3
Ripresa	3,975	182.9	1,007	113.4	1.3
ragione	5,214	239.9	1,395	157.1	1.3
FI	1,474	67.8	247	27.8	1.3
qualche	9,771	449.6	2,836	319.3	1.3
cose	7,075	325.5	1,997	224.9	1.3
presidente	10,198	469.2	2,984	336.0	1.3
giudizio	2,745	126.3	654	73.6	1.3
XVII	1,698	78.1	326	36.7	1.3
Partito	4,531	208.5	1,216	136.9	1.3
qualcuno	4,805	221.1	1,305	146.9	1.3

Figura 18: Comparazione parole-chiave tra corpus Uomini e corpus Donne in ParlamintIta

	ParlaMint-IT 2.1 (Italian parliament) : Women_Parlamintita		ParlaMint-IT 2.1 (Italian parliament) : Men_Parlamintita		
word	frequency	frequency/mill	frequency	frequency/mill	Score
donne	4,326	487.1	2,208	101.6	2.9
scuola	3,311	372.8	3,253	149.7	1.9
violenza	2,098	236.2	1,779	81.9	1.8
bambini	1,842	207.4	1,903	87.6	1.6
donna	1,167	131.4	997	45.9	1.6
scuole	1,706	192.1	1,882	86.6	1.6
ragazzi	1,223	137.7	1,153	53.0	1.6
formazione	1,735	195.4	2,023	93.1	1.5
Senatrice	1,014	114.2	889	40.9	1.5
istruzione	1,158	130.4	1,183	54.4	1.5
salute	2,477	278.9	3,352	154.2	1.5
povertà	1,138	128.1	1,203	55.3	1.5
minori	1,565	176.2	1,930	88.8	1.5
studenti	1,170	131.7	1,306	60.1	1.4
giovani	2,231	251.2	3,141	144.5	1.4
davvero	2,950	332.2	4,413	203.0	1.4
femminicidio	475	53.5	183	8.4	1.4

Figura 19: Comparazione parole-chiave tra corpus Donne e corpus Uomini in ParlamintIta

periodo privo di pandemie globali mostri come le parole chiave siano termini legati alla straordinarietà dell'evento pandemico.

	ParlaMint-IT 2.1 (Italian parliament) : Men_Parlamintita_Covid		ParlaMint-IT 2.1 (Italian parliament) : Men_Parlamintita		
lemma	frequency	frequency/mill	frequency	frequency/mill	Score
emergenza	2,103	959.4	5,214	239.9	1.6
sanitario	1,350	615.9	4,306	198.1	1.3
miliardo	1,845	841.7	8,755	402.8	1.3
ci	11,986	5468.0	85,846	3949.8	1.3
Conte	829	378.2	1,553	71.5	1.3
decreto	2,020	921.5	10,763	495.2	1.3
Italia	4,140	1888.7	27,374	1259.5	1.3
Ministro	2,429	1108.1	14,442	664.5	1.3
scuola	1,235	563.4	5,227	240.5	1.3
tutto	9,550	4356.7	70,734	3254.5	1.3
misura	1,916	874.1	11,007	506.4	1.2
pandemia	602	274.6	611	28.1	1.2
virus	578	263.7	706	32.5	1.2
ministro	1,712	781.0	10,117	465.5	1.2
opposizione	1,112	507.3	5,288	243.3	1.2
MES	516	235.4	542	24.9	1.2
decretoslegge	1,607	733.1	9,847	453.1	1.2
euro	1,954	891.4	12,929	594.9	1.2

Figura 20: Comparazione parole-chiave tra corpus Uomini post-pandemia e pre-pandemia

	ParlaMint-IT 2.1 (Italian parliament) : Men_Parlamintita_Covid		ParlaMint-IT 2.1 (Italian parliament) : Men_Parlamintita		
lemma	frequency	frequency/mill	frequency	frequency/mill	Score
emergenza	2,103	959.4	5,214	239.9	1.6
sanitario	1,350	615.9	4,306	198.1	1.3
miliardo	1,845	841.7	8,755	402.8	1.3
ci	11,986	5468.0	85,846	3949.8	1.3
Conte	829	378.2	1,553	71.5	1.3
decreto	2,020	921.5	10,763	495.2	1.3
Italia	4,140	1888.7	27,374	1259.5	1.3
Ministro	2,429	1108.1	14,442	664.5	1.3
scuola	1,235	563.4	5,227	240.5	1.3
tutto	9,550	4356.7	70,734	3254.5	1.3
misura	1,916	874.1	11,007	506.4	1.2
pandemia	602	274.6	611	28.1	1.2
virus	578	263.7	706	32.5	1.2
ministro	1,712	781.0	10,117	465.5	1.2
opposizione	1,112	507.3	5,288	243.3	1.2
MES	516	235.4	542	24.9	1.2
decretoslegge	1,607	733.1	9,847	453.1	1.2
euro	1,954	891.4	12,929	594.9	1.2

Figura 21: Comparazione parole-chiave tra corpus Donne post-pandemia e pre-pandemia

In alternativa, si possono affiancare due liste di parole che non sono state prodotte attraverso la funzionalità keywords, ma, come si è visto nella parte iniziale, semplicemente attraverso un calcolo delle frequenze. Così si sono prodotte due liste delle prime 10 parole più frequenti nel corpus Women_ParlamintIta_Covid e Men_ParlamintIta_Covid, come mostrano i grafici 1 e 2, il panorama sulle parole più utilizzate dai parlamentari maschili e femminili cambia consistentemente. Questo perché i valori qui non risultano da una comparazione, che in quanto tale calcola la lista in base a delle differenze di distribuzione, ma sono legati solamente ad un corpus.

4 Concordanze, collocazioni e distribuzione diacronica – uno sguardo al significato attraverso le parole nel contesto

In questa sezione invece ci occuperemo del significato di parole o concetti attraverso l'uso delle concordanze [25] e delle collocazioni [26]. Le concordanze sono uno strumento di

[25] Per concordanza si intende una serie di esempi di una data parola o frase che mostrano il contesto

[26] Le collocazioni sono quelle parole che si trovano nelle vicinanze di una specifica parola di ricerca. Lo studio delle collocazioni rivela aspetti del significato e dell'uso di una data parola nel suo contesto di utilizzo.

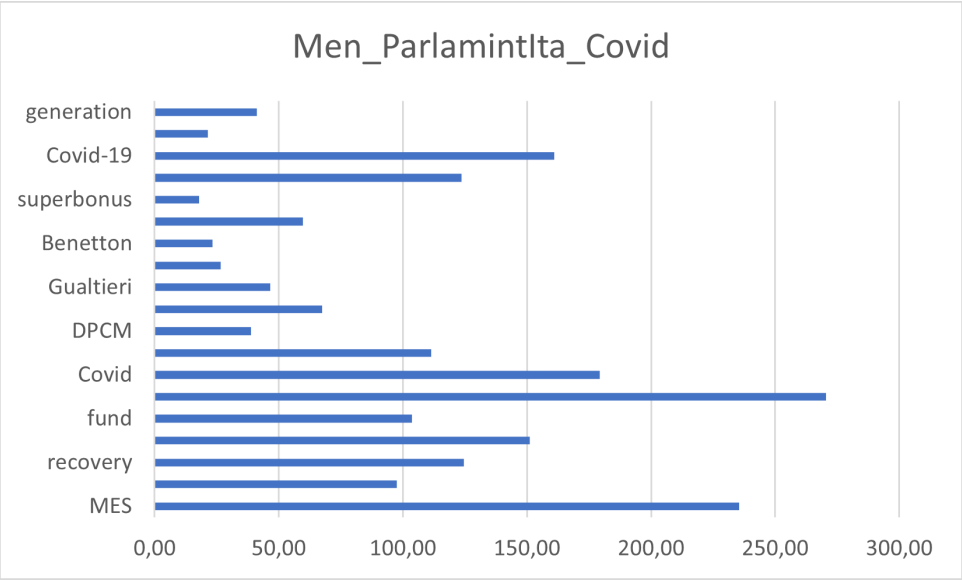


Grafico 1: Parole più utilizzate in Men_ParlamintIta_Covid per frequenza relativa

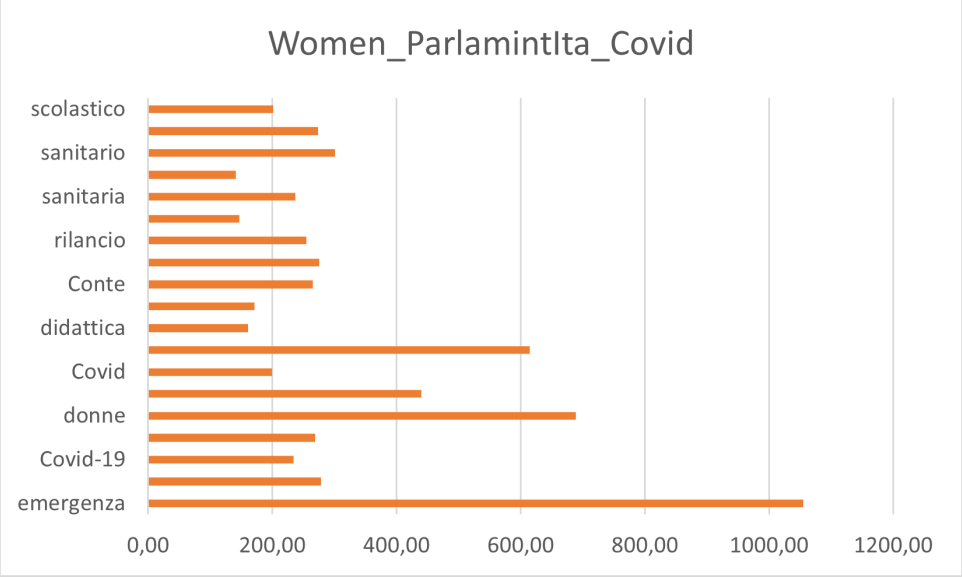
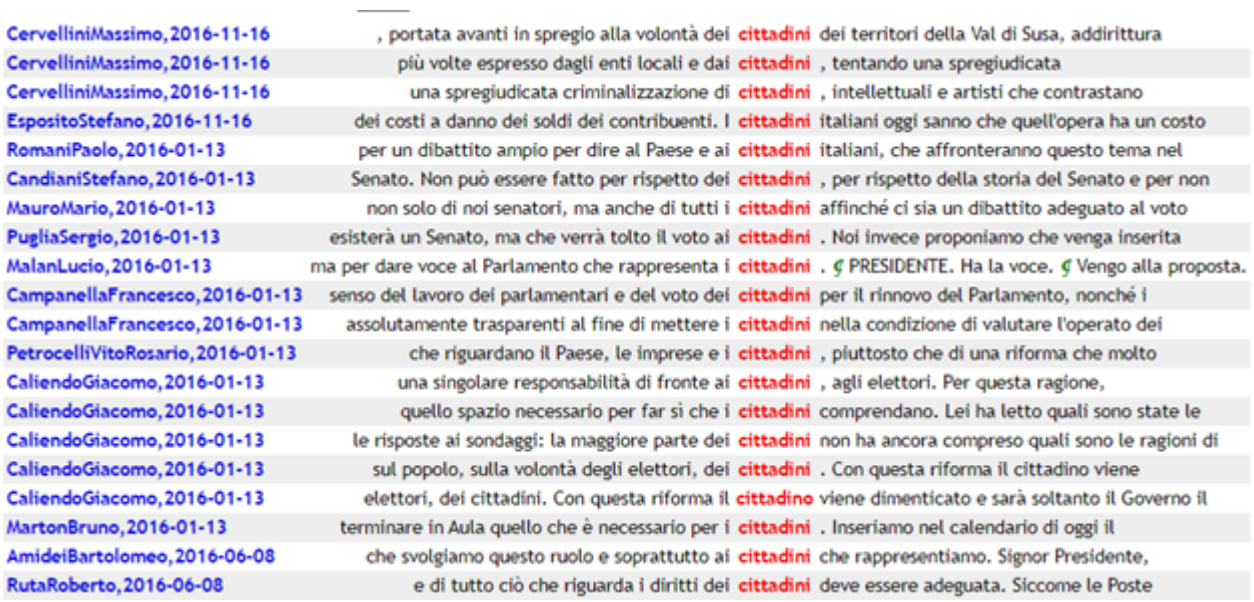


Grafico 2: Parole più utilizzate in Women_ParlamintIta_Covid per frequenza relativa

investigazione molto potente con una varietà di opzioni di ricerca. Le concordanze possono essere ordinate secondo criteri statistici e linguistici per ottenere il risultato desiderato: si possono trovare parole, frasi, tag, documenti, tipi di testo o strutture del corpus e mostrare i risultati nel contesto sotto forma di concordanza. La modalità di visualizzazione del testo è per frasi disposte verticalmente in serie, evidenziando una parola nodo al centro in maniera tale da sottolineare la presenza di pattern o fenomeni ricorrenti (22).



CervelliniMassimo,2016-11-16	, portata avanti in spregio alla volontà dei	cittadini	dei territori della Val di Susa, addirittura
CervelliniMassimo,2016-11-16	più volte espresso dagli enti locali e dai	cittadini	, tentando una spregiudicata
CervelliniMassimo,2016-11-16	una spregiudicata criminalizzazione di	cittadini	, intellettuali e artisti che contrastano
EspositoStefano,2016-11-16	dei costi a danno dei soldi dei contribuenti. I	cittadini	italiani oggi sanno che quell'opera ha un costo
RomaniPaolo,2016-01-13	per un dibattito ampio per dire al Paese e ai	cittadini	italiani, che affronteranno questo tema nel
CandianiStefano,2016-01-13	Senato. Non può essere fatto per rispetto dei	cittadini	, per rispetto della storia del Senato e per non
MauroMario,2016-01-13	non solo di noi senatori, ma anche di tutti i	cittadini	affinché ci sia un dibattito adeguato al voto
PugliaSergio,2016-01-13	esisterà un Senato, ma che verrà tolto il voto ai	cittadini	. Noi invece proponiamo che venga inserita
MalanLucio,2016-01-13	ma per dare voce al Parlamento che rappresenta i	cittadini	. ¶ PRESIDENTE. Ha la voce. ¶ Vengo alla proposta.
CampanellaFrancesco,2016-01-13	senso del lavoro dei parlamentari e del voto dei	cittadini	per il rinnovo del Parlamento, nonché i
CampanellaFrancesco,2016-01-13	assolutamente trasparenti al fine di mettere i	cittadini	nella condizione di valutare l'operato dei
PetrocelliVitoRosario,2016-01-13	che riguardano il Paese, le imprese e i	cittadini	, piuttosto che di una riforma che molto
CaliendoGiacomo,2016-01-13	una singolare responsabilità di fronte ai	cittadini	, agli elettori. Per questa ragione,
CaliendoGiacomo,2016-01-13	quello spazio necessario per far sì che i	cittadini	comprendano. Lei ha letto quali sono state le
CaliendoGiacomo,2016-01-13	le risposte ai sondaggi: la maggiore parte dei	cittadini	non ha ancora compreso quali sono le ragioni di
CaliendoGiacomo,2016-01-13	sul popolo, sulla volontà degli elettori, dei	cittadini	. Con questa riforma il cittadino viene
CaliendoGiacomo,2016-01-13	elettori, dei cittadini. Con questa riforma il	cittadino	viene dimenticato e sarà soltanto il Governo il
MartonBruno,2016-01-13	terminare in Aula quello che è necessario per i	cittadini	. Inseriamo nel calendario di oggi il
AmideiBartolomeo,2016-06-08	che svolgiamo questo ruolo e soprattutto ai	cittadini	che rappresentiamo. Signor Presidente,
RutaRoberto,2016-06-08	e di tutto ciò che riguarda i diritti dei	cittadini	deve essere adeguata. Siccome le Poste

Figura 22: Concordanza del lemma CITTADINO

Tuttavia, nonostante le sue potenzialità, le concordanze con grandi corpora come ParlaMint-IT possono corrispondere a così tanti risultati che analizzarli e interpretarli può essere un processo molto lungo. In questo caso l'analisi delle collocazioni (o *collocates*, in inglese) ci viene in aiuto. Esse rappresentano le parole che occorrono più frequentemente intorno ad una parola nodo; tale analisi è utile, poiché alcune parole hanno la tendenza ad occorrere con certe altre. Parafrasando [2](p.11), per conoscere davvero una parola, devi conoscere con chi si accompagna [2]. In questo senso, il significato di una parola è influenzato dalle parole con cui occorre. Una collocazione è una parola che occorre frequentemente con la parola nodo che si sta studiando. Bisogna differenziare il concetto di collocazione da quello di *"colligation"* con cui si indica un legame di co-occorrenza grammaticale come, per esempio, quello che lega i verbi con alcune preposizioni ('ANDARE' + 'con' o 'a'). Le collocazioni invece non sono motivate grammaticalmente, ma semanticamente. Per esempio, il verbo "causare" ha una tendenza ad avere delle collocazioni dal significato negativo (es. causare un danno). Questo viene definito dalla frequente co-occorrenza di questo verbo con parole dal significato principalmente negativo e non da regole grammaticali. Su NoSketchEngine è possibile studiare le collocazioni e visualizzare le concordanze. Per uno studio collocazionale, cliccare su search nella colonna a sinistra (23) e decidere che tipologia di ricerca effettuare:

- Simple – ricerca esattamente la stringa inserita
- Lemma – ricerca per lemma
- Phrase – una frase intera

- Word – ricerca di parole con uso di espressioni regolari
- Character – ricerca la stringa di caratteri anche all'interno di altre parole
- CQL[27] – corpus query language – una sintassi specifica per fare ricerche complesse e molto precise all'interno del corpus

Un esercizio potrebbe essere quello di cercare la parola "reddito" all'interno del corpus Parlimint-IT e di vedere il suo significato. In questo caso, volendo ricercare principalmente il reddito inteso come concetto politico-economico di flusso economico o di servizi ricevuto dai singoli, dalla collettività o da imprese, è preferibile evitare di fare una ricerca per lemmi, poiché il plurale "redditi" può riferirsi principalmente allo stipendio annuo percepito da un singolo. Per cui, selezionando Word in query type, si inserisce la parola "reddito" all'interno della stringa bianca. In questo caso, non verrà selezionato nessun sotto-corpus specifico perché l'obiettivo è quello di indagare l'uso di tale parole e di eventuali pattern legati a tale uso su un ampio campione di dati e quindi non c'è necessità di restringere il campo. Una ulteriore possibilità è quella di restringere il campo a quei contesti in cui appaiono determinate parole. Questo è possibile attraverso il filtro context che si trova sotto le stringhe di ricerche (24). In questo caso verrà lasciato bianco, ma si può invece essere interessati ai discorsi in cui si parla di "aumento" e quindi si può cercare "aumento" in una finestra comprensiva di 15 parole a sinistra e a destra del lemma.

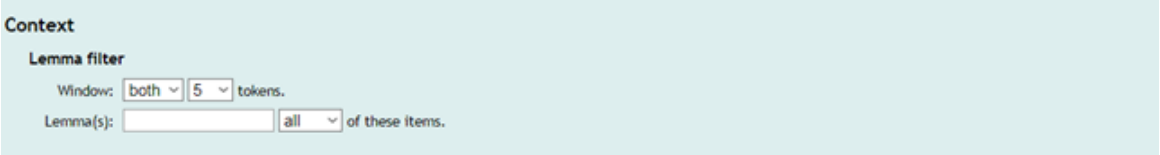


Figura 23: Selezione del contesto

Infine, nella sezione sottostante si può selezionare un metadato secondo l'annotazione spiegata nella sezione 2 oppure selezionare un sotto-corpus precedentemente creato. In questo esempio, l'interesse si rivolge alla totalità del corpus. Cliccando su make concordance si ottiene il risultato in (25).

Una delle possibilità è quella di scorrere e leggere attraverso i vari esempi. Naturalmente, già ad un primo colpo d'occhio sembra proprio che il discorso sul reddito sia, come ci si poteva aspettare, fortemente legato alla questione del reddito di cittadinanza, proposto dal Movimento 5 Stelle e divenuto legge durante la XVIII legislatura. Per esplorare una porzione più ampia del contesto, si può cliccare su una riga delle concordanze; nella parte bassa della schermata apparirà una finestra gialla 26 con la possibilità di aumentare la dimensione del testo visualizzato cliccando su ¡ previous e next ¿.

Inoltre, da una schermata di concordanze, varie opzioni sono possibili 27:

- Save – Salvare la ricerca
- Create subcorpus – Creare direttamente un sotto-corpus a partire da quei discorsi selezionati dalla nostra ricerca
- View options – Scegliere la modalità di visualizzazione (KWIC – per concordanze, o Sentence – per vedere la frase completa per ogni esempio)

[27] Al link seguente potete trovare un tutorial che mostra come utilizzare la CQL ¡https://sidih.si/cdn/120/1_EN_subcorpus.mp4¿

Figura 24: Concordanza per la parole "Reddito"

- Sort – Distribuire gli esempi in base alla prima parola a sinistra (left) o a destra (right) in ordine alfabetico o in base alle parole chiave (se più di una) in ordine alfabetico; in base alla data del discorso in ordine cronologico dal più recente al più vecchio (reference); o in modo casuale (shuffle).

- Sample – estrarre un campione random di frasi dalla ricerca

- Filter – filtrare nuovamente la ricerca in base alle varie annotazioni – evidenziare solo alcuni partiti o alcuni politici o alcune date, o selezionare solo le prime occorrenze della parola nel caso nel discorso ce ne fossero più di due

- Frequency – calcolo delle frequenze in base a Node forms – frequenza della parola di ricerca, Doc IDs – dell'ID del parlante o Text type – data del documento.

Tuttavia, come dicevamo precedentemente, potrebbe essere utile fare ordine e visualizzare la distribuzione delle parole che fanno da contesto alla parola nodo attraverso la funzionalità collocations, selezionando questa funzione dal menù di sinistra (penultima opzione). Dopo aver cliccato, comparirà il seguente riquadro, in cui potete scegliere diversi criteri di selezione (27).

Ci sono due cose da tenere sotto controllo: il numero di parole di distanza dalla parola nodo selezionato – es. calcolare le parole più frequenti entro 15 parole a sinistra e a destra dalla parola nodo e la tipologia di misura statistica che viene utilizzata per filtrare le collocazioni. Naturalmente le diverse misure statistiche privilegiano varie tipologie di associazione tra le parole, e danno quindi risultati diversi. Senza entrare troppo nel dettaglio [28], ci limiteremo a dire che misure come T-score e Log-Likelihood tendono ad evidenziare tali parole molto comuni e altamente frequenti all'interno di un dataset mentre MI (Mutual Information) o MI3 mettono in risalto parole molto rare che hanno una frequenza in generale molto bassa e che, laddove emergano dall'analisi collocazionale,

[28] Per approfondire, vedere [1]

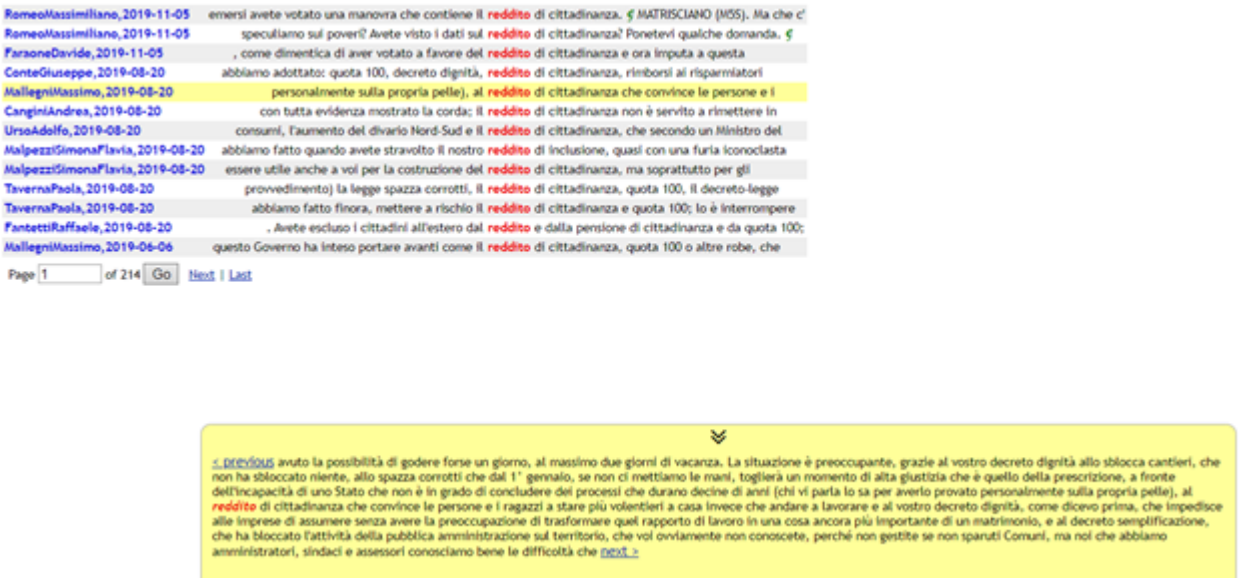


Figura 25: Visualizzazione del testo

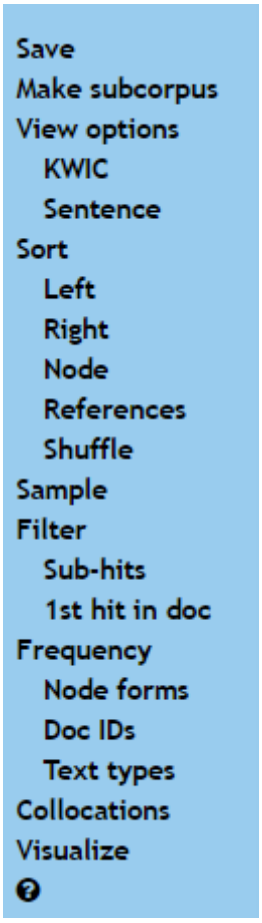


Figura 26: possibilità di esplorazione delle concordanze

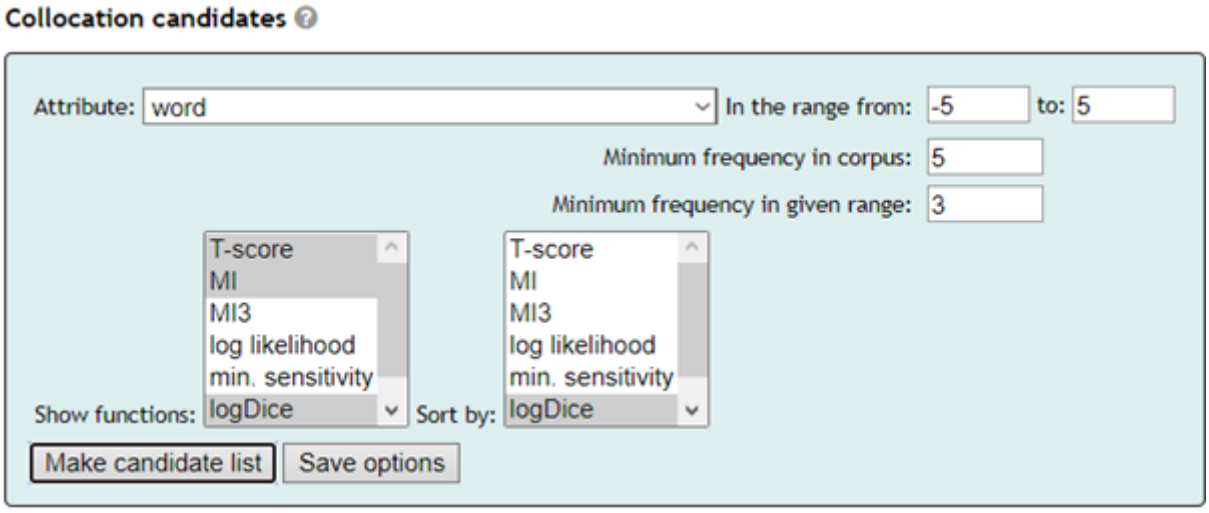


Figura 27: criteri di visualizzazione

spesso si trovano esclusivamente in associazione con la parola nodo[29]. In questo caso, ci si limiterà a selezionare una finestra di 10 parole a sinistra e a destra della nostra parola nodo, lasciando le altre informazioni invariate. Mentre come attribute, invece si selezionano i lemmi, poiché interessati al significato piuttosto che alla forma. Cosa notate? (cfr. 28)

Adesso si può tentare di ripetere questo calcolo analizzando però solamente le menzioni di reddito del M5S e poi della Lega. Come fare? Cliccate nella colonna a sinistra su concordance e poi su filter e, nella schermata che apparirà, nella sezione Speech.Speaker_Party_Name, selezionate Movimento 5 Stelle. A questo punto, cliccate su filter concordances e poi successivamente ripetete la stessa operazione per calcolare le collocazioni. Successivamente, partendo dalla schermata in Fig. 28, ripetete la stessa operazione di filtraggio, ma ora dovete compiere un passaggio in più, cioè deselezionare il filtro-Movimento 5 Stelle. Dopo aver cliccato sulla scritta concordance, dovete cliccare sulla stringa in Fig. 29 sulla parola 'reddito' e deselezionerete il filtro Movimento 5 Stelle. Dopo di questo, ripetete l'operazione di filtraggio appena visto per selezionare il partito Lega.

Le successive figure 30 e 31 sono le collocazioni della stessa parola ma in riferimento a due partiti differenti. Cosa ne pensate? Notate delle variazioni?

Infine, altra funzionalità è quella della distribuzione diacronica, cioè quella che ci permette di vedere la parola distribuita nel tempo. Attualmente, il corpus ParlaMint-IT ha una copertura diacronica relativa solo a due legislature che però corrispondono a 8 anni di discorsi politici. Tuttavia, questa funzionalità, con l'auspicio che in futuro il corpus venga allargato, potrebbe avere un impatto fondamentale soprattutto in termini di ricerca storica. Come si può analizzare la distribuzione storica di un termine?

Dopo aver nuovamente selezionato search e aver cercato la vostra parola secondo i parametri da voi scelti, dovrete prima calcolare le concordanze. A questo punto, sempre nella colonna a sinistra, cliccare su visualize. Le Figure 32 e 33, per esempio, mostrano rispettivamente la distribuzione del lemma "patrimoniale" e del lemma "vitalizio". Lo 0% corrisponde alle date più recenti mentre il 100% a quelle più antiche. Il grafico risultante, inoltre, può essere modificato in base alla granularity, che indica la grandezza degli

[29]Se non siete utenti esperti, lasciate le informazioni default che mettono in risalto le parole più frequenti da un punto di vista statistico.

		Cooccurrence count	Candidate count	T-score	MI	logDice
P | N	cittadinanza	889	3,169	29.808	11.937	12.408
P | N	quota	115	1,733	10.712	9.857	9.907
P | N	100	112	2,537	10.565	9.269	9.600
P | N	sostegno	155	5,482	12.418	8.626	9.370
P | N	garantito	55	1,098	7.405	9.451	9.098
P | N	basso	50	894	7.062	9.610	9.053
P | N	minimo	70	2,331	8.346	8.713	8.986
P | N	inclusione	42	808	6.471	9.505	8.842
P | N	reddito	89	4,248	9.401	8.194	8.823
P | N	povertà	62	2,341	7.852	8.532	8.808
P | N	pensioni	40	1,057	6.312	9.047	8.657
P | N	percettori	28	94	5.290	12.023	8.650
P | N	imposte	38	1,053	6.152	8.978	8.584
P | N	medio	37	979	6.071	9.045	8.579
P | N	capite	27	327	5.191	10.172	8.457
P | N	famiglie	91	6,344	9.491	7.647	8.448
P | N	pro	27	547	5.188	9.430	8.336
P | N	lavoratori	96	7,898	9.740	7.408	8.284
P | N	imposta	36	1,605	5.980	8.292	8.280
P | N	agricoltori	28	1,005	5.277	8.605	8.165
P | N	impresa	43	2,717	6.527	7.789	8.165
P | N	misura	59	5,045	7.634	7.353	8.061
P | N	disponibile	24	861	4.886	8.606	8.009
P | N	redistribuzione	19	257	4.354	10.013	7.991
P | N	imponibile	18	154	4.240	10.674	7.975

Figura 28: lista delle collocazione di 'reddito'

Figura 29: cruscotto delle collocazioni

		Cooccurrence count	Candidate count	T-score	MI	logDice
P | N	**cittadinanza**	673	3,169	25.937	12.437	12.310
P | N	**minimo**	95	2,331	9.737	10.056	9.795
P | N	**reddito**	121	4,248	10.985	9.539	9.514
P | N	**garantito**	48	1,098	6.922	10.157	9.436
P | N	**sostegno**	106	5,482	10.275	8.980	9.027
P | N	**povertà**	46	2,341	6.769	9.003	8.745
P | N	**quota**	37	1,733	6.071	9.123	8.705
P | N	**inserimento**	25	1,433	4.989	8.832	8.296
P | N	**100**	35	2,537	5.899	8.493	8.272
P | N	**beneficiari**	15	440	3.868	9.798	8.252
P | N	**1148**	11	14	3.316	14.325	8.247
P | N	**misura**	53	5,045	7.253	8.100	8.125
P | N	**dignità**	39	3,657	6.222	8.121	8.047
P | N	**dignitosa**	10	230	3.159	10.149	7.868
P | N	**beneficiario**	9	92	2.998	11.319	7.866
P | N	**percettori**	9	94	2.998	11.288	7.863
P | N	**introduzione**	23	2,066	4.779	8.183	7.862
P | N	**disponibile**	14	861	3.732	8.730	7.817
P | N	**fisiche**	10	355	3.157	9.523	7.745
P | N	**doppie**	8	69	2.827	11.564	7.722
P | N	**fasce**	11	558	3.310	9.008	7.702
P | N	**imposizioni**	8	137	2.826	10.574	7.645
P | N	**anticorruzione**	11	635	3.309	8.821	7.640
P | N	**pensione**	14	1,132	3.730	8.335	7.637
P | N	**attive**	11	642	3.309	8.805	7.634
P | N	**pensioni**	13	1,057	3.594	8.327	7.578
P | N	**universale**	12	897	3.454	8.448	7.570

Figura 30: collocazioni ‘reddito’ per il Movimento 5 Stelle

		Cooccurrence count	Candidate count	T-score	MI	logDice
P | N	**cittadinanza**	117	3,169	10.812	11.442	10.066
P | N	**percettori**	8	94	2.827	12.647	9.034
P | N	**basso**	10	894	3.158	9.720	7.977
P | N	**quota**	16	1,733	3.994	9.443	7.937
P | N	**autonomi**	7	564	2.642	9.870	7.885
P | N	**familiare**	10	998	3.158	9.561	7.866
P | N	**monoreddito**	3	33	1.731	12.742	7.806
P | N	**imposte**	8	1,053	2.823	9.162	7.489
P | N	**capite**	4	327	1.997	9.849	7.482
P | N	**percepito**	3	231	1.730	9.935	7.269
P | N	**100**	13	2,537	3.596	8.593	7.177
P | N	**pro**	4	547	1.996	9.107	7.103
P | N	**inclusione**	5	808	2.231	8.866	7.076
P | N	**pensioni**	6	1,057	2.443	8.741	7.070
P | N	**impresa**	12	2,717	3.453	8.379	6.976
P | N	**50.000**	3	383	1.729	9.206	6.961
P | N	**dipendente**	4	684	1.995	8.784	6.909
P | N	**integrare**	3	420	1.728	9.073	6.894
P | N	**agricoltori**	5	1,005	2.230	8.551	6.859
P | N	**imposta**	7	1,605	2.637	8.361	6.833
P | N	**produrre**	7	1,632	2.637	8.337	6.814
P | N	**medio**	4	979	1.993	8.267	6.564
P | N	**1.000**	3	646	1.727	8.451	6.546
P | N	**produce**	4	1,024	1.993	8.202	6.518
P | N	**fonte**	3	762	1.726	8.213	6.395
P | N	**agricoli**	3	797	1.725	8.148	6.352
P | N	**erogazione**	3	833	1.725	8.085	6.310
P | N	**riescono**	3	851	1.725	8.054	6.289
P | N	**disponibile**	3	861	1.725	8.037	6.277
P | N	**fini**	6	2,130	2.437	7.730	6.276
P | N	**prodotto**	8	3,194	2.813	7.561	6.186

Figura 31: collocazioni ‘reddito’ per la Lega

elementi in cui un dataset è suddiviso. In questo modo, è possibile selezionare i vari livelli di dettaglio (o di sintesi) dei dati raccolti. Maggiore è tale valore, maggiore è il numero di elementi in cui si suddivide un dataset e quindi meno sintetica è la visualizzazione – cioè maggiore il livello di dettaglio. Cosa potete dedurre dalle figure disposte qui sotto? Come è distribuito il lemma "patrimoniale"? Come è distribuito il lemma "vitalizio"?

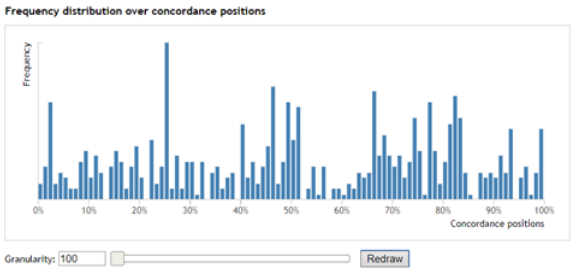


Figura 32: 'Patrimoniale'

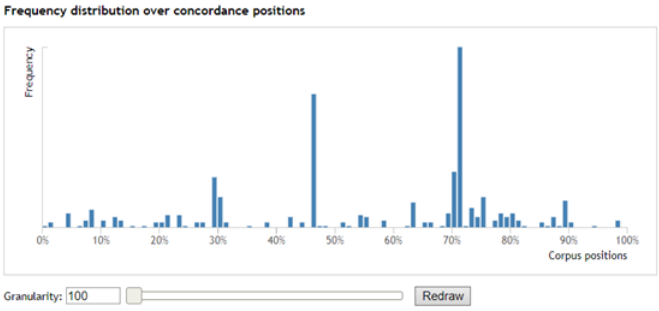


Figura 33: 'Vitalizio'

5 Note conclusive

Questo breve tutorial rappresenta una guida per i non-esperti che vogliano avvicinarsi allo studio dei dibattiti parlamentari grazie alla risorsa ParlaMint-IT inerente al progetto ParlaMint. Le possibilità di ricerca sono molteplici ma si è deciso di soffermarsi su quelle più importanti e che devono essere obbligatoriamente padroneggiate in vista di analisi più avanzate. Per approfondire la conoscenza dei corpora ParlaMint, uno strumento utile è anche il tutorial Voices of the Parliament, basato sui dati del parlamento sloveno[30].

Il progetto ParlaMint è di fondamentale importanza sia per la ricerca che per la didattica. Inoltre, esso può anche essere interpretato come un importante segnale di democrazia e uno strumento per rafforzarla. L'auspicio, quindi, è che tali archivi siano sempre più grandi e soprattutto liberamente accessibili.

6 Ringraziamenti

Tengo personalmente a ringraziare l'Istituto di Linguistica Computazionale "A.Zampolli" per il supporto fornitomi. Un ringrazimento alla Dott.ssa Francesca Frontini, alla Dott.ssa

[30] https://sidih.github.io/voices/index.html

Monica Monachini e alla Dott.ssa Valeria Quochi per i preziosi consigli. Un ulteriore ringraziamento va al Dott. Giacomo Morbiato e al Dott. Stefano Fortin per l'essenziale assistenza nella revisione della bozza finale del tutorial. Infine ringrazio Andrea Di Stefano e Giacomo Del Fante per aver dedicato il loro tempo come primi lettori della versione definitiva e aver fornito preziosi suggerimenti.

Riferimenti bibliografici

[1] V. Brezina. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press, 2018.

[2] J. R. Firth. *Papers in Linguistics, 1934-1951*. London: Oxford University Press, 1957.