I corpora del progetto Univers-ITA: Univers-ITA, Univers-ITA-ProUniv e Univers-ITA-ProGior

Home page del progetto: https://site.unibo.it/univers-ita/

1. Breve descrizione:

Il corpus **Univers-ITA** è composto da 2.137 testi redatti da studentesse e studenti iscritti, per l'a.a. 2020/21, al secondo anno di corsi di laurea triennale e magistrale a ciclo chiuso. Il corpus è rappresentativo per aree disciplinari (umanistica, scientifica, sanitaria e sociale) e geografiche (Nord, Centro e Sud + Isole). I testi sono stati redatti secondo una traccia comune, con uno stile formale e sono stati successivamente analizzati sia quantitativamente (individuando ad esempio il numero di farsi, di parole diverse ecc. per ogni testo), sia qualitativamente (attraverso l'annotazione manuale di tutti i tratti devianti rispetto a quanto prescritto dall'italiano normativo e scolastico). Le informazioni ricavate da questa doppia analisi sono state implementate nel corpus e rappresentano possibili chiavi di ricerca, al pari dell'ampio corredo di metadati sociobiografici ricavati dal questionario che ogni partecipante al progetto ha compilato. In questo modo il corpus restituisce una fotografia attendibile delle reali competenze di scrittura formale di studenti e studentesse, consentendo anche di individuare correlazioni sistematiche tra tratti linguistici e profili sociobiografici degli e delle scriventi. Il corpus consta di 810.715 parole (tokens). Cfr. anche https://site.unibo.it/univers-ita/it/corpora/univers-ita/.

Il corpus **Univers-ITA-ProUniv** è costituito da tesi, tesine, relazioni, ecc. nella versione *non* corretta dai docenti. Complessivamente, contiene 773 testi per 6.267.765 tokens. I testi sono stati ripuliti eliminando le sezioni dedicate ai riferimenti bibliografici e le citazioni molto lunghe (esterne al corpo del testo) e poi classificati secondo i seguenti metadati, utilizzabili come filtri di ricerca nel corpus: *Tipo di testo, Corso, Sede ateneo, Tipologia del corso, Anno accademico di redazione del testo, Area di nascita, Genere.* Il corpus può essere consultato in modalità bilanciata o non bilanciata (ovvero nella sua interezza). Infatti, impiegando gli stessi parametri di campionamento adottati per il corpus Univers-ITA, è stato creato a posteriori un sottocorpus rappresentativo per area geografica dell'ateneo e per area disciplinare dei corsi di studio. Questo sottocorpus ha dimensioni piuttosto ridotte, è infatti costituito da 254 testi e 2.578.072 parole. Tutti i testi sono stati poi lemmatizzati ed etichettati per parte del discorso. Cfr. anche https://site.unibo.it/univers-ita/it/corpora/univers-ita-prouniv.

Il corpus **Univers-ITA-ProGior** contiene articoli di carattere giornalistico scritti da studenti e studentesse ed estratti da blog, giornali universitari online e siti di informazione/opinione gestiti da

studenti e studentesse universitari di diversi atenei italiani. I testi raccolti sono classificati per argomento trattato (ad es. arte-cultura, economia-società, ecc.), per collocazione geografica dell'ateneo associato al sito (Nord, Centro, Sud e Isole) e per anno di redazione del testo (dal 2012 al 2021). Queste tre informazioni sono attualmente utilizzabili come filtri di ricerca per l'interrogazione del corpus. Tutti i testi sono stati poi lemmatizzati ed etichettati per parte del discorso. Il corpus è composto da 1.630 testi per un totale di 1.692.846 parole. Cfr. anche https://site.unibo.it/univers-ita/it/corpora/univers-ita-prouniv.

2. Come citarli:

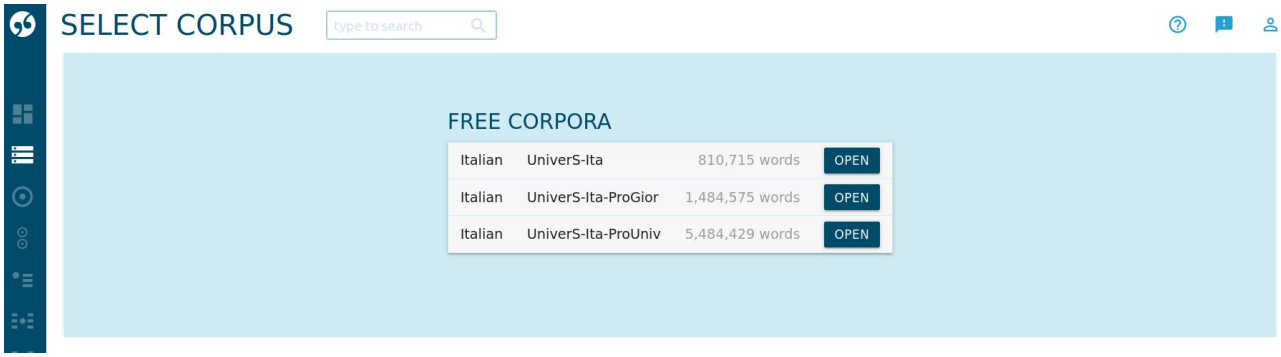
GRANDI, Nicola, BALLARÈ, Silvia, CHIUSAROLI, Francesca, GALLINA, Francesca, PASCOLI, Matteo, PISTOLESI, Elena; *Corpus Univers-ITA.* 2023, DOI: https://doi.org/10.60760/unibo/univers-ita

GRANDI, Nicola, BALLARÈ, Silvia, CHIUSAROLI, Francesca, GALLINA, Francesca, PASCOLI, Matteo, PISTOLESI, Elena; *Corpus Univers-ITA-ProUniv.* 2023, DOI: https://doi.org/10.60760/unibo/univers-ita-prouniv

GRANDI, Nicola, BALLARÈ, Silvia, CHIUSAROLI, Francesca, GALLINA, Francesca, PASCOLI, Matteo, PISTOLESI, Elena; *Corpus Univers-ITA-ProGior.* 2023, DOI: https://doi.org/10.60760/unibo/univers-ita-progior

3. Accesso ai corpora

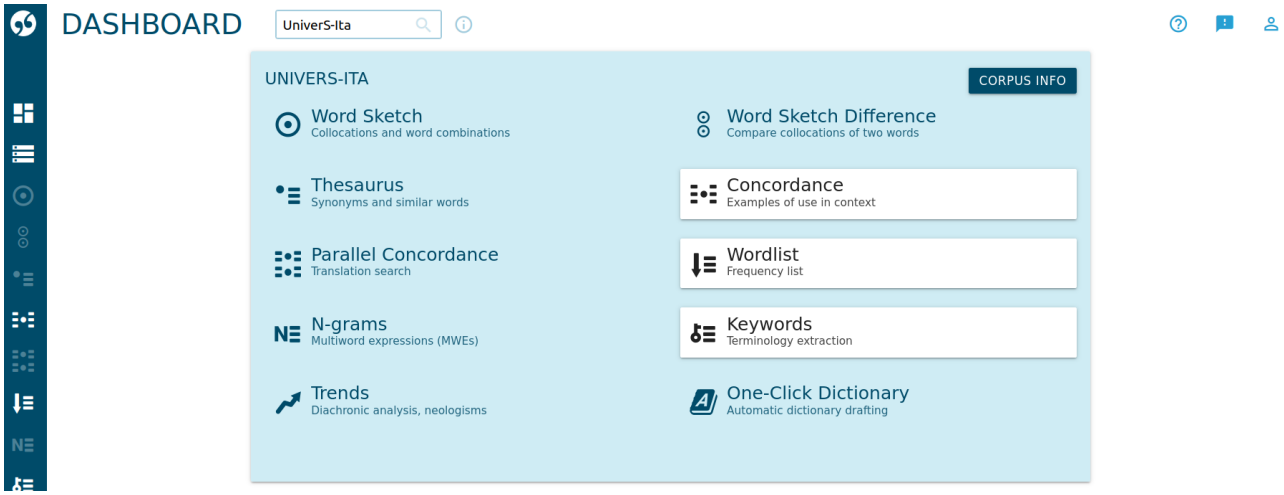
I corpora sono accessibili tramite l'URL: https://corpora.ficlit.unibo.it/CUSP



Nella schermata iniziale è possibile scegliere uno dei tre corpora: UniverS-Ita, UniverS-Ita-ProGior e UniverS-Ita-ProUniv.

4. Strumenti di NoSketchEngine

Una volta scelto il corpus, arriviamo alla schermata principale del corpus dove è possibile scegliere tra gli strumenti *Concordance*, *Wordlist* e *Keywords* di NoSketchEngine.

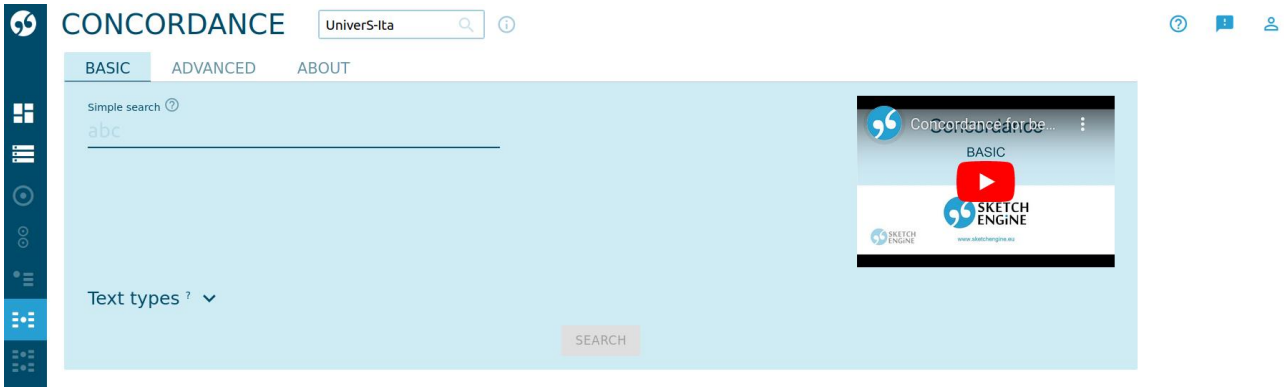


Concordance permette, evidentemente, di esplorare le concordanze nel corpus, *Wordlist* fornisce statistiche sul lessico del corpus e *Keywords* permette di cercare termini caratteristici di un corpus rispetto a un altro corpus di riferimento. Vedremo sotto come usare gli strumenti *Concordance* e *Wordlist*.

4.1. Concordance

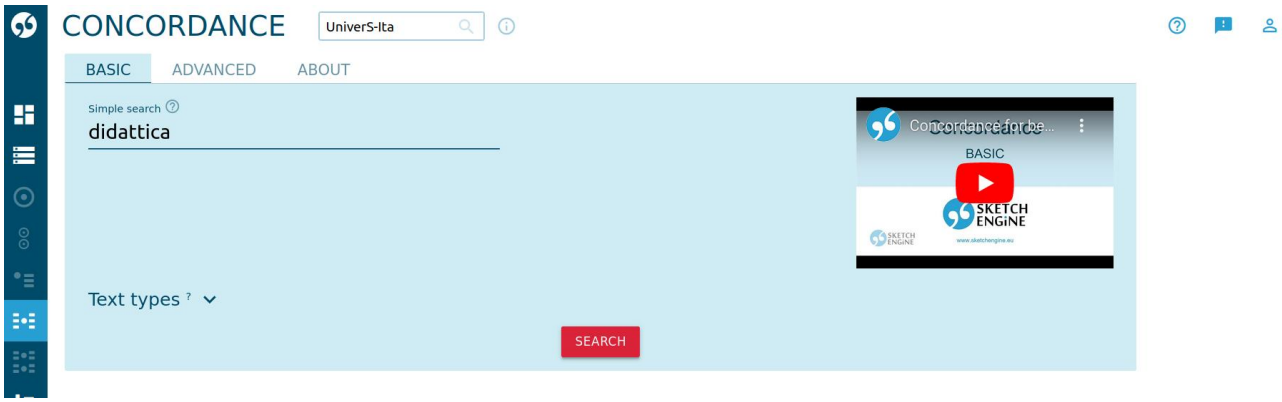
Attiviamo lo strumento *Concordance*, facendo clic sul relativo riquadro dopo aver scelto il corpus, oppure dal menu a scomparsa che si trova sulla sinistra della pagina.

Ricerca semplice



A questo punto, siamo nella sezione “Basic” della ricerca. Nel campo “Simple search” possiamo inserire una parola, di cui verranno cercate le occorrenze nel corpus. Se la parola inserita è la forma di riferimento di un lemma (es. un infinito, o un aggettivo maschile singolare), verranno trovate tutte

le forme flesse di quel lemma. Se inseriamo una serie di parole, o lemmi, SketchEngine cercherà nel corpus le occorrenze dove queste parole compaiono in sequenza.

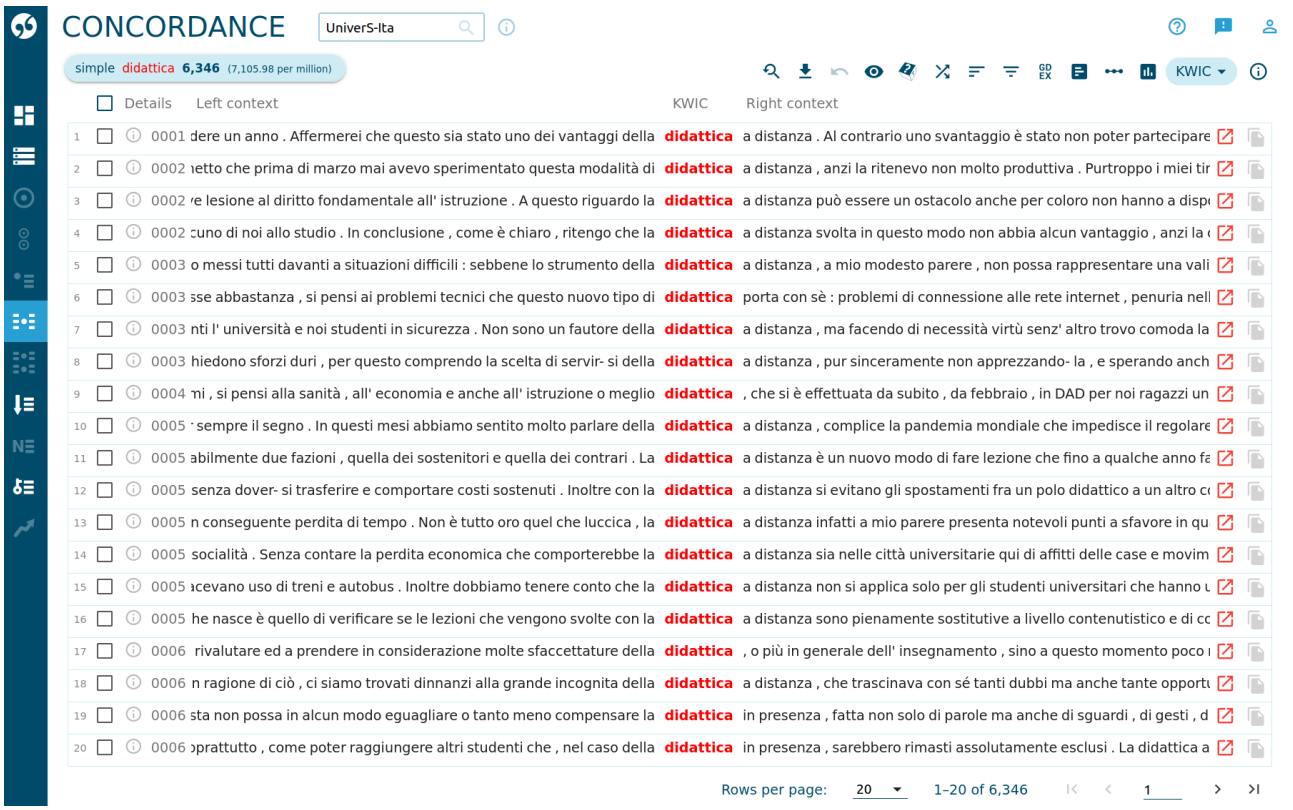


Notare che NoSketchEngine presenta sulla destra un video tutorial (in inglese) per l'uso dello strumento *Concordance*.

Dopo aver impostato la chiave di ricerca, in questo caso la parola "didattica", facciamo clic su "SEARCH".

4.1.1. Esplorazione dei risultati

Il sistema ci mostra la tabella con le occorrenze della parola (o sequenza di parole) cercata, che è mostrata in rosso nella colonna "KWIC" (*Key Word In Context*, ovvero Parola Chiave Nel Contesto). La finestra mostra 20 righe alla volta: nel navigatore in fondo alla pagina è possibile scorrere tutti i risultati.

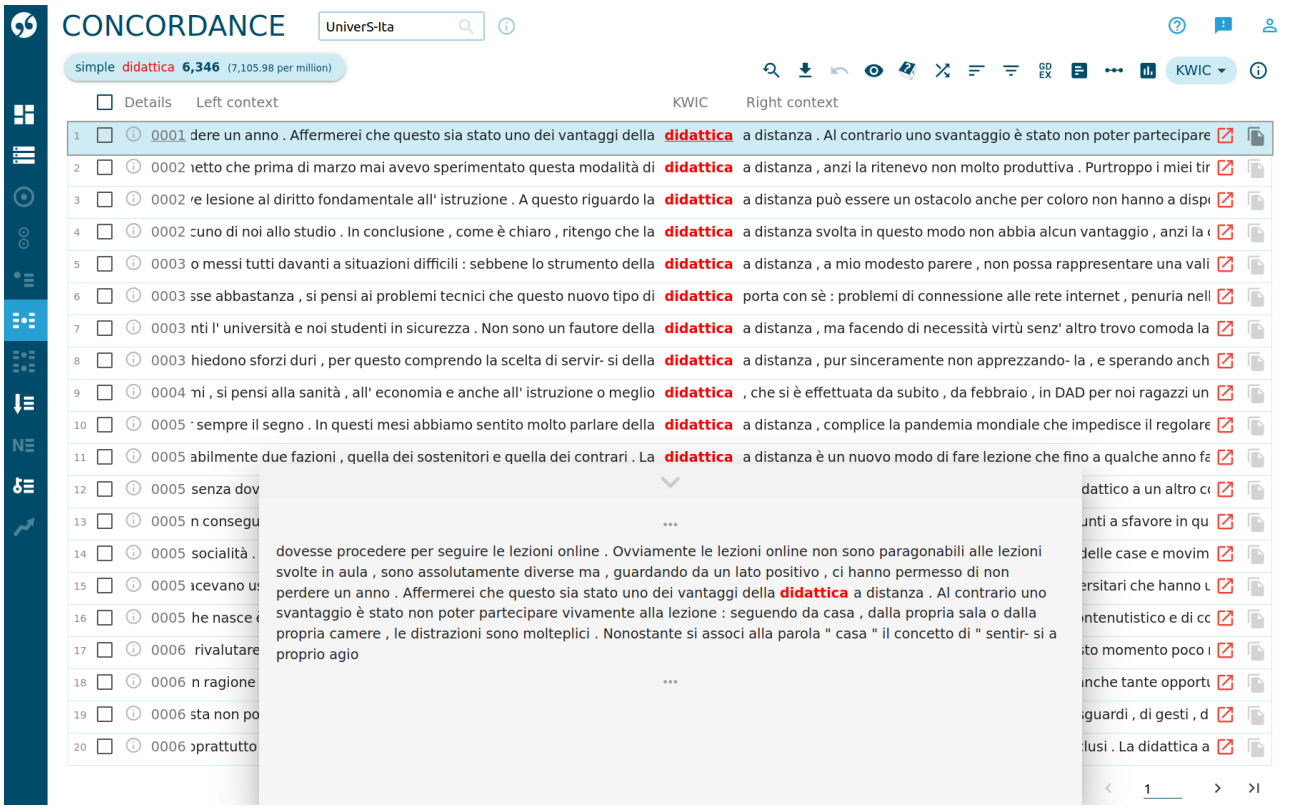


Facendo clic sul pulsante ⓘ all'inizio di ogni riga si accede alle informazioni sul documento relativo; **per il solo corpus Univers-ITA**, facendo clic sulla freccetta alla fine di ogni riga, e poi sul link che compare in basso nella finestra, si apre (in un nuovo pannello del browser) la scheda relativa al documento, contenente i dati relativi all'analisi quantitativa dei testi (es numero di frasi, numero di parole diverse, ecc.), alcune statistiche sulle annotazioni qualitative (per le quali si veda poco oltre, sezione 4.3) e un breve profilo sociobiografico dello/della scrivente ricavato dalle risposte fornite al questionario sociobiografico:

doc N	1	frasi	18	readability(all)	98.45	annotazioni:	
doc ID	5faba2da3f2d7dcd1ecc4e3c	lemmi	234	readability(base)	68.96	coerenza [COE]	2
sede	pisa	parole	296	readability(syntax)	99.22	lessico [LES]	5
corso	giurisprudenza	token	582	readability(lexicon)	52.02	morfosintassi [MFS]	6
		token(parola)	514			marcatezza [MRC]	0
		token(verbo)	123			ortografia [ORT]	1
		token(sostantivo)	100			punteggiatura [PUN]	6
		token(aggettivo)	42			registro [REG]	0
		token(avverbio)	59			sintassi [SIN]	5

1	**A quale Ateneo sei iscritto/a?** dipartimento di giurisprudenza
1/1e1	**presso quale corso di laurea?** laurea magistrale
3	**Hai frequentato altri corsi universitari prima di iscriverti a quello attuale?** no
4	**Sesso** femmina
5	**Quanti anni hai?** 20
6	**Dove sei nato/a?** italia
6/6e1:68ccb505	**Regione?** toscana
6/6e2	**Provincia?** pisa
7.1	**Da dove proviene tua MADRE?** italia
7.1/7e1e1:68ccb505	**Regione?** campania
7.1/7e1e2	**Provincia?** napoli
7.2	**Da dove proviene tuo PADRE?** italia
7.2/7e2e1:68ccb505	**Regione?** campania
7.2/7e2e2	**Provincia?**

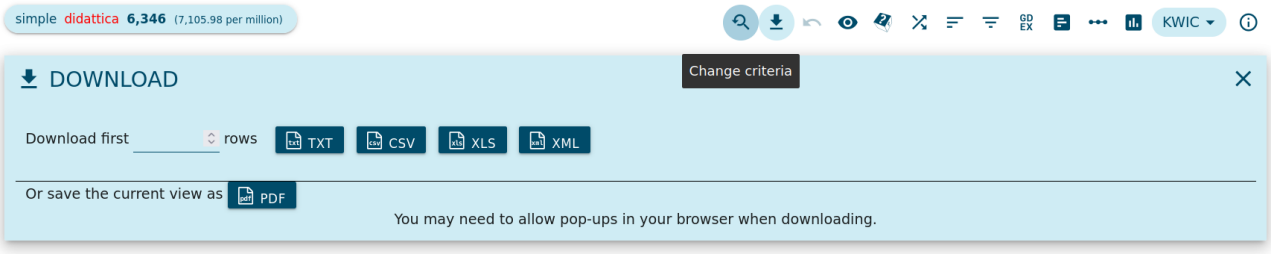
Facendo clic sulla keyword in rosso, si apre un riquadro che mostra un contesto più ampio rispetto a quanto mostrato nella riga della tabella.



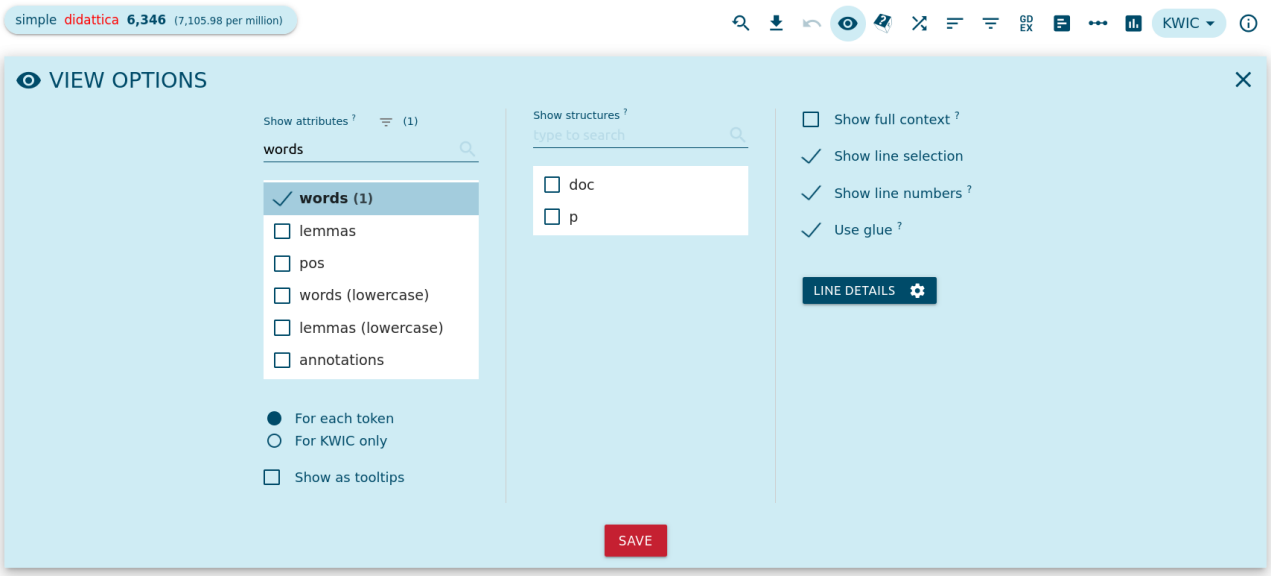
Sopra alla tabella delle concordanze, nella parte destra della finestra, abbiamo una fila di pulsanti. Spostando il mouse sopra le icone, viene mostrato il nome di ciascun comando. Tra questi segnaliamo:

Change criteria: per modificare i criteri della ricerca o eseguirne una nuova.

Download: per scaricare la tabella dei risultati.



View options: è possibile scegliere quali informazioni aggiuntive mostrare sotto ogni token nella tabella. In particolare, è possibile mostrare il lemma, la parte del discorso oppure le annotazioni qualitative.

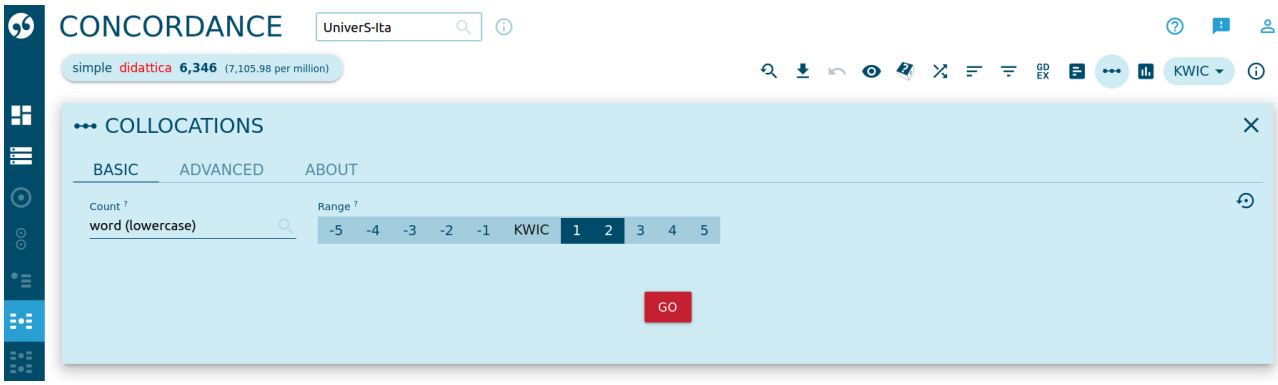


Nella figura seguente sono mostrati i codici delle parti del discorso. Per l'elenco di questi codici, vedere sotto nel paragrafo "Tagset".

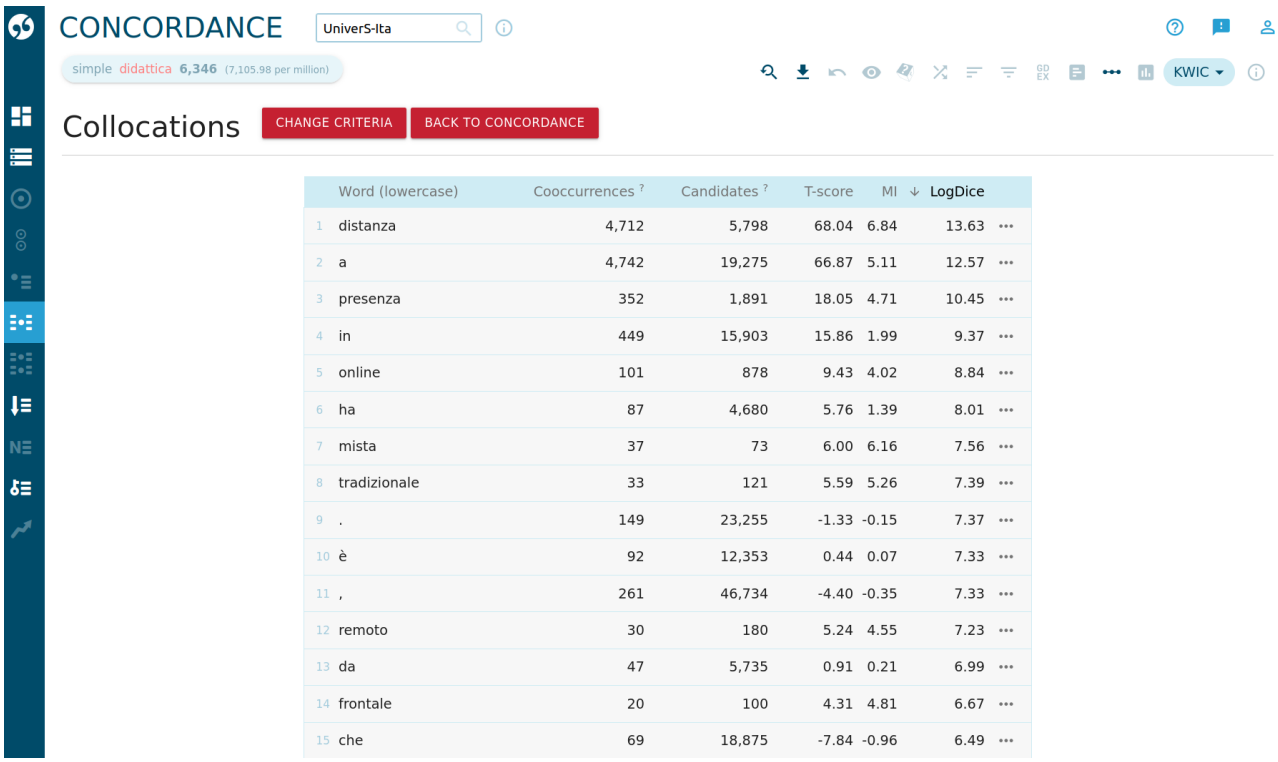
Nella figura seguente sono mostrate le annotazioni qualitative. Per la spiegazione di queste annotazioni, si rinvia al paragrafo 4.3. Annotazioni qualitative più oltre.

Sort: è possibile riordinare le occorrenze in base alla chiave di ricerca o al contesto sinistro o destro; normalmente i risultati vengono mostrati in base al numero di documento.

Collocations: per esaminare le collocazioni della keyword.



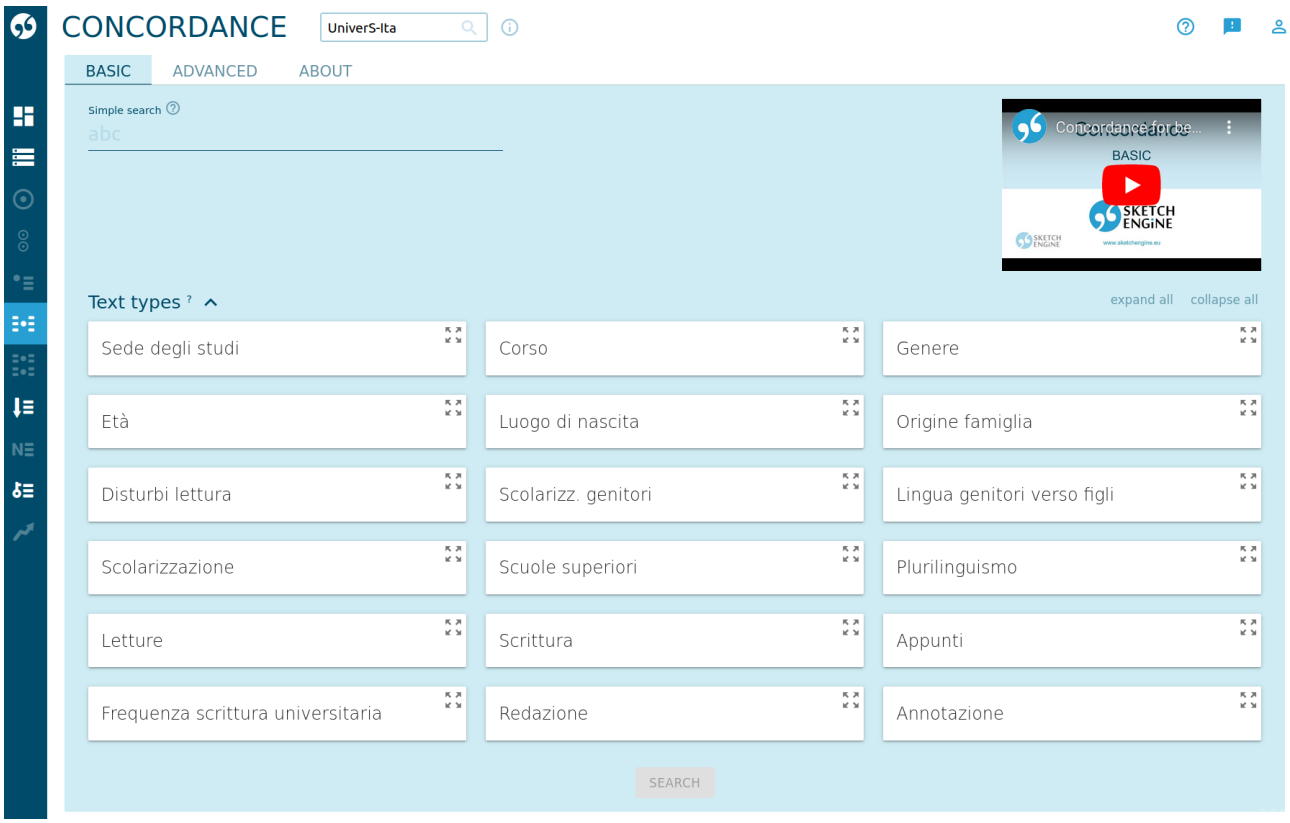
Nella figura precedente sono mostrate le impostazioni per vedere le collocazioni a destra della parola cercata ("didattica"), ignorando la differenza tra minuscole e maiuscole ("word (lowercase)").



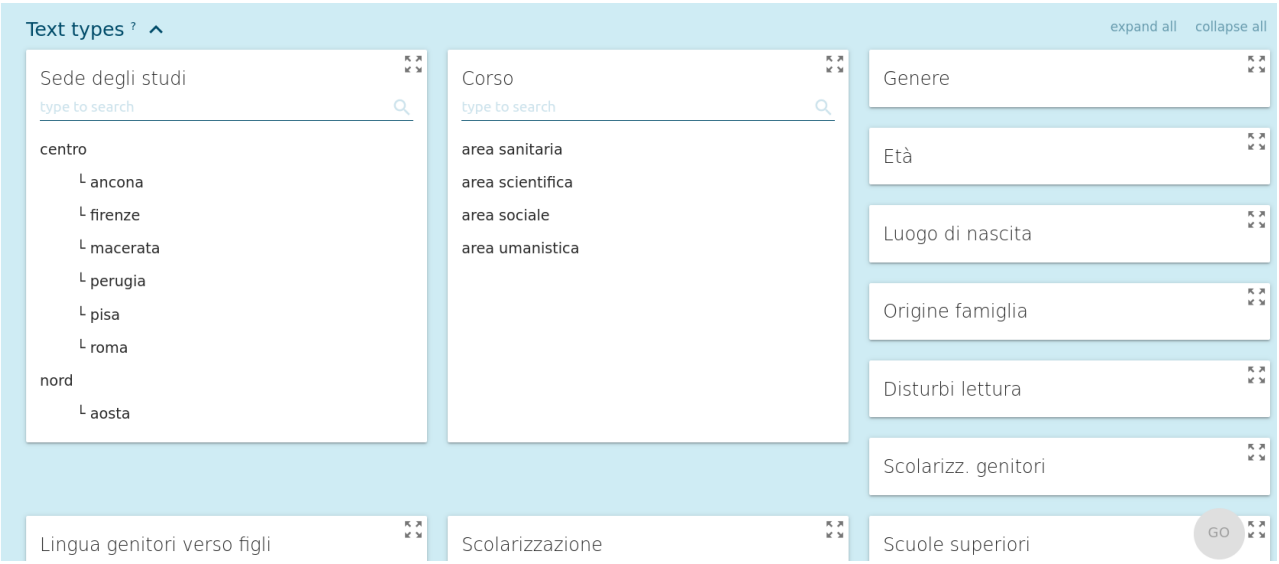
	Word (lowercase)	Cooccurrences	Candidates	T-score	MI	LogDice
1	distanza	4,712	5,798	68.04	6.84	13.63
2	a	4,742	19,275	66.87	5.11	12.57
3	presenza	352	1,891	18.05	4.71	10.45
4	in	449	15,903	15.86	1.99	9.37
5	online	101	878	9.43	4.02	8.84
6	ha	87	4,680	5.76	1.39	8.01
7	mista	37	73	6.00	6.16	7.56
8	tradizionale	33	121	5.59	5.26	7.39
9	.	149	23,255	-1.33	-0.15	7.37
10	è	92	12,353	0.44	0.07	7.33
11	,	261	46,734	-4.40	-0.35	7.33
12	remoto	30	180	5.24	4.55	7.23
13	da	47	5,735	0.91	0.21	6.99
14	frontale	20	100	4.31	4.81	6.67
15	che	69	18,875	-7.84	-0.96	6.49

4.1.2. Filtri di ricerca

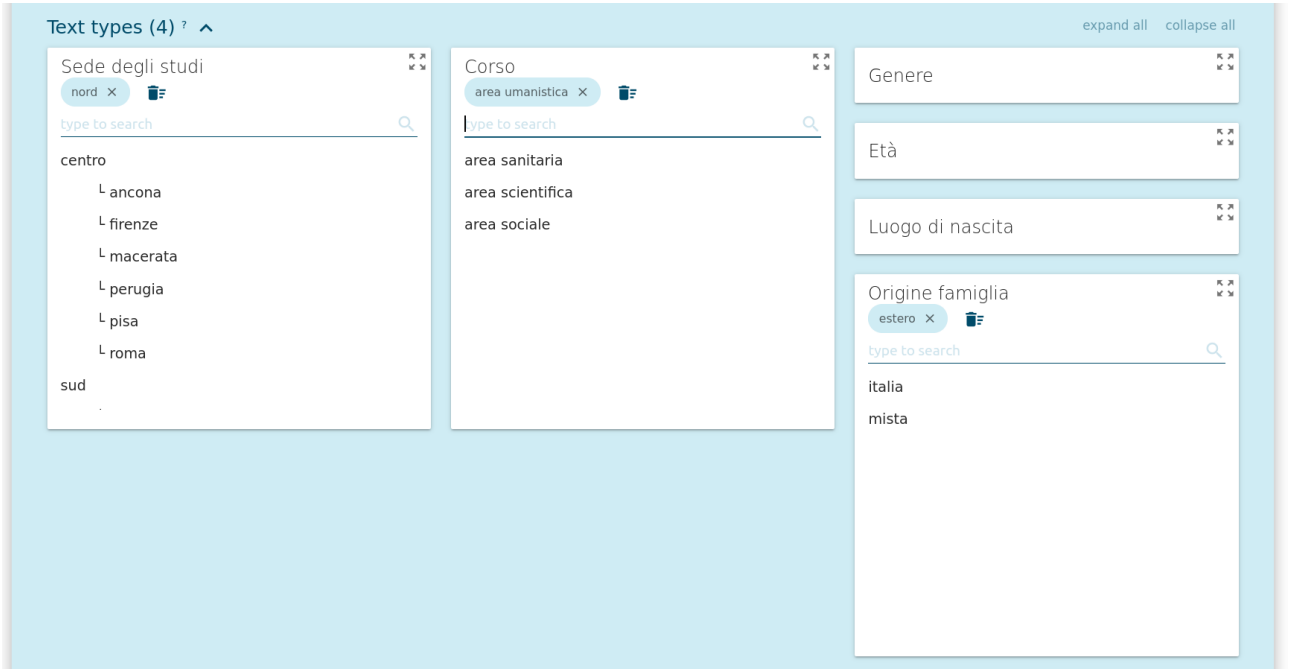
È possibile eseguire ricerche di concordanze su una parte del corpus, selezionata in base ai metadati (ricavati dalle risposte al questionario). Tornando allo strumento Concordance, si può aprire la sezione "Text types" e selezionare i criteri per filtrare una parte del corpus. **Per il solo corpus Univers-ITA**:



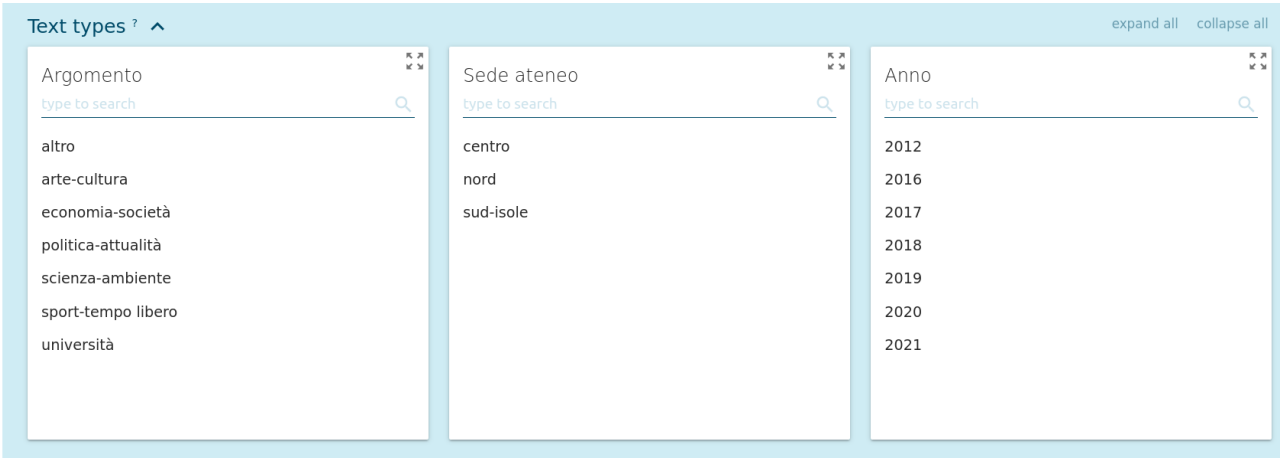
Per attivare i filtri, aprire un riquadro e selezionare la variabile desiderata:



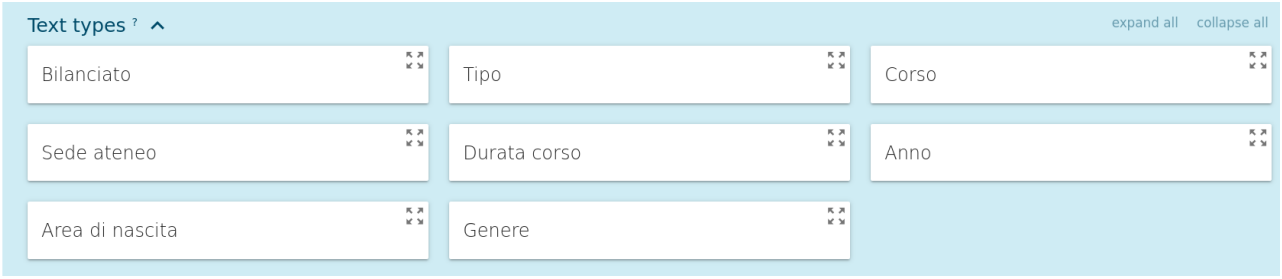
Per esempio, si può fare clic su "centro" per selezionare le produzioni degli studenti di università del centro Italia, oppure su una sede specifica (o più di una). È possibile selezionare più metadati simultaneamente (per esempio: applicare la ricerca alle produzioni di studenti iscritti a corsi di area umanistica nelle università del nord, con famiglia di origine straniera).



Per il corpus Univers-ITA-ProGior, sono disponibili i seguenti metadati:



Per il corpus Univers-ITA-ProUniv sono disponibili i seguenti metadati:

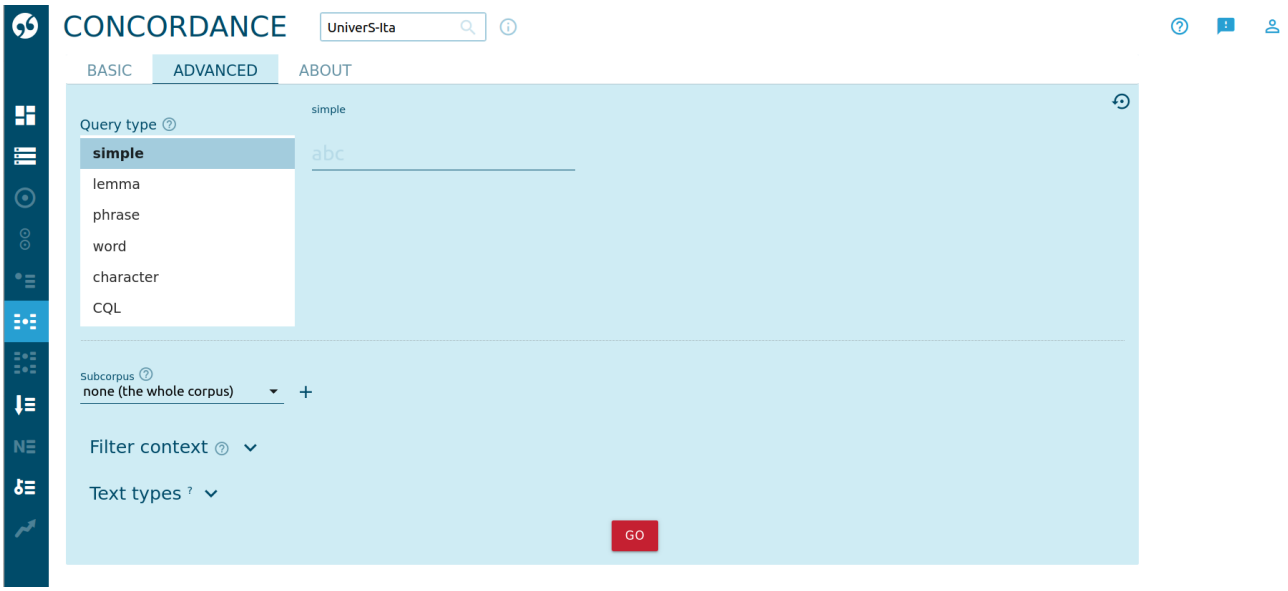


Per quanto riguarda il primo criterio, occorre precisare che il corpus Univers-ITA-ProUniv nel suo complesso non è bilanciato; per compiere studi quantitativi può essere desiderabile eseguire le

ricerche di concordanze specificando "Sì" nel criterio "Bilanciato": in questo modo si seleziona un sottocorpus bilanciato per sede ateneo e area disciplinare (matadato "Corso").

4.1.3. Ricerca avanzata

Attiviamo lo strumento Concordance e selezioniamo la sezione "Advanced".

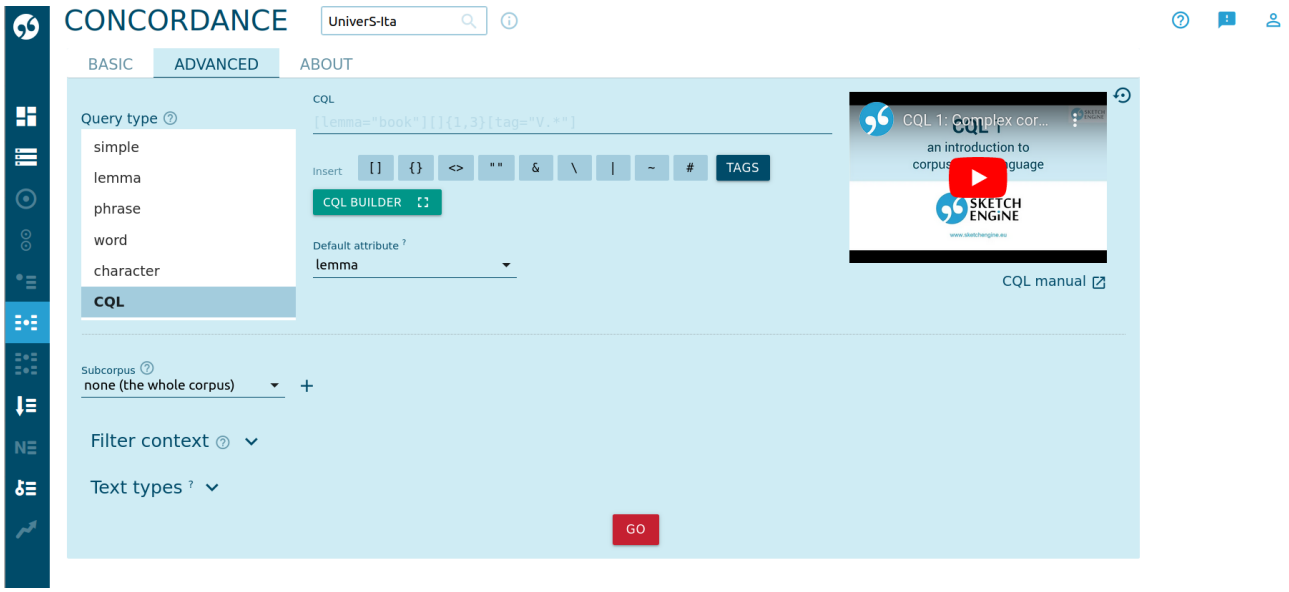


Nel menù in alto a sinistra ("Query type") possiamo scegliere il tipo di ricerca. "simple" corrisponde alla ricerca semplice che abbiamo visto finora.

Con "lemma", "phrase", "word" e "character" possiamo cercare esplicitamente tutte le forme del lemma dato, rispettivamente, o una sequenza esatta di parole, o la forma specifica di una parola, o anche un singolo carattere (per esempio, cercare nel corpus dove compare la lettera “ù”).

4.1.4. Corpus Query Language

Per la versione più potente e flessibile dell’interrogazione del corpus, scegliamo dal menu “Query type” la voce “CQL”. CQL sta per “Corpus Query Language” (Linguaggio di Interrogazione del Corpus), che definisce una sintassi precisa, per formulare le richieste, che dobbiamo rispettare pedissequamente.



In questa pagina NoSketchEngine presenta sulla destra un video tutorial (in inglese) per l'uso del linguaggio CQL e un link al relativo manuale online.

Nel campo CQL, dobbiamo specificare uno o più token, ciascuno con le caratteristiche desiderate. Ogni token si esprime con una coppia di parentesi quadre. All'interno delle parentesi quadre, mettiamo i criteri di selezione del rispettivo token. Se lasciamo vuote le parentesi quadre, significa che vogliamo un token qualunque nella rispettiva posizione.

[]

Un criterio di selezione consta di una variabile e del relativo valore che quella variabile deve avere. Variabile e valore devono essere separate da un segno di uguale e il valore deve essere tra virgolette.

[variabile="valore"]

Si può anche specificare che il valore della variabile deve essere diverso da quanto indicato: in questo caso facciamo precedere il segno di uguale da un punto esclamativo:

[variabile!="valore"]

Se inseriamo più criteri, questi vanno separati da "&" se vogliamo che siano tutti soddisfatti, o da "|" se vogliamo che ne sia soddisfatto almeno uno.

[var1="val1" & var2="val2"]

Le variabili che possiamo utilizzare sono:

word	una parola specifica, con distinzione tra maiuscole e minuscole;
lemma	tutte le forme di un lemma, con distinzione tra maiuscole e minuscole;
pos	il codice della parte del discorso (vedere sotto nel paragrafo "Tagset");
lc	una parola specifica, ignorando la differenza tra maiuscole e minuscole;
lemma_lc	tutte le forme di un lemma, ignorando la differenza tra maiuscole e minuscole;
ann	l'annotazione qualitativa (vedere sotto nel paragrafo "Annotazioni qualitative").

Specificando un valore, possiamo usare il punto come carattere jolly; per esempio, con:

[word="cas."]

il criterio sarà soddisfatto dai token *casa*, *case*, *casi*, *caso*, *cast*, eccetera. L'asterisco indica che un carattere può non esserci o essere ripetuto: per esempio, il criterio

[word="piant*i"]

sarà soddisfatto dai token *piani* e *pianti*, e il criterio

[word="da.*i"]

sarà soddisfatto dai token *dai*, *dati*, *danni*, eccetera.

Per trovare *le occorrenze della parola "il" o della parola "la"*, dovremo scrivere nel campo CQL:

[lc="il" | lc="la"]

Per trovare *le occorrenze della parola "lo" che siano pronomi clitici*:

[lc="lo" & pos="PC"]

si cerca in questo modo un token, che abbia la forma "lo", ignorando la distinzione tra maiuscole e minuscole, e che abbia l'attributo "pos" (parte del discorso) uguale a "PC" (pronome clitico).

Per trovare *le occorrenze della parola "didattica" non seguite da una preposizione (semplice o articolata)*:

[lc="didattica"][pos!="E.*"]

il secondo token non deve avere l'attributo "pos" iniziante per E.

Per trovare *le occorrenze della parola "didattica", seguita da un token qualunque, seguito a sua volta da una voce del verbo essere*:

[lc="didattica"] [] [lemma_lc="essere"]

Infine, si possono usare le parentesi graffe dopo un token per specificare il numero minimo e massimo di ripetizioni di token con le medesime caratteristiche:

[lc="didattica"] []{0,3} [lemma_lc="essere"]

tra "didattica" e la voce del verbo essere possono stare uno, due, tre o nessun token di qualunque tipo;

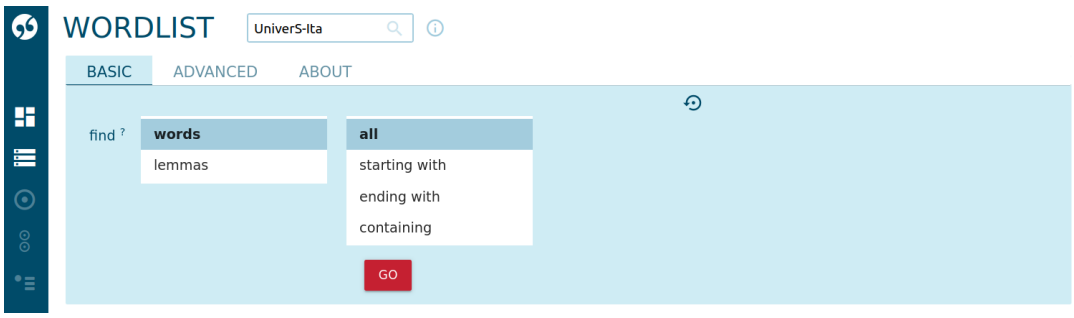
[pos="V.*"]{2,4}

cerca sequenze di due, tre, o quattro voci verbali consecutive, inclusi verbi ausiliari e modali.

4.2. Wordlist

Lo strumento Wordlist fornisce i dati sulla frequenza delle parole presenti nel corpus.

Attiviamo lo strumento Wordlist, facendo clic sul relativo riquadro dopo aver scelto il corpus, oppure dal menu a scomparsa che si trova sulla sinistra della pagina.



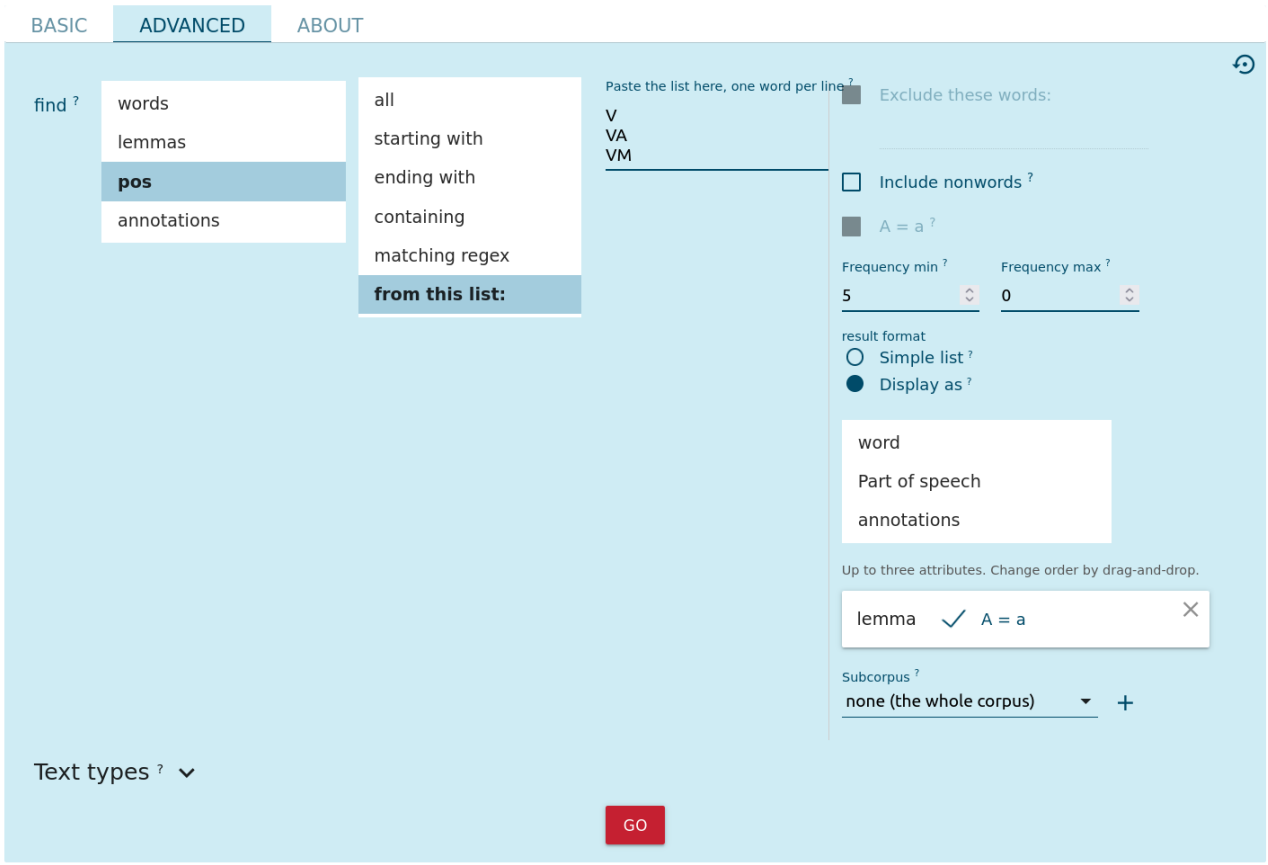
Nella sezione "Basic" possiamo scegliere se avere le statistiche sulle singole forme o sui lemmi. Nel menu di destra è possibile selezionare parole (o lemmi) che iniziano o finiscono in un certo modo oppure che contengono una data sequenza di caratteri.

	Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency
1	di	30,722	14	i	7,316	27	della	4,709	40	lezione	2,589
2	a	19,275	15	con	6,725	28	ha	4,680	41	sia	2,554
3	e	19,235	16	una	6,386	29	ci	3,875	42	essere	2,453
4	che	18,875	17	didattica	6,345	30	questo	3,851	43	lo	2,436
5	la	18,219	18	l'	6,265	31	o	3,741	44	tutti	2,386
6	in	15,903	19	più	6,064	32	ad	3,624	45	tempo	2,328
7	è	12,353	20	distanza	5,798	33	ma	3,484	46	seguire	2,299
8	per	12,036	21	da	5,735	34	come	3,331	47	se	2,258
9	non	11,276	22	studenti	5,368	35	del	3,173	48	questa	2,215
10	si	10,672	23	sono	5,107	36	dei	2,923	49	molto	2,211
11	un	10,458	24	anche	5,098	37	al	2,692	50	nel	2,146
12	il	10,274	25	lezioni	5,075	38	delle	2,688			
13	le	8,625	26	gli	4,791	39	alla	2,658			

Rows per page: 50 1–50 of 6,910 1

Tramite il pusante "View options", in alto, è possibile mostrare le frequenze relative (per milione di token) invece che quelle assolute.

Nella sezione "Advanced" è possibile stabilire dei criteri di ricerca; per esempio, possiamo vedere la lista delle frequenze dei verbi selezionando "pos" nel primo menu, "from this list" nel secondo, inserire i codici POS dei verbi (principali, ausiliari e modali) nel campo, e selezionare "display as: lemma".



Il risultato è nella figura seguente:

Part of speech (1,087 items)

	Lemma (Lowercase)	Frequency		Lemma (Lowercase)	Frequency		Lemma (Lowercase)	Frequency
1	essere	28,885	18	riguardare	915	35	continuare	573
2	avere	11,993	19	svolgere	914	36	frequentare	569
3	potere	8,093	20	perdere	899	37	volere	568
4	fare	3,319	21	rendere	892	38	creare	566
5	dovere	3,158	22	pensare	840	39	affrontare	556
6	seguire	2,753	23	sentire	772	40	utilizzare	554
7	permettere	1,890	24	registrare	768	41	diventare	550
8	stare	1,673	25	studiare	735	42	tornare	549
9	trovare	1,522	26	prendere	696	43	rimanere	504
10	venire	1,414	27	mancare	694	44	presentare	456
11	portare	1,216	28	credere	659	45	considerare	435
12	vivere	1,068	29	ritenere	637	46	ritrovare	433
13	riuscire	1,038	30	costringere	637	47	sapere	416
14	andare	1,029	31	risultare	629	48	cambiare	403
15	vedere	1,022	32	parlare	626	49	tenere	398
16	dire	1,009	33	mettere	604	50	cercare	393
17	dare	975	34	passare	600			

Rows per page: 50 1–50 of 50 1

Anche con lo strumento Wordlist possiamo limitare la ricerca a una parte del corpus, selezionando i criteri nella sezione "Text types" (vedi sopra).

4.3. Annotazioni qualitative del corpus Univers-ITA

Nel solo corpus Univers-ITA sono stati annotati manualmente tutti i tratti che configurassero una qualche forma di "devianza" o allontanamento rispetto a quanto prescritto dalla grammatica normativa o scolastica dell'italiano. Questi tratti si caratterizzano per gradi diversi di "devianza" o allontanamento dalla norma. Si hanno, cioè, tratti fortemente stigmatizzati che, tipicamente, occorrono in testi decisamente caratterizzati verso il basso in diafasia e diastratia (ad esempio realizzazioni ortografiche substandard come *o comprato* senza *h*), ma anche altri che si trovano con una certa frequenza in produzioni mediamente controllate di colti (come, ad esempio, le costruzioni marcate come le dislocazioni o il pronome *gli* generalizzato), benché non di rado condannati dalle grammatiche normative poiché estranei al vecchio standard più rigidamente codificato.

I codici delle annotazioni qualitative sono i seguenti:

COE – coerenza:

- Uso illogico dei connettivi
- Mancata esplicitazione delle relazioni logiche che intercorrono fra i contenuti espressi (giustapposizione)
- Contraddittorietà
- Frammentazione delle informazioni

LES – lessico:

- Povertà/eccessiva genericità lessicale
- Lessico improprio
- Ripetizioni
- Plastismi
- Violazione di collocazioni
- Malapropismi

MFS – morfosintassi:

- Mancato accordo per genere e numero
- Mancato rispetto della consecutio temporum
- Inadeguata gestione del riferimento (ad esempio, pronomi distanti dai loro antecedenti, pronomi che rimandano a referenti dotati di realtà concettuale anziché testuale)
- Reggenze preposizionali errate

MRC – marcatezza:

Frasi dislocate

Frasi scisse o pseudoscisse

Frasi a tema sospeso

ORT – ortografia:

Assenza/impiego scorretto dell'apostrofo

Uso dell'accento con forme verbali monosillabiche (*fa, sa, so*) e con la forma apocopata dell'avverbio *poco*

Non sono annotati evidenti errori di battitura, uso dell'accento grave al posto di quello acuto e viceversa, uso improprio delle maiuscole o minuscole.

PUN – punteggiatura:

Omissione dei segni interpuntivi

Sostituzione di un segno interpuntivo con un altro

Inserimento di segni interpuntivi in contesti incongrui

REG – registro:

Lessico non adeguato al contesto scritto sorvegliato mediamente formale

Uso di "gli" sovraesteso per "loro" e "le"

Uso del "tu" impersonale

SIN – sintassi e coesione:

Omissione della preposizione nella coordinazione di sintagmi

Mancanti o scorretti parallelismi

Gerundi assoluti

Omissioni argomentali, ad esempio: "ricominciare a recarsi in presenza"

Interruzione della continuità sintagmatica, ad esempio: "Basterebbe pensare alle famiglie che vivono, magari anche numerose, in un monolocale"; "Sono, infine, felice delle scelte fatte dai miei professori"

Poiché queste annotazioni possono riguardare una sequenza di token, il primo e l'ultimo token della sequenza avranno un codice aggiuntivo: per esempio, per l'annotazione "SIN" (errore di sintassi), avremo sul primo token le annotazioni "SIN" e "#SIN", sull'ultimo token le annotazioni "SIN" e "SIN#", e su tutti gli altri token della sequenza soltanto "SIN". Questo sarà utile per la ricerca avanzata delle concordanze con il linguaggio CQL, spiegata sopra; in molti casi può essere utile cercare solo i token iniziali delle annotazioni:

[ann="#SIN"]

4.4. Tagset

I tag (etichette) dei token relativi alla parte del discorso seguono lo standard ISST-TANL. L'elenco dei tag è sempre disponibile cliccando sul pulsante ⓘ in alto, seguendo il link "tagset"; è disponibile anche all'indirizzo https://corpora.ficlit.unibo.it/CUSP/tagset.pdf

A	aggettivo
AP	aggettivo possessivo
B	avverbio
BN	avverbio di negazione
CC	congiunzione coordinativa
CS	congiunzione subordinativa
DE	determinativo esclamativo
DI	determinativo indefinito
DQ	determinativo interrogativo
DR	determinativo relativo
DD	determinativo dimostrativo
E	preposizione
EA	preposizione articolata
FB	punteggiatura bilanciata (parentesi, virgolette, ecc.)
FC	punteggiatura di fine frase
FF	virgola e trattino
FS	punteggiatura di fine periodo
I	interiezione
N	numero cardinale
NO	numero ordinale
PD	pronome dimostrativo
PE	pronome personale
PI	pronome indefinito
PP	pronome possessivo
PQ	pronome interrogativo
PR	pronome relativo
PC	pronome clitico
RD	articolo determinativo
RI	articolo indeterminativo
S	nome comune
SA	abbreviazione
SP	nome proprio
T	predeterminativo (*tutto*, *entrambi*, ecc.)
VA	verbo ausiliario
VM	verbo modale

V	verbo principale
X	altro