

## LancsBox and Lancaster Stat Tool online: a survival guide

*A cura di Martina Lucia Bentivegna, Cristina Maya Rao, Stefano Angelo Rizzo.*

La prima cosa che serve per analizzare un corpus è assemblare un corpus.  
Sorpriendente, lo so.

Vi sono molteplici modi per farlo: potete scaricare e convertire una grande varietà di testi (in pdf, probabilmente) in formato testo (.txt), potete fare affidamento su un sito che raccolga per voi i testi delle pagine web più attinenti al vostro discorso (ad esempio, incrociando tre o più parole chiave su [SketchEngine](#)). Fatto sta che a un certo punto avrete una cartella con molti file .txt sul pc, pronto per essere elaborato da LancsBox. È estremamente importante che i molteplici file di testo si trovino dentro una cartella che caricheremo come corpus; usare un unico file di testo unificato o fargli caricare una grande quantità di file di testo presi singolarmente non funziona. O meglio, potete farlo, ma l'analisi avrà risultati sballati già alla tokenizzazione (parleremo a breve dei token).


Se volete che il vostro corpus sia organizzato in subcorpora (ad esempio, volete analizzare sia una collana di libri che i singoli volumi), vi conviene creare una cartella con dentro sottocartelle, ognuna con dentro il testo/i testi del subcorpora.  
È anche ravvisabile, meglio dirlo da subito, che la grandezza del corpus sia sul milione di parole. The more the merrier. L'unico caso in cui questa regola viene meno è nel caso in cui il campione di testi coincida con la popolazione totale dei testi (fatta breve: sono 10.000 parole, ma sono tutti i testi del genere. Esaminate cartoline, per caso?).

Il concetto di "parola" è molto relativo ed eccessivamente generico nell'analisi di corpora. Si distinguono in modo gerarchico i token, i type, ed i lemmas. Ci sarebbero anche i lexeme, ma sono unità così specifiche che ancora non risultano suscettibili all'analisi automatica, quindi fregiamocene allegramente.

I token contano ogni singola istanza della parola. Se nel testo c'è scritto 25 volte "parallelepipedo", avremo 25 token "parallelepipedo".

I type consistono di ogni parola presente almeno una volta nel testo, indipendentemente da quante volte compaia. Se nel testo c'è scritto 25 volte "lavandino", avremo un type "lavandino".

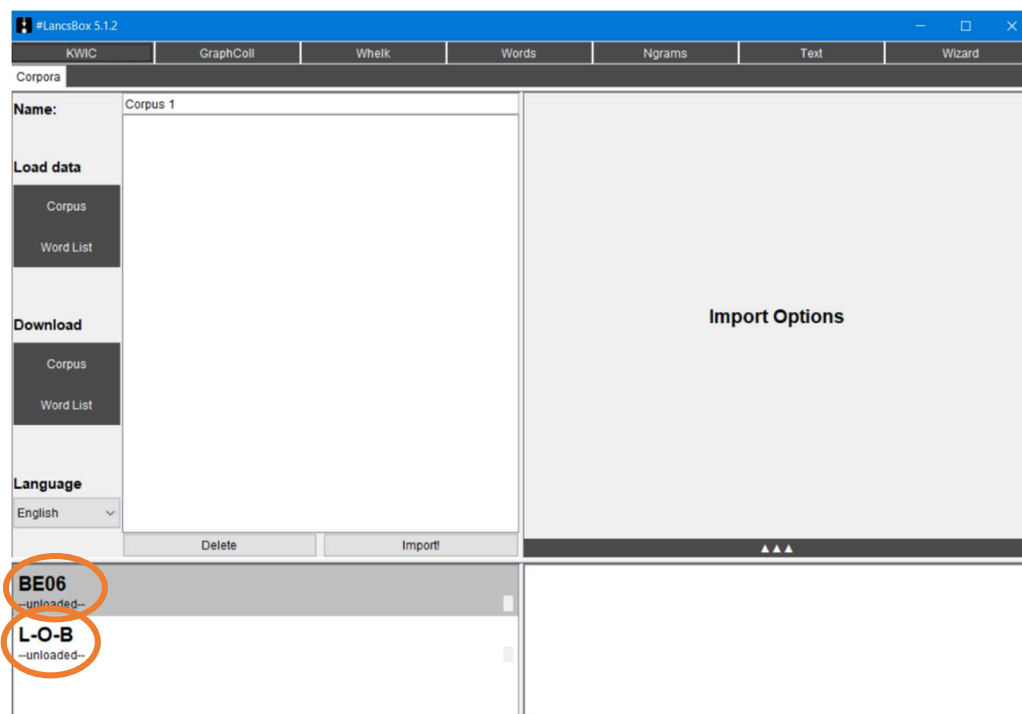
I lemmas sono le unità più primordiali e non declinate del testo, come verbi considerati all'infinito o i pronomi personali al di là del fatto che siano soggetto, oggetto etc. Nella frase "lui vede ciò che tu hai visto", si conterebbe come lemma una sola volta la coppia "vede-visto" come verbo all'infinito "vedere". Tutti i lemmas di questa frase sono dunque lui, vedere, ciò, che, tu.

I lemmas sono inoltre la chiave per l'interrogazione dei POS (part of speech) di un testo. Il POS di una parola indica la funzione grammaticale di quel sintagma: se è aggettivo, nome, verbo, articolo... Si potranno analizzare i POS selezionandoli nei vari tool (di solito nel riquadro  **Type** ).

Mo vi chiederete: a che mi serve sapere ciò? E la risposta è semplice. Di ogni corpus e sottocorpus i token sono l'unità minima per l'analisi del testo: è quante parole compongono il testo stesso, quindi molto probabilmente sarà il denominatore di ogni vostra assunzione sul testo. Di ogni corpus (e sub) i lemmas indicano il vocabolario usato dall'autore nella stesura dei testi, mentre i types indicano la varietà di quest'ultimo. Su questo, torneremo più avanti, ma capite bene che se un autore non fa altro che ripetere le stesse parole, bisogna vedere quanto sono vicini tra di loro i valori di lemmas fratto i token e di types fratto i token. Ergo, fatevi furbi: segnatevi e tenete sempre sottomano il numero di token, types e lemmas riconosciuti automaticamente da LancsBox ogni volta che carichi. Notate bene che il software potrebbe riconoscere cose diverse da quello che, ad esempio, riconosce l'analisi di SketchEngine durante la costruzione di un corpus.

LancsBox vi accoglierà con una schermata grigia, fredda e gelida. Non è per cattiveria, ma semplicemente non è compito del programma presentare i dati in modo user-friendly o in modo graficamente accattivante: per quello ci sono programmi come [Gephi](#). Il programma è un calcolatore statistico, questa guida vi aiuterà nel fargli masticare i dati corretti e ad interpretare cosa il programma ci darà come risultato.

[Scarichiamo](#) e avviamo dunque LancsBox.

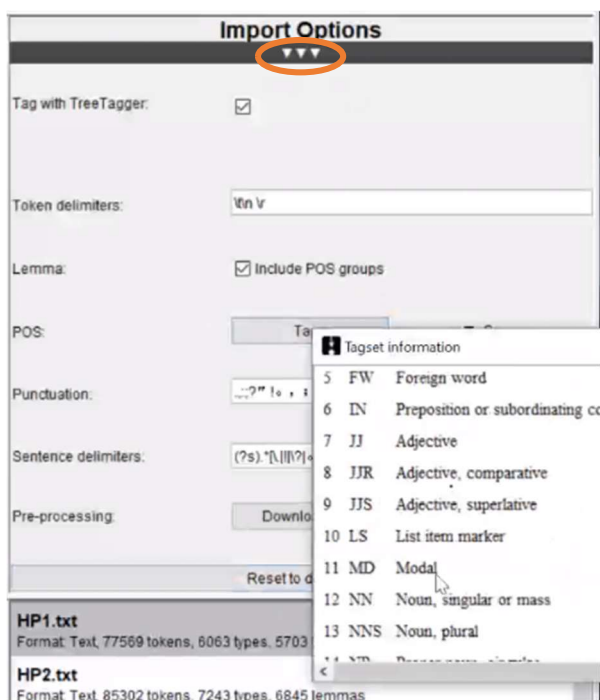


Bello eh? Se avete pensato un "sì", avete gusti strani.

Nella parte inferiore a sinistra troviamo i corpus che sono visti dal programma. Nel nostro screen abbiamo il **BE06** ed il **L-O-B**. Si trovano online, ad esempio su SketchEngine se non nei loro siti dedicati. Noterete che sotto hanno la scritta "-unloaded-": sta a significare che il programma sa che quelli sono corpus, ma non li sta ancora prendendo in considerazione per le analisi. Come fare per farli prendere in esame? Basta premere il pulsantino rettangolare nella sezione opposta al titolo del corpus.

Ultima nota, avendo citato nuovamente SketchEngine: se questo sito viene utilizzato per creare un corpus, allora conviene usare gli strumenti direttamente disponibili all'interno della piattaforma, almeno per le analisi più semplici. Nel passaggio dei dati dalla piattaforma a LancsBox, quest'ultimo potrebbe interpretare diversamente i dati, sfalsando (seppur di poco) alcune statistiche, conviene quindi sempre annotare il numero di token e type che vengono contati la fonte principale e fare dei confronti o delle proporzioni.

Come caricare il proprio corpus? Alla sezione "load data", clicca su "Corpus" se hai un'intera cartella di files di testo, oppure "Word List" per un singolo file di testo.



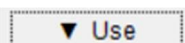
Appariranno nella parte inferiore a sinistra, e andranno poi "caricati" su

In alto a destra, trovi le "Import Options", puoi cliccare sui tre triangolini per vederle. Se clicchi su "tags" [ riquadro situato accanto alla voce POS ], vedrai come le part of speech verranno taggate, essenzialmente, con quale nome (esempi

Noun, JJ = Adjective).

Controllate che tra i separatori di parole figurino tutta la punteggiatura utile alla lingua del corpus. Ad esempio, di default la virgoletta (') non comparirà ma è necessario inserirla se i vostri testi sono in italiano. Non fare ciò ha come conseguenze che, ad

esempio "dell'arte" sia un token unico e non l'unione di "dell" e "arte", con tutto ciò che ne consegue a livello di conteggio. Troviamo anche una sezione di pre-processing con "downloads" degli utili strumenti opzionali utili a ripulire il corpus quando viene scaricato da internet o nel caso di dati potenzialmente sporchi. Scaricati, vanno usati con



e possono anche essere editati se ne avete le competenze in C#.

Andiamo di sopra. Partendo dalla sezione Words possiamo vedere tutte le parole dell'intero corpus. Se lo vogliamo confrontare, dobbiamo cliccare sui tre triangolini in basso, e ci permetterà di caricare un nuovo corpus. La cosa si noterà nella barra a destra, dove appariranno due pallini, indicanti il corpus 1 e il corpus 2. Se li trascini uno dentro l'altro, sulla sinistra apparirà il confronto tra i due, mentre sulla destra appariranno le tabelle con lockwords e keywords. Se clicchi su uno dei due pallini con il destro, apparirà una tabella con delle statistiche: complexity e lexical. In quest'ultimo possiamo trovare il numero dei files, i type, i tokens, la type-token ratio, la standardized type-token ratio (in teoria su 10.000, secondo il programma), e infine la moving average type-token ratio (con la media semi-mobile, cioè sono segmenti che si sovrappongono tra di loro).

**Dispersion** = valore della dispersione (o presenza) di un type nel testo: maggiore è il valore, minore è la presenza.

Se per caso vi spunta sempre 0, allora c'è un problema nella costruzione del corpus.

La dispersione deriva dal calcolo della distribuzione delle occorrenze, cosa incalcolabile se il corpus è assemblato come testo unico, ossia come un unico file (word, notepad...). Non fraintendete: in alcune circostanze unire tutto in unico file risulta comodo e snello, ma non per studi accurati sulla dispersione.

Può essere una soluzione dividere meglio i subcorpora o assicurarvi che più testi, appunto, non siano stati caricati come un unico file .txt.

Se si continua ad avere il problema, forse la colpa è di notepad: provate a usare notepad++, ottenibile gratuitamente al link [Downloads | Notepad++ \(notepad-plus-plus.org\)](http://notepad-plus-plus.org).

Ci sono due tipi di dati “diversi” nell’analisi testuale, che vanno trattati in modo diverso rispetto al normale: le stopwords e gli hapax. Le stopwords sono quelle parole in grande quantità, ma fondamentalmente inutili all’analisi testuale. Tipo gli articoli o le congiunzioni. Gli hapax sono quelle parole che vengono ripetute una sola volta nel corpus. Possono essere estremamente indicative tanto quanto inutili. Fatta breve, fanno impazzire i valori numerici delle analisi se non le trattiamo con un occhio di riguardo. Dopotutto, un autore può scegliere di usare quella parola per un motivo specifico, oppure per riempire un periodo con una parola che non dice nulla su di lui. Parleremo meglio di questi casi particolari mentre affronteremo lo strumento “Words”.

Prima di iniziare con i vari tool, bisogna sapere come cercare le varie parole. Ovviamente possiamo cercare la parola per esteso, ma questo non ci dà le sue variazioni, come il plurale o un avverbio derivato da quel lemma. Per questo Lancsbox ci viene incontro con espressioni come l’asterisco: cercare una radice di una parola seguita da un asterisco (ad esempio: parol\*) sarà sempre più fruttuoso. Se per qualche motivo vi interessano le desinenze, l’asterisco andrà messo all’inizio invece che alla fine. O se vi interessa che una parola contenga una certa sequenza di lettere... Beh avete capito, all’inizio e/o alla fine si possono mettere gli asterischi per delimitare o ampliare il campo di ricerca. Occasionalmente potremo cercare il POS della parola espandendo il menu contestuale, come avviene con lo strumento Whelk. Ricordate sempre, per i POS, che il programma è in inglese e che mastica l’inglese, non altre lingue.

**Strumento KWIC (KeyWord-In-Context):** produce la lista delle concordanze di una o più parole. Una volta cercata la parola d’interesse, ci dà la lista delle occorrenze in totale e in frequenza normalizzata. Inoltre, alla sezione “text”, mostra in quanti testi la parola appaia, se il corpus è formato da più subcorpora.

#LancsBox 5.1.2

KWIC GraphColl Whelk Words Ngrams Text Wizard

Corpora KWIC: pandemic X

Search

Search pandemic Occurrences 345 (1.05) Texts 5 Corpus Corpus 2 Context 7 Display Text

Index	File	Left	Node	Right
13	2016.bt	mainly by restrictions related to the covid-19	pandemic,	were a warning. The cost of Mr
14	2016.bt	new measures were introduced, though the coronavirus	pandemic	and a holiday weekend meant cross-Channel traffic
15	2016.bt	unknown that is taking place during a	pandemic	that has upended life around the world.
16	2016.bt	had was soon shattered by the Coronavirus	pandemic,	with the first UK cases reported on
17	2016.bt	all but ground to a halt. The	pandemic	has eaten away at the already emaciated
18	2016.bt	the last 12 months since the COVID-19	pandemic,	as millions have relied on them more
19	2016.bt	a time when concerns over a raging	pandemic	and new coronavirus variants have triggered fears
20	2016.bt	register huge growth in demand post COVID-19	pandemic	situation attributable to the demand for the
21	2016.bt	it had not been for the covid-19	pandemic,	I think we would be discussing little
22	2016.bt	Containers are cheap when there isn't a	pandemic,	right? Not all exports/imports can be carried
23	2016.bt	lower than 2019, due to the coronavirus	pandemic,	It meant factories were temporarily shut and
24	2016.bt	do not bear the brunt of the	pandemic	needs to be a top priority for
25	2016.bt	with another sovereign nation" should. "In a	pandemic,	in a health crisis, we want to
26	2016.bt	not a two-week break to solve the	pandemic,	it is likely that we will see

Lo strumento ci aiuta dunque a capire quante volte una parola viene scritta, in quali testi e in quali sezioni del testo (attraverso il contesto di parole a destra e a sinistra). Come preannunciato, possiamo ricercare una singola parola oppure una parola con uno o più asterischi (esempio: cerchiamo art\* = avremo termini come artistic, artist, arts, ma anche Arthur). In generale, il programma accetta tutte le espressioni regolari (regex), trovate nella guida ufficiale un recap generale a questo link: [LancsBox 4.5 search.pdf](#).

**Strumento GraphColl:** fa vedere i collocates di una certa parola, rappresentata al centro del grafico sulla sinistra. La rappresentazione grafica è significativa:

- la lunghezza delle linee esprime la forza dell'associazione (più sarà vicina, più forte sarà);
- il colore più o meno intenso dei pallini esprime la frequenza (più intenso significa più utilizzato)
- se in alto a sinistra, al posto di "Free" scegliamo la forma del grafico "Positional", ci verrà mostrato da che lato i collocates saranno posti rispetto alla parola ricercata, sinistra o destra. La posizione viene data anche dallo strumento "Span" (in alto a sinistra), dove si potrà scegliere quante parole prima e quante dopo, rispetto al nodo centrale, tenere in considerazione.

Se si vogliono confrontare due parole, basta cancellare la prima cercata e scriverne una seconda. Occhio, però: il programma potrebbe crashare. *Siete avvisati.*





e nodo e viceversa. Se il risultato sarà basso, vorrà dire che il collocato sarà attratto dal nodo, ma il nodo non sarà attratto dal collocato;

- **C1:** il risultato sarà la frequenza dei collocati in tutto il corpus;
- **O11:** il risultato sarà la frequenza del nodo sommato al suo collocato, nella finestra di collocazione;
- **T-Score:** evidenzia la frequenza della collocation;
- **Z-Score** evidenzia di quante deviazioni standard ogni valore di un campione si discosta dalla media;
- **Cohen's D:** una variabile tra 0 e 1 che si usa per indicare la differenza standardizzata tra due medie, le compara. Se non ci sono differenze, tenderanno a 0, mentre se ve ne sono molte, tenderanno a 1.

La Cohen's D è interessante perché rispetto ad altre misure, essa tiene in considerazione l'effect size (=che diverse grandezze di diversi corpora possano influire sui calcoli statistici). Per farla breve: se la grandezza conta, Cohen ne tiene conto.

- **Dispersion:** ci dice la distribuzione delle occorrenze all'interno del nostro corpus.
- **Juilland's D:** una comparazione di come sono distribuiti i fenomeni in relazione anche ad altri fenomeni linguistici nello stesso corpus. Il risultato sarà sempre un numero tra 0 (distribuzione estremamente irregolare) e 1 (distribuzione perfettamente uniforme).
- **Frequency:** dà quante volte una parola viene associata ad un collocato. Ne dà la position, il collocato (che è il collocato più utilizzato accanto alla parola ricercata), stat (è il valore della selected value), la frequenza del collocato associato alla parola che abbiamo cercato e la frequenza in generale della parola ricercata nel corpus.

Strumento "Threshold": permette di modificare manualmente la statistic value e la collocation frequency. Dipende da cosa cerchiamo: alcuni grafici potrebbero apparire "pieni" e dunque, al fine di sfoltirli, potremmo voler inserire un valore nel settore della collocation frequency che sia più alto (ad esempio 0.05) così da far apparire solo collocates più vicini.

**Si possono espandere i vari grafici fino a creare una collocations network (ma mi dispiace per il vostro povero computer):** è possibile cercare più termini nella barra di ricerca, facendo così generare per ogni parola il proprio grafico di collocati e poi al centro i collocati in comune ad ambo (o più) termini. Appaiono, sulla sinistra, anche gli shared collocates: è il numero di collocati in comune tra le parole cercate.

Ecco un esempio: da cosa capiamo che Hermione e Ron sono i sidekick di Harry nei primi tre libri della saga della Rowling?

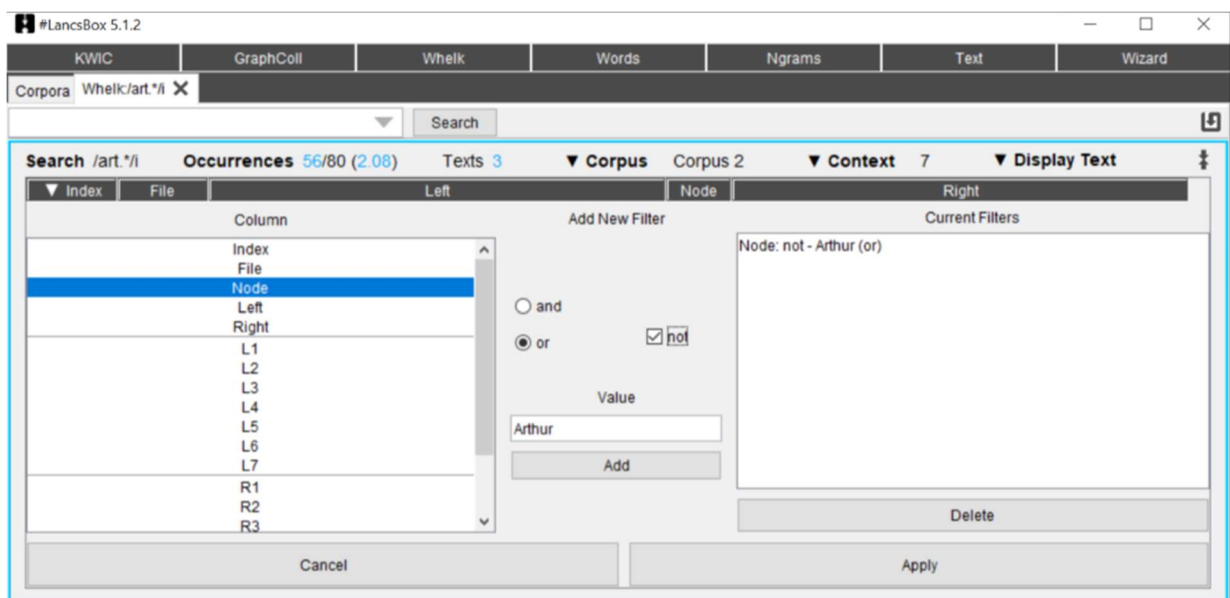




Una piccola legenda alla comprensione del grafico:

- la linea grande grigia è il valore medio;
- la lineetta sopra e quella sotto rappresentano i valori più alti e i valori più bassi del range preso in considerazione, quindi il contesto;
- la lineetta rossa rappresenta valori individuali (frequenza relativa delle variabili linguistiche nei testi presi in considerazione);
- i pallini presenti sulle linee dovrebbero essere i valori dove si trovano gli estremi

Per filtrare una ricerca, bisogna andare alla sezione "node". Lì troveremo un menu dove inserire a sinistra dove si colloca la nostra eccezione (a destra, a sinistra, nel nodo stesso...) e spuntare la casella "not", aggiungendo l'eccezione. Occhio a non confondere il tasto di conferma, apply, con quello delete! Si possono integrare diverse tipologie di eccezione.



**Strumento Words.** Abbiamo due tipi di frequenza: **assoluta**, (è il conteggio di tutti i token, nel corpus o nel testo, che appartengono ad un particolare type - ci permette di ordinare un elenco dalla parola più usata a quella meno frequente) e **relativa** (o normalized, si usa per confrontare più corpora - può essere considerata come la mean frequency, ovvero la media delle frequenze della parola in samples (campioni) ipotetici di x (base per la normalization) token del corpus). Usiamo la frequenza relativa per descrivere il numero medio di occorrenze di una variabile linguistica all'interno di tutto il corpus.

**Legge di Zipf (spiegata troppo buona assai):** in base alla frequenza assoluta della parola più frequente in un corpus, la frequenza assoluta della parola più frequente è il doppio della seconda. E la seconda, il doppio della terza. E così via, descrivendo una curva. È da notare che con enorme probabilità, le parole più ripetute - prima che la curva si normalizzi -

saranno stopwords assolutamente inutili all'analisi testuale. Ne abbiamo parlato prima. È bene dunque escluderle, così di trattare elementi interessanti. Alla fine della curva avremo gli hapax, le parole che compaiono solo una volta. Anche queste possono dare noie, ma non sono del tutto inutili come congiunzioni o articoli.

Come faccio a levare le stopwords dalla mia visualizzazione?

Le stoplist vanno create post analisi di frequenza e vanno inserite manualmente. Cliccando con il tasto destro su "Type" si apre la finestra che ci interessa.

Le stoplist si azzerano a ogni visualizzazione; quindi, conviene inserire più parole separate da un pipe | (con maiusc, lo stesso tasto dello slash \ sulla tastiera). Ancora meglio, fate un documento notepad pronto da copiare e incollare o da modificare. Tenete a mente che il programma ha un limite delle stopwords che filtra: non pensate di poter togliere mezzo corpus e saltare dritto a cosa vi interessa. Provate a integrare script di python per la pulizia testuale se non riuscite a fare pace con i limiti del codice di LancsBox.

### Le misure della dispersione:

- **Range:** descrive la dispersione delle frequenze, dalla più piccola alla più grande;
- **Range2:** dividendo il corpus in sezioni, ci permette di osservare la dispersione di un'occorrenza, cioè in quali è presente e in quali no. Dunque, Range2 è uguale al numero delle sezioni in cui è presente l'occorrenza;
- **Range Percentual:** esattamente come il calcolo del Range2, solo poi moltiplicato per 100 = è la percentuale;
- **Standard deviation:** ci dice la distanza tra le distribuzioni delle occorrenze e la media delle frequenze relative delle occorrenze (è il grafico con la linea e i pallini). Più piccola è la SD, più saranno distribuite le occorrenze rispetto alla distribuzione normalizzata della media;
- **Coefficient of variation:** descrive la quantità di variazione relativa alla frequenza media relativa. Più variazione c'è tra le frequences, più è irregolare la dispersion. Più il coefficiente di variazione è vicino allo 0, più è uniforme la distribuzione della parola (il perfetto contrario del Juilland's D). Il massimo valore di CV dipende dal numero di sezioni in cui è stato diviso il corpus, e per ottenere un valore percentuale corretto si prende in considerazione il CV massimo;
- **Juilland's D:** una comparazione di come sono distribuiti i fenomeni in relazione anche ad altri fenomeni linguistici nello stesso corpus. Il risultato sarà sempre un numero tra 0 (distribuzione estremamente irregolare) e 1 (distribuzione perfettamente uniforme);
- **Deviation of proposition:** confronta la distribuzione attesa di un'occorrenza nelle diverse sezioni del corpus con la sua effettiva distribuzione. La proporzione attesa è il numero dei tokens di ogni sezione fratta i tokens totali di tutto il corpus. La proporzione osservata è la frequenza assoluta di ogni sezione fratta la frequenza assoluta di tutto il corpus. A differenza della Juilland's D, il significato di 0 è relativo ad una buona distribuzione, mentre 1 è relativo ad una distribuzione estremamente irregolare;
- **Average Reduced Frequency:** coefficiente che valuta l'importanza delle occorrenze in un corpus tenendo conto della frequenza e della dispersione: la parola con una frequenza ed una distribuzione più uniformi sarà considerata la più importante. Per

calcolarla, ci servono: frequenza assoluta della parola, il numero totale dei tokens del corpus, e la posizione della parola nel corpus. Confronta il totale delle occorrenze rispetto a come si trovano fisicamente nel corpus con l'ipotetica distribuzione nel corpus (proposizione attesa o osservata): da questo riduce la frequenza normalizzata in base al fatto che questa risulti più vicina o più lontana rispetto all'ipotetica uniforme distribuzione. Se la frequenza assoluta dà un numero, il numero della ARF dovrà essere più vicino possibile alla frequenza assoluta. In soldoni: se la parola è in tutto il testo, è importante più di quanto sia la sua presenza in una porzione del testo.

**Come trovare keywords, lockwords (parole con simile occorrenza in due corpus presi in considerazione) e stopwords?** Basta trascinare i due "pallini" dei due corpus l'uno dentro l'altro, e troveremo le tre tabelle. Nota bene: se hai un corpus solo, basta che nei triangolini di sotto carichi lo stesso corpus che stai analizzando: spunterà una sola tabella, con le lockwords presenti in tutto il corpus.

**Strumento N-Grams:** sono agglomerati ricorrenti di parole. Se clicchiamo, permette un'analisi approfondita di bigrams e trigrams, che possono essere definiti come combinazioni contigue (situate nelle vicinanze) di type, lemmas e POS. Sono, in soldoni, le parole che hanno più probabilità di stare insieme. Se clicchiamo su Grams (in alto a destra), stabiliamo la grandezza dei Grams (possono essere da 0 a 10). Una volta cliccato "apply", avremo i nostri risultati, assieme alla loro frequenza (assoluta o relativa) e alle loro misure di dispersione. Se facciamo doppio click sinistro sul pallino rappresentante il nostro corpus, ci dice da quali subcorpora è formato. Se lo facciamo col destro, ci dà la tabella con frequency e lexical stats. Se tiriamo i pallini uno dentro l'altro, troviamo i lockgrams: sono come le lockwords, ma in n-grams, ovvero a coppie di 2, di 3 e così via. Possiamo scegliere, durante la nostra ricerca, la statistica. Possiamo trovare:

- Simple math
- LogLik
- %DIFF
- LogRatio
- Cohen' D

**Strumento Text:** strettamente collegato a Whelk, ci dà lo stesso tipo di ricerca fatta lì ma all'interno del testo, sottolineando la parola da noi cercata.

**Strumento Wizard:** produce ricerche e report da stampare sia in doc che in html. Combina i "poteri" degli altri strumenti, e permette di esportare i risultati ottenuti per le proprie relazioni.

## Lancaster Stat tool Online

Prima di iniziare, bisogna tenere a mente che le variabili possono essere: nominali, ordinali o variabili di scala (scale variable).

Una variabile nominale ha dei valori che rappresentano diverse categorie senza rispettare ordine o gerarchie: si sceglie un criterio qualitativo per raggruppare.

Una variabile ordinale rappresenta un raggruppamento dei casi in categorie distinte, con una gerarchia intrinseca.

Infine, una variabile a scala è una variabile quantitativa che può mostrare attraverso un valore in una scala la quantità di una particolare caratteristica.

Parliamo di **Stat tools online**.

In breve, abbiamo una moltitudine di mini-tool utili all'analisi statistica dei nostri dati.

Lancsbox è utile, ma non può fare il lavoro statistico per voi. Per questo possiamo andare su [Statistics in Corpus Linguistics: Lancaster Stats Tools online \(lancs.ac.uk\)](https://lancs.ac.uk/statistics-in-corpus-linguistics/lancaster-stats-tools-online). Il sito dà dei pdf

per spiegarvi ogni tool, esattamente dove c'è scritto **For help click [here](#)**, come intuibile.

Ovviamente quanto vi serve un tool dipende dalla vostra domanda di ricerca, non serve sapere cosa ogni tool faccia o come esso funzioni. Orientatevi in base a come sono divisi.

Il primo strumento utile è il Graph tool, in  **Introduction**.

Se avete dei dati in formato excel (foglio di calcolo), potete importarli per farvi ritornare un istogramma in caso di una variabile linguistica, mentre se si deve tener conto di più variabili linguistiche tornerà un boxplot.

Osservando le **Error Bars** 95% Confidence Intervals, possiamo notare la sovrapposizione o no dell'interquartile range nel grafico.

Se non avete idea di cosa sia l'interquartile range, non disperate. Si tratta di un range che divide tutti i dati in quartili (= 4 parti) e vede quanto sono lontani i quartini centrali: più lontano dagli estremi e più sono dispersivi, più sono vicini e più sono omogenei.

Lo **Scatterplot** segna le diverse variabili statistiche nel corpus, comparandole. I pallini mostrano come un singolo parlante usa entrambe le variabili, e può mostrare una regression line che passa sotto. (regression line = se la media dell'uso di quelle variabili va ad aumentare o si abbassa)

La **Line chart** mostra l'andamento di una variabile linguistica all'interno di uno o più corpus.

**Geomapping** ritorna i dati della geolocalizzazione in cui la variabile si verifica.

La **Stacked Barchart** restituisce come grafico a barre la variabile linguistica in termini di localizzazione, latitudine, longitudine e frequenza.

**Sparklines** ritorna l'andamento della frequenza di una o più parole come sparkline, ossia visualizzato come la lineetta classica che fa su e giù.

**Candlestick plot** ritorna la stessa cosa di sparklines, ma in formato candlestick.

Poi in **2 Vocabulary** troviamo il **Dispersion calculator**, nel quale inserire le frequenze relative separate da punti e virgola (;), le frequenze relative si trovano usando Lancsbox.

Troviamo, sempre in quella sezione, l'**ARF calculator**. Questo tool ci permette di calcolare l'Average Reduced Frequency, la parola con una frequenza e una dispersione più uniforme sarà da considerarsi la più importante.

In **3 Semantics and discourse** troviamo il **Collocation calculator**, che consente di visualizzare facilmente i 12 calcoli statistici più comuni circa le collocazioni (le association measures: MU, LL, LOGDICE, MI, Z-score, LOGRATIO, MI2, T-score, MINIMUM SENSITIVITY, MI3, DICE, DELTA P).

Troviamo anche l'**Agreement Calculator**, che calcola il p-value e l'inter-rater agreement measure. Ovviamente è necessario generare una legenda comune a chi valuta, prima di procedere con le valutazioni per l'inter-rater agreement.

Andando avanti, i prossimi tool sono in **4 Lexico-grammar**.

La Cross-tab consente la visualizzazione di un sommario di dati che si dividono in varie categorie o sub-categorie.

Il Categories comparison tool ci consente di verificare il chi-squared, chi-squared con Yates's correction, Log likelihood e il Fisher exact test di una lista di dati. Consultate il pdf per capire come inserire correttamente i dati.

Il tool per la Logistic Regression consente di costruire un modello (anche graduale) per l'analisi di variabili binarie.

In **5 Register variation** troviamo il Correlation calculator, che può essere usato per calcolare le correlazioni parametriche o non- (Pearson's o Spearman's correlation) o per produrre e prendere visione di matrici di correlazione. Troviamo anche un tool per la cluster analysis, evidenziando quindi i "gruppi" più importanti. Infine, troviamo un tool per la MD (Multi-Dimensional) analysis.

Arriviamo a **6 Sociolinguistics**. Il primo strumento che vediamo è "Group comparison tool instructions": viene usato per comparare due o più subcorpora, permette di effettuare dei test parametrici e non parametrici, inoltre è possibile fare dei test statistici su gruppi differenti o dei test statistici con misure ripetute. Per maggiori informazioni, consultate il PDF.

Il secondo tool che troviamo è quello della "Correspondence analysis", che permette di fare una analisi delle corrispondenze e di visualizzarla tramite un grafico.

Troviamo poi la "Mixed effect logistic regression", che permette di analizzare questo tipo di regression, inoltre consente di ottenere i valori dei modelli e di ottenere sia gli effetti randomizzati dei valori che quelli regolarizzati dal punto di vista statistico.

All'interno di **7 Change over time** troviamo tool che ci permettono di lavorare coi dati diacronici. Come primo strumento, troviamo il "Bootstrapping test", che effettua la "Mixed effect logistic regression", ma che permette di visualizzarne gli effetti in un breve tracciato. Troviamo poi il "Neighbour cluster", che permette di fare un'analisi VNC (Variability-based neighbour clustering), e permette di visualizzarne gli effetti in un triplo dendrogramma (nome difficile per un semplicissimo grafico ad albero che monitora l'uso del termine che stiamo cercando all'interno di diversi anni).

Lo strumento "Peaks and troughs" permette di applicare un modello generalizzato additivo a dati

linguistici, visualizzandone i risultati tramite un grafico con una curva a regressione non parametrica tra il 95 e il 99% della variabile Cis.

L'ultimo strumento della sezione è l'UFA, che viene utilizzato per fare l'analisi di picchi di variazioni di alcune parole nei corpus storici, comparando le collocation nel corso del tempo e categorizzandole come "costanti", "iniziali", "finali" e "transitorie". Fa visualizzare il loro diverso impiego con un grafico "peaks and troughs".

Siamo alla fine:  **Bringing everything together**. È anche un bel titolo. I tool che ci riguardano particolarmente sono "Effect size calculator" e il "Meta-analysis calculator".

Il primo ci permette di visualizzare l'ampiezza dell'effetto di un (certo tipo di) input, converte inoltre le variazioni degli effect size e calcola l'intervallo di confidenza (CI) di quei valori.

L'ultimo combina diversi tipi di analisi. Calcola l'effetto complessivo e il 95% dell'intervallo di confidenza. Ci ritorna anche il forest plot (la digi-evoluzione dei singoli diagrammi ad albero, a livello meta. Un po' come questa guida).