
Breve guida all'uso di alcuni software per l'analisi testuale ed il trattamento automatico del linguaggio (TAL)

Versione 1.4 del 18/09/2008

Triple

Tavolo di Ricerca sulla Parola e il Lessico
Research Desk on Word and Lexicon



Dipartimento di Linguistica, Università Roma Tre

a cura di **Valentina Efrati**
vefrati@uniroma3.it

Indice

PREMESSA	- 3 -
MONOCONC PRO.....	- 4 -
BREVE PRESENTAZIONE.....	- 4 -
LINEE GUIDA	- 4 -
<i>Avvio del programma</i>	- 4 -
<i>Il menu Help</i>	- 5 -
<i>Caricare un corpus</i>	- 6 -
<i>La frequenza</i>	- 6 -
<i>Ricerca di concordanze</i>	- 7 -
<i>Ricerca di collocati e collocazioni</i>	- 12 -
<i>Salvataggio, stampa e chiusura del programma</i>	- 13 -
WORDSMITH TOOLS.....	- 14 -
BREVE PRESENTAZIONE.....	- 14 -
LINEE GUIDA	- 14 -
<i>Installazione ed avvio del programma</i>	- 14 -
<i>Il menu Help</i>	- 15 -
<i>Caricare un corpus</i>	- 16 -
<i>I programmi</i>	- 17 -
<i>Ricerca di concordanze</i>	- 17 -
<i>Ricerca di collocati e collocazioni</i>	- 22 -
<i>Creare una wordlist con il programma Wordlist</i>	- 24 -
<i>Creazione di una lista di parole significative con il programma Keywords</i>	- 26 -
<i>I programmi di utilità Viewer e Aligner</i>	- 28 -
<i>Salvataggio, stampa e chiusura del programma</i>	- 29 -
THE SKETCH ENGINE.....	- 30 -
BREVE PRESENTAZIONE.....	- 30 -
LINEE GUIDA	- 31 -
<i>Avvio del programma</i>	- 31 -
<i>Ricerca di concordanze</i>	- 32 -
<i>Generare keywords</i>	- 37 -
<i>Usare la sezione Context</i>	- 38 -
<i>La funzione Word Sketch</i>	- 38 -
<i>La funzione Thesaurus</i>	- 42 -
<i>La funzione Sketch Difference</i>	- 43 -
GLOSSARIO	- 44 -
BIBLIOGRAFIA.....	- 48 -

PREMESSA

Questa guida nasce con lo scopo di fornire un ausilio all'uso dello strumento informatico per l'analisi di testi in formato elettronico. Il processo di memorizzazione e codifica dei dati è la più ovvia delle applicazioni dell'informatica nel settore umanistico, operazione preliminare a qualsiasi forma di trattamento automatico. Passo successivo alla memorizzazione e alla conseguente codifica del testo è l'organizzazione del materiale. Tra le molteplici possibili forme di strutturazione dei dati, sicuramente quella che consente di accostarsi agli obiettivi dell'analisi testuale in modo appropriato è l'organizzazione nella forma della base di dati, o *database*. Organizzare il materiale testuale nella forma della base di dati significa consentire operazioni di *information retrieval*, cioè di recupero dell'informazione ricercata, che superano il livello del mero riconoscimento dei dati informativi per stringhe di caratteri, cioè il *pattern recognition*, consentendo inoltre indagini testuali di varia natura. I software che realizzano analisi testuali sono molteplici. In questa sede si è deciso di trattarne solo alcuni che, a nostro avviso, offrono un ottimo compromesso tra ricchezza funzionale e praticità d'uso. Va detto, però, che pur presentando spesso caratteristiche differenti, la maggior parte dei software per l'analisi testuale possiede un sottoinsieme comune di funzionalità di base (come ad esempio la possibilità di realizzare [concordanze](#), [indici](#), [liste di frequenza](#), [collocazioni](#), ecc.); ciò permette di estendere la consultazione di questa guida come base anche per quei software di analisi testuale non citati esplicitamente.

Si fa presente, inoltre, che i software menzionati in questa guida sono tutti disponibili nel laboratorio TRIPLE del Dipartimento di Linguistica della Facoltà di Lettere e Filosofia dell'Università degli Studi Roma Tre.

MONOCONC PRO

OS: Windows (NT//98/2000/ME/XP)

Licenza individuale: \$85 (Disponibile Demo)

Breve presentazione

MonoConc è un programma interattivo per Windows molto agile e veloce. Il programma è di fascia medio/alta in quanto a prestazioni. Per quanto riguarda le ricerche di stringhe supporta le **espressioni regolari** e la ricerca di etichette (se il corpus è già etichettato). Il programma permette la personalizzazione dell'alfabeto di caratteri usato (per esempio le norme su cosa considerare confine di parola, ecc.), esegue **concordanze**, **liste di frequenza**, ricerche avanzate. Le maschere di lavoro sono elementari. Particolarmente versatile nella visualizzazione delle occorrenze, utilizza il formato *KWIC* oltre a quello per frase, elimina le occorrenze che non interessano, visualizza un contesto molto ampio, segnala marcatori particolari (*tags*) all'interno del contesto.

URL: <http://www.athel.com/mono.html>

Linee guida

Avvio del programma

Per avviare il programma fare doppio-click sul file MonoPro. Una volta avviato il programma, apparirà una semplice schermata come quella mostrata in Figura 1.

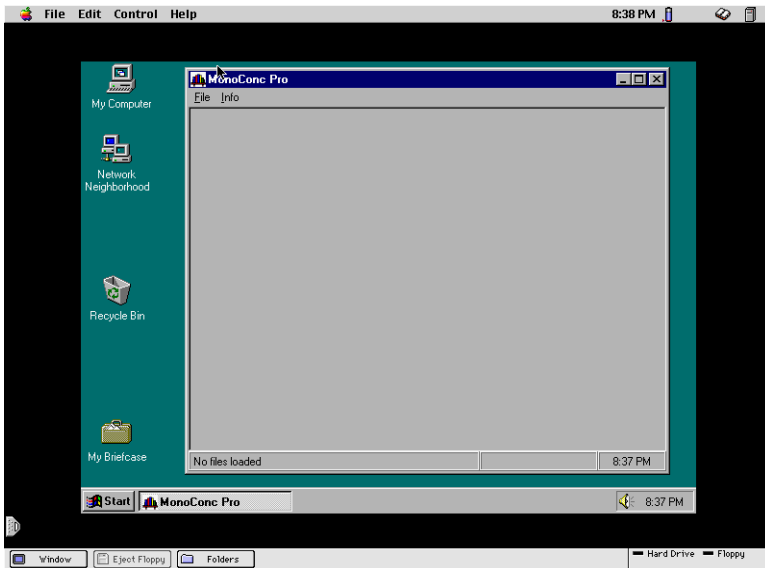


Figura 1 - Schermata principale di Monoconc Pro

Il menu Help

Il menu “help” si occupa di fornire alcune indicazioni di base sull’utilizzo di Monoconc. L’“help” è organizzato per argomenti. È sufficiente cliccare sul titolo dell’argomento di interesse, come si vede in Figura 2.

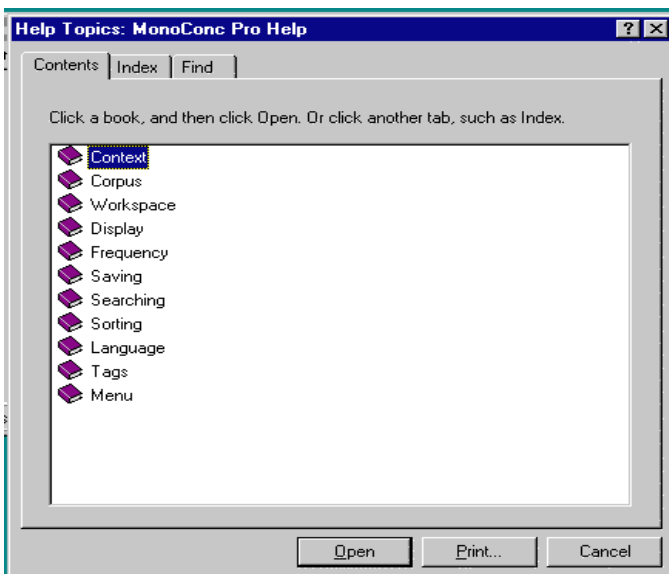


Figura 2 - Menu “Help”

Caricare un corpus

La prima operazione è quella di caricare un corpus nel programma. Una volta caricato un corpus appariranno alcuni elementi di menu relativi all'analisi e alla visualizzazione dei testi.

In Figura 3, al centro, è possibile vedere i testi del corpus caricati in Monoconc mentre in basso, a sinistra ci sono informazioni sul numero di file caricati e a destra informazioni sul numero di parole nel corpus. Occorre tenere presente che -una volta terminate le proprie ricerche- per scaricare definitivamente i testi dal programma non è sufficiente chiudere le relative finestre ma è necessario scegliere l'opzione "Unload corpus" dal menu "File".

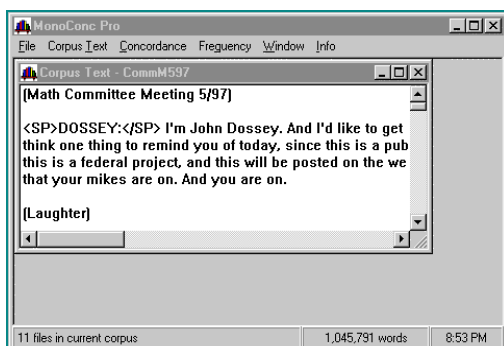


Figura 3 - Visualizzazione dei testi del corpus

La frequenza

Di solito, le parole più frequenti in un corpus sono anche quelle che danno maggiori indicazioni sulla natura del corpus (tranne le parole *troppo frequenti* che sono per lo più quelle dal significato grammaticale, tipicamente molto brevi).

Per trovare la distribuzione delle parole nell'intero corpus, basta scegliere "corpus frequency data" dal menu "Frequency" e selezionare "Frequency order" se si desidera visualizzarle in ordine di frequenza o "Alphabetical order" se si desidera visualizzarle in ordine alfabetico. Nel menu "Frequency" c'è il sottomenu "Frequency option" dal quale è possibile limitare la presentazione dei dati in tre modi: 1) settando il numero massimo di linee della *frequency list*; 2) settando il limite inferiore di frequenza accettabile (la frequenza minima accettata); 3) settando il limite superiore di frequenza accettabile (la frequenza massima accettata). Vedere la Figura 4.

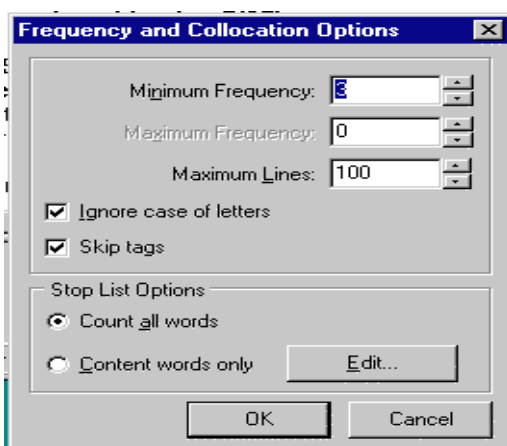


Figura 4 – Le opzioni del menu Frequency

Ricerca di concordanze

Uno strumento per la ricerca di concordanze è un programma che ricerca dei *patterns* nel testo sulla base di una *search query*. I vantaggi nell'uso di un programma per concordanze sono che esso permette di: i) trovare istanze di parole o stringhe; ii) trovare stringhe nel contesto di altre stringhe; iii) cercare *patterns* particolari per poi riadattare e raggruppare istanze simili per poter rivelare le loro proprietà.

Per realizzare delle *query* più articolate, dobbiamo ricorrere all'uso delle *Regular expressions*. Le *espressioni regolari* (in inglese abbreviate in *regexp*, *regex* o *RE*) servono per trovare corrispondenze di modelli (*patterns*) su stringhe e costituiscono uno strumento tanto difficile ed ostico (soprattutto all'inizio) quanto potente ed utile. L'operazione principale che utilizza una *espressione regolare* è il *matching*, cioè la verifica che una stringa appartenga all'insieme descritto dall'*espressione regolare*.

Selezionare "Search" dal menu "Concordance" oppure la sequenza di tasti ctrl-s, come mostrato in Figura 5.

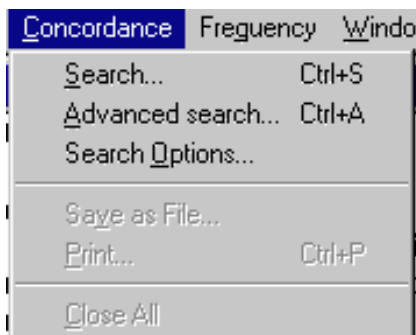


Figura 5 - Il menu “Concordance”

Apparirà una finestra (Figura 6) dove inserire un termine di ricerca (ad esempio **speak**) e premere successivamente il pulsante OK.

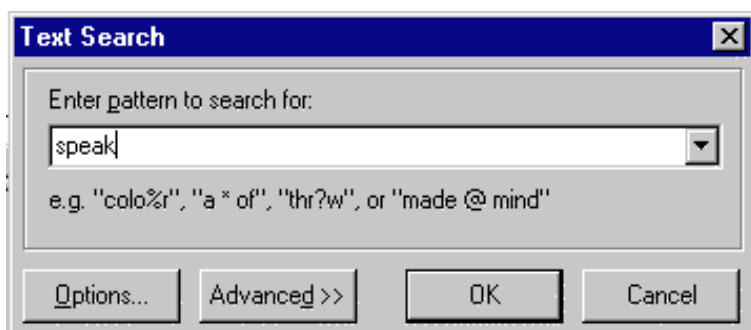


Figura 6 – Esempio di ricerca semplice in un testo

Cliccando sul pulsante “advanced” è possibile settare alcune impostazioni avanzate (Figura 7).

Per chiarirne l’uso vediamo un esempio: per cercare il *present perfect* (passato prossimo) in un corpus di inglese non annotato dobbiamo cercare ad esempio *has* o *have* seguito da una parola che finisce in *-ed*, quello che dobbiamo fare è scegliere “Search” ed inserire la seguente stringa: **ha?%*ed**. Se volessimo permettere che una parola opzionale occorra tra *have/has* e il *participle* dovremmo, allora, entrare nelle opzioni di ricerca (*Search Options*) ed impostare da 0 a 1 i valori del carattere jolly ‘@’, successivamente inserire la stringa **ha?%@*ed**.

Per effettuare ricerche più complesse bisogna affidarsi al potere espressivo delle **espressioni regolari**. Dal menu “Concordance” scegliere “Advanced Search” e selezionare il bottone “Regular expression”. Riprendendo l’esempio appena mostrato proviamo ad impostare una **query** sulla base del formalismo delle **espressioni regolari**; la **query** da usare è la seguente: **\bha[vs]e?\W\w+e[nd]\b** (per una corretta interpretazione del simbolismo usato, vedere voce “*Regular expressions*” del glossario).

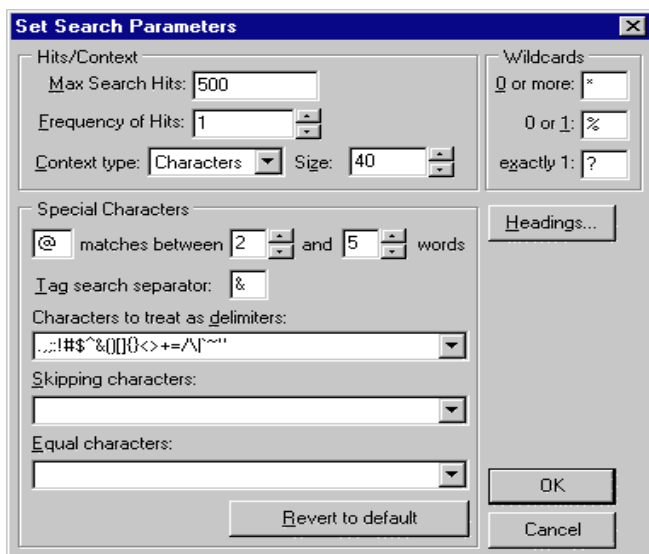


Figura 7 – Impostazione delle opzioni di ricerca

Dopo aver impostato tutti i parametri di interesse ed avviata la ricerca, si aprirà una finestra con i risultati delle concordanze (Figura 8).

Comunemente, l'istanza di parola cercata è centrata attorno ad una certa quantità di testo (cotesto destro e sinistro). Questo tipo di presentazione dei risultati è chiamato formato KWIC (*Key Word In Context*).

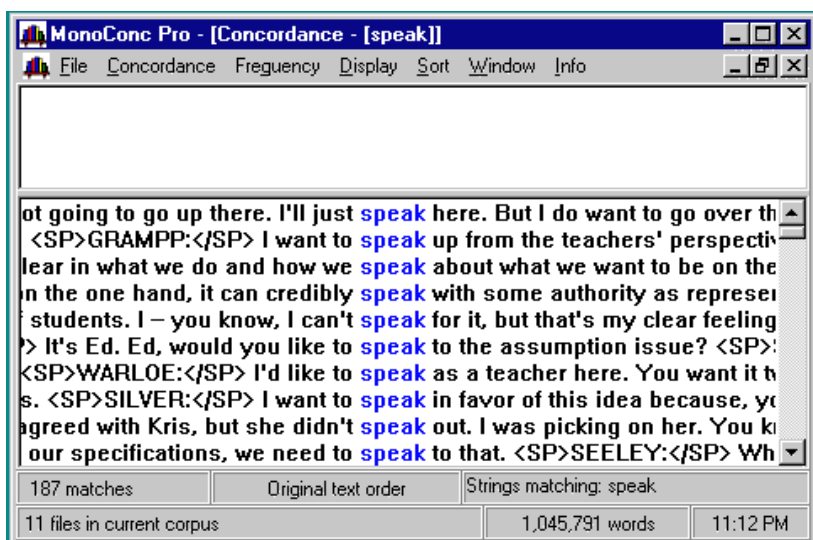


Figura 8 – Risultati delle concordanze nel formato KWIC

Una volta terminata la ricerca, è possibile applicare le potenzialità del programma per scoprire dei patterns nei risultati. Ad esempio un modo per trovare quali parole sono associate con la parola *speak* è quello di ordinare in ordine alfabetico le istanze della parola che segue quella ricercata. In questo modo le istanze di *speak to* appariranno una di seguito all'altra, allo stesso modo le istanze di *speak with*.

Per impostare l'ordinamento basta selezionare, dal menu "sort", tra le varie opzioni, quella desiderata, ad esempio *1st right* ordina in base alla prima parola a destra della parola cercata, *1st left* ordina in base alla prima parola a sinistra della parola cercata (Figure 9-10). La possibilità di alterare l'ordine originale di visualizzazione dei risultati consente di fare indagini più specifiche ed è molto utile soprattutto quando si vuole studiare ad esempio con quali parole si accompagna più di frequente la parola chiave in esame (ad esempio una visualizzazione con ordinamento di tipo *1st right* sulla parola chiave *speak* ci permette di osservare con quali preposizioni si accompagna, come *speak to*, *speak with*, eliminando le istanze che non ci interessano.

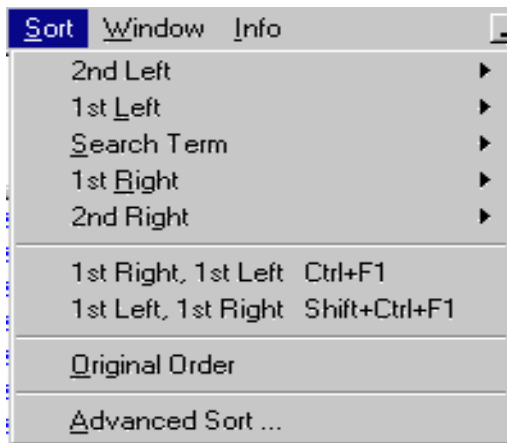


Figura 9 – Il menu "sort"

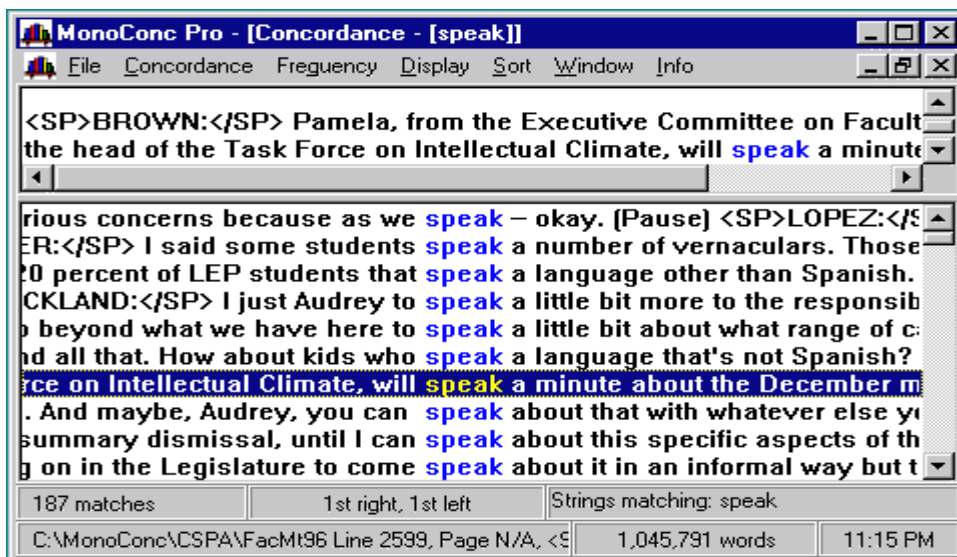


Figura 10 – I risultati ordinati

Per vedere un contesto più ampio è sufficiente cliccare sulla linea di concordanza di interesse ed esso apparirà in una finestra di contesto sopra quella dei risultati (Figura 10).

Spesso si ha bisogno di opzioni di ricerca più potenti della ricerca semplice per parola illustrata precedentemente; è il caso delle ricerche per “parti del discorso” (*Part of Speech, POS*). Si potrebbe ad esempio essere interessati alla ricerca di *patterns* del tipo “un verbo seguito da un pronome possessivo seguito dalla parola *way*”. Va ribadito però che una simile ricerca complessa nel corpus necessita obbligatoriamente di un corpus che sia annotato (*taggato*) con le POS e che il *tagset* che specifica quali siano le POS sia noto. In Figura 11 è mostrato come è possibile specificare una simile struttura da ricercare nel corpus.

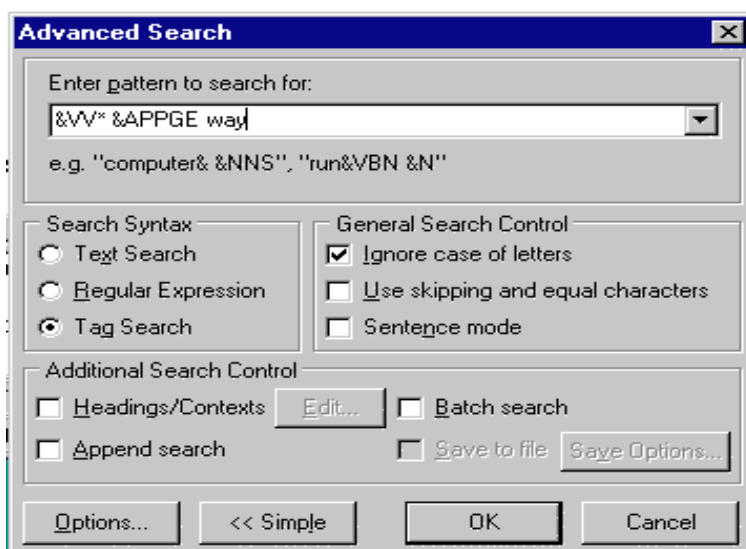
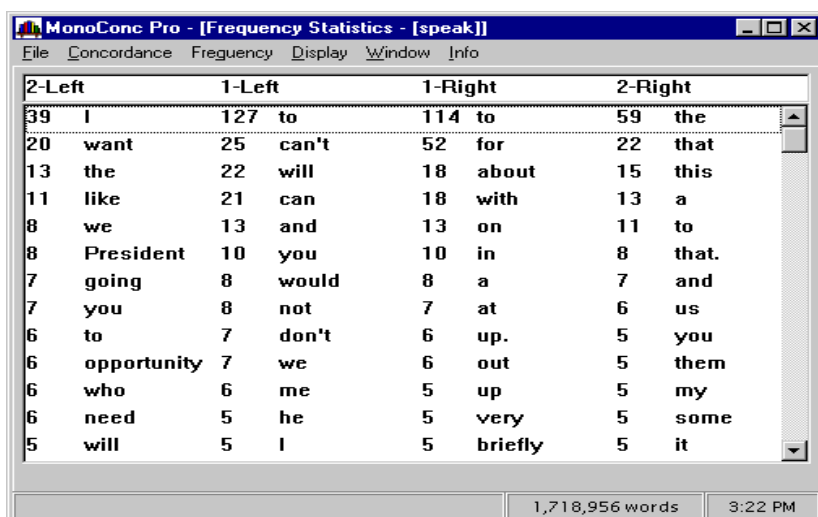


Figura 11 – Un esempio di ricerca annotata con le “parti del discorso” (POS)

Ricerca di collocati e collocazioni

MonoConc Pro è dotato di una gamma di statistiche di frequenza. I due tipi principali sono la *corpus frequency* e la *collocation frequency*. Il comando “corpus frequency data” crea una *wordlist* dell’intero corpus. Scegliendo, invece, “collocate frequency data” dal menu “Frequency” è possibile vedere i collocati della search word ordinati in base alla frequenza. Per “collocati” di una parola si intendono parole che appaiono insieme frequentemente. Il computo della *collocate frequency* è legato alla particolare parola cercata, infatti il menu “Frequency” appare solo dopo che una ricerca è stata eseguita. Le *collocations* (collocazioni) prodotte vengono graficamente rappresentate in 4 colonne, una colonna per ogni posizione circostante la parola chiave (2nd left, 1st left, 1st right e 2nd right). Le colonne mostrano i collocati in ordine discendente di frequenza.

Riprendendo l’esempio fatto precedentemente, una volta cercata la parola *speak* è possibile selezionare “collocate frequency data” dal menu “Frequency” e vedere a prima vista le parole più comuni che occorrono subito dopo la parola *speak*. Come si può vedere nella Figura 12, osservando i collocati di *speak*, la parola più comune dopo *speak* è *to* insieme ad una vasta gamma di altri collocati.



The screenshot shows a window titled "MonoConc Pro - [Frequency Statistics - [speak]]". The window contains a table with four columns: "2-Left", "1-Left", "1-Right", and "2-Right". The rows represent collocates for the word "speak", sorted by frequency. The first row shows "39 I", "127 to", "114 to", and "59 the". The second row shows "20 want", "25 can't", "52 for", and "22 that". The third row shows "13 the", "22 will", "18 about", and "15 this". The fourth row shows "11 like", "21 can", "18 with", and "13 a". The fifth row shows "8 we", "13 and", "13 on", and "11 to". The sixth row shows "8 President", "10 you", "10 in", and "8 that.". The seventh row shows "7 going", "8 would", "8 a", and "7 and". The eighth row shows "7 you", "8 not", "7 at", and "6 us". The ninth row shows "6 to", "7 don't", "6 up.", and "5 you". The tenth row shows "6 opportunity", "7 we", "6 out", and "5 them". The eleventh row shows "6 who", "6 me", "5 up", and "5 my". The twelfth row shows "6 need", "5 he", "5 very", and "5 some". The thirteenth row shows "5 will", "5 I", "5 briefly", and "5 it". At the bottom of the window, it displays "1,718,956 words" and "3:22 PM".

	2-Left	1-Left	1-Right	2-Right
39	I	127	to	59 the
20	want	25	can't	22 that
13	the	22	will	15 this
11	like	21	can	13 a
8	we	13	and	11 to
8	President	10	you	8 that.
7	going	8	would	7 and
7	you	8	not	6 us
6	to	7	don't	5 you
6	opportunity	7	we	5 them
6	who	6	me	5 my
6	need	5	he	5 very
5	will	5	I	5 some
			5	5 briefly
				5 it

Figura 12 – Tabella di frequenza dei collocati per la parola *speak*

Uno svantaggio della semplice visualizzazione mostrata in Figura 12 è che non è possibile calcolare la frequenza di collocazioni che constano di tre o più parole. Infatti, non

siamo in grado di dire dall'analisi della tavola dei collocati di *speak* quanto è frequente il pattern (la sequenza) *to speak to*. Per calcolare la frequenza di collocazioni con tre parole è necessario usare il comando “advanced collocation” nel menu “Frequency”(Figura 13) dove occorre impostare i valori per le tre posizioni occupate dalle parole.

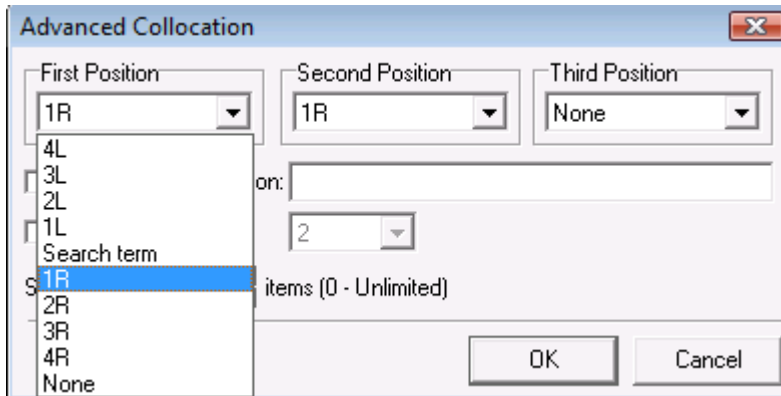


Figura 13 – Comando “advanced collocation” nel menu “Frequency”

Salvataggio, stampa e chiusura del programma

Infine, è possibile salvare e stampare i risultati delle ricerche. Per uscire dal programma basta scegliere “Exit” dal menu “File” oppure digitare la sequenza *ctrl-q*. Va sottolineato però che quando si chiudono le finestre di lavoro il corpus rimane caricato, quindi per scaricarlo bisogna andare nel menu “File” e scegliere “unload corpus”.

WORDSMITH TOOLS

OS: Windows 98/2000/XP

Licenza individuale: 75 € (Disponibile Demo)

Breve presentazione

Wordsmith Tools è un pacchetto completo e flessibile per l'elaborazione di concordanze, *liste di frequenza* e interrogazioni complesse. Si tratta di una serie integrata di programmi a 32-bit, che lavora sotto sistema operativo Windows 98/2000/XP. Gli strumenti principali sono Concord, Keywords e Wordlist, cui si aggiungono alcune utilità [*alignment* (allineamento), analisi dei caratteri del testo, individuazione di coppie minime, ecc.], tutti con un'interfaccia user-friendly. Il pacchetto permette di selezionare le diverse lingue (con i rispettivi principi di ordinamento alfabetico), riconosce i principali sistemi di *tagging* (annotazione), permette ricerche sui *tags*, permette la predisposizione di *stop lists* e *lemma lists*, di calibrare i contesti delle concordanze, l'attivazione/disattivazione della *case-sensitivity* e molto altro.

Complessivamente si tratta di uno dei pacchetti più versatili, ricchi e poco costosi a disposizione ed è inoltre largamente usato in studi scientifici.

URL : <http://www.lexically.net/wordsmith/>

Linee guida

Installazione ed avvio del programma

Lanciare il file di installazione: il file richiede di indicare una cartella nella quale copiare i file di lavoro dell'installazione (p.es. C:\Documenti\Wsm5): questa cartella serve solo per l'installazione, alla fine potrà essere eliminata.

Indicare una cartella (o crearla) e lasciare che il programma copi i file.

Aprire la cartella indicata e cliccare sul file setup.exe. Apparirà una maschera che chiederà di indicare dove installare il programma. L'opzione predefinita è c:\wsmith. Si consiglia di mantenere questo percorso. Se si desidera cambiare il percorso è necessario installare il programma in una cartella con un percorso simile a quello indicato, cioè in una cartella che sia subordinata di un solo livello rispetto al drive (p.es D:\wsmith o E:\cartella e mai cose del tipo C:\Documenti\MiaCartella\WordSmith 5 ecc.). Eliminare la cartella che serve per l'installazione (non servirà più).

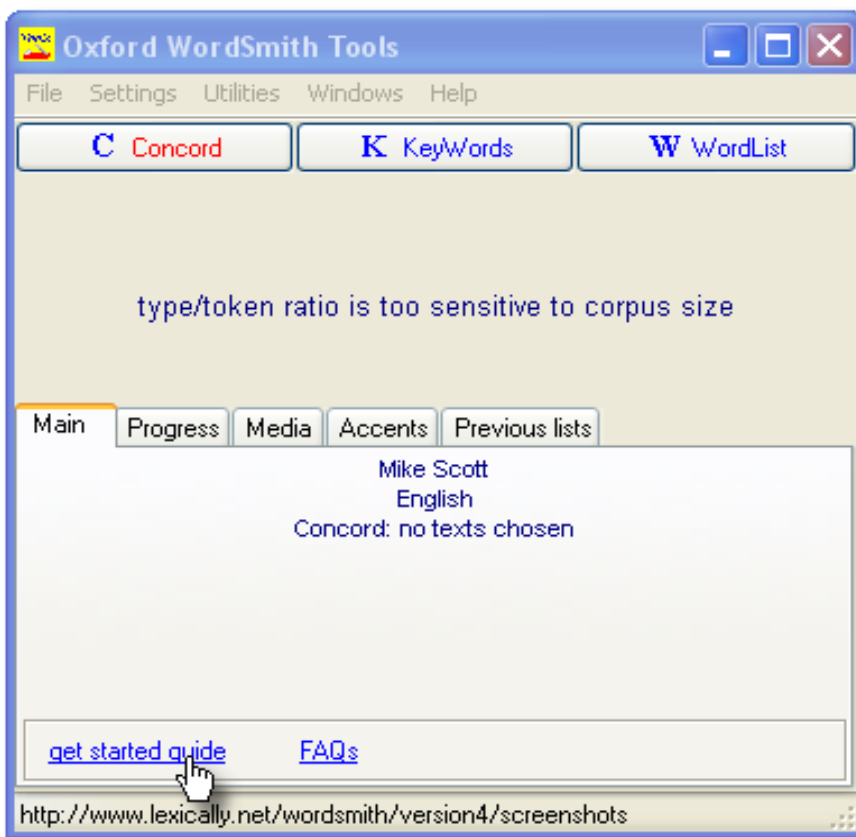


Figura 1 – Schermata principale di WordSmith Tools

Il menu Help

Il menu “Help” si occupa di fornire alcune indicazioni di base sull’utilizzo di Word Smith Tools. L’“help” è organizzato per argomenti. È sufficiente cliccare sul titolo dell’argomento di interesse, come si vede in Figura 2.

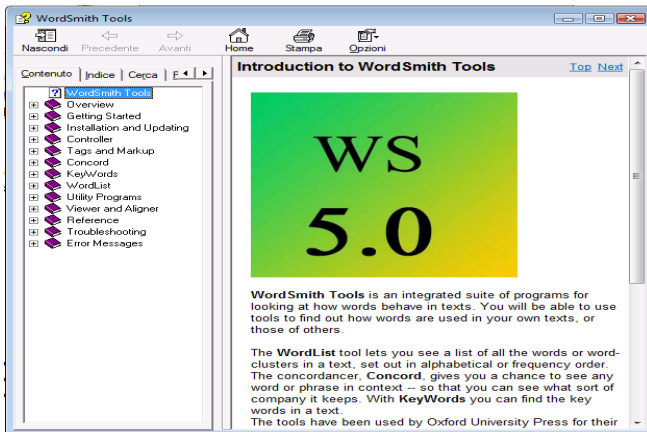


Figura 2 – Menu “Help”

Caricare un corpus

La prima operazione è quella di caricare un corpus nel programma. Per scegliere i file cliccare dal menu “File” sull’opzione “Choose texts” ed apparirà una schermata come quella mostrata in Figura 3. Nella colonna di sinistra si possono vedere i file disponibili mentre nella colonna di destra andranno trascinati i file selezionati da analizzare con WordSmith.

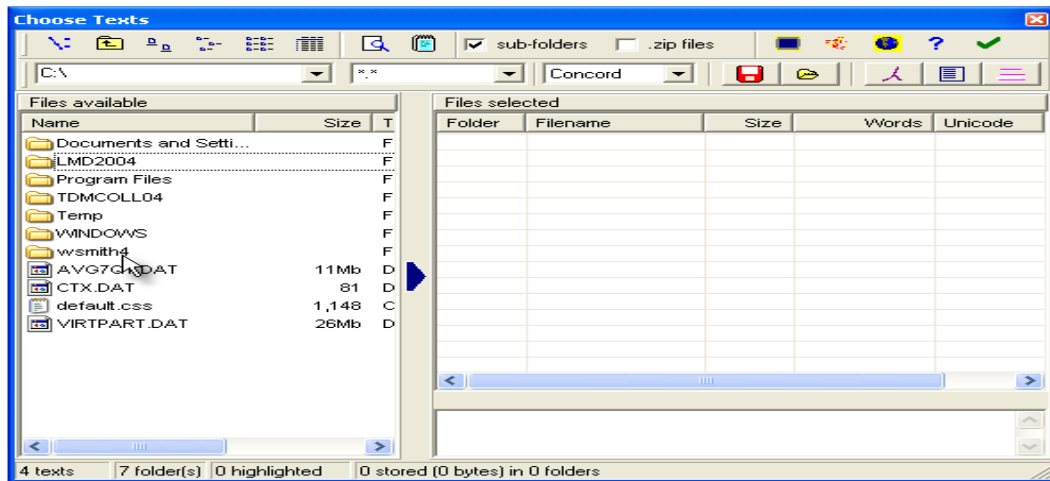



Figura 3 – Scelta dei file da caricare in WordSmith

Come si può osservare in Figura 4, è possibile vedere i testi del corpus caricati; in basso a sinistra ci sono informazioni sul numero di file caricati e sul numero di cartelle esplorate. Per terminare premere il bottone verde  in alto a destra oppure chiudere la finestra.

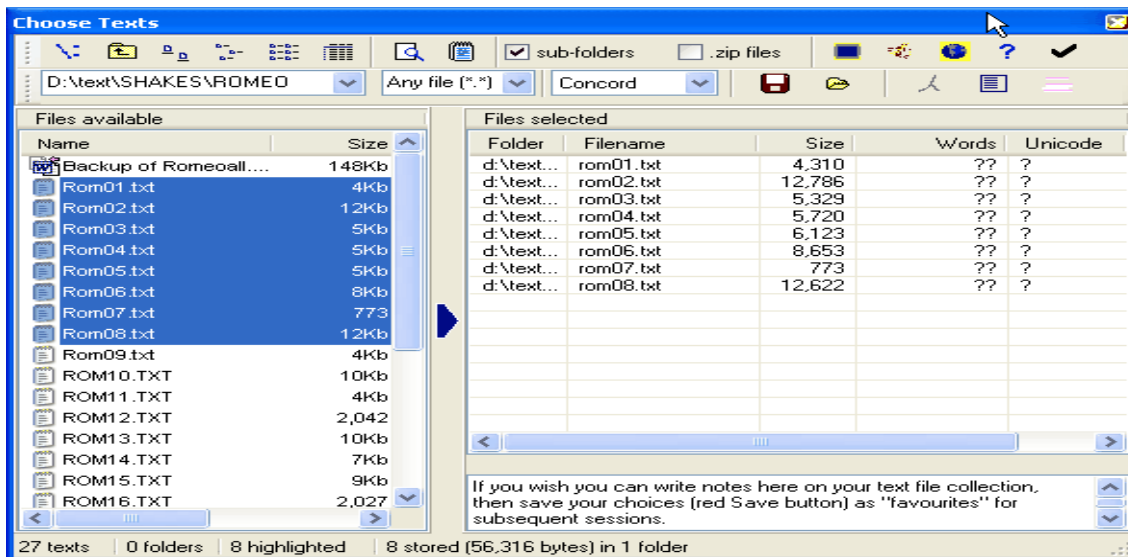


Figura 4 – Esempio di selezione e aggiunta di file

I programmi

WordSmith è dotato di tre programmi principali:

- **Concord** che permette di generare concordanze usando file in formato dos, ansi e solo testo, e permette inoltre di trovare *collocations*, *patterns* e *clusters*;
- **Keywords** che permette di individuare le parole-chiave dei testi del corpus;
- **Wordlist** che produce *liste di frequenza*, con la possibilità di condurre alcune analisi statistiche, lemmatizzazioni, comparazioni tra liste, ecc.

Il pacchetto permette di selezionare le **diverse lingue** (con i rispettivi principi di ordinamento alfabetico), riconosce i principali sistemi di *tagging* (annotazione), permette ricerche sui *tags* e permette la predisposizione di *stop lists* e *lemma lists*, di calibrare i contesti delle concordanze, l'attivazione/disattivazione della *case-sensitivity* e molto altro.

Ricerca di concordanze

Per creare una nuova concordanza è sufficiente cliccare sul bottone “concord” e, dalla nuova finestra che si è nel frattempo aperta, scegliere dal menu “File” l’opzione “new”. Cliccare sul bottone “choose Texts now” e scegliere i testi da esaminare. Cliccare su “search word” per

inserire la parola o le parole da cercare. Il procedimento appena descritto è mostrato nelle Figure 5-8.

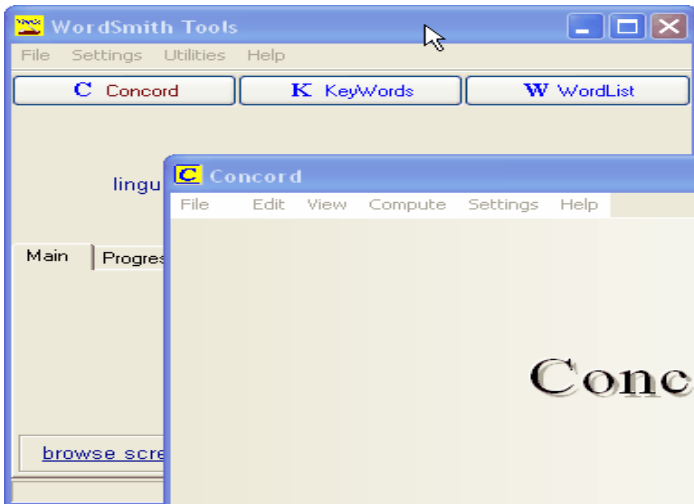


Figura 5 – Selezione del programma Concord

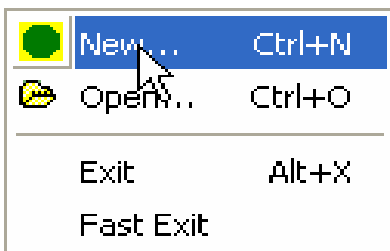


Figura 6 – Apertura dei file di testo

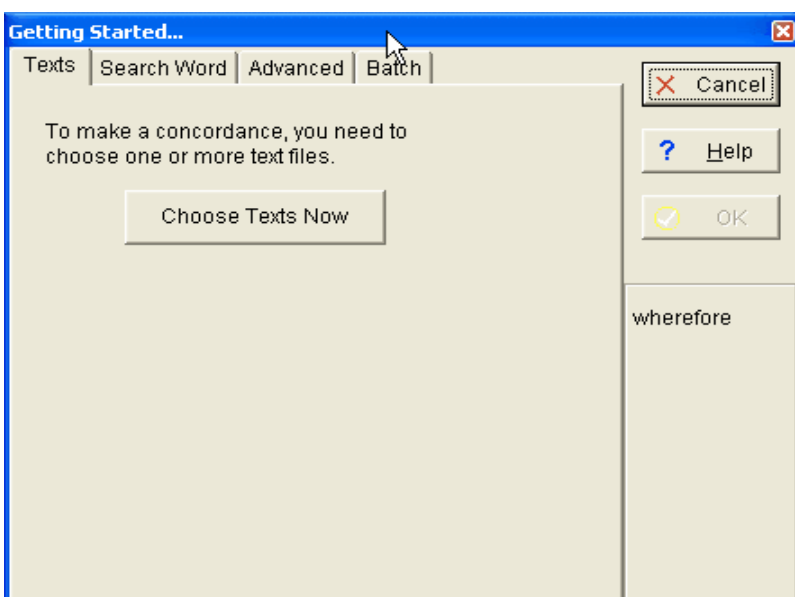


Figura 7 – Scelta dei file di testo

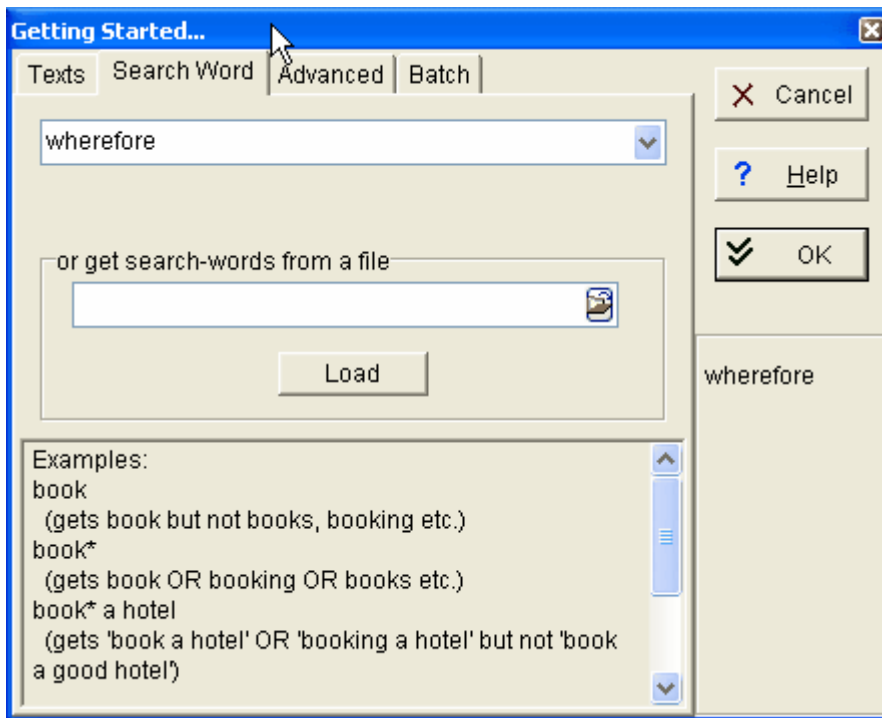


Figura 8 – Scelta della/e parola/e da ricercare

Un esempio di concordanza è mostrato in Figura 9. È possibile vedere come, per ogni occorrenza della parola cercata, viene riportata l'informazione sul testo nel quale la parola occorre e sul numero di parole contenute in quel testo.

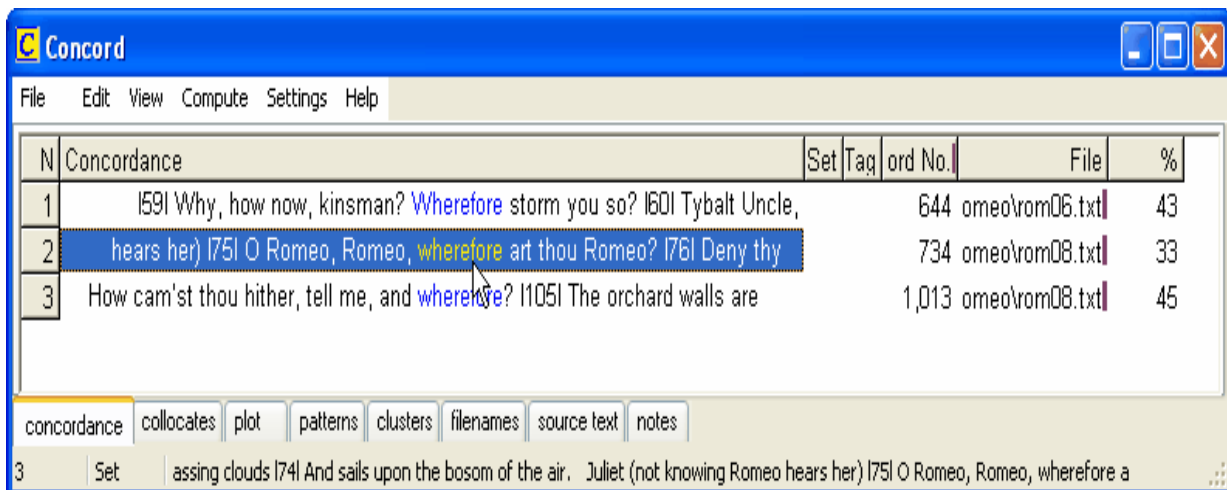


Figura 9 – Esempio di concordanza

Per vedere il testo relativo ad una *entry* basta fare doppio click sulla riga in questione e, come si può vedere in Figura 10, il *cotesto* per quella riga si espande.

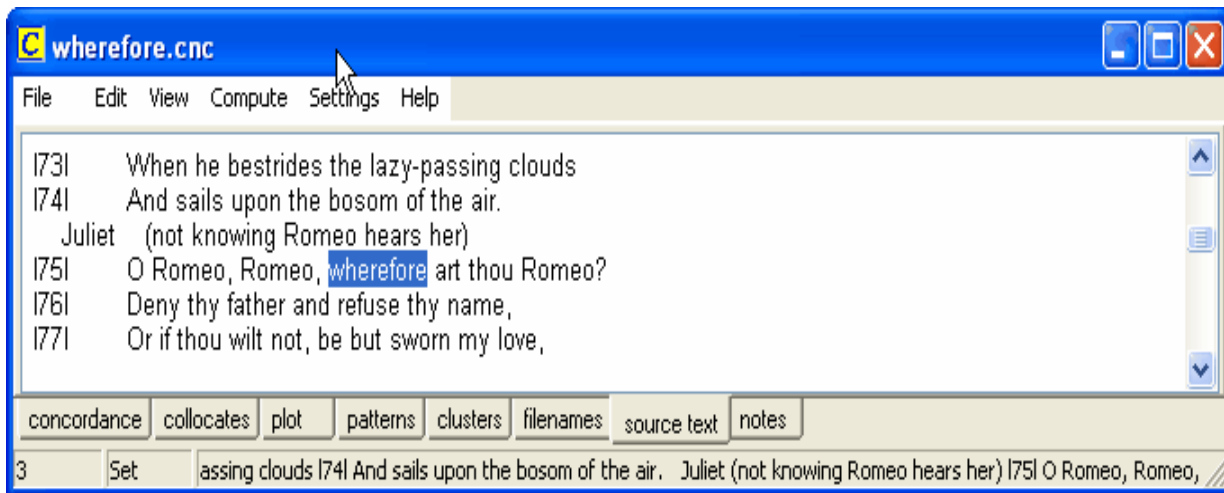


Figura 10 – Espansione del contesto di una occorrenza della parola ricercata

Dal menu “Settings” scegliendo “Customise” è possibile settare alcune impostazioni (Figure 11-12).

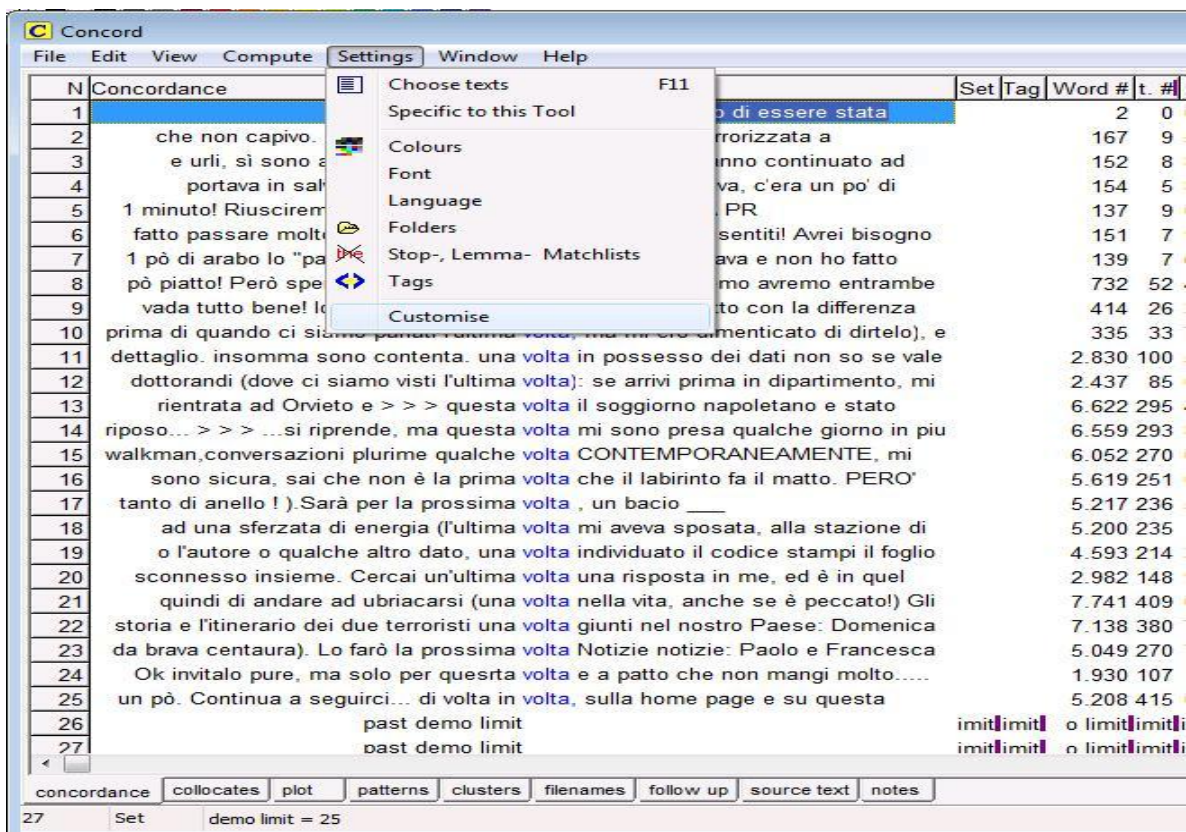


Figura 11 – Menu “Settings” (Impostazioni) del programma Concord

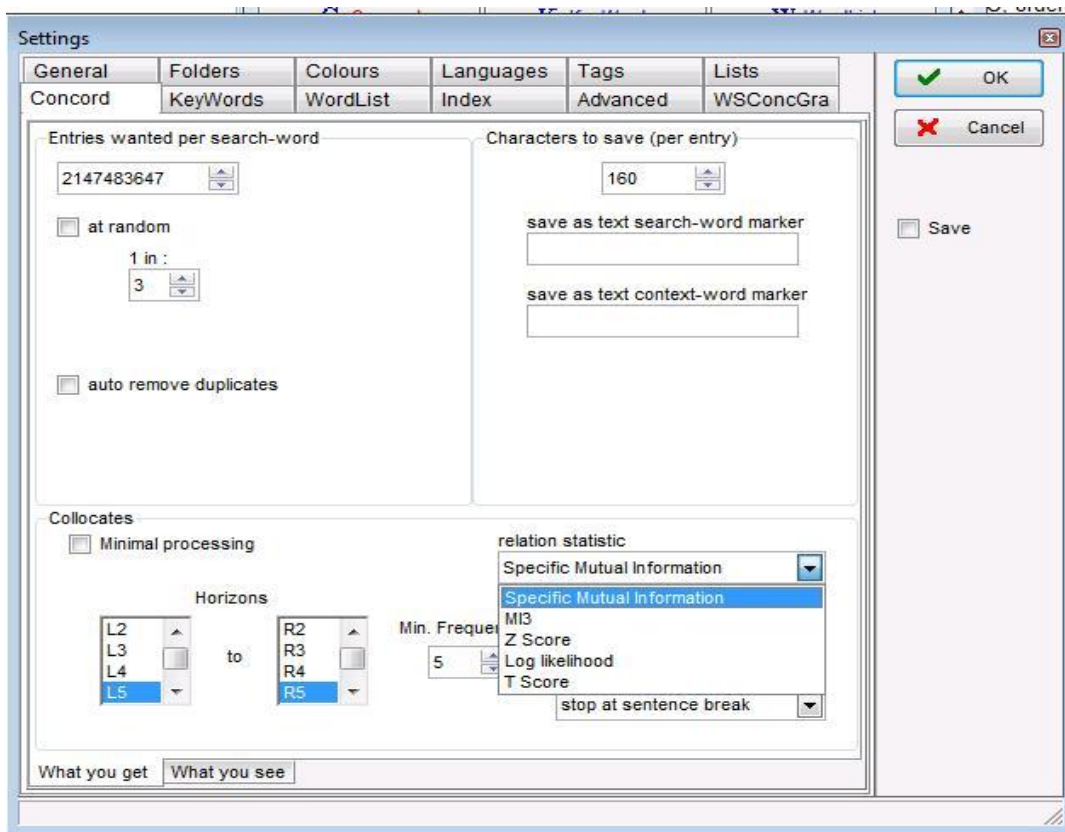


Figura 12 – Impostazioni delle opzioni di ricerca

È possibile analizzare corpora annotati (con le “parti del discorso”).

Scegliere “Choose” → “Adjust Settings” ed impostare i campi secondo le proprie esigenze, inserendo le informazioni sui **tags** da considerare o da ignorare.

È possibile operare una concordanza su una parte del discorso. Ad es. usando il BNC (British National Corpus) si può esprimere:

<w PRP> at <w ATO> the <w AJO> great <w N2> houses

Se invece si desidera vedere tutte le preposizioni in un testo selezionato dal BNC si dovrà scrivere: <w PRP>* (Figura 13).

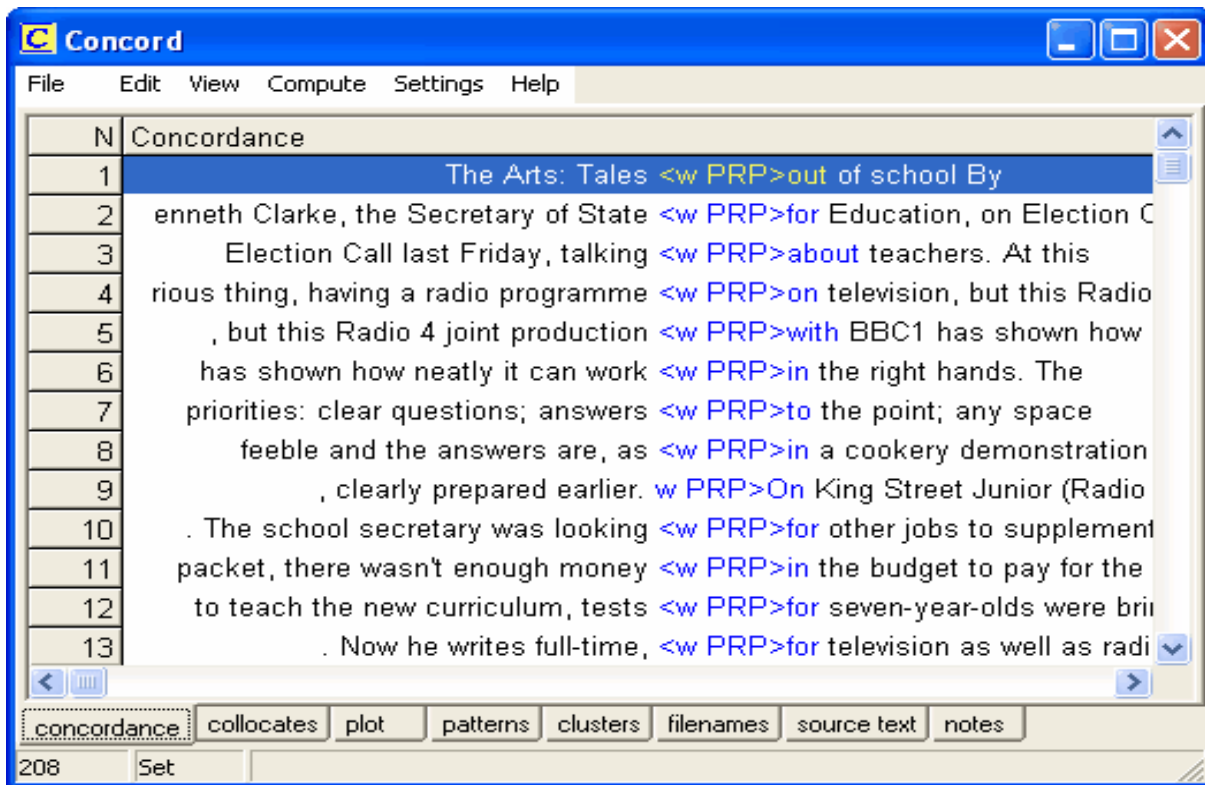


Figura 13 – Esempio di Concordanza basata sulle parti del discorso

Ricerca di collocati e collocazioni

WordSmith offre la possibilità di vedere i collocati di una *word search* ordinati ad esempio in base alla frequenza. In Figura 14 è mostrato un esempio in cui ci sono i collocati della parola inglese *ago* calcolati usando la sezione scritta del BNC. Sono presenti circa 17.000 istanze di *ago* e il collocato con la maggiore frequenza (9000 volte insieme ad *ago*) è *years*.

N	Word	With	elation	Total	tal Left
1	AGO	ago	0.000	16,785	47
2	YEARS	ago	0.000	9,033	8,936
3	A	ago	0.000	6,967	4,608
4	THE	ago	0.000	6,352	1,615
5	WAS	ago	0.000	2,951	1,183
6	OF	ago	0.000	2,949	1,345
7	AND	ago	0.000	2,740	623
8	TO	ago	0.000	2,506	679
9	IN	ago	0.000	2,263	826
10	TWO	ago	0.000	2,160	2,031
11	THAT	ago	0.000	1,801	722
12	IT	ago	0.000	1,695	668
13	I	ago	0.000	1,694	413
14	LONG	ago	0.000	1,591	1,527
15	MONTHS	ago	0.000	1,383	1,367
16	HE	ago	0.000	1,372	240
17	HAD	ago	0.000	1,312	442
18	THREE	ago	0.000	1,187	1,110
19	SOME	ago	0.000	1,123	983
20	FEW	ago	0.000	1,084	1,039
21	YEAR	ago	0.000	1,066	980

2,871 Type-in AGO

Figura 14 – Collocati per la keyword *ago*

Il problema è capire quanto ogni collocato di *ago* presente nella lista sia effettivamente strettamente collegato con esso (parole come *a*, *the*, *was* possono dirsi collocati di *ago*?).

Per ovviare a questo problema dal menu principale di Concord scegliere “Compute” -> “Mutual information” e selezionare una *wordlist* da usare come confronto; ciò permette sulla base di analisi sulla rilevanza statistica dei collocati di ottenere risultati significativi (Figura 15).

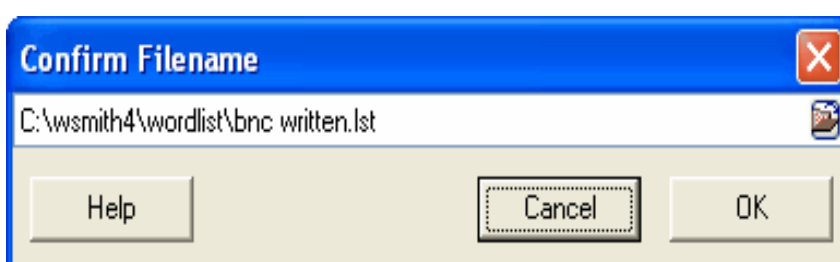


Figura 15 – Wordlist di riferimento

Quello che si ottiene è la lista in Figura 16 dove i *top items* della lista riflettono proprio la tendenza di *ago* ad accompagnarsi con intervalli temporali e numeri.

N	Word	With	Relation	Total	tal Left
1	AGO	ago	12.403	16,785	47
2	HENSLEY	ago	10.631	5	1
3	AEONS	ago	9.879	11	8
4	FORTNIGHT	ago	9.336	121	121
5	YEARS	ago	9.218	9,033	8,936
6	MOONS	ago	8.840	13	12
7	WEEKS	ago	8.754	1,047	1,029
8	SEASONS	ago	8.548	81	81
9	MILLENNIA	ago	8.512	9	9
10	MONTHS	ago	8.387	1,383	1,367
11	MOMENTS	ago	8.367	179	178
12	UNTHINKABLE	ago	8.128	18	15
13	DECADE	ago	7.939	165	164
14	COUPLE	ago	7.697	360	342
15	TWENTY	ago	7.658	405	387
16	CENTURIES	ago	7.592	126	123
17	TEN	ago	7.521	485	468
18	FIFTY	ago	7.500	133	127
19	TH	ago	7.485	10	0
20	MOOTED	ago	7.471	5	5
21	EIGHTEEN	ago	7.466	54	50
22	INCEPTION	ago	7.427	9	8
23	HUNDRED	ago	7.343	250	241
24	FIFTEEN	ago	7.342	97	95

Figura 16 - Collocati per *ago* risultanti dal confronto con una *wordlist* di riferimento

Creare una wordlist con il programma Wordlist

Il programma Wordlist produce liste di parole ordinate alfabeticamente e per frequenza (Figura 17), oltre a tavole statistiche riassuntive (Figura 18); la lista può essere prodotta per un solo file o per più file, in quest'ultimo caso si può decidere se optare per una lista unica o per più liste.

N	Word	Freq.	%	Texts	%	emmas
1,928	WHAT	56	0.51	7	87.50	
1,929	WHEN	17	0.15	6	75.00	
1,930	WHENCE	1		1	12.50	
1,931	WHERE	16	0.14	5	62.50	
1,932	WHEREFORE	3	0.03	2	25.00	
1,933	WHEREIN	1		1	12.50	
1,934	WHEREIN	1		1	12.50	

frequency alphabetical statistics filenames notes

2,021 Type-in WHEREFORE

Figura 17 – Lista di parole

N	0	1	2	3
text file	overall	vrom01.txt	vrom02.txt	vrom03.txt
file size	56,316	4,310	12,786	5,329
tokens (running words) in text	11,073	689	2,532	1,062
tokens used for word list	10,088	668	2,295	957
types (distinct words)	2,021	355	687	397
type/token ratio (TTR)	20.03	53.14	29.93	41.48
standardised TTR	37.41	*	35.80	*
standardised TTR std.dev.	53.90	*	45.40	*
standardised TTR basis	1,000	1,000	1,000	1,000
mean word length (in characters)	4.14	5.00	4.15	4.15
word length std.dev.	1.97	2.59	1.95	1.89
sentences	171	13	39	13
mean (in words)	58.99	51.38	58.85	73.62
std.dev.	73.01	53.71	86.88	89.40
paragraphs	133	3	31	9
mean (in words)	75.85	222.67	74.03	106.33
std.dev.	120.58	281.95	110.33	99.61
headings				
mean (in words)	*	*	*	*
std.dev.	*	*	*	*
sections	8	1	1	1
mean (in words)	1,261.00	668.00	2,295.00	957.00
std.dev.	754.96	*	*	*

frequency alphabetical statistics filenames notes

2,021 Type-in WHEREFORE

Figura 18 - Statistiche

Il comando “make a wordlist” indicizza il testo alfabeticamente e per frequenza e propone un quadro riassuntivo di informazioni statistiche riguardanti le caratteristiche linguistiche del file indicizzato ((*standardised*) *type/token ratio*, lunghezza media delle parole, delle frasi, dei paragrafi, ecc.).

La *type/token ratio* è una misura della varietà lessicale di un testo; poiché è fortemente dipendente dalle dimensioni del corpus, viene offerta anche in versione “standardizzata” (la misurazione finale è ottenuta dividendo i testi in parti uguali di 1000 parole, misurando la *type/token ratio* per ciascuno e poi facendo la media - l’utente può definire le dimensioni delle sezioni in cui viene suddiviso il testo). WordSmith calcola la *type/token ratio* moltiplicando i *types* per cento e dividendo il risultato per il numero di *tokens*.

Due *wordlists* possono essere paragonate fra di loro, o una *wordlist* può essere paragonata con un corpus di riferimento per evidenziare i termini che contraddistinguono il testo sotto esame rispetto ad un altro testo più lungo o ad un insieme di testi utilizzati come *benchmark*.

Una *wordlist* può essere fatta sia per singole parole, sia per sequenze di parole (di solito 2-4), cambiando le impostazioni del programma (“settings wordlist”, attivando “clusters” e scegliendo il numero di parole che i *clusters* devono contenere).

Creazione di una lista di parole significative con il programma Keywords

Per determinare quali sono le parole più significative di un testo occorre prima avere a disposizione (salvate su pc) due *wordlists*: una, ad esempio, del testo che si sta analizzando (il nostro corpus “specialistico”), l’altra di un testo che si ritiene abbia un qualche rapporto di paragonabilità con il testo in questione, ma se ne distingua per una o più caratteristiche lessicali che si desidera mettere in evidenza (il corpus “di riferimento”) (Figura 19).

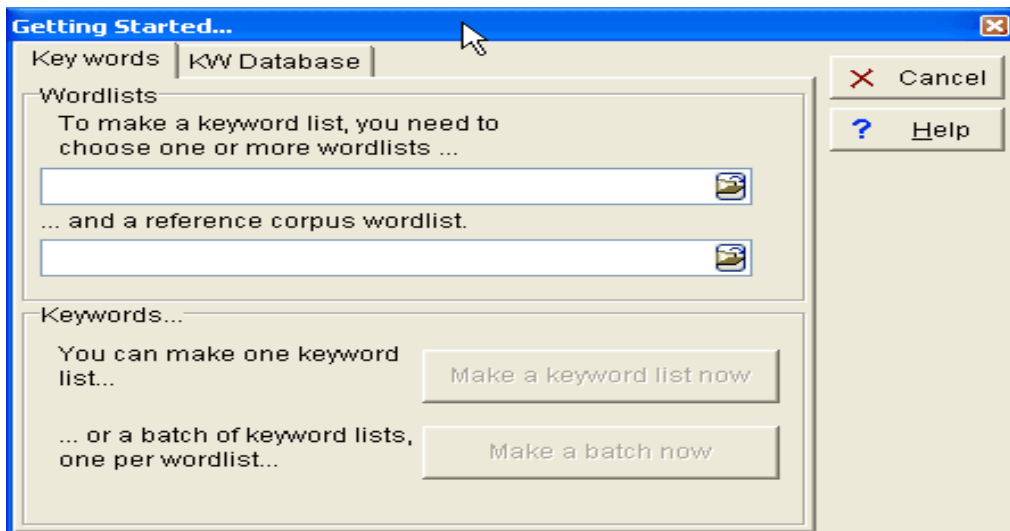


Figura 19 – Creare una *keyword*

Nella finestra di dialogo di Keywords si scelgono le due *wordlists* desiderate, quindi il programma procede alla comparazione (Figura 20).

N	Key word	Freq.	%	. Freq.	RC. %	eyr
1	ROMEO	130	1.17	311		.83
2	BENVOLIO	49	0.44	4		86
3	THOU	74	0.67	753		84
4	JULIET	74	0.67	1,126		79
5	CAPULET	33	0.30	14		54
6	THEE	44	0.40	630		47
7	MERCUTIO	26	0.23	22		40
8	THY	37	0.33	632		38
9	MONTAGUE	28	0.25	134		36
10	LOVE	72	0.65	22,224	0.02	34
11	TIS	29	0.26	423		31
12	CAPULET'S	17	0.15	0		30
13	TYBALT	17	0.15	4		28
14	SAMSON	22	0.20	158		26
15	NURSE	36	0.33	3,175		26

Figura 20 – Esempio di Keywords

La lista di parole ottenuta può essere utilizzata per accedere alle concordanze relative a ogni singola parola e al testo intero. Si possono paragonare anche *wordlists* di *clusters* (ad esempio per digrammi).

È possibile modificare i modelli statistici utilizzati da WordSmith per il calcolo delle *keywords*. Questi modelli possono essere visualizzati selezionando il comando “Settings” -> “customise” e poi scegliere “keywords” dal menu del programma Keywords di WordSmith

Tools. Le *keywords* vengono ordinate secondo un indice di “*keyness*” (anche noto come “Importance” o “Aboutness”: non è importante il loro valore assoluto, quanto piuttosto l’ordinamento che si realizza (Figura 21).

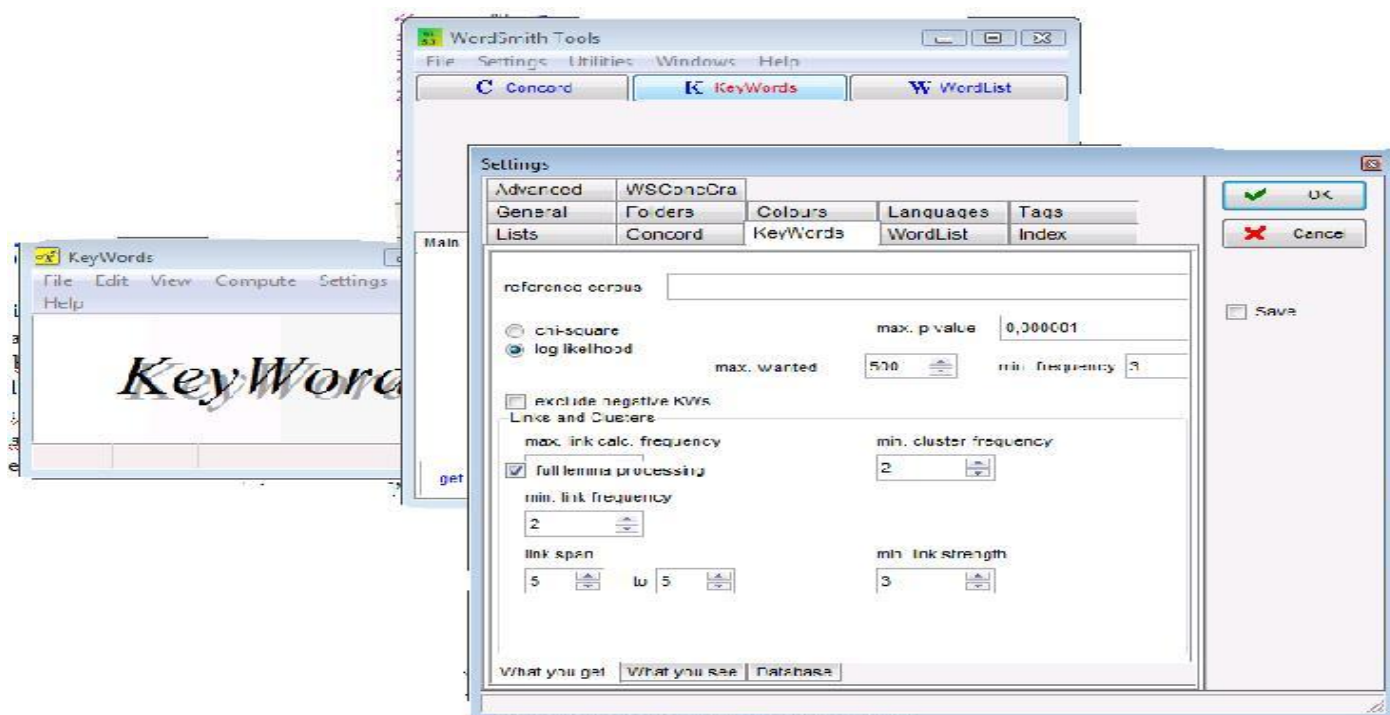


Figura 21 – Impostazioni per le *keywords*

I programmi di utilità Viewer e Aligner

Questo strumento è utile soprattutto per allineare due testi paralleli (per approfondimenti sul concetto di allineamento vedere la relativa voce *alignment* del glossario). Il corpus allineato può poi essere salvato come testo e utilizzato nelle concordanze per evidenziare le scelte operate e/o le diversità fra i testi (ad esempio un corpus allineato di traduzioni dello stesso brano).

Occorre notare però che WordSmith Tools non ha un *parallel concordancer* vero e proprio, quindi cerca la parola specificata come in un corpus monolingue, ma grazie

all'allineamento di *Aligner*, il termine equivalente nell'altra lingua apparirà nelle vicinanze, e sarà quindi di facile reperimento.

Per unire due parti che sono state separate nell'allineamento, cliccare sulla parte da spostare e portarla (“drag and drop”) alla fine della frase a cui appartiene. Per separare due parti che sono state erroneamente unite cliccare su “un-join”, selezionare la parola alla fine della quale si vuole operare la cesura; la parte finale sarà automaticamente spostata nello slot successivo. Il testo salvato in formato *.vwr* può essere riaperto (*aligner* -> *open saved dual text*) e modificato successivamente. Per essere aperto con altri programmi deve però essere ritrasformato in *.txt*.

Salvataggio, stampa e chiusura del programma

Infine, è possibile salvare e stampare i risultati delle ricerche. Per uscire dal programma basta scegliere “Exit” dal menu “File” oppure digitare la sequenza *alt-x*.

THE SKETCH ENGINE

OS: Non richiede installazione

Licenza individuale: 55.25 € annuali (Disponibile Demo)

Breve presentazione

Sketch Engine è uno strumento di analisi di corpora che prende come input un corpus in una qualsiasi lingua e dei *patterns* grammaticali corrispondenti (ovviamente dipendenti e strettamente legati alla lingua scelta) ed è in grado di generare concordanze, *word sketches* per le parole della lingua in esame ma anche *similar words* e *sketch differences*.

Gli *word sketches* sono pagine, generate automaticamente, basate sul corpus caricato che riassumono il comportamento collocazionale e grammaticale di una parola.

Furono usati per la prima volta nel Macmillan English Dictionary e presentati all'Euralex 2002.

Inizialmente gli *word sketches* erano disponibili solo per l'inglese. Successivamente con lo sviluppo di Sketch Engine si è potuto creare uno strumento di analisi di corpora che prende come input un corpus in una qualsiasi lingua e dei *patterns* grammaticali ed è, quindi, in grado di generare *word sketches* per le parole nella lingua scelta.

Sketch Engine genera anche un Thesaurus e le *sketch differences* che specificano somiglianze e differenze tra quasi-sinonimi (ovvero specifica per due parole legate semanticamente quale "comportamento" condividono e in cosa, invece, differiscono), particolarmente utili per i lessicografi interessati alle differenziazioni tra quasi-sinonimi.

URL: <http://www.sketchengine.co.uk/>

Linee guida

Avvio del programma

Sketch Engine è disponibile soltanto sul sito web <http://www.sketchengine.co.uk/> (Figura 1).

LEXCOM
Lexical
Computing

Sketch Engine

user: Valentina E

Preloaded Corpora

Language	Name	Tokens [?]	
Chinese	Chinese GW, simpl	706 427 624	info
Chinese	Chinese GW, trd	706 428 333	info
English	British National Corpus	111 244 375	info
English	ukWaC	2 035 621 120	info
French	French web corpus	126 850 281	info
German	deWaC	1 644 785 836	info
Italian	itWaC	1 909 535 984	info
Japanese	JpWaC	409 384 405	info
Russian	Russian Web Corpus	187 965 822	info
Spanish	Spanish web corpus	116 900 060	info

[more corpora >>](#)

User Corpora

- [Corpus Builder](#) - create corpora from your own texts
- [WebBootCaT](#) - create domain specific corpora from the web

Setup

- [Change password](#)
- [Logout \(experimental\)](#)

Trac

- [Sketch Engine documentation](#)
- [Create new ticket](#) - report bugs, request support or new features

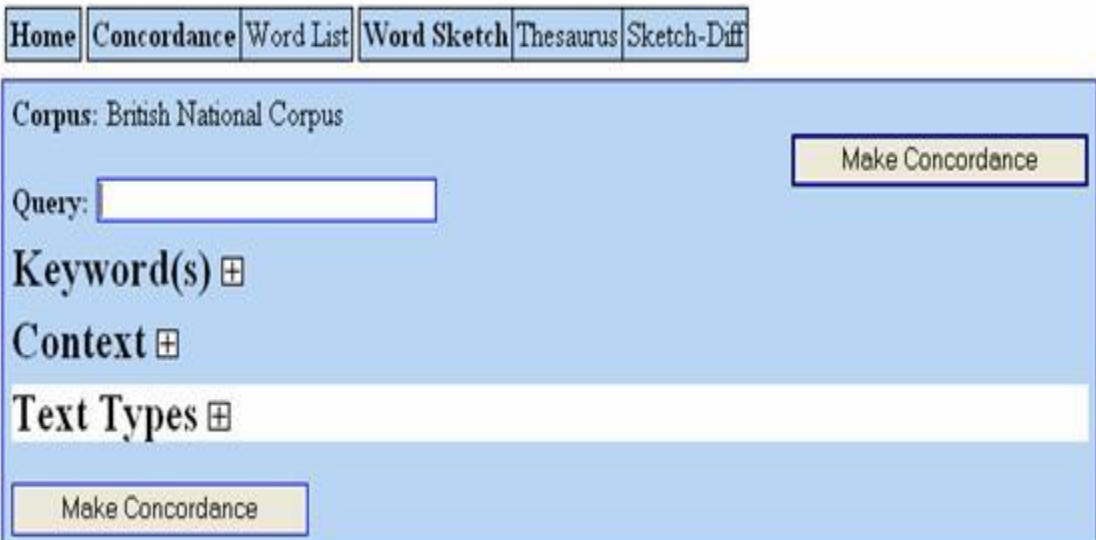
Figura 1 – Pagina principale di Sketch Engine

Da questa pagina è possibile selezionare il corpus da esaminare. Sono presenti corpora in diverse lingue; ad esempio per l'inglese vi sono il *British Academic Spoken English Corpus* (BASE), il *British Academic Written English Corpus* (BAWE), il *British National Corpus* ed il *ukWaC*,

per il francese il *French Web Corpus*, per l'italiano l' *itWac* ecc. Gli esempi presenti nella guida sono tratti dal *British National Corpus (BNC)*.

Ricerca di concordanze

I sei tasti lungo la parte superiore della pagina (Figura 2) permettono di navigare attraverso i vari strumenti messi a disposizione dal programma, o di tornare all'inizio (Home). Per effettuare una *query* bisogna inserire il termine da ricercare nella casella di testo accanto al campo "Query" (*query box*) e cliccare sul pulsante "Make Concordance".



The screenshot shows the Sketch Engine interface for searching concordances. At the top, there is a navigation bar with six buttons: Home, Concordance, Word List, Word Sketch, Thesaurus, and Sketch-Diff. Below this, the main area has a light blue background. It starts with "Corpus: British National Corpus" on the left and a "Make Concordance" button on the right. Below that is a "Query:" label followed by an empty text input box. Underneath are three expandable sections: "Keyword(s) ⊕", "Context ⊕", and "Text Types ⊕". At the bottom left of the main area is another "Make Concordance" button.

Figura 2 – Ricercare concordanze con Sketch Engine

Se, come il BNC, il corpus scelto è lemmatizzato, i termini saranno confrontati sia con il lemma che con la parola. Per fare delle ricerche più specifiche, cliccare sul "+" al lato di Keywords, di Context o di Text Types, dove sono disponibili più opzioni di ricerca.

Un esempio di risultato di ricerca di concordanze è mostrato in Figura 3:

Home	Concordance	Word List	Word Sketch	Thesaurus	Sketch-Diff	Corpus: British National Corpus Hits: 1098 conc description
View options	Sample	Filter	Sort	Frequency	Collocation	

Page 1 of 55 [Go](#) [Next](#) | [Last](#)

A05 their magical meaning , and the tramps who **haunt** them -- comes from the striking poem Lud
A08 transcendental . As if it had ever been there . All **haunted** by the sense of a sacred space , a sacred
A08 art and culture . But the metaphor goes on **haunting** us , he wrote . Merely to deny it , even
A08 of Kafka 's , which have never ceased to **haunt** me : Where does the strength come from
A0F mumbled to myself . Memories started to **haunt** me again . Memories of my childhood , memori
A0G thinnings , for all these are favourite **haunts** of many harmful insects . *</p>Make life*
A0K . This fear of the mob has continued to **haunt** the executive , who saw that control could
A0K of self can help link the tribalism which **haunts** the police defensiveness to an understanding
A0P sensitivity and commitment , which was to **haunt** Leonard throughout his life . Names , like
A0P something else . The responsibility was going to **haunt** him for years to come . The die had been
A0P the folk-singers ' is to himself .) It **haunted** , it tyrannised their lives , not least
A11 style with a series of runs over its old **haunts** from Peterborough to York . A pity so many
A11 from their traditional Midland main line **haunts** by the mid-1980s , and the arrival of Class
A15 . *</p><p>A ' Brownie ' or friendly spirit **haunts** the island , and his consent to land is*
A18 Dostoevsky novel , with no children in it but **haunted** by the toys of absent innocence and peace

Figura 3 – Risultato di una ricerca di concordanze

È possibile espandere il contesto di ogni riga di concordanza, cliccando sul nodo di ricerca della riga stessa. Il contesto apparirà in un pannello in basso dove è anche possibile espandere ulteriormente il contesto sia a destra che a sinistra, rispettivamente *expand left* e *expand right* (Figura 4).

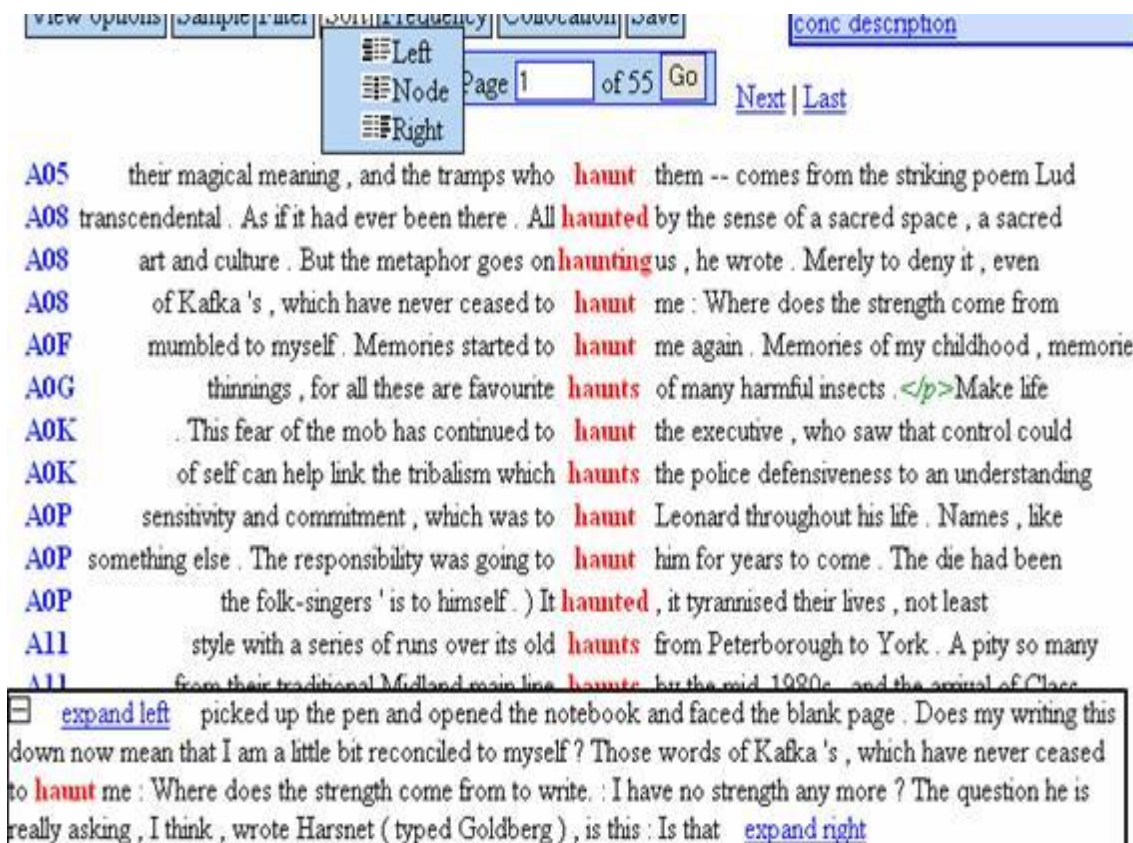


Figura 4 – Espansione del contesto

I bottoni che si vedono nella seconda riga in alto in Figura 3 permettono di lavorare sulle concordanze. Sono presenti le seguenti opzioni:

View options: permette di alternare fra la visualizzazione standard *KWIC* della concordanza (che compare per default) e la visualizzazione completa di frase e permette anche di visualizzare nuove schermate per cambiare la vista della concordanza in vari modi.

Riassumendo alcune funzioni sono:

- la colonna “Attribute”, che permette di cambiare la visualizzazione di default (nella quale soltanto il testo è visibile nella riga della concordanza) con un certo numero di visualizzazioni alternative in cui è possibile vedere i *tags* che identificano le “parti del discorso” (*Parts of Speech, POS*), le forme lemmatizzate ed altri campi di informazioni. La funzione è raramente necessaria in lessicografia, ma può essere utile per scoprire perché una riga inattesa del corpus

è stata confrontata con la *query*, poiché a volte la causa è una *lemmatizzazione* o un'assegnazione di *parti del discorso* errata;

- la colonna “Structure”, che permette di cambiare la visualizzazione di default per mostrare i *tags* di inizio e fine di strutture quali le frasi, i paragrafi ed i documenti. Di nuovo, anche questo è improbabile sia necessario in lessicografia tradizionale;
- la colonna “References”, che serve ad indicare il tipo di informazioni per quanto riguarda i testi originali che compaiono (in azzurro) a sinistra della riga della concordanza. Per default vi è un identificatore del documento da cui è presa la riga di concordanza a cui fa riferimento. Altri campi di informazione sui documenti del corpus possono essere selezionati e per quelli sarà visualizzato il valore che la riga della concordanza ha per quel campo;
- il box “page size” permette di specificare una dimensione di pagina più lunga per la visualizzazione: per default, ogni pagina delle concordanze contiene 20 righe, ma è possibile aumentare fino a 500 righe (questo però rallenterà un po’ il recupero della concordanza).

Sample: utile se si sta cercando un item molto frequente. Permette di generare un campione scelto a caso dalle righe del corpus. Cioè se ad esempio cercate *play* (verbo) e decidete che non desiderate analizzare 37.632 occorrenze, utilizzate il tasto “Sample” per ridurle ad un numero trattabile.

Sort: permette di effettuare un ordinamento di diverso tipo. L’ordinamento è importante perché rappresenta un modo veloce di scoprire dei *patterns*.

Frequency: permette di osservare due tipi di informazioni di frequenza per quanto riguarda il termine di ricerca:

- *Multilevel frequency distribution* (distribuzione di frequenza multilivello): mostra la frequenza di ogni forma di un dato lemma. Per vedere come questo funziona, immaginiamo di

effettuare una concordanza per il verbo *forge*: quando è pronta la concordanza andare su “frequency” all'opzione *Multilevel frequency distribution*. Il primo livello mostra le frequenze delle forme *forge, forged, forging* e *forges*. Il secondo e terzo livello permettono delle ricerche più complesse: per esempio se selezionate “secondo livello” e “1R” (posizione di una parola alla destra della parola di ricerca) vedrete quali parole compaiono in questa posizione e quanto frequente è ciascuna di queste parole.

- *Text type frequency distribution*: mostra come il termine di ricerca è distribuito nei testi all'interno del corpus. Potete trovare, per esempio, che una parola come *police* compare sensibilmente più spesso nei testi giornalistici che in altri tipi di testo. Questo è uno strumento potenzialmente utile che potrebbe mostrare - per esempio - che un termine medico particolare non si limita solo a testi medico-specialistici.

Collocation: permette di compilare liste di parole che co-occorrono frequentemente con il termine di ricerca (i suoi collocati). Va sottolineato però che dove gli *word sketches* sono disponibili producono un resoconto più sofisticato dei collocati e vanno quindi preferiti.

Il box blu in alto a destra sempre in Figura 3 mostra quale corpus si sta utilizzando e quanti risultati sono stati trovati per la ricerca effettuata.

Generare keywords

Cliccando su “Keyword(s)” appaiono le seguenti opzioni, visibili in Figura 5:

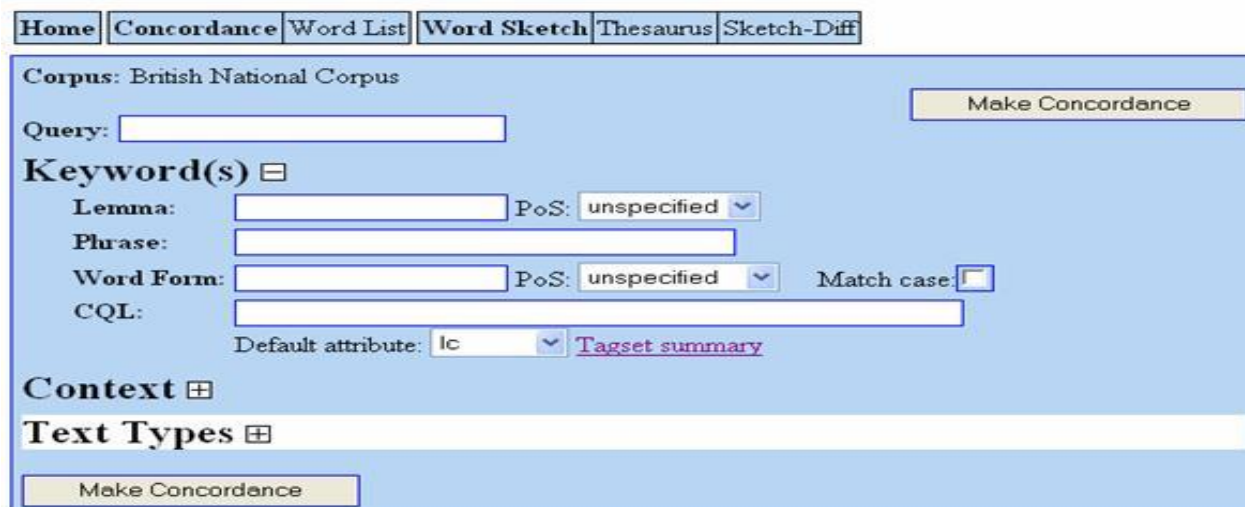


Figura 5 – Generare Keywords

Nel box “Lemma” è possibile inserire un lemma ed una particolare *parte del discorso* (sempre se si assume che il corpus sia lemmatizzato e annotato).

Nel box “Phrase” è possibile inserire qualsiasi espressione multi-parola.

Nel box “Word Form” è possibile ricercare una specifica forma di parola e specificare di quale *parte del discorso* debba far parte (nel caso di omografi).

Nel box “CQL” è possibile eseguire delle interrogazioni complesse utilizzando il *Corpus Query Language* descritto nel *Corpus Querying and Grammar Writing* presente nella homepage di Sketch Engine.

Il link *tagset summary* fornisce dettagli sui *tags* delle *parti del discorso* usati per annotare il corpus e quindi necessari per effettuare delle *queries* sulle parti del discorso.

Usare la sezione Context

Cliccando sul “+” accanto al box Context appaiono le opzioni visibili in Figura 6.

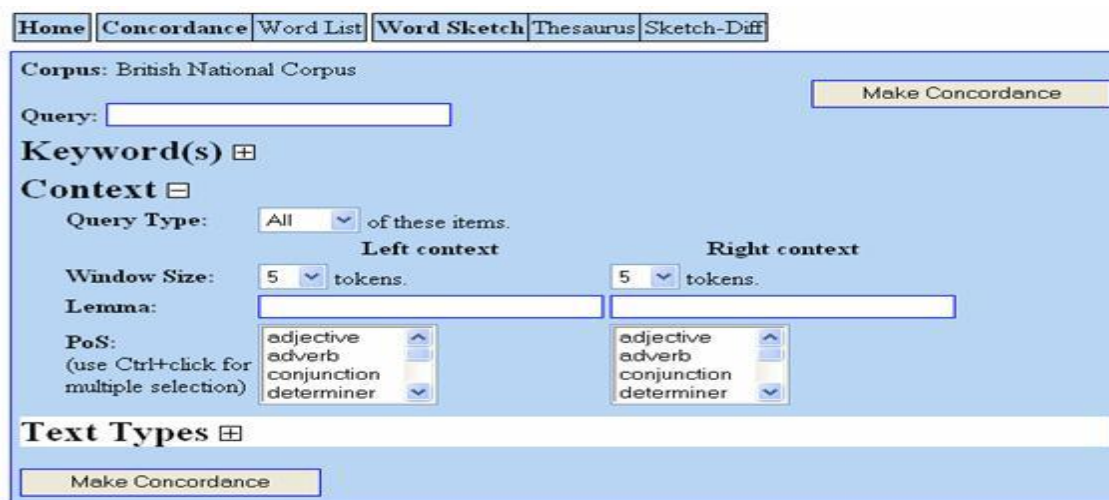


Figura 6 – Effettuare una ricerca in un contesto ben definito.

È possibile specificare il contesto destro e/o sinistro (con relativa parte del discorso) della parola di ricerca.

Cliccando sul “+” accanto a “Text Types”, invece, è possibile limitare la ricerca ad una parte del corpus, per vedere ad esempio come una parola si comporta in un dato linguaggio settoriale o genere testuale.

La funzione Word Sketch

La funzione Word Sketch costruisce riassunti automatici del comportamento grammaticale di una parola in un corpus.

Per studiare il comportamento di una parola, piuttosto che osservarla in una finestra arbitraria di testo, la funzione Word Sketch permette di osservare le relazioni grammaticali in cui la parola si manifesta. Per l’inglese è stato usato un repertorio di 27 relazioni grammaticali tra le quali si trovano, ad esempio per i nomi: *object_of*, *subject_of*, *modifier*, *and/or*, *adj_subject_of*, ecc.

La funzione Word Sketch restituisce una lista di collocazioni per ogni relazione grammaticale in cui la parola si manifesta. È possibile per ogni collocazione vedere i contesti nei quali la parola radice e il suo collocato occorrono insieme.

Per consentire alla funzione Word Sketch di classificare i lemmi, essa deve sapere per ogni parola del testo qual è il lemma corrispondente. Sketch Engine NON supporta questo processo. Esistono strumenti a disposizione dei linguisti per sviluppare lemmatizzatori per alcune lingue.

Se non si dispone di un lemmatizzatore, anche se con risultati non ottimali, è possibile applicare la funzione Sketch Engine alle forme di parola.

Gli *Word Sketches* sono pienamente integrati con le concordanze (se ad esempio l'utente clicca sulla parola *toast* nella lista di oggetti con alta rilevanza nel Word Sketch per il verbo *spread*, verrà restituita una concordanza dei contesti dove *toast (n)* occorre come oggetto di *spread (v)*).

Per realizzare *Word Sketches* significativi, Sketch Engine presuppone che l'input sia annotato (taggato). L'obiettivo è la decisione della corretta classe di parole per ogni parola nel corpus (es: determinare se un'occorrenza di *toasts* è un'occorrenza di un nome plurale o un'occorrenza di un verbo alla terza persona singolare presente). Un sistema di annotazione (*tagger*) presuppone una analisi linguistica della lingua in oggetto che dà luogo ad un insieme di categorie sintattiche della lingua detto anche *tagset*.

Al fine di identificare le relazioni grammaticali tra le parole, Sketch Engine deve sapere come trovare le parole unite da una relazione grammaticale nella lingua in esame. Ci sono due possibilità:

- 1°. Il corpus in input è stato già analizzato sintatticamente (*parsed*). In questo caso l'informazione su *quali* istanze di parola stanno in *quali* relazioni grammaticali con *quali* altre istanze di parola è inclusa (integrata) nel corpus.
- 2°. Il corpus in input viene caricato in Sketch Engine senza essere stato prima analizzato sintatticamente. In questo caso si usa il

processo di identificazione di istanze di relazione grammaticale supportato da Sketch Engine.

Ora vediamo come usare la funzione Word Sketch: cliccando su “Word Sketch” si apre la “Word Sketch entry form” mostrata in Figura 7.

The screenshot shows a web interface with a navigation bar at the top containing buttons for Home, Concordance, Word List, Word Sketch, Thesaurus, and Sketch-Diff. Below this is the 'Word Sketch Entry Form' section. It contains several input fields and a button:

- Corpus: British National Corpus
- Lemma: [text input field]
- Part of speech: [dropdown menu showing 'noun']
- Sort grammatical relations:
- Minimum frequency: [text input field with value 5]
- Minimum salience: [text input field with value 0.0]
- Maximum number of items in a grammatical relation: [text input field with value 25]
- [button: Show Word Sketch]

Figura 7 – “Word Sketch entry form”

Bisogna inserire un lemma nel record “Lemma” e la sua *part of speech* (parte del discorso) scegliendola nel menu a tendina accanto al campo “Part of speech”. Sono disponibili *Word Sketches* per nomi, verbi e aggettivi ma non per altre classi di parole; inoltre la loro creazione dipende anche dalla presenza di una quantità sostanziale di dati: infatti se si tenta di creare un *Word Sketch* per un item molto raro si riceverà un messaggio dove vi si dice che non c’è alcuno *Word Sketch* disponibile. Di solito c’è bisogno di alcune centinaia di istanze di parola per ottenere un utile *Word Sketch*.

La Figura 8 mostra un esempio di una parte dello *Word Sketch* per la parola *challenge*:

challenge British National Corpus freq = 6243

[change options](#)

object of 1816 2.7	subject of 352 1.0	a modifier 2230 2.6	n modifier 974 1.3	modifies
pose 92 8.69	face 67 6.47	biggest 49 7.95	heparin 16 8.61	Anneka
relish 17 7.74	confront 8 6.01	greatest 53 7.93	Rolex-jackie 6 7.63	trophy
mount 42 7.6	lay 7 3.3	serious 72 7.35	commuter 10 7.5	cup
face 155 7.59	come 25 2.59	intellectual 26 7.33	celebrity 10 7.35	bid
present 131 7.23	involve 5 2.01	formidable 16 7.3	steel 30 7.25	final
meet 199 7.21	start 5 1.83	daunting 10 6.94	leadership 26 6.45	server
resist 20 6.69	begin 6 1.78	exciting 21 6.94	city 118 6.45	initiative
withstand 7 6.43	become 7 1.32	larval 9 6.93	promotion 17 6.3	match
tackle 14 6.29	go 7 0.29	direct 46 6.75	Merseyside 5 6.08	tour
constitute 17 6.08		legal 50 6.56	title 40 6.06	cash
evade 5 6.01	adj subject of 84 1.1	blind 12 6.41	silk 8 5.89	series
accept 58 6.0	likely 9 3.13	toughest 6 6.31	milk 14 5.84	project
overcome 11 5.77		major 69 6.25	Stewart 8 5.79	scheme
counter 6 5.76		ultimate 11 6.24	dollar 7 5.48	area
eniov 35 5.69		fundamental 16 6.19	wine 10 4.73	programme

Figura 8 – Parte del *word sketch* per la parola *challenge*

Ogni colonna mostra le parole che tipicamente si uniscono con *challenge* in particolari relazioni grammaticali (*gramrels*). La maggior parte di queste relazioni sono evidenti. Per esempio la relazione *object_of* elenca i verbi che prendono tipicamente *challenge* come oggetto, presentati nell'ordine di importanza statistica piuttosto che di frequenza grezza. È possibile in qualunque momento commutare il “Word Sketch mode” e il “Concordance mode” (ovvero passare dalla visualizzazione delle relazioni grammaticali alla visualizzazione delle concordanze per una parola scelta) e questo è sicuramente un modo utile di ottenere più informazioni su una combinazione particolare di parole. Quindi, se ad esempio si desidera osservare il contesto in cui occorre la stringa “pose + *challenge*”, basta cliccare semplicemente sopra il numero vicino a *pose* nella lista *object_of* (92) e sarà direttamente visibile lo schema di concordanza che mostra tutte le occorrenze di questa combinazione.

La funzione Thesaurus

La funzione Thesaurus elenca, per ogni aggettivo, nome o verbo, le altre parole più simili ad esso nel loro uso nella lingua. Si accede a questa funzione cliccando sopra il tasto “Thesaurus” dalla homepage, o nella barra superiore presente in ogni pagina ed inserendo la parola di interesse.

Poter disporre di un insieme molto grande di istanze di relazioni grammaticali ci permette di avere una ricca rappresentazione del lessico di una lingua; in tal modo è possibile andare al di là dell’osservazione del comportamento delle parole, un lemma alla volta, e usare queste istanze di relazioni per mostrare dei *patterns* attraverso gruppi di parole (se ad esempio troviamo la coppia di istanze di relazioni grammaticali <object, drink, beer>, <object, drink, wine>, possiamo usarla come prova del fatto che *beer* e *wine* siano nella stessa categoria del Thesaurus).

Sketch Engine costruisce un Thesaurus sotto forma di un insieme dei “vicini più vicini” per ogni parola usando la matematica per calcolare le somiglianze.

Un Thesaurus sviluppato in questo modo tratto dal BNC è consultabile su <http://wasps.itri.bton.ac.uk>.

Spesso osservando una voce del Thesaurus ci si chiede: che cosa rende quelle parole così simili? Oppure, in cosa differiscono? La somiglianza è basata sulla condivisione di triple (*shared triples*) (ad esempio *wine* e *beer* “condividono” la tripla <obj, drink, ?>). Allora ciò che due parole hanno in comune sono le *shared triples* che hanno un’alta rilevanza per entrambe le parole. Anche la differenza tra due quasi-sinonimi (*near-synonyms*) può essere identificata come una tripla che ha un’alta rilevanza per una parola ma nessuna occorrenza (o bassa rilevanza) per l’altra.

In tal modo, come è possibile produrre “Word Sketch” è possibile anche produrre “Sketch Difference” come illustrato nel prossimo paragrafo.

La funzione Sketch Difference

La funzione *Sketch Difference* è un modo molto accurato di confrontare due parole simili: essa mostra i *patterns* e le combinazioni che i due item hanno in comune ed anche quei *patterns* e quelle combinazioni che sono più tipiche di una parola piuttosto che di un'altra. Cliccando su una qualsiasi parola in un'entrata del Thesaurus elaborato apparirà una schermata che evidenzia lo *Sketch Difference* fra le due parole. Alternativamente è possibile cliccare sopra il bottone "Sketch Difference" sulla homepage o sulla parte superiore di qualsiasi pagina e ciò porterà alla schermata principale dello "Sketch Difference". Se, ad esempio, si fosse interessati a confrontare la parola *intelligent* con la parola *clever* si otterrebbe lo "Sketch Difference" di Figura 9.



Figura 9: "Sketch Difference" per le parole *clever* e *intelligent*

Lo schema è diviso in tre parti principali: la prima parte mostra i *common patterns* (cioè quelle combinazioni dove *clever* ed *intelligent* si comportano in maniera molto simile), la seconda e la terza parte mostrano i *patterns* solo per *clever* ed i *patterns* solo per *intelligent*.

Glossario

Alignment (Allineamento): Un file di allineamento è un file di lavoro che serve per elaborare una coppia di file che siano l'uno la traduzione dell'altro. Lo scopo del file di allineamento è quello di isolare nei due testi tutti i segmenti e accoppiare ciascun segmento alla sua traduzione esatta così da creare delle unità di traduzione.

Collocations (Collocazioni): Combinazioni di parole caratterizzate da una particolare frequenza d'uso, ossia dalla preferenza per l'occorrenza congiunta dei suoi componenti.

Concordance (Concordanze): Lista delle occorrenze di una *keyword* (parola chiave) nel testo, ciascuna presentata nel suo contesto linguistico. Esse permettono di esplorare l'uso di una parola nei singoli "habitat" linguistici in cui occorre.

Co-text (Cotesto): L'insieme degli elementi linguistici (generalmente la quantità standard è definita utilizzando un criterio basato sul numero dei caratteri e delle parole ortografiche) che, in ogni singolo esempio, co-occorrono con la parola chiave oggetto di studio.

Frequency list (Lista di frequenze di parola): Lista di tutte le parole nel corpus a cui è associato un peso in base alla loro frequenza.

Keyword (Parole chiave): Termini di ricerca che gli studiosi utilizzano per fare le proprie indagini nel corpus.

KWIC (Formato KeyWord In Context): È il formato standard in cui vengono presentate le concordanze di una forma lessicale specifica detta parola chiave o *keyword*. Le concordanze in formato *kwic* contengono tante righe quante sono le occorrenze della parola chiave nel testo. Ciascuna riga è centrata sulla parola chiave, che è preceduta e seguita da un numero prefissato di caratteri di contesto. I numeri all'inizio di ogni riga identificano la riga di testo in cui si trova una data occorrenza della parola.

Lemmatization (Lemmatizzazione): Consiste nel ricondurre ogni parola del testo al relativo esponente lessicale o *lemma*.

Pattern (Modello o schema): È un modello di stringa che specifica i criteri che devono soddisfare le stringhe da individuare nel testo.

Part of speech, POS (Parte del discorso): Categoria lessicale o sintattica (ed i tratti relativi) di una data parola.

Query (Interrogazione): Interrogazione del sistema automatico utilizzato per analizzare il testo in esame per effettuare ricerche nel corpus.

Regular expression, regex, ER (Espressioni regolari): Strumento utile per le esplorazioni avanzate dei dati testuali. Le ER permettono di esprimere criteri di ricerca complessi e articolati velocizzando il processo di estrazione ma anche migliorando la qualità dei dati estratti sia in termini di accuratezza che di generalità. Le ER sono una notazione algebrica che permette di definire in maniera formale e rigorosa degli schemi (*patterns*) di stringhe. In quanto definisce uno schema di stringa, una ER rappresenta l'insieme delle stringhe che corrispondono a tale schema. Il processo di identificazione di eventuali stringhe che soddisfino uno schema viene detto *pattern matching*. Quando un programma verifica se in un file di testo esistono stringhe di caratteri conformi allo schema specificato da una ER, il file viene generalmente analizzato riga per riga: se una riga contiene una stringa che corrisponde al *pattern*, il programma restituisce una risposta altrimenti passa alla riga successiva, fino a quando non arriva alla fine del file.

- **Sintassi delle espressioni regolari**

- La ER più elementare è quella costituita da un solo carattere (*/a/* - specifica il *pattern* che è soddisfatto dal carattere *a*);
- Le ER sono *case-sensitive* (*/a/* è soddisfatta solo da *a* non da *A*);

- Una *ER* particolare è `//`: questa corrisponde alla stringa vuota, ovvero la stringa composta da zero caratteri;
- Una *classe di caratteri* –costituita da un insieme di caratteri racchiusi tra parentesi quadre– è una *ER* che viene soddisfatta da uno qualsiasi dei caratteri indicati all'interno delle parentesi (`/[la]/` - il carattere *l* o il carattere *a*);
- Il *trattino* (`-`) specifica un intervallo di caratteri (`/[a-z]/` - una qualsiasi lettera minuscola dell'alfabeto);
- Una classe di caratteri in cui la parentesi di apertura sia immediatamente seguita dal carattere `^` è soddisfatta da qualsiasi carattere diverso da quelli specificati nella classe (`/[^A-Z]/` - qualsiasi carattere che non sia una lettera maiuscola). Se `^` non compare in prima posizione all'interno della classe non viene, invece, interpretato come negazione;
- Il carattere *backslash* (`\`) è utilizzato per segnalare che il carattere che segue ha un'interpretazione speciale, diversa da quella usuale. Ad esempio alcuni caratteri speciali rappresentano caratteri di controllo, come il ritorno a capo (`\n`), la tabulazione (`\t`), il *line feed* (`\r`);
- Un carattere speciale è il carattere *jolly* (`.`) che corrisponde a qualsiasi carattere escluso il ritorno a capo;
- La disgiunzione tra stringhe di caratteri viene espressa tramite l'operatore di *alternativa* (`|`) (`/il|la/` - *il* o *la*);
- I *moltiplicatori* sono operatori che permettono di specificare quante volte il carattere che li precede deve comparire nel *pattern*. I moltiplicatori più comuni sono i seguenti:
 - `?` : zero o una occorrenza del carattere precedente
 - `*` (*Kleen star*) : zero o più occorrenze del carattere precedente
 - `+` : una o più occorrenze del carattere precedente
- Le *ancore* sono caratteri speciali che indicano la posizione precisa nella riga di testo in cui deve comparire la stringa che soddisfa la *ER*. Le quattro *ancore* principali sono:

- **^**: inizio della riga di testo
- **\$**: fine della riga di testo
- **\b**: confine di token
- **\B**: qualsiasi punto che non sia un confine di token

Tag (Etichetta di marcatura): Sequenza di caratteri visibili in un testo secondo una convenzione standard e intercalata nel testo secondo precise regole di combinazione. Tale etichetta serve a rappresentare l'informazione strutturale del testo al quale è applicata.

Tagger (Annotatore): Disambiguatore morfo-sintattico, strumento automatico che consente di annotare porzioni estese di testo.

Tagging (Annotazione): L'annotazione linguistica di un testo consiste nella codifica di informazione linguistica associata al dato testuale. Essa permette di rendere esplicita, interpretabile ed esplorabile dal computer la struttura linguistica implicita nel testo. Tipicamente, l'annotazione avviene in relazione ai tradizionali livelli di descrizione linguistica: morfologia, sintassi, semantica, pragmatica.

Tokenizing (Tokenizzazione): Passo preliminare di qualsiasi elaborazione computazionale del testo. **Tokenizzare** un testo significa dividere le sequenze di caratteri in unità minime di analisi dette *token* (parole, punteggiatura, date, numeri, sigle, ecc.). I *token* possono essere anche unità strutturalmente complesse ma sono comunque assunte come unità di base per i successivi livelli di elaborazione (morfologico, sintattico, ecc.).

Wordlist (Lista di parole): Lista di parole ampiamente ricorrenti.

Bibliografia

Monoconc

A brief introduction to the basic use of Monoconc Pro: <http://www.athel.com/tour.html>

Wordsmith Tools

Online manual in English, Oxford WordSmith Tools 5.0:
<http://www.lexically.net/downloads/version5/html/index.html>.

Step by step guide to Wordsmith:
http://www.lexically.net/wordsmith/step_by_step/index.html

Sketch Engine

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004), *The Sketch Engine*, In Proceedings of the Eleventh EURALEX International Congress, 105-116.

Sketch Engine documentation: <http://trac.sketchengine.co.uk/wiki/SkE/DocsIndex>