

Brown Corpus: caratteristiche dell'American English e sua variazione diacronica tra gli anni 1961-1991

Descrizione dei corpora

L'analisi qui presentata del **Brown Corpus** (1961) è stata compiuta mediante un confronto diacronico con il **Frown Corpus** (1991), tenendo in considerazione che entrambi appartengono alla cosiddetta "Brown Family of corpora", ovvero sono compilati seguendo i medesimi criteri, risultando di conseguenza comparabili.

Il Brown Corpus è un corpus di inglese americano scritto, pubblicato presso la Brown University nel 1964, composto da 500 testi di circa 2000 parole ciascuno per un totale di circa 1.000.000 di parole. I testi di cui si avvale sono libri e periodici pubblicati negli USA nel 1961, suddivisi in 15 categorie: press reportage, press editorial, press reviews, religion, skill and hobbies, popular lore, belles-lettres (biografie, memorie, etc), miscellaneous (testi governativi e istituzionali), learned (testi accademici), fiction, mystery and detective fiction, science fiction, adventure and western, romance and love story, humour.

Il Frown Corpus (The Freiburg-Brown corpus of American English) rispetta gli stessi criteri, ma si avvale di testi pubblicati nel 1991-92.

Lavoro preliminare

Per rientrare nel limite di 1 milione di parole su SketchEngine senza dover effettuare una selezione, dunque un ulteriore "sampling" del corpus abbiamo riunito i file .txt del Frown Corpus in un unico file .xml, successivamente pulito attraverso le seguenti operazioni di sostituzione. Abbiamo eliminato i tag `<0_>poem</0_>`, `<0_>table</0_>` et similia e sostituito `<0_>picture</0_>` e Fig con il tag `<figure/>`. I tag del tipo `<*_>a-grave</*_>` o `<*_>gamma</*_>` hanno ricevuto un trattamento differenziato, a seconda del contesto e della parte di testo contenuta nel tag: nei casi simili al primo esempio si è cercato di conservare il senso della parola sostituendo il testo riportato sopra in corsivo con la lettera più vicina a quella attesa (es. "i" per "naive", "a" per "chateau"), con l'unica eccezione dei nomi propri, per i quali è stata semplicemente eliminata la lettera, non costituendo essi un focus per il nostro studio né un rischio di falsa omografia. Nei casi simili al secondo esempio riportato, dunque con parole riguardanti simboli matematici, si è preferito sostituire la parte di testo interessata con il tag `<symbol/>`.

Questa operazione ci ha permesso di ridurre considerevolmente il numero di parole, preso conto del fatto che nel calcolo delle words Sketch Engine non computa i tag.

In più ogni file del Frown Corpus è stato taggato con l'attributo "register" corrispondente (press reportage, press editorial, fiction general, etc.), al fine di facilitare la lettura dei dati sul software.

Differenze lessicali

Si è deciso di indagare intanto le eventuali **differenze lessicali** tra i due corpora.

Utilizzando lo strumento *Keywords*, si è proceduto manipolando i criteri fino a stabilire i seguenti: *focus on*:10 — *minimum frequency*: 100 — *attribute*: lemma.

In questo modo si sono escluse parole rare nel reference corpus (Frown) quali terminologie scientifico-matematiche, a nostro parere meno significative per uno studio diacronico, e si è concentrata l'attenzione sulle parole frequenti nel linguaggio comune, per osservare le differenze di frequenza e utilizzo (attraverso lo strumento *Concordance*, come si vedrà più avanti). Si specifica che in ogni caso lo strumento fornisce come output le parole tipiche del focus corpus, ciò su cui i parametri selezionati vanno ad incidere è la sola visualizzazione delle stesse e il loro ranking. Sketch Engine utilizza infatti il *simple math method* che per il calcolo dello score di “keyness” (quanto una parola è “chiave” di un corpus) prende in considerazione il rapporto tra la somma delle frequenze relative nel focus corpus e una costante k e la somma delle frequenze relative nel reference corpus con la stessa costante. Il valore attribuito al parametro *focus on* corrisponde proprio a questa k .

Significativo osservare la presenza nel Brown Corpus di termini legati al contesto socio-culturale americano degli anni 60:

Lemma	Brown	Frown	score
Negro	145.45	17.89	5.6

Tabella1: Keywords Brown□Frown

Ma anche parole legate al contesto storico e bellico (sono gli anni della guerra fredda) quali:

Lemma	Brown	Frown	score
Kennedy	141.20	24.71	4.4
Corps	97.82	28.12	2.8
Peace	168.41	72.43	2.2
Army	233.70	146.56	1.5

Tabella 2: Keywords Brown□Frown

Ma altrettanto significativa è la presenza di parole legate alla religione, quali:

Lemma	Brown	Frown	score
spirit	190.53	71.58	2.5
christ	99.52	37.49	2.3
church	374.35	188.32	1.9
lord	90.16	47.72	1.7

Tabella3: Keywords Brown□Frown

Invertendo nel calcolo i due corpora e prendendo quindi il Frown come *focus corpus* e il Brown come *reference corpus*, nel primo troviamo parole legate al contesto storico degli anni 90 quali:

Lemma	Brown	Frown	score
Clinton	2.55	269.27	22.2
Bush	20.41	372.37	12.6

Tabella 4: Keywords Frown□Brown

Nonché una serie di parole legate all'innovazione tecnologica, alla finanza e all'ambiente:

Lemma	Brown	Frown	score
Environmental	5.95	86.92	6.1
Computer	15.31	122.70	5.2
Investor	14.46	103.11	4.6
Debt	20.41	93.73	3.4
Technology	36.57	105.66	2.5
Television	42.53	100.55	2.1

Tabella 5: Keywords Frown□Brown

E infine, parole legate al genere e alla sessualità:

Lemma	Brown	Frown	score
Gender	2.55	92.03	8.19
Male	47.63	182.35	3.3
Sexual	50.18	161.05	2.8
Woman	381.91	1097.52	2.8

Tabella 6: Keywords Frown□Brown

Differenze grammaticali

Quest'ultimo punto ci ha portato a voler indagare l'uso dei **pronomi personali** nei due corpora, in particolare di "he/she", per osservare se ci sia stato un cambiamento nella rappresentazione del genere femminile. Per fare questo si è utilizzato lo strumento *Wordlist* □ *From this list* □ *Frequency per million* implementato per ogni subcorpus.

Di seguito le frequenze normalizzate e la distribuzione nei subcorpora.

Pron.	Brown	Fiction	Hum.	B.L.	Press	Skills.	Relig.	Pop.	Learn.	Misc.
He	8336	17957	9139	8923	6192	2310	5198	6230	2122	2388
She	2541	7217	3683	1324	962	347	303	2810	371	14

Tabella 7: Frequenze relative "he" e "she" nel Brown Corpus (le frequenze relative sono arrotondate)

Pron.	Frown	Fiction	Hum.	B.L.	Press	Skills	Relig.	Pop.	Learn.	Misc.
He	6620	14453	8717	6892	2215	6064	2476	2215	1343	1191
She	3586	8778	12890	3357	1545	1838	536	1545	380	174

Tabella 8: Frequenze relative “he” e “she” nel Frown Corpus (le frequenze sono arrotondate)

Dal suddetto confronto emerge un generico aumento negli anni dell’uso di pronomi femminili, praticamente in tutti i subcorpus eccetto “Popular Lore”.

Questa considerazione unitamente alla precedente sulle parole legate al genere, risulta coerente con il periodo storico cui afferiscono i testi del Frown, considerato che i cosiddetti “gender studies” nascono proprio negli USA tra gli anni 70 e 80.

Molta letteratura è stata prodotta sull’aumento del **livello di informalità** nell’utilizzo della lingua inglese (e non solo). Esso si concretizza, ad esempio, in fenomeni come l’incremento nell’uso delle abbreviazioni e delle varianti più informali di uno stesso significato. Per l’osservazione di questi processi diacronici prendiamo tre esempi:

- Il paragone tra le frequenze relative della negazione abbreviata “n’t” è stato eseguito con lo strumento *Concordance* all’interno del quale è stata implementata la formula CQL [word=“.*n’t”]; i risultati dimostrano che nella registrazione della lingua risalente al 1991 la quantità di negazioni abbreviate supera in termini relativi quella registrata nei testi del 1961 (Frown=2531.62 per 1 milione di token, Brown=1787.06 per 1 milione di token). Questa leggera differenza tuttavia non è sufficiente ad affermare che c’è stato un aumento delle negazioni contratte, poiché il test *chi quadro* restituisce un valore di 1, che per un grado di libertà pari a 1 come il nostro significa che esiste il 50% di possibilità che le variabili siano indipendenti e che il risultato sia dovuto al caso.
- L’aumento dell’uso della forma contratta di “going to”- ovvero “gonna”- è stato osservato attraverso lo strumento wordlist (lasciato con la sua impostazione di default, dunque su find:words) a cui è stato aggiunto il parametro from this list; i risultati (Frown=25.56 per 1 milione di token, Brown=14.46 per 1 milione di token) dimostrano un’elevata presenza nel Frown rispetto al Brown. Per quanto riguarda la significatività statistica valgono gli stessi risultati riportati sopra per “n’t”.
- “Shall” è un verbo modale usato per esprimere intenzionalità nel futuro, dare suggerimenti o asserire che qualcosa succederà o dovrà succedere con certezza, in particolare usato alla prima persona singolare e plurale. Come verbo modale è tuttavia caduto in disuso e il suo uso è limitato a documenti formali. Utilizzando lo stesso metodo di ricerca adoperato per “gonna” è emerso che esso compare nel Brown circa 228 volte per 1 milione di token e nel Frown solo 124. Anche in questo caso vi è il 50% di possibilità che la differenza osservata sia dovuta al caso.

Conclusioni

Da questo sguardo generale possiamo concludere che esiste una differenza tra il Brown e il Frown Corpus e dunque che all'interno dell'utilizzo dell'inglese americano è riscontrabile una variazione diacronica, in particolare tra gli anni 1961 e 1991-92. La differenza è stata osservata e calcolata secondo i metodi statistici applicati a una comparazione tra corpora creati appositamente per essere facilmente giustapposti.

Per iniziare si guardi all'aumento della frequenza di utilizzo di termini relativi ai mass-media e a problemi di carattere sociale e ambientale su cui solo recentemente si è iniziato a porre l'attenzione; al contrario, nel Brown Corpus è possibile ricostruire una sorta di dominio semantico della guerra presumibilmente per il contesto della Guerra Fredda, di termini religiosi che appaiono meno frequentemente nel Frown coerentemente con una inevitabile secolarizzazione della società, e di parole ad oggi marcatamente offensive quali "negro".

Andando avanti nel ripercorrere lo studio sopra esposto, si sono misurate importanti differenze nelle prassi grammaticali. Non solo il pronome personale femminile segna di star scalando lentamente le vette della rappresentatività, coerentemente alle nuove consapevolezza di cui sopra, ma è possibile indagare ulteriormente anche altri fenomeni. L'aumento delle abbreviazioni e la decrescita dell'utilizzo di termini formali lasciano supporre che la deriva verso l'informalità, per cui tanta opinione pubblica incolpa i social network, risale a ben prima della loro nascita.

Le differenze lessicali e grammaticali evidenziate nella presente ricerca restituiscono un quadro di come gli eventi storici e le evoluzioni del pensiero possano influenzare una lingua, e confermano la tendenza sempre più dilagante all'informalità.