SSIT

Scuola Superiore per Interpreti e Traduttori

sede di Pescara

Analisi dei Corpora

Luglio 2020

Barbara Mennella

e.mail: info@scuolainterpreti.it

Web: https://www.scuolainterpretionline.com

Segreteria Studenti: 08527754 - 3274915165

Un **Corpus**, al plurale **Corpora,** nella sua accezione generale è una raccolta ordinata e completa di opere o di autori, ma nel gergo della linguistica ha assunto il significato di insieme di testi, ordinati in file, e trattati in modo da essere gestibili e interrogabili attraverso alcuni software che ne permettono di osservare l'uso di una lingua e di verificarne tendenze generali su base statistica. Possiamo considerare i **Corpora** come dei ***database*** formati da testi accomunati da alcune caratteristiche, ad esempio la lingua, il periodo storico, la provenienza, per ambiti settoriali (letteratura, filosofia, scienze, ecc...) selezionati e organizzati per facilitare le analisi linguistiche.

A seconda del tipo di analisi testuale o linguistica possono essere suddivisi in token, vale a dire unità minime in cui è suddivisibile il testo. Le parti del discorso contenute nei testi sono inoltre classificate in base alla loro funzione cioè con lemmatizzazioni e annotazioni di vario genere. In questo modo i corpora possono essere analizzati ed elaborati da appositi programmi informatici che consentono di consultare, ricercare, filtrare e generare statistiche del loro contenuto.

Spesso nel settore della traduzione si parla anche di **Corpora paralleli**, che possono essere bilingue o multilingue, vale a dire formati dai testi originali e dalle relative traduzioni in una o più lingue.

Attraverso strumenti di analisi quantitativa e statistica possiamo indagare le regolarità linguistiche che emergono dai testi e che costituiscono la base per la descrizione della struttura del linguaggio. La disciplina che studia il linguaggio attraverso i **Corpora** è nota come **Linguistica dei Corpora**. Oltre che dagli studiosi di linguistica, i corpora sono usati frequentemente nell'ambito della lessicografia, della traduzione o dell'insegnamento delle lingue.

Rivestono un'importanza fondamentale nella linguistica contemporanea essendo utilizzati per selezionare lemmi in base alla loro frequenza d'uso, per identificare le costruzioni tipiche in cui una parola occorre e per coglierne le sfumature di senso in base ai contesti. La creazione e l'utilizzo di un **Corpus** si rivela molto interessante per il traduttore in quanto può verificare la frequenza d'uso di una parola o di una serie di parole in un certo contesto.

I traduttori specializzati in alcuni settori molto tecnici, come il campo medico o scientifico, possono crearsi un corpus selezionando un certo numero di testi di fonti autorevoli (riviste mediche, articoli scientifici, etc.) e poi caricarli su dei software detti ***concordancer***, cioè strumenti per consultare un corpus e visualizzare le parole (ed eventualmente le sue possibili varianti) nel loro contesto.
A differenza delle memorie di traduzioni, il cui utilizzo si rivela interessante più per i traduttori tecnici che per i traduttori letterari, l'utilizzo dei corpora appare utile per qualsiasi tipo di traduzione, soprattutto per la traduzione attiva.

Sono ben note le difficoltà da parte di un traduttore non madrelingua a produrre un buon testo nella lingua di arrivo che sia corretto e completo, l'analisi dei Corpora è un aiuto indispensabile per i traduttori che accettino lavori in attiva, pur consapevoli dei rischi e dei propri limiti linguistici, a cui cercano rimedio in svariati modi.

Cosa sono i Concordancer e come funzionano

Sono programmi per la ricerca e l'analisi linguistica in un corpus, si usano per verificare le relazioni tra le parole di un testo e dare accurate informazioni circa il modo in cui sono utilizzate nel loro contesto.

Può essere usato anche per l'analisi di corpora paralleli (cioè testi bilingue) o anche come aiuto all'interno di un sistema di traduzione assistita come i CAT Tools per analizzare le memoria di traduzione e verificare come una parola o parte di una frase sono state tradotte in contesti simili. Un **Concordancer**, di norma, da la possibilità di cercare combinazioni di parole all'interno di una gamma specificata o anche solo parti di parole. Una volta terminato, propone una lista di parole ordinate alfabeticamente, inserite in una frase di contesto, o altri dati di carattere linguistico. Questi dati possono essere usati in vario modo, come, per esempio, per studiare le collocazioni (gli abbinamenti frequenti di parole), verificare l'uso delle preposizioni o più semplicemente la frequenza delle occorrenze di una data parola.

All’interno di un **Corpora** il testo deve essere rappresentativo per essere utile e significativo. Esistono vari metodi per valutare la **rappresentatività di un campione**, ma la maggior parte delle valutazioni più accurate si basano sulla ricchezza del vocabolario, misurata come numero di parole diverse presenti nel corpus.

L’estensione di un corpus è la sua ampiezza, che influenza il grado di rappresentatività di un campione testuale.
L’analisi all’interno di un corpora può essere di due tipi:

Analisi Statistica

- Analisi condotte più volte e ripetibili
- Usate in maniera standardizzata
- Comparabilità dei risultati

Analisi Dinamica

- Analisi di tipo diacronistico
- Maggiore difficoltà di distribuzione e trattamento
- Necessità di un corpus ben formato e ampio

Esistono vari tipi di **Corpora**, le differenze sono fondamentali per il tipo di utilizzo che se ne vuole fare.

Corpus specialistico	**Corpus generale o di riferimento**
orientato alla descrizione di una particolare varietà del linguaggio o ad un ristretto dominio applicativo ✓ analisi della terminologia biomedica, ecc. ✓ linguaggio infantile ✓ linguaggio sportivo, economico, ecc. ✓ linguaggio patologico	trasversale rispetto alle diverse varietà di un linguaggio plurifunzionale, orientato a rappresentare tutti gli aspetti caratteristici proponendosi come risorsa di riferimento per la descrizione di un argomento e può essere organizzato in vari sotto-corpora specializzati per varietà.

- ✓ **corpus sincronico:** descrive un particolare stadio del linguaggio (i testi appartengono tutti ad una stessa finestra temporale)
- ✓ **corpus diacronico:** descrive il mutamento linguistico (i testi appartengono a diverse finestre temporali)
- ✓ **corpus monolingue:** contiene testi di una sola lingua
- ✓ **corpus bi/plurilingue:**
  - **corpus parallelo** – lo stesso testo originale è presente e tradotto in più di una lingua, uno di seguito all'altro.
  - **corpus allineato** – ciascuna frase o parola della lingua L1 è esplicitamente collegata col suo traducente nella lingua L2
  - **corpus comparabile** – testi in più lingue (non tradotti) appartenenti alle stesse tipologie (ciascuna lingua è rappresentata da testi diversi)
- ✓ **corpus di scritto:** solo testi di linguaggio scritto
- ✓ **corpus di parlato:** solo trascrizioni di linguaggio parlato
- ✓ **corpus misto:** testi scritti e trascrizioni di parlato (in proporzioni variabili)
- ✓ **speech database:** campioni di linguaggio parlato in forma di segnale acustico (più eventualmente la trascrizione ortografica)
- ✓ **corpus multimediale:** testi scritti, video, parlato in forma di segnato acustico, ecc.

Le applicazioni dello studio dei corpora va ben al di là del campo traduttivo

Dizionari	• Individuazione accezioni delle parole • Incidenza termini nell'uso corrente • Definizione casi d'uso delle parole
Grammatiche	• Maggiore aderenza agli usi correnti della lingua • Individuazione regole d'uso della lingua
Trattamento automatico	• Realizzazione parser, tagger e lemmatizzatori statistici • Traduzione automatica più accurata
Didattica	• Realizzazione testi adeguabili alle esigenze degli studenti • Organizzazione insegnamento della lingua

Tra i più comuni **Concordancer**, usati in ambito linguistico, ricordiamo:

- **AntConc**
- **ApSIC Xbench**
- **BootCat**
- **Cor-pusEye**
- **GlossaNet**
- **MonoConc**
- **WordSmith**

Una lista completa di software che analizzano i Corpora per trasformarli in informazioni statistiche utili è disponibile sul sito web https://www.kdnuggets.com/software/text.html.

Alcuni sono progettati appositamente per chi intende realizzare articoli da pubblicare online su blog, siti o portali, mentre altri estrapolano informazioni dai feedback dei consumatori a fini di marketing. La maggior parte di questi software è a pagamento, tuttavia è disponibile una versione di prova per testarne le funzionalità e capire se e in quale misura può essere utile al lavoro del traduttore professionista.