A2. Analisi della distribuzione di frequenza di un testo: istruzioni

Linguistica e Filologia Digitale (Simone Ciccolone)

a.a. 2019/2020

Quest'attività laboratoriale è **opzionale**. Gli studenti che intendono sostenere l'esame da frequentanti possono scegliere di completare, **oltre all'attività A1 (obbligatoria), una qualsiasi di queste attività:**

- **A2. Analisi della distribuzione di frequenza di un testo**
- **A3. Creazione di un documento TEI di (una parte di) un testo**
- **A4. Trascrizione in ELAN di un minuto di parlato dialogico**

Modalità di consegna

Gli studenti sono invitati a consegnare un **elaborato scritto con il resoconto sintetico** (anche in forma di appunti) dello svolgimento dell'attività laboratoriale (come descritta di seguito), **insieme ai materiali elaborati nel corso dell'attività**. Il resoconto dovrà contenere, oltre a un brevissimo commento sulle scelte e le procedure adottate, una discussione ragionata sulla distribuzione di frequenza osservata nel testo, ad es. presentando commenti in merito ai tipi più frequenti o al numero di *hapax*.

I materiali, preferibilmente in formato elettronico, potranno essere consegnati:

- **tramite la piattaforma di e-learning (elearning.unica.it)**, caricando i file nell'attività **"A2. Consegna del resoconto dell'attività"**;
- per **e-mail a simone.ciccolone@unica.it**;
- durante gli **orari di ricevimento**.

Risultato atteso

L'attività laboratoriale A2 riguarda l'analisi della distribuzione di frequenza di un testo: a partire dai materiali elaborati per l'attività A1, lo studente dovrà generare due liste, una delle occorrenze e una dei tipi. Tramite l'inserimento di queste due liste in un foglio di calcolo, dovrà calcolare la frequenza di ogni *type* e produrre una rappresentazione grafica appropriata della distribuzione di frequenza (profilo rango-frequenza e/o spettro di frequenze). Successivamente, dovrà analizzare e commentare tale distribuzione, anche tramite osservazioni mirate della lista di parole e delle frequenze. L'analisi può essere ulteriormente approfondita tramite la lemmatizzazione e/o la categorizzazione anche di parte dei dati, al fine di proporre riflessioni di linguistica quantitativa più articolate.

I materiali da consegnare dovranno quindi includere: (1) le tabelle con la lista di occorrenze e la lista di frequenza dei types; (2) i grafici realizzati per rappresentare la distribuzione di frequenza; (3) il resoconto scritto con l'analisi.

Procedura

1. Creazione della lista di occorrenze

Si consiglia di partire dai materiali prodotti nel corso dell'attività laboratoriale A1. Alla fine della fase 3 dell'attività A1, avete prodotto un file di testo tokenizzato: ogni riga corrisponde a una singola parola del testo, e le parole sono disposte nell'ordine in cui compaiono nel testo di partenza.

Da questo testo tokenizzato possiamo produrre la **lista di occorrenze**, semplicemente ordinando alfabeticamente tutte le righe. Lo possiamo fare tramite la funzione "Sort lines" di SublimeText (menù `Edit > Sort lines`).

Occorre tener conto di un dettaglio: la presenza di **maiuscole e minuscole**. Conviene riportare tutto il testo in minuscola, per evitare conteggi distinti della stessa parola in base alla forma grafica (si verifichi che non si creino problemi con i nomi propri: prima della tokenizzazione, o almeno prima dell'ordinamento alfabetico, occorre verificare che eventuali nomi propri omografi di nomi comuni siano distinguibili, e dove necessario univerbati; ad es. "Maestro Ciliegia" > "Maestro_Ciliegia"). Per convertire tutto il testo in minuscola, usare la funzione "convert case": `Edit > Convert case > Lower case`.

Ora selezionate tutto e copiate la lista di occorrenze in un foglio di calcolo (Excel, Calc, Numbers o Google Sheets). Rinominate il foglio come "tokens" o "occorrenze".

2. Creazione della lista di tipi

A partire dalla lista di occorrenze, create una lista contenente solo forme diverse in modo univoco, o tramite SublimeText o tramite il foglio di calcolo (funzione "Remove duplicates").

In SublimeText, usate la funzione "permute lines": `Edit > Permute lines > Unique`. A questo punto avrete un testo contenente una singola riga per ogni forma grafica diversa. Se non avete risolto prima la questione delle maiuscole, vi troverete qui dei duplicati della stessa parola, una con iniziale maiuscola e l'altra minuscola. Questo può andar bene se le maiuscole rimaste corrispondono ai nomi propri, altrimenti dovrete ripetere la procedura dopo aver corretto la lista di occorrenze al punto precedente.

Ora selezionate tutto e copiate la lista di tipi nello stesso documento di calcolo, ma su un "foglio" diverso, che rinominerete come "types" o "tipi".

3. Calcolo delle frequenze

Ora, nel foglio "tipi", inserite nella seconda colonna una formula per calcolare il numero di occorrenze corrispondenti al type:

```
=CONTA.SE('Tokens!A:A';A2)
```

Il primo campo della formula **"CONTA.SE"** contiene l'intervallo della tabella "Tokens" in cui sono riportate tutte le occorrenze. Il secondo campo indica la forma da cercare (*verbatim*) nell'intervallo. La formula restituisce il conteggio di celle all'interno dell'intervallo selezionato (campo 1) che corrispondono al criterio indicato (campo 2), ovvero: il numero di occorrenze per il *type* nella prima colonna. Copiate questa formula su tutta la colonna e per tutti i *types* della tabella.

N.B.: In ogni tabella, inserite sempre nella prima riga le intestazioni relative ai dati contenuti in ogni colonna, per facilitarne la lettura.

4. Rappresentazione grafica

Potete scegliere se rappresentare la distribuzione di frequenza con il profilo rango/frequenza, con lo spettro di frequenze o con entrambi.

4.1. Profilo rango/frequenza

A questo punto, **ordinate la tabella dei tipi in ordine decrescente in base alla frequenza**: in questo modo visualizzerete in alto i tipi più frequenti. All'interno di gruppi di parole con la stessa frequenza (una **classe di frequenza**) l'ordine non è importante (dovrebbero mantenere l'ordine alfabetico della lista di partenza; se preferite mantenere l'ordine di comparsa nel testo e fare riflessioni al riguardo, andate alla **sezione 6.1**).

Aggiungete la colonna con il **rango**: inserite il numero 1 in corrispondenza della prima riga, e la formula `"= C2 + 1"` nella seconda riga. Copiate la formula su tutte le righe successive (la formula dovrebbe automaticamente adattarsi, sostituendo C2 con C3 alla terza riga di contenuto e così via).

Ora potete rappresentare graficamente la distribuzione di frequenza tramite il **profilo rango/frequenza**: selezionate le colonne "frequenza" e "rango" e aggiungete un grafico a dispersione; regolate i parametri avendo cura di avere il rango sull'asse X (in ordine crescente) e la frequenza sull'asse Y; applicate le scale appropriate come opportuno.

4.2. Spettro di frequenze

Per creare uno **spettro di frequenze**, occorre innanzitutto produrre una tabella che conteggi il numero di tipi diversi per ogni classe di frequenza (ovvero, che conti quante righe sono presenti per ognuno dei valori presenti nella colonna delle frequenze). Lo si può fare o applicando la formula "CONTA.SE" oppure, più efficacemente, creando una **tabella pivot**.

Selezionate la tabella con i tipi e le rispettive frequenze. Usate la funzione "Crea tabella pivot" (in Google Sheets è nel menù "Dati", in Excel è nel pannello "Inserisci") per creare la tabella pivot su un nuovo foglio. Selezionate il campo "frequenza" come campo di riga e il campo con i *types* come valori: essendo in formato stringa, la tabella pivot dovrebbe automaticamente calcolare il conteggio di righe con lo stesso valore nella colonna delle frequenze - ovvero, conteggiare il numero di tipi all'interno di ogni classe di frequenza del testo.

Ora potete produrre un grafico (a dispersione, o anche a linea spezzata) partendo dalla selezione dell'intestazione di riga e della colonna con il conteggio di tipi per classe di frequenza: la classe di frequenza deve comparire sull'asse X (possibilmente in ordine crescente) e la *type frequency* sull'asse Y.

5. Analisi della distribuzione di frequenza

Una volta prodotte la lista di types ordinati per frequenza e le rappresentazioni grafiche, potete procedere alla discussione dei dati relativi alla distribuzione di frequenza. La vostra analisi dovrà tener conto di più aspetti tra quelli indicati di seguito:

1. parole ad alta frequenza: quali sono, che caratteristiche hanno etc.;
2. hapax: numero di hapax, rapporto con il numero totale di types etc.;
3. andamento zipfiano: rapporto numerico tra le frequenze, proporzione tra hapax ed elementi più frequenti, concentrazione del numero di occorrenze etc.;
4. parole-funzione e parole-contenuto: quali parole-contenuto sono più frequenti, posizione delle parole grammaticali etc.

Non è necessario dilungarsi molto nella discussione né presentare un testo formalmente compiuto o definitivo: anche questa parte del resoconto può essere in forma di appunti, benché più articolato e con qualche argomentazione.

6. Integrazioni opzionali dell'analisi

Chi volesse può proseguire l'attività laboratoriale approfondendo l'analisi con alcune osservazioni aggiuntive. Se ne propongono qui alcune, a titolo di spunto e di eventuale attività di esercitazione autonoma.

6.1. Dinamica del "vocabolario"

Potrebbe essere interessante osservare la quantità di tipi diversi all'interno del testo in modo dinamico, ovvero osservando di quanto aumenta il "vocabolario" (qui: il numero di types) nel corso del testo. Ad esempio, potremmo contare il numero di types nelle prime 1000 parole, nelle prime 2000 parole e così via.

Per farlo, occorre partire dal testo tokenizzato, prima dell'ordinamento alfabetico dei tokens. Copiate la lista di occorrenze non ordinata su un foglio di calcolo e aggiungete un numero progressivo nella seconda colonna. Per sicurezza, è meglio che i numeri non siano calcolati, per evitare che le formule cambino il risultato quando viene cambiato l'ordinamento delle righe: scrivete 1 e 2 nella prima riga, selezionate le due caselle, poi trascinate in basso così come si fa per copiare una formula (se preferite usare un calcolo tramite formula, potete copiare la colonna e incollare poi solo i valori, mantenendo così la giusta numerazione eliminando le formule).

Dalla lista di occorrenze, non ordinata, create la lista di types (tramite SublimeText, o ancor meglio all'interno del foglio di calcolo, con "rimuovi duplicati" o tramite una tabella pivot). Inserite il calcolo della frequenza dei types, ordinate la lista dei types per frequenza decrescente e aggiungete anche il rango. Aggiungete infine una colonna in cui riportate, nella tabella dei types, il primo valore del numero sequenziale assegnato alla prima occorrenza di ogni type. Provate a rappresentare la crescita del vocabolario nel vostro testo tramite questi valori inseriti nelle liste di parole generate.

6.2. Lemmatizzazione e frequenza lessicale

La lemmatizzazione del testo permette di fare osservazioni più pertinenti a livello linguistico, tenendo conto delle diverse forme della parola e aggregando più correttamente i dati. Potete lemmatizzare anche solo la porzione di types a più alta frequenza del vostro testo: a partire dalla tabella dei types ordinata in ordine decrescente di frequenza, cominciate a lemmatizzare i tipi più frequenti, fino ad arrivare al 25-40% del numero totale di occorrenze o a types con frequenza tra 7 e 10. In questo modo dovreste riuscire a lemmatizzare piuttosto rapidamente una parte consistente del vostro testo. Riprovate a fare i calcoli relativi alla distribuzione di frequenza, usando però come categorie non i types ma i lemmi.

6.3. Parti del discorso e *type-token ratio*

Così come per la lemmatizzazione, l'etichettatura per parti del discorso ci permette di fare osservazioni quantitative più pertinenti da un punto di vista linguistico. Per procedere a questo tipo di analisi, seguite i passi descritti nella sezione precedente e categorizzate per parti del discorso la porzione di types più frequenti. Provate ad analizzare la distribuzione di frequenza anche in base a questo nuovo livello di informazioni aggiunto, e a calcolare il *type-token ratio* per le diverse categorie lessicali.

Simone Ciccolone (4/12/2019)