Tavole statistiche

Presentazione

Sommario

1. Introduzione

Illustriamo qui sinteticamente le diverse statistiche elaborate sulla base della lemmatizzazione del corpus. Richiamiamo preliminarmente alcuni termini utilizzati in seguito.:

- **forma-occorrenza** (ingl. **token**) o semplicemente **forma**: qualunque *parola* nel senso informatico, cioè qualunque sequenza di caratteri alfanumerici (al massimo 26) compresa tra spazi bianchi o altri delimitatori (segni di interpunzione) che occorra nei testi:
- **forma-tipo** (ingl. **type**): il **tipo** di forma-occorrenza cui si riconducono le diverse forme-occorrenze costituite dalla stessa sequenza di caratteri e occorrenti nei testi analizzati;
- **lemma** (detto anche **forma di citazione** del vocabolo o, anche, **lessema**): la configurazione convenzionale che assume una voce del dizionario, al quale si riconducono (nelle lingue flessive come l'italiano) le diverse forme-tipo delle parti del discorso variabili e, ovviamente, le forme-tipo delle forme cosiddette invariabili (che possono anche conoscere, in realtà, variazion di tipo eufonico, come ad esempio *o* e *od*, *e* e *ed* ecc.; come di consueto il lemma ha la forma del singolare per i sostantivi, del singolare maschile per gli aggettivi, dell'infinito per i verbi;
- **lemmatizzazione**: assegnazione delle forme-tipo (flesse o invariabili) a uno dei lemmi del dizionario,
- **polirematica**: gruppo di parole (al massimo di 67 caratteri) che ha un significato unitario, non desumibile da quello delle forme che lo compongono, comune nell'uso corrente (*veder rosso, essere al verde*) e nei linguaggi tecnico-specialistici (*motore a scoppio, particella elementare*);
- **delimitatori**: caratteri (generalmente segni di interpunzione) che, anche insieme allo spazio bianco, separano le diverse forme occorrenti in un testo;
- **categoria grammaticale:** ciascuna delle classi morfologico-sintattiche in cui il dizionario e la grammatica ripartiscono gli elementi del discorso;

- **marca d'uso:** informa sul grado di utilizzazione di un lessema (basso uso, alto uso, comune, ecc.) o sul particolare ambito d'uso (letterario, tecnico-specialistico ecc.);
- **forme frequenti**: forme-tipo più spesso incontrabili in testi italiani (tipicamente articoli, preposizioni, congiunzioni, pronomi, avverbi), che normalmente interessano circa il 50% di tutte le forme-occorrenza presenti nei testi stessi. Per alleggerire la gestione del data base e soprattutto il compito dell'operatore della lemmatizzazione manuale, la lemmatizzazione delle forme frequenti è stata effettuata automaticamente dal programma, senza quindi coinvolgere l'operatore; nei casi in cui l'assegnazione automatica delle forme frequenti è tuttavia risultata non univoca, i lemmi individuati sono stati assegnati a una categoria grammaticale fittizia, denominata *categoria zero* o *categoria multipla*.

Seguono ora le illustrazioni di tre tipi di tavole statistiche:

- tavole e dati relativi al corpus nel suo insieme (**Statistiche generali forme-lemmi** e **Statistiche omografi**);
- tavole e dati forniti per ciascuno dei cento romanzi del corpus (**Statistiche generali per opere** e **Statistiche dettagliate per opere**);
- tavole e dati calcolati per ogni singolo lemma (**Indici di specificità, dispersione, uso)**.

Prima di passare all'illustrazione dei tre tipi di tavole, diamo conto qui delle diverse formule utilizzate nei calcoli.

1. Misura della leggibilità

Indice Gulpease

$$89-(Lp:10)+(3\times Fr)$$

dove

Lp è il totale delle lettere del testo x 100 diviso per il totale delle parole (occorrenze) del testo;

Fr è il totale delle frasi del testo x 100 diviso per il totale delle parole del testo;

89 e 3 sono parametri numerici, fissi per ciascuna singola lingua, assunti per consentire che i valori di leggibilità oscillino tra un minimo di 0 e un massimo di 100.

2. Misura della ricchezza lessicale

Indice di Guiraud

$$V/\sqrt{N}$$

dove

V è il numero totale dei lemmi del testo

N è il totale delle forme occorrenti nel testo

3. Analisi delle specificità

Indice TFIDF (Term Frequency – Inverse Document Frequency)

$$w_{t,i} = TF_{t,i} * IDF_t$$

dove:

$\mathbf{T_t}$ è il t-esimo termine del lessico di una lingua (lemmario del corpus, nel nostro caso)

$\mathbf{TF_{t,i}}$ è la frequenza del termine $\mathbf{T_t}$ nel documento $\mathbf{D_i}$ (romanzo i-esimo)

$\mathbf{IDF_t}$ è la frequenza inversa dei documenti del corpus contenenti il termine $\mathbf{T_t}$

quindi:

$$w_{t,i} = TF_{t,i} * \log (N / n_t)$$

dove:

N è il totale dei documenti del corpus

$\mathbf{n_t}$ è il numero di documenti del corpus in cui compare il termine $\mathbf{T_t}$

$\mathbf{IDF_t}$ è il $\log (N / n_t)$

L'indice rileva la specifica rilevanza di ciascun determinato lessema in un particolare documento o in una sezione di un testo e può individuare le parole caratterizzate da valori alti dell'indice, cioè le parole chiave di un testo.

.

Indici di dispersione e uso

In un campione di testi diviso in sezioni o in un corpus di parti distinte si dice *dispersione* il grado di presenza di un lessema nelle diverse sezioni o parti. La dispersione è massima per un lessema presente in ogni sezione o parte, è minima per un lessema presente solo in una. La misura della dispersione integra in modo significativo quella della semplice frequenza grezza: a

parità di frequenza, un lessema con maggiore dispersione appare linguisticamente più importante di un lessema con minor dispersione. Una formula di calcolo della dispersione *D* è stata elaborata da Alphonse Juilland, tenendo conto della significatività statistica della presenza di un lessema nelle varie parti di un corpus.

Coefficiente *D* di Juilland

$$D = 1 - \frac{v}{\sqrt{n-1}}$$

dove:

$$v = \frac{\sigma}{\bar{x}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$ (deviazione standard)

i è il romanzo i-esimo del corpus

n è il totale dei documenti del corpus

$$F = \sum_{i=1}^{n} x_i$$

$\bar{x}$ = **F / n**

L' occorrenza normalizzata x_i del lessema in oggetto nel testo i-esimo si calcola dividendo il numero di occorrenze del lessema per il numero totale di occorrenze di tutti i lessemi nel testo in questione e moltiplicando il risultato, convenzionalmente, per 10.000.

La frequenza media $\bar{x}$ si calcola dividendo il numero totale di occorrenze normalizzate del lessema in oggetto in tutto il corpus per il numero totale dei documenti del corpus.

Indice d'uso

Il grado di effettiva rilevanza di un lessema nell'uso di una lingua si misura integrando la misura della sua frequenza assoluta con la sua dispersione nei testi. La formula è:

U = F * D

dove:

U è l'uso

F è la frequenza assoluta del lessema

D è la misura della sua dispersione

Passiamo ora, come anticipato, a illustrare i diversi tipi di tavole.

2. Statistiche generali forme-lemmi

Sezione L x F

Il prospetto rappresenta la distribuzione di frequenza dei lemmi per forme e viceversa.

All'inizio del foglio sono riportati alcuni dati del corpus nel suo insieme:

- il numero totale di lemmi distinti
- il numero totale di forme-tipo distinte
- il numero totale di occorrenze di forme (parole) in associazione con i lemmi

Descrizione dei due quadri del prospetto:

a. distribuzione di frequenza dei **lemmi per forme-tipo**: numero totale di lemmi (e relativa percentuale) che compaiono, nel corpus, in associazione con 1 sola forma, con 2 forme, con 3 forme ecc.; per ogni classe compare inoltre il numero totale di parole corrispondenti e la relativa percentuale rispetto al totale.

Sono indicate le seguenti classi:

1	lemmi in associazione con una sola forma
2	lemmi in associazione con 2 forme
3	lemmi in associazione con 3 forme
4	lemmi in associazione con 4 forme
5	lemmi in associazione con 5 forme
6-10	lemmi in associazione con un numero di forme compreso tra 6 e 10
11-20	lemmi in associazione con un numero di forme compreso tra 11 e 20
21-30	lemmi in associazione con un numero di forme compreso tra 21 e 30
31-40	lemmi in associazione con un numero di forme compreso tra 31 e 40
>40	lemmi in associazione con un numero di forme superiore a 40

b. distribuzione di frequenza delle **forme-tipo per lemmi**: numero totale di forme (e relativa percentuale) che compaiono, nel corpus, in associazione con 1 solo lemma, con 2 lemmi, con 3 lemmi ecc.; per ogni classe compare inoltre il numero totale di parole corrispondenti e la relativa percentuale rispetto al totale.

Sono indicate le seguenti classi:

1	forme in associazione con un solo lemma
2	forme in associazione con 2 lemmi
3	forme in associazione con 3 lemmi
4	forme in associazione con 4 lemmi
5	forme in associazione con 5 lemmi
6	forme in associazione con 6 lemmi
7	forme in associazione con 7 lemmi
8	forme in associazione con 8 lemmi
9	forme in associazione con 9 lemmi
>9	forme in associazione con un numero di lemmi superiore a 9

Sezione Lemmi

Il prospetto rappresenta in forma sintetica la distribuzione di frequenza dei lemmi per classi di occorrenza e per numero di opere (romanzi).

All’inizio del foglio sono riportati alcuni dati del corpus nel suo insieme:

- il numero totale di opere
- il numero totale di lemmi distinti
- il numero totale di occorrenze di forme (parole) in associazione con i lemmi.

Descrizione dei due quadri del prospetto:

a. distribuzione di frequenza dei **lemmi per classi di occorrenza**: numero totale di lemmi (e relativa percentuale) che compaiono, nel corpus 1 sola volta, 2, 3, 4, 5, da 6 a 10 volte ecc.; per ogni classe compare inoltre il numero totale di parole corrispondenti e la relativa percentuale rispetto al totale;

b. distribuzione di frequenza dei **lemmi per numero di opere**: numero totale di lemmi (e relativa percentuale) che compaiono, nel corpus, in 1 sola opera, in 2, 3 ... 10 opere, in un numero di opere da 11 a 20 ecc.; per ogni classe compare inoltre il numero totale di parole corrispondenti e la relativa percentuale rispetto al totale; dal valore indicato in corrispondenza dell’ultima classe (>= 100) si può ricavare, nel nostro caso, il numero di lemmi (239) presenti in tutte le opere del corpus.

Sezione <u>Forme</u>

Il prospetto rappresenta in forma sintetica la distribuzione di frequenza delle forme-tipo per classi di occorrenza e per numero di opere (romanzi).

All’inizio del foglio sono riportati alcuni dati del corpus nel suo insieme:

- il numero totale di opere
- il numero totale delle forme
- il numero totale di occorrenze di forme (parole)

Descrizione dei due quadri del prospetto:

a. distribuzione di frequenza delle **forme per classi di occorrenza**: numero totale di forme (e relativa percentuale) che compaiono, nel corpus, 1 sola volta, 2, 3, 4, 5, da 6 a 10 volte ecc.; per ogni classe compare inoltre il numero totale di parole corrispondenti e la relativa percentuale rispetto al totale.

b. distribuzione di frequenza delle **forme per numero di opere**: numero totale di forme (e relativa percentuale) che compaiono, nel corpus, in 1 sola opera, in 2, 3 ... 10 opere, in un numero di opere da 11 a 20 ecc.; per ogni classe compare inoltre il numero totale di parole corrispondenti e la relativa percentuale rispetto al totale; dal valore indicato in corrispondenza dell’ultima classe (>= 100) si può ricavare, nel nostro caso, il numero di forme (279) presenti in tutte le opere del corpus.

Sezione <u>Lemmi Cat</u>

Il prospetto rappresenta la distribuzione per categoria grammaticale dei lemmi, con la relativa percentuale rispetto al totale.

Per ogni categoria compare inoltre il numero totale di occorrenze corrispondenti ai lemmi indicati e la relativa percentuale rispetto al totale, oltre al numero medio di occorrenze per lemma.

Sezione <u>Lemmi SubCat</u>

Il prospetto riporta con maggior dettaglio gli stessi dati della sezione precedente, con le informazioni articolate per sotto-categoria grammaticale (ad esempio, il dato relativo alla categoria *sostantivo* qui è distinto per sost. maschile, sost. femminile, sost. maschile plurale ecc.).

Sezione Lemmi Lun

Il prospetto rappresenta la distribuzione dei lemmi per lunghezza (con totali e percentuali relativi ai lemmi distinti ed alle occorrenze), con valori da 1 a 26 caratteri.

Viene anche fornito il numero totale di caratteri di tutti i lemmi individuati e delle relative occorrenze, oltre al numero medio di caratteri dei lemmi e delle loro occorrenze.

Sezione Graf L Lun

Rappresentazione grafica della distribuzione dei lemmi per lunghezza (Sezione "Lemmi Lun").

Sezione Pivot L Lun

Tabella di servizio per la costruzione del grafico nella Sezione "Graf L Lun".

Sezione Forme Lun

Il prospetto rappresenta la distribuzione delle forme-tipo per lunghezza (con totali e percentuali relativi alle forme distinte ed alle occorrenze), con valori da 1 a 26 caratteri.

Viene anche fornito il numero totale di caratteri di tutte le forme individuate e delle relative occorrenze, oltre al numero medio di caratteri delle forme e delle loro occorrenze.

Sezione Graf F Lun

Rappresentazione grafica della distribuzione delle forme-tipo per lunghezza (Sezione "Forme Lun").

Sezione Pivot F Lun

Tabella di servizio per la costruzione del grafico nella Sezione "Graf F Lun".

Sezione Stats x Marca

Il prospetto rappresenta la distribuzione per marca d'uso dei lemmi, con la relativa percentuale rispetto al totale.

Per ogni marca compare inoltre il numero totale di occorrenze corrispondenti ai lemmi indicati e la relativa percentuale rispetto al totale, oltre al numero medio di occorrenze per lemma.

Per il significato delle marche d'uso si faccia riferimento alla *Tabella di decodifica* qui in appendice. .

Nota: "**Nessuna**" marca sta a indicare (1) che i lemmi corrispondenti non presentano nel dizionario di riferimento una marca univoca, bensì marche distinte nel corpo della relativa descrizione, a

seconda delle diverse accezioni dei lemmi stessi; (2) che, nel caso di lemmi non previsti dal dizionario, nell'assegnazione i lemmatizzatori non hanno indicato la marca d'uso.

Sezione Stats x Marca Cat

Il prospetto rappresenta la distribuzione dei lemmi per marca d'uso e per categoria grammaticale, con i relativi totali per ciascuna marca d'uso e per ciascuna categoria.

Nota: per la decodifica delle marche d'uso e per il significato di "**Nessuna**" marca valgono le stesse considerazioni della sezione precedente.

Sezione Stats x Marca Cat Occ

Il prospetto è analogo a quello della sezione precedente, ove in luogo del numero dei lemmi sono riportati i valori delle corrispondenti occorrenze, con i relativi totali per ciascuna marca d'uso e per ciascuna categoria.

Sezione Stats x ES

Il prospetto rappresenta la distribuzione per lingua dei lemmi marcati "ES", con la relativa percentuale rispetto al totale.

Per ogni lingua compare inoltre il numero totale di occorrenze corrispondenti ai lemmi indicati e la relativa percentuale rispetto al totale, oltre al numero medio di occorrenze per lemma.

Nota: "<>" sta a indicare che ai lemmi corrispondenti, pur marcati "ES", non è stata associata alcuna lingua in fase di assegnazione (si tratta generalmente di lingue non ben individuate oppure di lingue inconsuete non trovate nell'elenco proposto).

Sezione Stats x DI

Il prospetto rappresenta la distribuzione per dialetto dei lemmi marcati "DI", con la relativa percentuale rispetto al totale.

Per ogni dialetto compare inoltre il numero totale di occorrenze corrispondenti ai lemmi indicati e la relativa percentuale rispetto al totale, oltre al numero medio di occorrenze per lemma.

Nota: "<>" sta a indicare che ai lemmi corrispondenti, pur marcati "DI", non è stato associato alcun dialetto in fase di assegnazione.

Sezione Stats x RE

Il prospetto rappresenta la distribuzione per regionalismo dei lemmi marcati "RE", con la relativa percentuale rispetto al totale.

Per ogni regionalismo compare inoltre il numero totale di occorrenze corrispondenti ai lemmi indicati e la relativa percentuale rispetto al totale, oltre al numero medio di occorrenze per lemma.

Nota: "<>" sta a indicare che ai lemmi corrispondenti, pur marcati "RE", non è stato associato alcun regionalismo in fase di assegnazione.

3. Statistiche omografi

Si tratta di un'unica sezione con due prospetti simili.

Il primo prospetto rappresenta la distribuzione delle forme tipo in base alla loro assegnazione, effettuata dall'elaboratore, ad un numero variabile di lemmi a causa della loro omografia (forme cioè che presentano la stessa grafia ma con etimo, significato e talvolta pronuncia diversi).

A seconda dei casi, infatti, una forma tipo può corrispondere a "zero" lemmi (non è stato quindi trovato nel dizionario nemmeno un lemma a cui assegnarla), ad un solo lemma (forma "non ambigua"), ovvero ad un numero di lemmi maggiore di uno; in quest'ultimo caso il successivo intervento dell'operatore ha provveduto al suo disambiguamento.

Nel prospetto si trovano pertanto le diverse situazioni (fino a 12 lemmi potenziali), in cui è indicato sia il numero delle forme (e relativa percentuale sul totale), sia le corrispondenti occorrenze (e relativa percentuale sul totale).

Il secondo prospetto, con le stesse caratteristiche del primo, indica il numero di forme (e corrispondenti occorrenze) che comprendono, tra i potenziali lemmi a cui sono assegnabili, verbi pronominali; ogni qualvolta, infatti, l'infinito di una forma verbale preveda anche la versione pronominale, questa viene inclusa tra le scelte proposte all'operatore, aumentando di conseguenza il numero di possibili omografi per quella forma.

4. Statistiche generali per opere

Tutte le sezioni contengono dati in forma comparata per i cento i romanzi oggetto dell'analisi.

Sezione Stats F

Il prospetto riporta i seguenti dati per ognuno dei romanzi elaborati (in **grassetto** i titoli delle colonne):

- il numero totale di occorrenze di forme (**Parole**)
- il numero totale di **Forme- tipo** distinte
- il numero totale di **Pagine**
- il numero medio di parole per forma (**Par/For**)
- il numero totale di hapax di forme tipo (**Hapax F**)
- la percentuale di hapax rispetto al totale delle forme tipo (**%HF/F**)
- la percentuale di hapax rispetto al totale delle parole (**%HF/P**)
- il numero totale di occorrenze di poliremathiche (**Occ Polir**)
- il numero totale di **Poliremathiche**
- il numero medio di occorrenze di poliremathiche per poliremathica (**Occ/Poli**)
- il numero medio di parole per poliremathica (**Par/Poli**)
- il numero totale di hapax di poliremathiche (**Hapax P**)
- la percentuale di hapax rispetto al totale delle poliremathiche (**%HP/Pol**)
- la percentuale di Hapax rispetto al totale delle occorrenze di poliremathiche (**%HP/Occ**)
- il numero totale di poliremathiche *Origine* (**Poli Orig**)
- il numero medio di occorrenze di poliremathiche per poliremathica *Origine* (**Occ/Pol.Or.**)
- la percentuale di poliremathiche *Origine* rispetto al totale (**Orig %**)

In fondo alla tabella, oltre ai totali (**Totali corpus**), sono anche indicati il numero delle **Forme frequenti** e, per differenza, il numero delle **Altre forme**, con relative percentuali sul totale, in termini sia assoluti sia di occorrenze.

Nota: le poliremathiche *Origine* sono poliremathiche (in genere preposizionali, sostantivali o aggettivali) nella forma riportata nel dizionario di riferimento che, ai fini dell'elaborazione, sono state "moltiplicate" per renderle riconoscibili in modo automatico; ad esempio, *accanto a* ha generato *accanto al, accanto alla, accanto agli* ecc. Il termine "poliremathiche" viene quindi qui applicato all'insieme ottenuto dopo questa operazione.

Sezione Stats L

Il prospetto riporta i seguenti dati per ognuno dei romanzi elaborati (in **grassetto** i titoli delle colonne):

- il numero totale di occorrenze di forme (**Parole**)
- il numero totale di **Forme tipo** distinte
- il numero totale di caratteri, esclusi i segni di interpunzione, relativi a tutte le occorrenze di forme (**Lettere**)
- il numero totale di **Frasi** nel testo
- il numero totale di frasi a "zero verbi" nel testo, in cui cioè non compaiono né verbi né participi passati (**Frasi 0V**)

- la percentuale di frasi a “zero verbi” rispetto al totale delle frasi (**% 0V**)
- il numero medio di parole per frase (**Par**/**Fra**)
- indice **Gulpease** (di “leggibilità”)
- il numero totale di **Lemmi** distinti
- il valore percentuale (**%L/F**) del numero dei lemmi (Lemmi) rispetto al numero delle forme (Forme)
- il numero totale di hapax di Lemmi (**Hapax L**)
- la percentuale di hapax rispetto al totale dei Lemmi (**%HL/L**)
- indice **Guiraud** (di “ricchezza lessicale”)

In fondo alla tabella sono riportati i totali in tutto il “corpus” (“**Totali Corpus**”).

Sezione Chart

Tabella di servizio per la preparazione dei grafici.

Sezione Chart Gulp

Rappresentazione grafica dell’indice Gulpease per i romanzi del “corpus”, dalla sezione “Stats L”.

Sezione Pivot Gulp

Tabella di servizio per la costruzione del grafico contenuto nella sezione “Chart Gulp”.

Sezione Chart Guiraud

Rappresentazione grafica dell’indice Guiraud per i romanzi del “corpus”, dalla sezione “Stats L”

Sezione Pivot Guiraud

Tabella di servizio per la costruzione del grafico contenuto nella sezione “Chart Guiraud”.

5. Statistiche Dettagliate per Opere

Tutte le sezioni contengono dati in forma comparata per i cento i romanzi oggetto dell'analisi.

Si tratta in particolare di cinque terne di prospetti, relative a dati statistici sui lemmi per ogni singola opera.

Il primo elemento di ogni terna contiene dati sul numero dei lemmi, il secondo dati sulle occorrenze dei lemmi ed il terzo i valori percentuali delle occorrenze dei lemmi rispetto ai totali.

Ogni prospetto è organizzato con un'opera per ogni riga e i valori dell'elemento distintivo della terna nelle colonne.

Le terne di prospetti e le relative descrizioni sono riportate di seguito:

- **Tipo Cat Lem/Occ/Occ %** – distribuzione delle categorie grammaticali per ogni opera;

- **Marca Lem/Occ/Occ %** – distribuzione delle marche d'uso per ogni opera;

- **ES Lem/Occ/Occ %** – distribuzione delle lingue per opera.

- **DI Lem/Occ/Occ %** – distribuzione dei dialetti per opera.

- **RE Lem/Occ/Occ %** – distribuzione dei regioalismi per opera.

6. Statistiche per Polirematiche

Sezione Classi

Il prospetto rappresenta in forma sintetica la distribuzione di frequenza delle polirematiche per classi di occorrenza e per numero di opere (romanzi).

All'inizio del foglio sono riportati alcuni dati del corpus nel suo insieme:

- il numero totale di opere
- il numero totale di polirematiche distinte
- il numero totale di occorrenze di polirematiche

Descrizione dei due quadri del prospetto:

a. distribuzione di frequenza delle **polirematiche per classi di occorrenza**: numero totale delle polirematiche (e relativa percentuale) che compaiono, nel "corpus", 1 sola volta, 2, 3, 4, 5 volte ecc. Per ogni classe compare inoltre il numero totale di occorrenze di polirematiche corrispondenti e la relativa percentuale rispetto al totale.

b. distribuzione di frequenza delle **polirematiche per numero di opere**: numero totale delle polirematiche (e relativa percentuale) che compaiono, nel corpus, in 1 sola opera, in 2, 3 ... 10 opere, in un numero di opere da 11 a 20 ecc. Per ogni classe compare inoltre il numero totale di occorrenze di polirematiche corrispondenti e la relativa percentuale rispetto al totale. Dal valore indicato in corrispondenza dell'ultima classe (>= 100) si può ricavare, nel nostro caso, il numero di polirematiche (solo 2) presenti in tutte le opere del "corpus".

Sezione Poli Cat

Il prospetto rappresenta la distribuzione per categoria grammaticale delle polirematiche, con la relativa percentuale rispetto al totale.

Per ogni categoria compare inoltre il numero totale di occorrenze corrispondenti alle polirematiche indicate e la relativa percentuale rispetto al totale, oltre al numero medio di occorrenze per polirematica.

Sezione Poli Marca

Il prospetto rappresenta la distribuzione per marca d'uso delle polirematiche, con la relativa percentuale rispetto al totale.

Per ogni marca compare inoltre il numero totale di occorrenze corrispondenti alle polirematiche indicate e la relativa percentuale rispetto al totale, oltre al numero medio di occorrenze per polirematica.

Per il significato delle marche d'uso si faccia riferimento alla relativa *Tabella di decodifica*.

Nota: "**Nessuna**" marca sta a indicare che le polirematiche corrispondenti non presentano nel dizionario di riferimento una marca univoca.

Sezione Poli Lun

Il prospetto rappresenta la distribuzione delle polirematiche per lunghezza (con totali e percentuali relativi alle polirematiche ed alle occorrenze), con valori da 3 a 46 (caratteri).

Sezioni Tipo Cat Poli e Tipo Cat Occ

Le due sezioni contengono dati in forma comparata per i cento i romanzi oggetto dell'analisi.

Si tratta di due prospetti relativi alla distribuzione delle categorie grammaticali delle polirematiche per per ogni singola opera.

Il primo contiene dati sul numero delle polirematiche, il secondo dati sulle occorrenze delle polirematiche.

Ogni prospetto è organizzato con un'opera per ogni riga e le diverse categorie grammaticali nelle colonne.

7. Indici di specificità, dispersione, uso

Sono stati calcolati gli indici di specificità, dispersione ed uso per tutti i lemmi del corpus.

Per quanto riguarda il primo, si propone una sezione ridotta (**Indici specificità 50**), contenente i primi 50 lemmi per ogni singola opera, ordinati per valori decrescenti dell'indice TFIDF (Term Frequency – Inverse Document Frequency).

Gli indici di dispersione ed uso sono invece riscontrabili nelle tre liste complete che offrono i lemmi ordinati per valori decrescenti di frequenza assoluta, dispersione e uso.