

LE PAROLE DEL RAV

L'analisi testuale per la
metavalutazione

GLOSSARIO PER L'ANALISI TESTUALE





INVALSI - Via Ippolito Nievo 35, Via Marcora 18/20, 00153 Roma.

Le parole del Rav. L'analisi testuale per la metavalutazione

A cura di:

Donatella Poliandri, Ughetta Favazzi, Monica Perazzolo, Isabella Quadrelli, Emanuela Vinci

Il Glossario è stato redatto da Ughetta Favazzi e Emanuela Vinci

Alla riflessione sull'autovalutazione e valutazione delle scuole, e al progetto PON Valu.E, hanno contribuito in questi anni: Fabio Alivernini, Mattia Baglieri (consulente), Paola Bianco, Roberta Cristallo (consulente), Nicoletta Di Bello, Graziana Epifani, Stefano Famiglietti (responsabile settore Web), Ughetta Favazzi, Brunella Fiore (assegnista), Francesca Fortini, Michela Freddano, Letizia Giampietro, Filippo Gomez Paloma (consulente), Angela Litteri, Beba Molinari (consulente), Lorenzo Mancini, Sara Manganelli, Daniela Marinelli, Flora Morelli, Enrico Nerli Ballati, Monica Perazzolo, Donatella Poliandri (responsabile dell'Area di Ricerca INVALSI - Innovazione e Sviluppo e del progetto PON Valu.E), Elisabetta Prantera, Isabella Quadrelli (consulente), Maria Ranieri (consulente), Sara Romiti, Simone Russo, Stefania Sette, Consuelo Torelli (assegnista), Emanuela Vinci.

Caso (case)	Nel corpus oggetto di analisi a ciascun caso corrisponde un codice meccanografico (RAV). Quando si parla di percentuale di casi si fa riferimento alla percentuale di RAV in cui è contenuta una parola o categoria
Corpus testuale	Un corpus di testi è una qualsiasi raccolta di frammenti testuali fra loro confrontabili (documenti, verbali, domande aperte, resoconti di <i>focus group</i> , interviste libere, raccolta di lettere o canzoni, etc.). Il corpus è l'insieme dei testi sui quali si vuole e si deve effettuare l'analisi, è un insieme ragionato di testi che corrispondono ad un obiettivo, allo scopo per cui verrà analizzato (Bolasco, 1999).
Frammento	Ogni frammento, o unità di contesto, costituisce il livello elementare per sfogliare la collezione. A sua volta esso come ogni insieme è composto da unità elementari (parole) (Bolasco, 2013). I frammenti possono avere lunghezza variabile in base al tipo di materiali raccolti nella collezione. In questo studio ogni frammento è composto dalle parole contenute in ciascun campo aperto per l'individuazione dei punti di forza, di debolezza e per la descrizione delle motivazioni.
Hapax	Termini o parole che appaiono nei frammenti di testo una sola volta (occorrenza pari a 1).
Keyword in context	Si tratta dell'analisi testuale più semplice da implementare. In italiano è nota come analisi delle concordanze e consiste nel collocare una specifica parola selezionata nel contesto in cui essa viene utilizzata. Una volta selezionata una parola si identificano i co-testi (destra e sinistra) con l'obiettivo di disambiguare il reale significato di ogni occorrenza di un termine.
Keyword retrieval	Questa funzione di WordStat consente di recuperare qualsiasi documento, paragrafo o frase contenente una parola chiave specifica, una combinazione di parole chiave o di rilevarne l'assenza.
Multiword	Segmento ripetuto di particolare rilevanza nel contesto dell'analisi, ad esempio Prove Invalsi, Curricolo verticale.
Occorrenza (token)	Ogni volta che una parola appare o ricorre in un corpus definisce una sua occorrenza, che viene anche chiamata <i>token</i> (Bolasco, 2013 p.52)
Partizione	Raggruppamento di frammenti (una partizione è ad esempio costituita da tutti i frammenti relativi alla descrizione dei punti di forza per una specifica area o sub-area).
Segmento ripetuto	Una qualsiasi sequenza di parole adiacenti nel testo di varia lunghezza.
Stop list	Si tratta di una lista di esclusione che consente di rimuovere dall'analisi testuale alcune forme grafiche presenti nei testi, ma ritenute poco significative ai fini delle analisi, come ad esempio congiunzioni, preposizioni, articoli. Il software QDAMiner WordStat permette di definire una <i>stop list</i> personalizzata: le parole in essa contenute non vengono inserite nel computo delle analisi implementate.
Topic extraction	La procedura di Topic extraction presente in WordStat è finalizzata a individuare la struttura tematica nascosta di una raccolta di testo. La principale procedura statistica utilizzata per l'estrazione di argomenti

	in WordStat è un'analisi fattoriale con rotazione Varimax al fine di estrarre un numero ridotto di fattori.
Type	Il <i>type</i> è una qualunque forma grafica (parola) che appare nel corpus, ovvero una sequenza di caratteri compresa tra due separatori (Bolasco, 2013). Per separatori si intendono ad esempio gli spazi, le virgole, ecc. L'insieme delle forme grafiche viene rappresentato dalla lista a cui è associato il numero di occorrenze con cui è presente nel corpus (frequenza).
Type/token	Il quoziente <i>type/token</i> è un indice di ricchezza lessicale che indica la varietà di parole differenti contenute in un testo. Si calcola dividendo il numero delle parole diverse usate in un testo (<i>type</i>) per il numero delle parole complessive di quel testo (<i>token</i>) moltiplicato per 100. Un rapporto <i>type/token</i> superiore al 20% caratterizza un corpus testuale non sufficientemente ricco dal punto di vista lessicale per effettuare un'analisi quantitativa.
TF • IDF	L'indice è l'acronimo di " <i>Term Frequency-Inverse Document Frequency</i> " e rappresenta una formula matematica utilizzata per calcolare l'importanza di una parola inserita in un documento in rapporto ad altri documenti simili. Dal punto di vista matematico si pondera la frequenza di termine per la frequenza inversa del termine nel documento in esame. Tale ponderazione attribuisce rilevanza alla frequenza ma anche al fatto che le parole o le categorie siano concentrate in pochi documenti piuttosto che in molti. I termini con un valore elevato dell'indice <i>tf-idf</i> sono rilevanti in quanto rappresentativi del suo contenuto; diversamente se si presentano in più documenti è meno discriminante.