Qualcosa di nuovo sulla lettura. Nuove prospettive di conoscenza con i big data

STEFANO BANDERA
Social media analyst, Roma
mr.stefanobandera@gmail.com

GIOVANNI CARUSO
Psicologo, Roma
giovanni_caruso@hotmail.com

CHIARA FAGGIOLANI
Dipartimento di Scienze documentarie, linguistico-filologiche e geografiche, Università di Roma la Sapienza,
chiara.faggiolani@uniroma1.it

ANDREA RICCI
Centro Nazionale per i Trapianti, Roma
andrea.ricci@iss.it

DOI: 10.3302/2421-3810-201601-084-1

Introduzione

La *lettura* è l'oggetto di studio del progetto di ricerca che si presenta in queste pagine. Lettura intesa come «ciò che succede quando leggiamo»[1]. A quel «succedere quando» sono connessi diversi aspetti: le motivazioni soggiacenti, le sue modalità e i suoi tempi, il cosa si legge e il piacere che se ne ricava[2], la sua socialità, infine.
Queste dimensioni – il sospetto è che non siano soltanto queste – insieme definiscono il suo significato: esse non sono il background, non sono qualcosa di esterno, per intenderci, sono dentro l'esperienza di lettura.
Il progetto PERCE.READ (La percezione della lettura in Italia nel contesto del *social reading*), nato all'interno della cornice istituzionale del Master in *Data science* dell'Università degli studi di Roma Tor Vergata, si propone di studiare il "contesto" in cui avviene la lettura oggi, farne emergere le "connessioni" con altre pratiche, in definitiva definirne il "significato", a partire dalle possibilità offerte da quello scenario tecnologico e prima ancora conoscitivo in rapidissima evoluzione che va sotto il nome di *big data*, ovvero le grandi masse di dati presenti sul web, lasciate più o meno volontariamente dai lettori.
La nostra convinzione è, infatti, che la lettura si stia profondamente riconfigurando non solo per il «suo migrare dal testo/libro gutenberghiano al testo/libro digitale» ma anche in «uno spazio "altro", il Web, i cui segni frusciano tra macchine e menti delle persone, e tra mac-

Per tutti i siti web l'ultima consultazione è stata effettuata il 12 maggio 2016.
Il contributo rispecchia globalmente le opinioni dei quattro autori: tuttavia, l'introduzione, le conclusioni e i paragrafi 3 e 5 sono stati scritti da Chiara Faggiolani, il paragrafo 1 da Giovanni Caruso, il paragrafo 2 da Andrea Ricci, il paragrafo 4 da Stefano Bandera.

[1] Antoine Compagnon – autore con Roland Barthes della voce "lettura" per l'*Enciclopedia* Einaudi (Roland Barthes - Antoine Compagnon, *Lettura*, in *Enciclopedia*, vol. 8, Torino, Einaudi, 1979, p. 176-199) – definisce così la lettura. Cfr. Antoine Compagnon, *Il demone della teoria: letteratura e senso comune*, Torino, Einaudi, 2000, p. 173 (tit. or. *Le Démon de la théorie: littérature et sens commun*, Paris, Èditions du Seuil, 1998). Sulle questioni definitorie di lettura si veda la panoramica offerta da Luca Ferrieri, *La lettura spiegata a chi non legge: quindici variazioni*, Milano, Editrice Bibliografica, 2013, p. 20-30 (versione epub). Qui l'Autore ricorda come, nonostante esistano importantissimi contributi e risultati nel campo della teoria della lettura, quello che «ancora oggi ci manca è l'unificazione dei diversi conseguimenti ottenuti dalla scienza letteraria, dalla sociologia, dalla pedagogia e psicologia, dall'estetica, dalla semiotica, dall'ermeneutica, dall'etica, dal decostruzionismo, dalla comparatistica, per non citare che alcuni dei filoni più fecondi che si sono occupati di ricerca sulla lettura. In termini fisici potremmo dire che manca una teoria unificata della lettura», cfr. Ivi, p. 32.

[2] Cfr. Roland Barthes, *Il piacere del testo*, traduzione di Lidia Lonzi, Torino, Biblioteca Einaudi, 1975, (tit. or. *Le plaisir du texte*, Paris, Èditions du Seuil, 1973).

chine ed altre macchine, nascosti nella loro invisibile forma digitale»[3].
Il gruppo di ricerca che ha lavorato al progetto PERCE.READ si caratterizza per la sua spiccata interdisciplinarità: la competenza nel trattamento statistico dei dati strutturati e non strutturati, la conoscenza dei *social media*, delle policy relative alla *privacy* e all'estrazione informatica dei dati, unite alla conoscenza dei processi cognitivi implicati nella lettura sono state fondamentali per la progettazione, la rilevazione, l'analisi e l'interpretazione dei dati nel particolare contesto degli studi sulla lettura.
Si ritiene utile sottolineare tale aspetto perché accostandosi a studi di questo tipo la commistione di diversi approcci metodologici e interpretativi appare essere una *conditio sine qua non.*
Per ragioni di spazio e per meglio rispondere agli obiettivi della sezione che ci ospita, ovvero presentare ricerche innovative e *in progress*, ci soffermeremo essenzialmente sul metodo utilizzato nel nostro studio ed è per questo che essenzialmente è di tracce lasciate dai *lettori* che parleremo.

Il cervello che legge

La lettura è un processo cognitivo molto complesso rispetto al quale è necessario aprire una breve parentesi utile alla contestualizzazione della nostra ipotesi di lavoro[4].
Alla voce "lettura" del *Dizionario di Psicologia* di Umberto Galimberti, leggiamo che si tratta di un «processo di acquisizione informativa che consente, previo riconoscimento delle combinazione segniche che costituiscono le parole del linguaggio scritto, l'associazione del significante (segno) al significato (senso)»[5]. Il riconoscimento della parola scritta, quindi, avviene dapprima a livello percettivo-visivo tramite l'identificazione delle singole lettere che la compongono all'interno di un processo influenzato dal grado di familiarità della forma della parola che è stato acquisito con l'esercizio. Segue poi il livello semantico di decodifica del significato delle parole nel quale, in sostanza, si passa da una rappresentazione visiva ad una rappresentazione interiore del suono e del significato in base a quel "lessico interno" che si acquisisce a partire dalla propria esperienza linguistica.
A livello neuronale si può quindi affermare che sussistono a due differenti reti cerebrali[6]:
- la rete cerebrale del suono, comprendente le regioni superiori del lobo temporale, la corteccia frontale inferiore e la corteccia precentrale dell'emisfero sinistro e, nei casi in cui vi sia compatibilità fra lettere viste e suoni uditi, la regione del *planum* temporale che viene specificamente attivata;
- la rete cerebrale del significato, comprendente la parte posteriore della circonvoluzione temporale media, il lato ventrale anteriore del lobo temporale e la parte triangolare della regione frontale inferiore dell'emisfero sinistro e, nei casi in cui due o più parole condividono lo stesso significato, la regione temporale media sinistra, anch'essa specificamente attivata.

È, dunque, grazie all'attivazione della regione occipito-temporale ventrale sinistra che riusciamo a riconoscere le lettere e le parole: tale regione, secondo Stanislas Dehaene, si sarebbe riciclata a tale funzione[7]. Secondo questa tesi, nota come "riciclaggio neuronale", il nostro cervello, grazie alla sua innata plasticità che gli ha consentito di adattarsi all'ambiente circostante, per poter acquisire le competenze necessarie alla lettura nell'attuale contesto culturale ha dovuto convertire vecchie funzioni che avevano una utilità nel nostro passato evolutivo.
Inoltre, la scoperta dei "neuroni specchio"[8] ha implicazioni importanti nell'esperienza di lettura. Molto sinteticamente, si può dire che grazie a queste particolari cellule,

[3] Cfr. Chiara Faggiolani - Maurizio Vivarelli, *Leggere in rete: la lettura in biblioteca al tempo dei big data*, in *Bibliotecari al tempo di Google: profili, competenze, formazione.* Relazioni del Convegno delle Stelline, Milano 17-18 marzo 2016, Milano, Editrice Bibliografica, 2016, p. 101-126.

[4] Per un approfondimento sulla scienza della lettura dal punto di vista delle neuroscienze si rimanda a Maryanne Wolf, *Proust e il calamaro: storia e scienza del cervello che legge*, Milano, Vita e Pensiero, 2009 (tit. or. *Proust and the squid: the story and science of the reading brain,* New York, Harper, 2007). In esergo l'A. riporta una citazione tratta da *The symbolic species* di Terrence Deacon che recita: «Sapere come qualcosa ha avuto origine è spesso il migliore indizio su come essa funziona». Questa è la chiave di lettura del volume. Si veda anche Stanislas Dehaene, *I neuroni della lettura*, Milano, Raffaello Cortina Editore, 2009 (tit. or. *Les neurones de la lecture*, Paris, Ed. Odile Jacob, 2007). Si veda anche Federica Fioroni, *Neuroscienze e lettura*, «Enthymema», 2013, n. 8, p. 223-229 <http://dx.doi.org/10.13130/2037-2426/3039>.

[5] Umberto Galimberti, *Dizionario di Psicologia*, Torino, UTET, 1999, p. 541.

[6] Joseph T. Devlin - Helen L. Jamison - Paul M. Matthews - Laura M. Gonnerman, *Morphology and the internal structure of words*, in *Proceedings of the national academy of science USA*, 41 (2004), n. 101, p. 14984-14988 < http://www.pnas.org/content/101/41/14984.full >.

[7] S. Dehaene, *I neuroni della lettura*, cit.

[8] Cfr. Vittorio Gallese - Luciano Fadiga - Leonardo Fogassi - Giacomo Rizzolatti, *Action recognition in the premotor cortex*, «Brain: a journal of neurology», 119 (1996), n. 2, p. 593-609 <http://dx.doi.org/10.1093/brain/119.2.593>. Cfr. Giacomo Rizzolatti - Luciano Fadiga - Vittorio Gallese - Leonardo Fogassi, *Premotor cortex and the recognition of motor actions,* «Cognitive brain research», 3 (1996), n. 2, p. 131-141, <http://brain-mind.med.uoc.gr/sites/default/files/S_C3_A3_Rizzolatti_1996.pdf>. Cfr. Giacomo Rizzolatti - Laila Craighero, *The mirror neuron system*, «Annual review of neuroscience», 27 (2004), p. 169-192, <http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.27.070203.144230>.

quando si osserva qualcuno eseguire un'azione, all'attivazione delle aree visive il nostro sistema motorio si attiva come se noi eseguissimo le azioni che osserviamo. In sostanza, vedere un'azione significa simularla. Ciò vale non solo quando si vede compiere un'azione ma anche quando se ne sente parlare o quando se ne legge. In tal senso i neuroni specchio sono ritenuti alla base dei processi di coinvolgimento emotivo ed empatico.
Considerato che la letteratura ricrea un mondo di emozioni e di esperienze e sia le emozioni sia le esperienze dei personaggi letterari abitano il mondo immaginario del romanzo, il meccanismo di simulazione ci aiuta ad attraversare quel mondo fittizio permettendo di capire e, in parte, rivivere le emozioni dei protagonisti e le loro vicissitudini[9]. Ai fini della nostra analisi, questo consente di poter avanzare ipotesi interpretative più strutturate sull'esperienza di lettura.

Cosa sappiamo dei lettori e della lettura

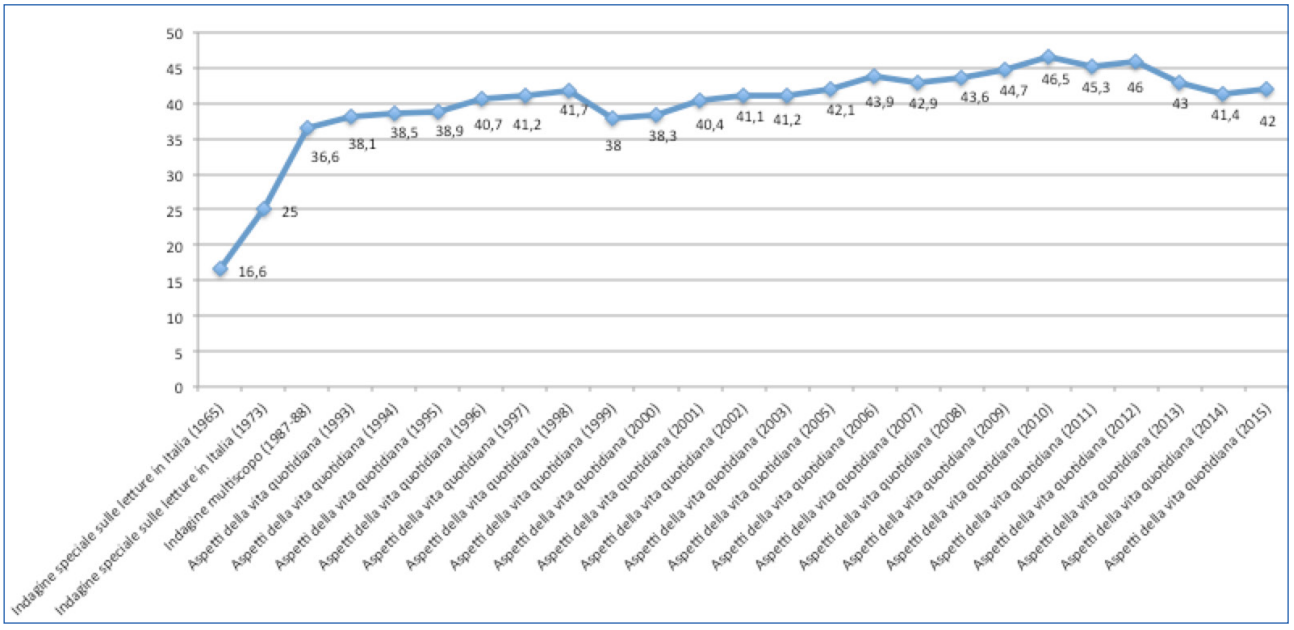
Le ricerche sull'esperienza di lettura delle neuroscienze sono ancora agli inizi.
L'attenzione degli studiosi, degli editori e più in generale degli attori della filiera del libro negli ultimi anni si è concentrata soprattutto su due aspetti:
- le tecnologie e il loro impatto: dalle trasformazioni del web 2.0 alle evoluzioni dei *device* dedicati, dagli standard (pdf o epub), al *digital lending* ecc.[10]
- il numero dei lettori: «la lettura è sparita dall'orizzonte. È la "grande assente" in tutte le riflessioni di oggi sulla letteratura. – osserva Emanuele Trevi – Si parla molto, semmai, del numero dei lettori»[11].

Le ricerche sui lettori[12] tendono a concentrarsi essenzialmente sugli aspetti quantitativi dei comportamenti e delle scelte: sappiamo come segmentarli, in base a quali variabili, sappiamo quanti libri leggono, quali generi amano, quale è il canale che preferiscono, quanto spendono[13].
Molta meno attenzione è stata dedicata al connubio tra le diverse questioni: indagando tra le pieghe della statistica ufficiale non è affatto facile intercettare il cambiamento delle pratiche di lettura alla luce delle trasformazioni soprattutto tecnologiche in corso.
Usando le parole di Roberto Casati: «ci sono oggi più risposte (tecnologiche) che domande (sociali), e dovremmo quindi cominciare a metterci in cerca delle buone domande»[14].
Le stime sulla lettura dal 1993 provengono dall'indagine annuale *Aspetti della vita quotidiana* che dal 2001 rileva anche dati riguardanti l'uso delle nuove tecnologie: personal computer e Internet[15].
Nel 2015 in Italia i lettori – secondo l'Istat le persone di sei anni e più che dichiarano di aver letto almeno un libro nel corso dell'anno per motivi non scolastici o professionali – sono stati il 42,0% della popolazione[16].
Dopo il forte incremento verificatosi negli ultimi trent'anni del Novecento – la percentuale è passata dal 16,6 del 1965 al 40,7 del 1996 – e dopo un lento ma progressivo aumento della quota di lettori registrato a partire dal 2000, che ha raggiunto il picco massimo nel 2010 (46,5%), negli ultimi quattro anni si è manifestata una altrettanto lenta e progressiva inversione di tendenza: la quota di persone che dichiarano di aver letto almeno un libro nel tem-

[9] Vittorio Gallese - Hannah Wojciehowski, *How stories make us feel: toward an embodied narratology*, «California italian studies», 2 (2011), n. 1, <http://escholarship.org/uc/item/3jg726c2>.
[10] Per una panoramica completa e approfondita sugli aspetti connessi al libro elettronico e alla testualità digitale il punto di riferimento è Gino Roncaglia, *La quarta rivoluzione. Sei lezioni sul futuro del libro*, Roma-Bari, Laterza, 2010.
[11] Cfr. Emanuele Trevi, *La lettura nella società postmoderna*, «Italianieuropei», 10 (2010), n. 2, <http://www.italianieuropei.it/la-rivista/item/1597-la-lettura-nella-società-postmoderna.html>.
[12] *Lettore* è per Istat – e quindi, per gli istituti di ricerca e per i ricercatori che si occupano del tema – colui che legge almeno un libro nell'anno di riferimento. A proposito della definizione di lettore commenta Luca Ferrieri: «L'intero apparato statistico della lettura si regge su una creatura mostruosa e ridicola, o mostruosamente ridicola: il lettore made-in-Istat, l'infelice lettore di almeno un libro, che fa *pendant* con il terribile lettore di un solo libro. L'Istat infatti, e a ruota tutti i ricercatori e gli istituti di ricerca statistica e sociologica, considerano lettore colui che ha letto almeno un libro nell'anno di riferimento». Cfr. L. Ferrieri, *La lettura spiegata a chi non legge* cit., p. 89. Si veda a questo proposito Marino Livolsi, *Almeno un libro: gli italiani che (non) leggono*, Scandicci, La nuova Italia, 1986.
[13] Per un approfondimento relativo a tutte le questioni connesse alla non lettura (e alla lettura) si veda Giovanni Solimine, *L'Italia che legge*, Roma-Bari, Laterza, 2010.
[14] Cfr. Roberto Casati, *Contro il colonialismo digitale*, Roma-Bari, Laterza, 2014, p. 19 (e-book).
[15] Ogni anno Istat pubblica il rapporto *La produzione e la lettura di libri in Italia*, che offre una panoramica dell'offerta e della domanda di libri nel nostro paese, partendo dai dati raccolti in due diverse indagini: l'*Indagine sulla produzione libraria* e l'indagine campionaria sulle famiglie *Aspetti della vita quotidiana*. Dal 13 gennaio 2016 sono disponibili i risultati relativi all'anno 2015: Istat, *La lettura in Italia: anno 2015*, 2016, <http://www.istat.it/it/archivio/178337>.
[16] *Ibidem*.

po libero nell'arco dei 12 mesi precedenti l'intervista è scesa dal 46% nel 2012, al 41,4% nel 2014, per poi invertire lievemente il trend ed arrivare al 42,0% del 2015 (Fig. 1).
Solo il 13,7% dei lettori sono "forti", leggono, cioè, in media almeno un libro al mese. Sono lo zoccolo duro del mercato editoriale: da soli essi acquistano circa il 40% dei libri venduti in Italia[17]. Non sono i lettori forti ad aver smesso di leggere libri: appassionati e fedeli, leggono tanto e nel tempo[18].
La quota più significativa di lettori in Italia è rappresentata dai cosiddetti "deboli", coloro che leggono al massimo tre libri in un anno. Rappresentano il 45,5% dei lettori: quasi 11 milioni di persone che mostrano un rapporto saltuario e fragile con la lettura. Quasi la metà dei lettori totali. Sono prevalentemente maschi – il 49,3% dei lettori maschi non legge più di 3 libri in un anno – con al più la licenza media (52,4%), e residenti nel Sud del paese (60.2%).
È opinione largamente condivisa che i comportamenti di lettura siano condizionati da numerosi fattori di natura ambientale, culturale, sociale, familiare.
Al fine di meglio comprendere lo scenario della lettura in Italia sopra descritto, per prima cosa abbiamo ritenuto opportuno approfondire i dati aggregati pubblicati dall'Istat nel rapporto *La produzione e la lettura di libri in Italia: anno 2015*[19], per quantificare l'effetto ovvero il peso delle variabili che maggiormente influenzano la lettura in Italia. In particolare abbiamo analizzato le tre variabili che la letteratura in materia definisce fortemente incidenti sui comportamenti di lettura: il titolo di studio, il sesso e l'età.
Dopo aver disaggregato in maniera semi-automatica il dato Istat abbiamo eseguito una regressione logistica ordinata[20] per valutare l'effetto delle singole co-variate sulla variabile di esito, ovvero il numero di libri letti. Tutte le variabili disponibili erano di tipo categoriale codificate come segue:
- Sesso: 0: maschio, 1: femmina;
- Classe età: 0: >65; 1: 45-64; 2; 25-44; 3: 06-24;
- Titolo di studio: 0: Licenza ele-

Fig. 1: Andamento della lettura dal 1965 al 2015 in relazione alle diverse fonti Istat. Valore espresso in %



[17] Cfr. I risultati dell'indagine *L'Italia dei libri 2011-2013*, commissionata dal Centro per il libro e la lettura a Nielsen. Il Centro è un istituto autonomo del Ministero dei beni e delle attività culturali e del turismo ed ha il compito di divulgare il libro e la lettura in Italia e di promuovere all'estero il libro, la cultura e gli autori nazionali (http://www.cepell.it).

[18] Sottolinea Luca Ferrieri che la quota dei lettori forti «da sola, garantisce l'acquisto di più della metà dei libri editi in Italia. Il "fattore LF" (lettori forti) rappresenta il vero "caso italiano"». Cfr. L. Ferrieri, *La lettura spiegata a chi non legge* cit. p. 92.

[19] Cfr. Istat, *Aspetti della vita quotidiana: anno 2015,* 2016,
<http://www.istat.it/it/files/2016/01/Tavole_lettura_2015.zip?title=La+lettura+in+Italia+-+13%2Fgen%2F2016+-+Tavole.zip>.

[20] Cfr. J. A. Anderson, *Regression and ordered categorical variables (with discussion),* «Journal of the Royal Statistical Society. Series B (Methodological)», 46 (1984), n. 1, p. 1-30; Cfr. Rollin Brant, *Assessing proportionality in the proportional odds model for ordinal logistic regression*, «Biometrics», 46 (1990), n. 4, p. 1171-1178.

mentare; 1: Licenza media; 2: Diploma superiore; 3: Laurea e post-laurea.

Come variabile di *outcome* è stato utilizzato il numero di libri letti nel corso dei 12 mesi che hanno preceduto l'intervista (secondo la codifica 0; 1-3; 4-6; 7-11; >12). La variabile di *outcome* è, dunque, di tipo categoriale ordinata.
L'analisi è stata eseguita utilizzando il software STATA v.10.1[21].
In Tabella 1 vengono riportati i risultati della regressione logistica ordinata.
Tutte le variabili analizzate evidenziano un effetto statisticamente significativo sulla lettura, confermando quanto sottolineato dalle indagini Istat.
Dall'analisi multivariata risulta evidente come il titolo di studio (odds Ratio = 1,97 Sta. Err = 0,018) e il sesso (odds Ratio = 1,9, Sta. Err = 0,032) siano le variabili che quantitativamente influenzano di più la lettura. Il sesso femminile e un elevato titolo di studio sono entrambi fattori che favoriscono la lettura.
Anche l'età influenza la lettura in maniera inversamente proporzionale – i soggetti giovani sono quelli che leggono di più – ma il suo effetto è quantitativamente meno rilevante.
La quota di lettori nel nostro paese è superiore al 50% della popolazione solo tra gli 11 ed i 19 anni mentre la fascia di età in cui si legge di più è quella tra i 15 e i 17 anni (53,9%).
L'aspetto più preoccupante è che sono proprio i più giovani ad aver smesso di leggere libri, ovvero la fascia d'età che ha sempre registrato (e tuttora registra) la percentuale di diffusione più alta della lettura. Nel 2015 rispetto all'anno precedente la quota di lettori è, infatti, diminuita dal 44,6% al 44,0% per i ragazzi tra i 6 ed i 10 anni; dal 53,5% al 52,1% per quelli tra gli 11 ed i 14 anni; dal 51,7% al 50,3% per giovani tra i 18 ed i 19 anni.
Solo la fascia di età tra i 15 e i 17 anni mostra un aumento della percentuale di lettori dal 51,1% al 53,9% divenendo la fascia di età che legge di più.
La lettura di libri viene praticata soprattutto dalle persone con un titolo di studio più elevato: leggono tre laureati su quattro (il 75,0% nel 2015 e il 74,9% nel 2014), ma la proporzione si riduce a uno su due fra chi ha conseguito al più il titolo di diploma superiore (50,2% nel 2015 e il 51,1% nel 2014).
Mentre la correlazione positiva tra lettura e titolo di studio o condizione professionale dovrebbe apparire abbastanza scontata, ciò che meriterebbe un diverso approfondimento è proprio il numero di laureati o dirigenti che dichiara di non prendere mai un libro tra le mani nel corso dell'anno.
Nel 2015 il 25% dei nostri concittadini laureati e il 49,8 % dei diplomati dichiara all'Istat di non aver letto neppure un libro nel tempo libero nei 12 mesi precedenti all'intervista.

Tab. 1 – Risultati della regressione logistica ordinale

```
. ologit  outcome2 eta2 sex2 studio2, or

Iteration 0:   log likelihood = -69394.326
Iteration 1:   log likelihood = -65409.432
Iteration 2:   log likelihood = -65347.712
Iteration 3:   log likelihood = -65347.656
Iteration 4:   log likelihood = -65347.656

Ordered logistic regression                     Number of obs   =      57276
                                                LR chi2(3)      =    8093.34
                                                Prob > chi2     =     0.0000
Log likelihood = -65347.656                     Pseudo R2       =     0.0583

------------------------------------------------------------------------------
    outcome2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        eta2 |   1.185337   .0100377    20.08   0.000     1.165826    1.205175
        sex2 |   1.898521   .0323301    37.65   0.000     1.836201    1.962956
     studio2 |    1.97904    .017727    76.21   0.000     1.944599    2.014091
-------------+----------------------------------------------------------------
       /cut1 |   1.860206   .0227679                      1.815582    1.904831
       /cut2 |   2.865708   .0246574                       2.81738    2.914036
       /cut3 |   3.696889   .0266381                      3.64468    3.749099
       /cut4 |   4.542385   .0297931                      4.483991    4.600778
------------------------------------------------------------------------------
```

[21] Prodotto dall'azienda statunitense StataCorp, <http://www.stata.com/company>.

Se, nel complesso, il livello di istruzione influisce in misura rilevante sui livelli di lettura (la quota di lettori oscilla tra un valore massimo del 75,0% fra i laureati ed un minimo del 25,7% per chi possiede al più la licenza elementare), osservando più nel dettaglio il fenomeno attraverso un confronto generazionale si rileva che anche tra le persone con un titolo di studio superiore la propensione alla lettura è andata diminuendo nel corso del tempo. I laureati con più di 45 anni leggono, infatti, in proporzione di più rispetto alle persone più giovani con equivalente livello d'istruzione.
«Evidentemente – come sottolinea il primo *Rapporto sulla promozione della lettura in Italia* – non basta saper leggere per diventare lettori – e aggiunge – in Italia, più ancora che in altri paesi industrializzati, si manifesta una forte discrepanza fra la crescita dei livelli di alfabetizzazione e i tassi di lettura nel tempo libero. Infatti, dobbiamo constatare che da qualche decennio crescita dell'istruzione e crescita della lettura viaggiano a velocità differenti»[22].

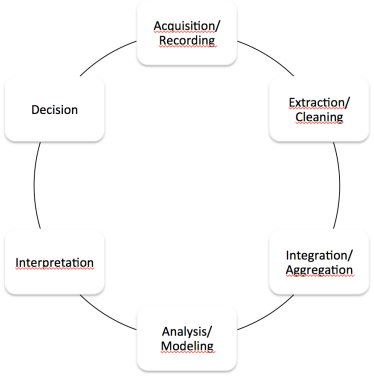
Nuove prospettive di conoscenza con i big data

Un aspetto decisivo che pare opportuno sottolineare è che nelle indagini Istat sopra descritte il rapporto con la lettura rilevato è assolutamente quantitativo, ma l'oggettività dei dati poggia su una domanda iniziale che possiamo definire del tutto soggettiva: gli intervistati si autodefiniscono lettori in base alla loro personale percezione di cosa sia la lettura e il libro[23]. Il sospetto è che l'utilizzo dei media digitali oltre ad aver modificato le abitudini di lettura stia cambiando anche la percezione e il significato attribuito a questa pratica.
È esattamente questo aspetto che abbiamo cercato di approfondire attraverso l'utilizzo dell'approccio *big data*.
Una definizione puntuale di questo fenomeno non è ancora stata formalizzata. In questa sede si fa riferimento a quella più accreditata:

> *Big data* is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making[24].

Oggi abbiamo a disposizione ingenti quantità di dati, di vari tipi e con diverso grado di qualità, che facciamo fatica ad inquadrare in categorie predefinite perché molto diversi da quelli a cui siamo stati abituati e con i quali abbiamo interagito finora. Grandi quantità di dati complessi che forniscono informazioni su fenomeni altrimenti difficilmente osservabili in modo diretto[25].
Attraverso l'uso delle applicazioni ICT per le nostre attività quotidiane lasciamo in modo più o meno volontario "briciole digitali": lasciamo una traccia nei *social network* a cui partecipiamo, nelle *query* che poniamo ai motori di ricerca, nei *tweet* che inviamo e riceviamo[26]. Queste tracce, se opportunamente analizzate, consentono di registrare e approfondire i comportamenti individuali e collettivi, i desideri, le opinioni, le relazioni tra le persone, la percezione dei fenomeni.
Il processo di gestione e analisi dei dati è costituito da sei passaggi (Fig. 2): (1) acquisizione, (2) estrazione, (3) integrazione, (4) analisi, (5) interpretazione, (6) decisione[27].
Nella fase di acquisizione (1), si selezionano i dati, che vengono filtrati e puliti per ridurne la possibile mancanza di accuratezza. In questa fase vengono generati eventuali metadati associati ai dati (ad esempio, come i dati sono stati acquisiti, da quale fonte ecc.).
Poiché i dati acquisiti non saranno tipicamente già nel formato richiesto per l'analisi, durante la fase di estrazione (2) occorre trasformare i dati, normalizzarli, pulirli per migliorarne la veridicità.
È in questa fase, ad esempio, che se si vuole procedere ad una analisi

Fig. 2: Processo di gestione e analisi dei *big data*



[22] Forum del libro, *Rapporto sulla promozione della lettura in Italia,* 2013, p. 9, <http://www.governo.it/DIE/attivita/rapporto_promozione_lettura.pdf>.
[23] La domanda presente nel questionario è la seguente: «Ha letto libri negli ultimi dodici mesi? Consideri solo i libri letti per motivi non strettamente scolastici o professionali».
[24] Cfr. Mark Beyer - Douglas Laney, *The Importance of Big Data: a definition*, Stamford (CT), Gartner, 2012.
[25] Per approfondimenti si veda anche Ian Ayres, *Super crunchers*, New York, Random House, 2008; Cfr. Chris Snijders - Uwe Matzat - Ulf-Dietrich Reips, "*Big data": big gaps of knowledge in the field of Internet science,* «International journal of Internet science», 7 (2012), n. 1, p. 1-5, <http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf>.
[26] Sulla costumizzazione della nostra esperienza online a partire dai dati lasciati più o meno involontariamente si veda Eli Pariser, *Il filtro: quello che Internet ci nasconde*, Milano, Il Saggiatore, 2012.
[27] Non sempre è così evidente il confine tra un passaggio e l'altro, ma questa suddivisione è funzionale ai fini della trattazione. Per una panoramica si veda *Challenges and opportunities with Big Data: a community white paper developed by leading researchers across the United States*, 2012, <http://www.purdue.edu/discoverypark/cyber/assets/pdfs/BigDataWhitePaper.pdf>.

di *text mining* i dati acquisiti da fonti diverse verranno integrati ad esempio in un unico *corpus* (3), secondo le logiche del software di analisi che il ricercatore avrà deciso di utilizzare.
Le fasi appena esaminate (acquisizione, estrazione e integrazione) possono essere inglobate nella più generale fase di "preparazione dei dati", che rappresenta un momento cruciale nella gestione dei *big data*.
Nella fase di analisi (4) i dati vengono esplorati per estrarre l'informazione ricercata. Tale esplorazione richiede l'adozione di metodologie che differiscono da quelle tradizionalmente usate per l'analisi statistica di piccoli campioni, e che comprendono tecniche di *data mining*, *text mining*, *machine learning*.
La successiva fase di interpretazione (5) richiede la conoscenza dell'ambito di riferimento dei dati stessi. Solo la conoscenza del contesto, della provenienza dei dati può consentire l'identificazione dei *pattern* di interesse. La medesima conoscenza approfondita del contesto di riferimento è determinante anche nell'ultima fase di decisione (6) finalizzata all'utilizzo delle informazioni ricavate in modo efficace e mirato.
Tornando all'obiettivo specifico del progetto PERCE.READ – ovvero indagare la percezione della lettura attraverso le tracce lasciate in rete dai lettori – abbiamo ritenuto opportuno applicare una analisi di *opinion mining*. Con questa espressione – sovente utilizzata come sinonimo di *sentiment analysis* – si fa riferimento ad un metodo che, raccogliendo e analizzando in tempo reale le reazioni degli utenti o i trend su un qualsiasi fenomeno a partire dai contenuti presenti nei *social network*, attraverso particolari tecniche di *text mining*, definisce l'opinione positiva o negativa, l'intensità di tale opinione, l'emotività con la quale è stata espressa e la sua rilevanza da parte del pubblico[28].
Particolarmente congeniale al nostro obiettivo è sembrata la campagna di promozione della lettura #ioleggoperché[29] promossa dall'Associazione italiana editori nel 2015 che, oltre ad una serie di eventi "reali" nelle piazze, ha previsto anche una intensa attività sui principali *social network*, generando una grande attività di conversazione su libri e lettura. La campagna chiedeva, infatti, ai lettori di esprimere e raccontare la propria passione per la lettura e condividerla attraverso l'utilizzo delle principali piattaforme social con l'utilizzo dell'*hashtag* #ioleggoperché.
Per questa ragione è sembrato particolarmente coerente con gli obiettivi del nostro progetto utilizzare i testi prodotti nel periodo che va da marzo a settembre 2015 in occasione di questa campagna come fonte principale per la costruzione del *corpus* oggetto dell'analisi di *opinion mining*.

L'estrazione dei dati e il processo di data cleaning

Prima di entrare nel merito del processo di estrazione dei dati dal web, che ha consentito di costruire il *corpus* della nostra analisi, sembra utile aprire una brevissima parentesi sull'utilizzo dei *social network* nel nostro paese.
Il caso italiano è, infatti, abbastanza curioso. In rapporto agli altri stati dell'Unione europea, l'Italia registra una delle percentuali più basse di penetrazione di Internet tra la popolazione: secondo i dati Eurostat, nel 2014 ancora un italiano su tre non ha utilizzato la rete[30].
A fronte di questa bassa penetrazione si registra, al contrario, un utilizzo massiccio dei *social network* in linea con gli altri stati europei, con una penetrazione che arriva a toccare quasi la metà della popolazione italiana (46%). Complessivamente sono 28 milioni gli italiani che hanno utilizzato almeno una volta i *social media* nell'arco del 2014, con 22 milioni di utenti che ormai vi accedono tramite dispositivo mobile[31].
Avere coscienza di questi numeri è importante perché l'universo di rifermento per la raccolta dei testi per l'analisi di *opinion mining* è stato proprio lo spazio virtuale del web e in particolare i canali del cosiddetto web 2.0 tra cui siti di *blogging* e *social media*[32].
La fase di *text mining* è stata preceduta dalla definizione e dalla messa a punto del *corpus* testuale tramite l'utilizzo di operatori booleani per la definizione delle chiavi di ricerca e tramite tecniche di *web scraping*[33], in grado di recuperare contenuti sulle principali piattaforme web.
Il processo integrato di definizione della *keyword* di ricerca, di esplorazione o *crawling*, e di estrazione

[28] Cfr. Andrea Ceron - Luigi Curini - Stefano Iacus, *Social media e sentiment analysis: l'evoluzione dei fenomeni sociali attraverso la rete*, Milano, Springer, 2013.
[29] Per informazioni dettagliate sulla campagna si rimanda al sito ufficiale <http://www.ioleggoperche.it/it/home/>.
[30] Cfr. Eurostat, *Internet usage by individuals in 2014*, 2015, <http://ec.europa.eu/eurostat/statistics-explained/index.php/Information_society_statistics_-_enterprises>.
[31] We Are Social, *Digital social & mobile in 2015: We Are Social's compendium of global digital statistics*, 2015, <http://wearesocial.com/uk/special-reports/digital-social-mobile-worldwide-2015>.
[32] James Governor - Dion Hinchcliff - Duane Nickull, *Web 2.0 architectures: what entrepreneurs and information architects need to know*, Sebastopol (CA), O'Reilly Media, 2009.
[33] Con il termine *web scraping* si intendono tutte quelle tecniche informatiche di estrazione dati da un sito web per mezzo di un software. Cfr. Ryan Mitchell, *Web scraping with Python: collecting data from the modern web*, Sebastopol (CA), O'Reilly Media, 2015.

dei dati è stato eseguito tramite il software proprietario Tracx[34].
Tracx è una piattaforma di *social media analytics* sviluppata per il settore commerciale in grado di recuperare e rielaborare dati web provenienti da fonti diverse in un *range* temporale limitato, al contrario dei principali software *open source* che richiedono una più lenta elaborazione e un accesso alle API (Application Programming Interfaces) di siti web più limitato.
La lettura, oggetto della ricerca, si legava ad una serie di *keyword* pertinenti quali: "lettura", "libro", "leggere". Se tali termini erano da un lato inclusi nel *topic* di analisi, dall'altro lato risultavano essere troppo generali e poco adatti per il recupero di testi online specifici per l'approfondimento del tema della percezione.
Abbiamo, quindi, ritenuto opportuno non partire da tali termini ma utilizzare parole chiave più specifiche, scelte tra gli *hashtag* di campagne di promozione della lettura quali #ioleggoperché, che oltre ad avere un'estensione semantica più ristretta, erano funzionali al nostro scopo di rilevazione dell'opinione.
Alla scelta di questi termini è seguita un'analisi qualitativa tramite *tag cloud* sui contenuti già raccolti con le chiavi di ricerca generali individuate in precedenza. L'obiettivo era osservare se tra le sequenze di parole più ricorrenti ve ne fossero alcune che rimandassero al nostro tema di ricerca, così da poterle includere nella *query* finale. Espressioni e *hype* emersi sono stati: "mi piace leggere", "#23 aprile"[35], "amo leggere".
La chiave di ricerca risultante è stata ottenuta tramite la combinazione dei diversi operatori logici OR, AND, NOT. Si è inoltre proceduto legando alcuni termini o espressioni tra loro con una distanza massima di tre parole, in modo da evitare contenuti fuori tema[36].

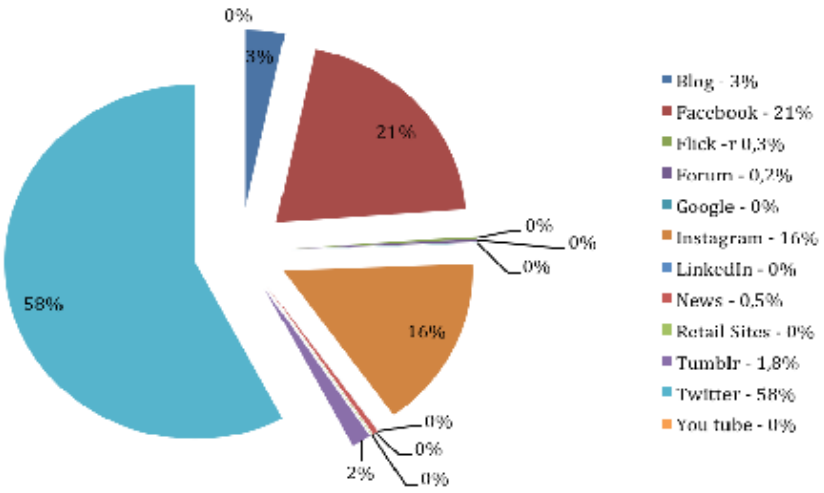
> #ioleggoperché OR [io leggo perché]~3 OR ioleggoperché OR ioleggodifferente OR #ioleggodifferente OR [io leggo differente] OR [piace leggere perché]~3 OR #librialpiede OR [libri al piede] OR librialpiede OR #libriinvaligia OR [libri in valigia] OR libriinvaligia OR #vivalalettura OR [viva la lettura] OR vivalalettura OR #amoleggere OR [amo leggere]~3 OR #vivalalalettura OR vivalalettura OR [viva la lettura] OR (#23aprile AND leggo) OR (#23aprile AND lettura) OR (#23aprile AND leggere) OR (#23aprile AND "libro") .

La fase di *crawling* o esplorazione, avvenuta tramite l'accesso alle API[37], presenti su siti e piattaforme web, ci ha consentito di recuperare contenuti quali post, articoli, *tweet* da siti di news, *social network*, blog e forum entrati poi a far parte del *corpus* definitivo. Occorre, inoltre, specificare che non tutti i siti web consentono l'accesso ad applicazioni terze e *spider*: alcuni come Facebook hanno cambiato di recente la propria policy in tale senso, mentre altri forniscono un ordine di priorità di accesso ai diversi *spider* in seguito ad accordi commerciali.
Nella successiva fase di *scraping* sono stati analizzati i metadati, scaricati e importati in un database locale. Nel nostro caso i contenuti scaricati tramite *scraping* hanno portato in un primo momento alla costituzione di un *corpus* costituito da un totale di 47.827 record, il 58% dei quali estratti da Twitter, il 21% da Facebook, il 16% da Instagram e il restante 5% da blog e altri *social media* (Fig. 3).
Nell'esplorare i dati all'interno del dataset abbiamo individuato un problema di duplicazione dei contenuti, successivamente eliminati, dovuto a due fattori:
1) un errore automatico del software di *scraping* (843 item). Questo è un errore che avviene di frequente, soprattutto su medie e grandi moli di dati, per via delle difficoltà di dialogo tra software e API del sito richiamato;
2) la presenza in rete, e in particolare sui *social network*, di bot (abbreviazione di robot) in grado di inviare contenuti in modo automatico – nel nostro caso, 1.281 *tweet* identici prodotti da un bot su Twitter in pochi minuti – e produrre altre azioni sociali[38].

Questa prima operazione di *data cleaning* ha portato, quindi, alla definizione del *corpus* testuale, in totale 45.703 record, da importare nel software di analisi statistica del testo IRaMuTeQ[39] per la successiva fase di *text mining*.

[34] Per informazioni più dettagliate su Tracx si rimanda al sito web della piattaforma <http://www.tracx.com/>.
[35] Il 23 aprile è la giornata mondiale del libro e del diritto d'autore.
[36] La distanza tra parole è definita dal simbolo ~.
[37] Jemima Kiss, *The Nutshell: a beginners' guide to APIs*, «The Guardian»,December 14,2007, <http://www.theguardian.com/media/pda/2007/dec/14/thenutshellabeginnersguide>.
[38] Lutz Finger, *Do evil: the business of social media bots*, «Forbes», February 17,2015, <http://www.forbes.com/sites/lutzfinger/2015/02/17/do-evil-the-business-of-social-media-bots/#66eb761e1104>.
[39] IRaMuTeQ (<http://www.iramuteq.org/>) è un software *open source* per l'analisi statistica di dati testuali, basato sul software R – IRaMuTeQ sta, infatti, per *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* – e sul linguaggio di programmazione Python. Permette di effettuare le seguenti analisi: statistiche di base, analisi delle specificità e analisi fattoriale delle corrispondenze, classificazione gerarchica discendente, analisi delle similitudini, nuvola di parole.

Fig. 3: Fonte dei testi estratti (in %)



Dopo una prima analisi del *corpus*, ci siamo resi conto che *hashtag*, URL e indirizzi mail sporcavano eccessivamente l'analisi, tanto da renderla scarsamente informativa. Abbiamo intrapreso quindi ad una seconda attività di *data cleaning*, questa volta lavorando direttamente sul file .txt, attraverso l'utilizzo di editor di testo avanzati e delle seguenti *regular expression*:

TROVA TUTTE LE STRINGHE CHE INIZIANO CON HTTP FINO ALLO SPAZIO SUCCESSIVO
(https?:\/\/[^\s]+)
(www[^\s]+)
_^(?:(?:https?|ftp)://)(?:\S+(?::\S*)?@)?(?:(?!10(?:\.\d{1,3}){3})(?!127(?:\.\d{1,3}){3})(?!169\.254(?:\.\d{1,3}){2})(?!192\.168(?:\.\d{1,3}){2})(?!172\.(?:1[6-9]|2\d|3[0-1])(?:\.\d{1,3}){2})(?:[1-9]\d?|1\d\d|2[01]\d|22[0-3])(?:\.(?:1?\d{1,2}|2[0-4]\d|25[0-5])){2}(?:\.(?:[1-9]\d?|1\d\d|2[0-4]\d|25[0-4]))|(?:(?:[a-z\x{00a1}-\x{ffff}0-9]+-?)*[a-z\x{00a1}-\x{ffff}0-9]+)(?:\.(?:[a-z\x{00a1}-\x{ffff}0-9]+-?)*[a-z\x{00a1}-\x{ffff}0-9]+)*(?:\.(?:[a-z\x{00a1}-\x{ffff}]{2,})))(?::\d{2,5})?(?:/[^\s]*)?$_iuS
TROVA TUTTE LE E-MAIL:
\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b
TROVA TUTTI GLI HASHTAG
(\B#\w\w+)
(?:(?<=\s)|^)#(\w*[A-Za-z_.\?]+\w*)
questa Ë migliore
TROVA TUTTE LE STRINGHE CHE INIZIANO CON @
(?:(?<=\s)|^)@(\w*[A-Za-z_.\?]+\w*)

Dopo l'eliminazione di *hashtag*, URL e indirizzi mail abbiamo ottenuto il *corpus*[40] definitivo composto da 44.928 UCI (unità di contesto iniziali), 37.694 forme grafiche e 1.650.864 occorrenze, con un 46,01% di *hapax* (17.344) ovvero le parole che ricorrono una sola volta, come mostra il diagramma di Zipf[41] in Fig. 4.

Uomini e donne che leggono: come cambia il significato attribuito alla lettura

La lettura è un passatempo decisamente femminile. Le indagini Istat dimostrano da sempre come le donne abbiano una maggiore propensione alla lettura già a partire dai 6 anni di età: nel 2015, ad esempio, complessivamente il 48,6% delle femmine e solo il 35% dei maschi hanno letto almeno un libro nel corso dell'anno[42]. Per questa ragione è sembrato particolarmente interessante osservare cosa si nasconde dietro le conversazioni sulla lettura di uomini e donne attraverso l'analisi di specificità del linguaggio femminile *vs* maschile[43].

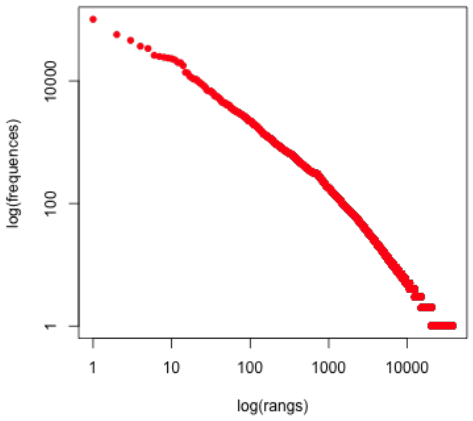
[40] In una logica di analisi automatica del testo per *corpus* si intende una collezione di testi o frammenti, che chiameremo unità di contesto iniziali (UCI), fra loro coerenti e pertinenti per essere studiate sotto un qualche punto di vista. I testi che costituiscono il *corpus* devono essere prodotti in condizioni di enunciazione simili e devono avere caratteristiche confrontabili in merito alla ricchezza del vocabolario e alla lunghezza. Cfr. Sergio Bolasco, *L'analisi automatica dei testi: fare ricerca con il text mining*, Roma, Carocci, 2013.

[41] Secondo la legge di Zipf, in ogni lessico specifico esiste una relazione fondamentale tra frequenza (f) e rango o ordine d'apparizione nel discorso/testo (r). Dove c=costante. Quindi f*r=c. In ogni testo (e in ogni lingua) la frequenza d'uso delle parole non segue la distribuzione normale ma si configura secondo una curva in cui poche parole hanno una frequenza d'uso molto elevata. Il grafico permette quindi di osservare il decremento proporzionale della frequenza di ogni parola in funzione dell'ordine di apparizione o rango collocando alla fine della curva, quindi, gli *hapax*, che hanno frequenza 1.

[42] Istat, *La lettura in Italia: anno 2015*, 2016, <http://www.istat.it/it/archivio/178337>.

[43] Si definisce specifica o caratteristica di un testo ogni parola o espressione sovra/sotto utilizzata rispetto ad una norma di riferimento (ad esempio il valore medio o il valore assunto in un modello di riferimento). Ogni specificità positiva (sovra-utilizzo) di una parola o di una espressione equivale ad un uso superiore a quello 'atteso'. Ogni specificità negativa di un termine, equivale ad un sotto-utilizzo (o rarità del termine, fino ad arrivare anche alla sua totale assenza) rispetto al valore 'atteso'. Essa consente di identificare, su apposite tabelle di contingenza prodotte dal software, quei termini che contraddistinguono e caratterizzano ogni singola parte in cui il *corpus* viene suddiviso, dal punto di vista di una delle variabili interessanti. Nel nostro caso l'analisi di specificità è stata condotta in base alle variabili genere, *social media* di provenienza e periodo

Fig. 4: Diagramma di Zipf



In questo paragrafo presentiamo in estrema sintesi qualche risultato ottenuto.

L'estrazione dei dati dal web, come evidenziato nei paragrafi precedenti, necessita di grande accortezza nel processo di *data cleaning*. Per più della metà dei testi estratti (53%), infatti, il sesso non era specificato (ND): il 34% dei testi era riconducibile a donne e solo il 13% a uomini. Per questa ragione è stata esclusa la modalità ND[44]. L'analisi di specificità ha evidenziato una sovra-rappresentazione delle forme verbali nelle donne rispetto agli uomini.

Sono state, quindi, prese in considerazione le azioni sovra-rappresentate (ovvero >2 per le donne) o sotto-rappresentate (ovvero <2 per gli uomini). Nel merito dei significati, è stato osservato che le azioni che le donne esprimono parlando di lettura sono riferibili essenzialmente a 6 ambiti:

1) emozioni: è stato osservato un vissuto emozionale molto forte connesso a polarità tendenzialmente positive: "piangere" (+ 6,07), "sorridere" (+5,11), "ridere" (+ 2,38), "commuovere" (+ 2,09). Non mancano però le emozioni negative: la "paura" (+4,39) è l'emozione specifica più peculiare del linguaggio femminile insieme alla "felicità" (+4,35);

2) empatia/partecipazione: "aiutare" (+3,59), "incentivare" (+3,44), "condividere" (+3,36), "raccontare" (+2,93), "partecipare" (+2,75), "immedesimare" (+2,36), "coinvolgere" (+2,25);

3) scoperta/evasione: "scoprire" (+8,44), "viaggiare" (+3,94), "immergere" (+ 3,61), "sognare" (+2,98), "fingere" (+2,74), "partire" (+2,68), "evadere" (+2,62), "sfuggire" (+2,09), "dimenticare" (+2,07). L'evasione si lega ad un altro universo di senso molto denso, che è quello attinente la famiglia, estremamente presente nel linguaggio specifico delle donne, ma anche all'universo della resistenza e della salvezza;

4) resistenza/salvezza: "bastare" (+4,62), "resistere" (+3,54), "sopravvivere" (+2,91), "mancare" (+2,20), "accettare" (+2,08), "respirare" (+2,04);

5) sentimento: "piacere" (+13,94), "amare" (+13,34), "sentire" (+11,56), "adorare" (+7,08), "innamorare" (+2,8401), "appassionare" (+2,48);

6) poter fare: "fare" (+16,91), "potere" (+11,19), "accadere" (+8,31), "riuscire" (+6,64), "diventare" (+6,60).

La lettura è percepita come la chiave per evadere dai problemi del quotidiano e per ritagliarsi uno spazio proprio e intimo, ricco di emozioni e magia. Sembrerebbe che le donne nella lettura possano trovare emozioni, sentono di poter fare tutto e ambiscono a vivere le vite più diverse, evadendo dai problemi del reale, esplorando dimensioni nuove e magiche attraverso il coinvolgimento empatico con i personaggi e le loro vicissitudini. La lettura, in tal senso, sembrerebbe essere simbolizzata come nutrimento interiore, non tanto per la mente quanto per il cuore. Anche gli aggettivi, infatti, denotano un universo di senso positivo in cui domina l'aspetto emozionale: "felice" (+7,93), "meraviglioso" (+7,35), "fantastico" (+6,54), "magico" (+3,19), "perfetto" (+2,85).

Rispetto alla specificità del linguaggio maschile è stata osservata una maggiore eterogeneità dei contenuti caratterizzati dall'assenza di forme troppo specifiche e facilmente riferibili a precisi contesti e ambiti di vita. Per le azioni connesse alla lettura, infatti, è parsa interessante la presenza di azioni riconducibili all'esternazione delle emozioni – "assistere" (+ 4,78), "vedere" (+3,23), "ballare" (+3,06), "celebrare" (+2, 88) – piuttosto che all'interiorizzazione delle stesse, come osservato per le donne.

Questa ipotesi ha assunto più corpo con l'esplorazione delle altre forme grammaticali, con particolare attenzione ai nomi – la forma maschile per eccellenza. Due gli universi di senso più densi che, per fornire una sintesi efficace, abbiamo definito la "saga" con una sovra-rappresentazione dei ruoli che fanno riferimento all'ambito dell'azione fisica e del combattimento – "cavaliere" (+8,15), "padrone" (+5,54), "teatrale" (+5,32), "potente" (+5,00), "nobile" (+5,00), "culmine" (+4,81), "re" (+3,86), "conquista" (+ 3,82) – e il "musical" che apre alla dimensione della poli-sensorialità – "concerto" (+ 2,78), "coreografico"

temporale. Queste ultime due hanno consentito di osservare come il linguaggio è cambiato con il passare del tempo e quali sono le caratteristiche specifiche del linguaggio nei differenti *social media*. Per motivi di spazio non è possibile rendere conto dei risultati emersi.

[44] Dall'analisi di specificità dell'intero *corpus*, risultava evidente come la modalità degli ND utilizzasse un lessico più tecnico fortemente diverso da quello utilizzato dagli uomini e dalle donne. La sensazione è che laddove manchi l'indicazione del sesso nel profilo dell'utente che ha inviato il testo, questo possa ricondursi al profilo di un ente istituzionale e non di un privato cittadino. Per questa ragione abbiamo ritenuto opportuno eliminarlo, pur essendo consapevoli della minore robustezza statistica dell'analisi.

(+2,74), ”theatre” (+2,74), ”danzatore” (+2,74), ”musicale” (+2,54), ”ballo” (+2,48). In generale, gli uomini sono attratti dalle emozioni forti e dall’effetto sorpresa: azione, movimento, sensazioni forti sono ciò che amano della lettura.
L’analisi di specificità per sesso ha quindi evidenziato profonde differenze nella percezione della letteratura da parte dei due generi. Questi dati, se relazionati con le statistiche sulla lettura, spiegano che la differenza non è soltanto quantitativa ma profondamente legata alle motivazioni soggiacenti e al significato, anche emotivo, attribuito a questa pratica.
Già questi dati sono sufficienti a farci ipotizzare che una campagna di promozione della lettura per essere efficace debba prevedere diversi stili e linguaggi per i due sessi, utili a sollecitare i diversi universi di senso coinvolti.
Tale differenza sembra sostanziarsi nella ricerca per generi letterari. Negli uomini è interessante la sovra-rappresentazione della parola ”fumetto” (+2,33), genere sovente considerato minore ma evidentemente interessante ed amato dal genere maschile; per le donne sono sovra-rappresentati i generi ”fantasy” (+2,39) e ”romance” (+2,58).

Conclusioni

Avere tanti dati a disposizione (*big data*) non significa necessariamente avere un alto potenziale informativo. Chi si avvicina a studi di questo genere deve sapersi muovere con molta cautela ed estremo rigore poiché diverse sono le difficoltà che si possono incontrare strada facendo, soprattutto nella fase di *scraping* e di *data cleaning*, come messo in evidenza nei paragrafi precedenti.
Questi aspetti, spesso (ma non sempre) evidenziati dalla letteratura sul trattamento dei *big data*, si possono pienamente e consapevolmente cogliere solo “sporcandosi le mani con i dati”. È dal mettere le mani in pasta che possono emergere interessanti suggestioni e utili indicazioni rispetto alle potenzialità degli strumenti e alle criticità dei vari step del processo di ricerca.
Il contesto della lettura è fatto di tante cose: informazioni – per esempio le informazioni disponibili sui libri o le recensioni scritte dai lettori – di mezzi – i *social network* o i media più tradizionali – di persone – gli autori, gli editori, i lettori stessi – di istituzioni – le scuole e le biblioteche più di tutti – e di relazioni che intercorrono tra queste.
Progetti come PERCE.READ, finalizzato a comprendere le dinamiche soggiacenti i comportamenti di lettura e la sua percezione dimostrano come sia oggi fondamentale poter integrare una conoscenza essenzialmente deduttiva – come è quella della statistica ufficiale, che rimane comunque punto di riferimento imprescindibile per qualsiasi studio sulla lettura – con quella induttiva prodotta dall’analisi di *big data*. Tale integrazione, riferendosi comunque a dati testuali, può essere utilmente implementata da approcci basati su logiche inferenziali abduttive utili a formulare ipotesi interpretative più attendibili. Si fa riferimento, ad esempio, all’analisi emozionale del testo[45], una metodologia elaborata in ambito psicoanalitico e psico-sociologico che, ponendo il *focus* sul piano affettivo-simbolico delle parole, cerca di rintracciare emozioni entro le produzioni discorsive e testuali per conoscere e intervenire nelle relazioni sociali[46].
Infatti, se attraverso l’induzione da sola ci si affiderebbe esclusivamente alla scoperta casuale, facendo affidamento soltanto sul ragionamento deduttivo non si potrebbero scoprire cose nuove.
In un contesto di grande trasformazione come quello attuale, questo tipo di approccio integrato può essere utile agli editori e a tutti gli attori della filiera del libro per individuare proposte creative e innovative capaci di rispondere in modo efficace alle sfide lanciate in maniera sempre più stringente dalla complessità del contesto in cui operano.
La rete e le piattaforme in cui la lettura si sta riconfigurando – per motivi di spazio non si è parlato di *social reading*, ma anche questo è un tema all’ordine del giorno – sono un ambiente il cui effetto collaterale principale è proprio la raccolta di dati ai quali sarebbe impossibile accedere in altro modo. È di questo ambiente che il progetto PERCE.READ ha fatto tesoro, con l’obiettivo di fornire interessanti suggestioni e utili indicazioni per lo sviluppo di un nuovo filone di ricerche sulla lettura.

[45] Per approfondimenti cfr. Renzo Carli - Rosa M. Paniccia, *Psicologia della formazione*, Bologna, Il Mulino, 1999; Idd., *Il colloquio come testo: l’analisi emozionale del testo*, in *Oltre l’intervista: il colloquio nei contesti sociali*, a cura di Giancarlo Trentini, Torino, ISEDI, 2000; Idd., *L’analisi emozionale del testo: uno strumento psicologico per leggere testi e discorsi*, Milano, Franco Angeli, 2002.
[46] Per un esempio di ricerca-intervento eseguita con la metodologia dell’analisi emozionale del testo cfr. Rosa M. Paniccia - Cecilia Sesto, *Una ricerca-intervento con le Biblioteche comunali romane come luogo di convivenza nella città: attese di bibliotecari e clienti a confronto,* «Rivista di psicologia clinica», 9 (2014), n. 1, p. 266-289, <http://www.rivistadipsicologiaclinica.it/ojs/index.php/rpc/article/view/456>.

ABSTRACT

Il tema di questo articolo è la lettura intesa come «ciò che succede quando leggiamo». A quel «succedere quando» sono connessi diversi aspetti: le motivazioni soggiacenti, le sue modalità e i suoi tempi, il cosa si legge e il piacere che se ne ricava, la sua socialità, infine. Queste dimensioni definiscono il significato attribuito alla lettura: esse non sono il background, non sono qualcosa di esterno, per intenderci, sono dentro l'esperienza di lettura.

Le ricerche in questo ambito, invece, tendono a concentrarsi essenzialmente sugli aspetti quantitativi dei comportamenti e delle scelte dei lettori: sappiamo come segmentarli, in base a quali variabili, sappiamo quanti libri leggono, quali generi amano, quale è il canale che preferiscono, quanto spendono. Indagando tra le pieghe della statistica ufficiale non è facile intercettare il cambiamento delle pratiche di lettura alla luce delle trasformazioni soprattutto tecnologiche in corso.

Sulla base di queste premesse l'articolo presenta il progetto di ricerca PERCE.READ (La percezione della lettura in Italia nel contesto del social reading), finalizzato a studiare il "contesto" in cui avviene la lettura oggi, farne emergere le "connessioni" con altre pratiche, in definitiva ridefinirne il "significato", soprattutto alla luce delle trasformazioni tecnologiche in corso. Le grandi masse di dati presenti sul web (big data), lasciate più o meno volontariamente dai lettori, sono il "mezzo" utilizzato per farlo.

La ricaduta auspicata: trasformare questi dati in informazioni utili agli attori della filiera del libro per individuare proposte creative e innovative capaci di rispondere in modo efficace alle sfide lanciate in maniera sempre più stringente dalla complessità del contesto in cui operano, soprattutto sul fronte della promozione della lettura.

SOMETHING NEW ABOUT READING. NEW PERSPECTIVES OF KNOWLEDGE WITH BIG DATA

This paper is about reading, considered as «what happens when we read». Different aspects are related to this «to happen when»: the underlying motivations, its mode and its timing, what we read and how satisfied we are and finally its sociability.

These dimensions define the present meaning of reading: they are not only the background of reading, they are not something external, they are the reading experience.

Research in this field in Italy tends to focus on quantitative aspects of behaviour and choice of readers: we know how to segment them, the principal variables, we know how many books they read, what genres they love, which is the channel they prefer and how much they spend.

If we try to investigate into the folds of official statistics, it is not easy to pick up the change of the reading practices in the contest of technological transformations.

On these premises, this paper presents PERCE.READ (Perception Reading in Italy in the contest of social reading), a research project aimed at studying the "context" in which the reading takes place today, brings out the "connections" with other practices. All this is useful for redefining the meaning of reading.

The great quantity of data on the web (big data), left more or less voluntarily by the readers, are the means used for these analyses.

The awaited impact is to turn big data into useful information for the publishing industry. This sector could acquire creative and innovative proposals capable to responding effectively to the challenges of the context in which it operates, especially in terms of reading promotion.