La risorsa di Italiano Standard ad alta variabilità linguistica per misurare la peculiarità di un corpus

Giovanni De Gasperis[1], Pasquale Pavone[2], Sergio Bolasco[3]

[1]Università degli studi dell'Aquila & Embeds – giovanni.degasperis@univaq.it
[2]Università degli studi di Modena e Reggio Emilia – pasquale.pavone@unimore.it
[3]Università di Roma La Sapienza – sergio.bolasco@uniroma1.it

Abstract
In the automatic analysis of texts, the added value due to the availability of statistical-linguistic resources is indisputable, both for the grammatical tagging of the forms of a corpus, and for the extraction of contents according to their over / under use with respect to the occurrences of a frequency lexicon for identifying the peculiar language. To this end, a corpus is built that is able to estimate the frequency of the so-called Italian Standard as a set of various linguistic typologies. This resource, usable in the TaLTaC software, is of such size as to lend itself to multiple use, both as a whole and in its individual types, each measurable in itself. The first part of the work describes the composition of the lexicon obtained from the corpus. In the second, the resource is tested with respect to a collection of tweets on Russia's war in Ukraine, measuring its specific thematic peculiarity.

Keywords: Standard Italian; frequency lexicons; peculiarity; war in Ukraine

Riassunto
Nell'analisi automatica dei testi è indiscutibile il valore aggiunto dovuto alla disponibilità di risorse statistico-linguistiche, sia per il tagging grammaticale delle forme di un corpus, sia per l'estrazione di contenuti in funzione del loro sovra/sotto uso rispetto alle occorrenze di un lessico di frequenza per l'individuazione del linguaggio peculiare. A tal fine si costruisce un corpus in grado di stimare la frequenza del cosiddetto Italiano Standard come insieme di varie tipologie linguistiche. Questa risorsa, utilizzabile nel software TaLTaC, è di dimensioni tali da prestarsi a un utilizzo plurimo, sia nel suo insieme, sia nelle sue singole tipologie, ciascuna misurabile di per sé. Nella prima parte del lavoro si descrive la composizione del lessico ottenuto dal corpus. Nella seconda si sperimenta la risorsa rispetto a una raccolta di tweets sulla guerra della Russia in Ucraina, misurandone la peculiarità tematica specifica.

Parole chiave: Italiano Standard; lessici di frequenza; peculiarità; guerra in Ucraina

1. Introduzione

Le caratteristiche della lingua comune vs speciale, o le differenze fra linguaggi specialistici vs settoriali (Gualdo, 2021) contraddistinguono molti studi dei linguisti negli ultimi decenni. La discussione rimanda spesso agli studi di sociolinguistica di Berruto (1987) sulle dimensioni diafasica e diamesica, che si incrociano nel continuum del linguaggio (ibid.: 21). Infatti, l'asse diamesico che va dal parlato allo scritto - e più in particolare dal parlato-parlato (dialoghi), e via via al parlato-scritto (interviste), allo scritto-parlato (discorsi in pubblico), allo scritto-scritto (letteratura) si incrocia con l'asse diafasico che va dall'italiano-informale (gergale o trascurato) via via fino allo italiano-formale (burocratico, tecnico-scientifico, aulico). Questi livelli sono espressioni dell'ampia variabilità della lingua, come insieme di linguaggio "comune" e linguaggi "specialistici" (settoriali e/o tecnico-scientifici). Ricostruirne un campione significativo - ad alta "bio-diversità" linguistica - può essere molto utile, dal punto di vista metodologico, come risorsa esterna per i processi di analisi automatica dei testi. L'idea alla base di questo lavoro è quella di costruire una risorsa statistico-linguistica[1] - che potremmo convenzionalmente definire di Italiano Standard (d'ora in poi istd) - di dimensioni tali da prestarsi a un utilizzo plurimo: sia come insieme delle suddette tipologie, sia come sub-lessico specifico di una o più di queste, ciascuna misurabile di per sé. Ciò è possibile in quanto la raccolta delle singole fonti è sufficientemente ampia per esprimere una stabilità della frequenza delle parole considerate, come si può osservare in tabella 1. La risorsa si applicherà, nel software TaLTaC, come lessico di frequenza su un corpus di tweets, al fine di estrarne la peculiarità specifica dei suoi contenuti, a seconda della tipologia prescelta.

2. Le fonti utilizzate per costruire la risorsa

La raccolta di testi compresi nel corpus, che chiameremo "abcd", composto per costruire la risorsa di istd è riconducibile a 8 diversi "Generi" della lingua, che qui di seguito, sommariamente, descriviamo secondo le loro 112 fonti utilizzate: A1 - Parlato: varie tipologie di lingua parlata (corpus del lessico lip di De Mauro *et al*., 1993); corpus per il nuovo vocabolario di base (nvdb) di De Mauro[2]; trasmissioni radio e Tv; focus groups; copioni di film, teatro e serie tv; interviste di vario tipo su argomenti quali: giovani, medicina, sanità, psicologia, scuola, cittadini, mass media; interviste pubblicate sui giornali a personaggi e celebrità;

[1] Per riferimenti ad archivi, corpora e lessici di frequenza vedi le fonti disponibili all'Accademia della Crusca in *https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228*; per un esempio di risorsa multilingue, "illimitata" e perennemente attualizzata cfr. Navigli e Ponzetto, 2012.

[2] Chiari e De Mauro, 2010, p. 29-31.

B2 - Tweets: una metà su temi vari, l'altra metà su tre argomenti: Unione Europea, petrolio in Basilicata, violenza alle donne;
B3 - Comunicazione Mediata dal Computer: e-mail, pagine Facebook, news group, forum di insegnanti, reclami di cittadini, annunci di lavoro, sms da Repubblica;
C4 - Narrativa: letteratura italiana, letteratura straniera, 100 romanzi vincitori del Premio Strega, raccolte di poesie, diari di viaggio, storie di vita, racconti per ragazzi di Rodari, componimenti di alunni delle elementari;
D5 - Stampa: articoli di Repubblica (annata 1988, stralcio delle annate utilizzate nel lessico rep90[3], articoli degli anni 2020 e 2021, fascicoli Robinson di cultura), La Stampa, Corriere della Sera, Sole 24 Ore, Le Monde Diplomatique, L'Osservatore Romano; articoli su tematiche specifiche (lavoro, immigrazione, elettrosmog, Libia, Covid);
E6 - Politica, Economia, Religione: discorsi parlamentari, discorsi programmatici di governo, lessico economico-finanziario (lef[4]), discorsi di Confindustria, lessico bancario, cronache di giornate borsistiche, encicliche papali, omelie;
E7 - Linguaggio Tecnico-Scientifico: testi di linguistica, statistica, informatica, didattica, geografia, enogastronomia e turismo;
E8 – Linguaggio giuridico, Report e Documenti: sentenze delle Corti di Appello, dei Conti e della Cassazione; delibere dell'agcm; Costituzioni Italiana ed Europea, Diritti dell'uomo; relazioni annuali del cnr, rapporti del censis, isfol e inea; bandi di gara di appalti, cartelle cliniche.
Gli 8 Generi, declinati nelle 112 fonti, nel loro insieme rappresentano la suddetta "bio-diversità" e possono raggrupparsi in 5 Tipi: A) Parlato, B) Web, C) Narrativa, D) Stampa, E) Linguaggi specialistici (politico, tecnico-scientifico, giuridico), come evidenziato dalla lettera che precede l'enumerazione dei Generi. Per la loro consistenza si veda la Tab. 1.
Il corpus "abcd", risultato della raccolta dei testi, produce un file di 911 mb in codifica utf8, formato da 60.000 frammenti (porzioni di testo), per un totale di 147,4 milioni di occorrenze, con un vocabolario di 1.535.060 forme grafiche.

3. La costruzione di risorse statistico-linguistiche dell'Italiano Standard

L'insieme di parole appartenenti al nostro lessico di Italiano Standard (istd_Forme) è limitato alle forme grafiche di ogni categoria grammaticale con almeno 5 occorrenze, previa una pulizia del vocabolario di base prodotto dal parsing del corpus "abcd" secondo i seguenti criteri: eliminazione di forme

[3] Bolasco, 2013, p. 261-264.
[4] Canzonetti, 2004, p. 337-350.

grafiche di numeri in cifre, di parole straniere (eccetto quelle correnti nella lingua italiana e presenti nella risorsa linguistica di TaLTaC, ad es. week, woman), di forme grafiche più lunghe di 25 caratteri (ad esempio, parole strillate nel web: NOOOOOOOOOOOO… oppure stringhe di caratteri qualsiasi disposti in sequenza: ARCpwdse823GEzxFbSOgC3kwceQtL…). Al contrario, sono comprese nella risorsa alcune parole con grafia errata, poiché riflettono errori assai frequenti (usuali nel web o nel parlato trascritto) come le parole terminanti con accento (perchè, ventitrè, quest'ultima anche nella forma ventitre). Infine il lessico istd_Forme contiene 600 parole senza categoria grammaticale[5] dovute a espressioni inusuali (vossignoria, contessina), parole tecniche dialettali (tajarin, bonet), forme arcaiche o poetiche (potea, facea), forme tronche (tien), parole con trattino (socio-professionale, salva-Italia, tam-tam). Un'altra risorsa di Italiano Standard, derivabile dalla precedente, è relativa ai verbi espressi come lemmi. Infatti, data l'elevata flessionabilità di un verbo (44 forme flesse in media per l'italiano), per considerare il suo peso in occorrenze è opportuno cumulare le sue flessioni con almeno 2 occorrenze. Faranno parte di questa risorsa - denominata istd_Verbi - i lemmi verbali con almeno 5 occorrenze[6]. Sempre in Tab. 1 è riportata la consistenza quantitativa totale e delle sue parti.

3.1. Estensione della risorsa ISTD_Forme alle Named Entities

Seguendo i principi dell'Italiano dell'Uso di De Mauro[7] sono state incluse nel lessico istd_Forme alcune entità nominali (Named Entities), anche "multiwords", quali: Nomi propri [Giovanni, Maria Teresa], Toponimi [Italia, New York], Celebrità [Draghi, Putin], Sigle e Acronimi [UE, Covid], Società [Alitalia, Buitoni], Giornali [Repubblica, Stampa]). Le Named Entities sono identificabili sia in fase di normalizzazione in TaLTaC[8], sia grazie all'alta frequenza nel corpus "abcd" di forme grafiche con iniziale Maiuscola. Questa scelta ha prodotto l'inserimento di oltre 20.000 termini con almeno 10 occorrenze. Peraltro, sono state escluse le forme con iniziale Maiuscola di parole comuni identificate grammaticalmente, in quanto hanno una quantità di occorrenze trascurabile rispetto alla stessa forma minuscola (generalmente compresa tra l'1% e il 5% di quest'ultime).

[5] L'assenza di categoria grammaticale è dovuta al fatto che tali forme non esistono nel dizionario di TaLTaC, utilizzato per il tagging grammaticale.

[6] Si noti che considerato un lemma verbale, ad es.<*capire*>, le sue flessioni in *ISTD_Forme* sono 51 (fino a 5 occ.), mentre nella risorsa *ISTD_Verbi* sono 63 perché concorrono al lemma anche le flessioni a soglia 2 occ.

[7] De Mauro (1999), Vol. I, Introduzione § 25 e § 30, pp. XXXVII, XXXIX, XL.

[8] Grazie al dizionario CUCS (Bolasco, 2013, p. 79-81).

Tabella 1. Consistenza di occorrenze - secondo gli 8 Generi e i 5 Tipi - del Corpus ABCD e delle Risorse ISTD_Forme e ISTD_Verbi

		CORPUS "*ABCD*"		RISORSE "*ISTD*"			
		FORME GRAFICHE totali *		FORME GRAFICHE totali **		LEMMI dei VERBI ***	
GENERE	TIPO	GENERE	TIPO (mln)	GENERE	TIPO (mln)	GENERE	TIPO (mln)
1_Parlato	A_Parlato	11.472.228	11,5	10.592.711	10,6	1.682.931	1,7
2_Tweets	B_Web	4.577.516	15,5	3.276.375	12,1	400.748	1,5
3 - Comunicazione Mediata dal Computer		10.962.014		8.882.641		1.102.855	
4_Narrativa	C_Narrativa	37.409.101	37,4	34.660.343	34,7	5.031.991	5,0
5_Stampa	D_Stampa	50.537.610	50,5	46.148.136	46,1	5.151.021	5,1
6_Linguaggio Politico & Economico & Ecclesiale	E_Linguaggi Specialistici	14.686.818	32,4	14.027.042	29,9	1.536.422	2,7
7_Linguaggio Tecnico-Scientifico		8.528.383		7.736.022		637.281	
8_Linguaggio Giuridico & Reports & Documenti		9.192.848		8.085.811		575.203	
		147.366.518		133.409.081		14.435.521	
		** valori dovuti a 1.535.060 forme totali*		*** valori dovuti a 175.980 forme con almeno 5 occ*		**** valori dovuti a 5.512 lemmi con almeno 5 occ*	

3.2. Il continuum delle diverse tipologie linguistiche

Una misurazione dell'intreccio dimensionale diam/diaf del lessico istd_Forme si può ottenere considerando una matrice di dati con in riga i primi 47mila records della risorsa[9] e in colonna le loro sub-occorrenze per Generi e Tipi. Applicando un'analisi in componenti principali (acp) si ottiene il risultato illustrato nel Grafico 1, in cui sulla seconda componente principale si osserva l'ordinamento A-B-C-D-E dei 5 Tipi, fedelmente descritti dai loro Generi, in un continuum[10] dal linguaggio più informale (tweets) a quello più formale (linguaggio giuridico). L'intreccio 3_cmc con 4_narr è giustificato dalla presenza in cmc di parlato-scritto. Il risultato d'insieme dell'acp prova la coerenza della raccolta dei testi del corpus.

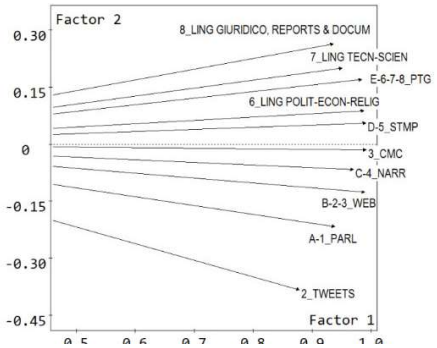

Figura 1. Rappresentazione sul piano fattoriale degli 8 Generi e dei 5 Tipi secondo l'articolazione diafasico-diamesica del continuum della lingua

[9] Le parole con almeno 100 occorrenze, pari alle fasce della legge di Zipf [Alte + Medie + primi decili delle Basse frequenze], (Bolasco, 2013, p. 209-211) assicurano nel nostro caso una copertura del 88% del corpus.

[10] Berruto (2005), p. 127-132.

4. Applicazione

Uno fra i principali obiettivi su come utilizzare le due risorse dell'ISTD è lo studio della peculariità[11] del linguaggio di un corpus sottoposto ad analisi. Questo avviene calcolando uno scarto standardizzato fra le occorrenze normalizzate del corpus e quelle della risorsa, assunta a modello. Nella fattispecie s'intende qui studiare il linguaggio di un corpus di tweets raccolti nei primi giorni di guerra della Russia in Ucraina (24-28 febbraio 2022). L'estrazione dei tweets è stata fatta attraverso le seguenti parole chiave: guerra-Ucraina, Nato-Russia, civili, invasione e profughi, popolazione e armi per un totale di 61.730 messaggi (solo "post"). Il corpus, TW_Guerra, di 11MB, analizzato con TaLTaC 4.0, è stato sottoposto alla risorsa ISTD_forme con l'obiettivo di misurare la peculiarità dei tweets rispetto a due tipologie: "Narrativa" e "Stampa", evidenziandone contenuti differenti. Per ottimizzare la descrizione del risultato, per ciascuna tipologia si riportano i contenuti estratti, citando le radici delle parole selezionate, dai valori più elevati dello scarto standardizzato via via a decrescere. Questa scelta (a parità di spazio) massimizza i contenuti, poiché si cumulano le varianti del sing/plur, femm/masc e delle flessioni verbali. Naturalmente ai primi posti di queste graduatorie rispetto alla risorsa totale o in qualunque sua tipologia - al di là dei nomi dei protagonisti Ucraina, Russia, Putin, Zelensky, Mosca, Kiev, Donbass, Biden, Macron, Cremlino - troviamo il lessico della guerra, che è scontato:

<ucrain, russ,[12] bombard, missil, arm, sanzion, invas, profug, rifug, milit, civil, pace, nuclear, mondial, conflitt, attacc, esercit, aiut, vittim, propagand, internaz … >.

Al contrario, gli scarti maggiori (sempre decrescenti) rispetto alla Narrativa sono relativi a:

- figure (popol, person, uomin, donne, bambin, gent, cittad, famigl)
- situazioni (casa, piazz, foto, immagin, pandem, solidar, uman, accord, appell, folli, resist, assed, aggress, crimin, difend, nazis, sostegn, negozia, diplomaz, Onu, america, ripud, territor, pazz, colpi, lavor, cuore, preghier, risch, occident, accogl, strad, ritard, escal, violen, pericol, protest, responsab, tavolo, allea, regime, nemic, …)
- percezione della condizione di vita (consegu, caus, ragion, rispost, problem, comunit, democr, interess, neutra, ideal, diffic, amic, sicurezz, …).

Considerando invece la peculiarità rispetto alla Stampa, questa risulta più sensibile a:

- figure (bambin, soldat, innocen)

[11] Bolasco (2013), p. 135-141.

[12] Ad esempio gli scarti decrescenti sono: *ucrain* |*i* 718, |*a* 291, |*o* 167, |*e* 122; *russ* |*i* 93, |*o* 47, |*e* 45, |*a* 38.

- toponimi (UE, Odessa, Bielorussia, Polonia, Crimea, Europa, Nato, Italia, Romania, Afghanistan, Finlandia, Occidente, Slovacchia, …)
- azioni e atteggiamenti (scapp, ripud, accogl, invad, fake, inermi, fuga, aiut, video, media, negozia, ucci, ammazz, crepare, massacr, genocidio, pazzo, dittator, pagliaccio, vergogn, fucili, esplos, tank, truppe, resist, ipocris, emergenz, ...)
- linguaggio sboccato (cazz, incazz, merd, schif, culo, caga, coglion, stronz, freg, fott, puttan, …). Quest'ultimo aspetto comprende circa 300 forme, di cui 104 a soglia 3 per un totale di 3.076 occorrenze.

Per quanto riguarda le forme "originali"[13] un esempio per tutte sono i "derivati" di Putin (6.420 occorrenze e peculiarità 383,16 rispetto al totale della risorsa)[14] quali: #Putin (2.246 occ), #PutinWarCriminal (186), #PutinHitler (113), putinian, filo-putin*, putinist, anti-putinian e altre 150 forme grafiche, aventi tutte il valore convenzionale 999.999 dello scarto "non calcolabile".

5. Aspetti computazionali per la misurazione della peculiarità

Il calcolo dello scarto è stato codificato in TaLTaC 4.0[15] in linguaggio Python puro, facendo a meno dell'utilizzo delle librerie di estensione numeriche, contando esclusivamente sull'utilizzo ottimizzato di indici dinamici implementati tramite dizionari e liste connesse ai valori delle variabili di classificazione dei frammenti. Lo scarto standardizzato è calcolato a valle del tagging grammaticale deterministico, confrontando le occorrenze relative della forma nel corpus TW_Guerra a quelle della risorsa statistico-linguistica istd corrispondente alla tipologia prescelta. Il confronto avviene solo se le forme grafiche del corpus dei tweets hanno la stessa categoria grammaticale espressa nella risorsa linguistica dell'italiano (diztal) o nel dizionario di normalizzazione (cucs) delle Named Entities. In particolare, sono state implementate ulteriori ottimizzazioni riguardanti la normalizzazione tramite cucs e le estrazioni delle MultiWord tramite un efficiente algoritmo di individuazione delle concordanze, usando come pivot la prima forma grafica e determinando le forme grafiche successive tramite gli indici di posizionamento dei token nei frammenti. Quest'ultima scelta ha evitato di eseguire ricerche inefficienti sull'intero vocabolario o l'intero corpus. In questo modo si sono potuti ottenere in tempi ragionevoli i risultati delle analisi riportate nei paragrafi precedenti.

[13] Ossia le forme presenti nel corpus e non nella risorsa, che esprimono il massimo della peculiarità.

[14] La parola *Putin* ha un valore dello scarto dalla **Narrativa** pari a 8.036,13 e dalla **Stampa** pari a 504,27.

[15] Per l'evoluzione di TaLTaC cfr. Bolasco (2010), Bolasco e De Gasperis (2017), e il sito https://www.taltac.com

Riferimenti bibliografici

Berruto G. (1987). *Sociolinguistica dell'italiano contemporaneo*. La Nuova Italia Scientifica.
Berruto G. (2005). *Fondamenti di sociolinguistica*. Ed. Laterza.
Bolasco S. (2010). Taltac 2.10 Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi. Led Edizioni Universitarie.
Bolasco S. (2013). *L'analisi automatica dei testi*. Carocci Editore.
Bolasco, S. e Gasperis, G. D. (2017). Taltac 3.0. a multi-level web platform for textual big data in the social sciences. In *Data Science and Social Research* (pp. 97-103). Springer, Cham.
Bolasco S. e Pavone P. (2010). Automatic Dictionary and Rule-Based Systems for Extracting Information from Text. In Palumbo, F. *et al*. (eds.) "*Data Analysis and Classification*" 6th Conference of the Classification and Data Analysis, Springer Publ., pp. 189-198.
Canzonetti A. (2004). *La variabilità dei poliformi nel lessico di frequenza del linguaggo economico-finanziario*. Annali del Dipartimento di Studi Geoeconomici Linguistici Statistici Storici per l'Analisi Regionale 2003-2004. Pàtron Editore.
Chiari I. e De Mauro T. (2010). The new basic vocabulary of Italian: problems and methods. Statistica Applicata – *Italian Journal of Applied Statistics*, vol. 22 (1): 23-37.
De Mauro T., Mancini F., Vedovelli M. and Voghera M. (1993). *Lessico di frequenza dell'italiano parlato*. Etaslibri.
De Mauro T. (1999). *Grande Dizionario Italiano dell'Uso*. UTET, 8 Volumi.
Gualdo R. (2021). *Introduzione ai linguaggi specialistici.* Carocci Editore.
Navigli R. e Ponzetto S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, (193): 217-250.