

ELENA PEPPONI

IL CORPUS MIGRAZIONI NEL PROGETTO *LISDIGIO*:  
CONCEZIONE, DESIGN, SVILUPPO E  
PRIME INDAGINI LINGUISTICHE

1. INTRODUZIONE

All'interno del PRIN 2022 *LiSDiGio* (*Lingua e storia della discriminazione sui giornali italiani*) una delle direttrici tematiche è stata quella che ha riguardato le persone migranti o provenienti dall'estero per motivi politici, sociali, ambientali. L'Italia è infatti tradizionale terreno di migrazione di popoli sin dagli albori della sua storia. L'intento del progetto PRIN nella sua totalità è quello di mappare la discriminazione perpetrata sulla stampa italiana dagli anni Sessanta agli anni Novanta del Novecento a danno di alcune categorie sociali ritenute fragili: non ci si poteva pertanto esimere dal considerare le migrazioni come uno dei terreni più fertili per analizzare un certo tipo di narrazione discriminante.

Per realizzare questa ricerca è stato necessario costruire tre diversi corpora che mettessero insieme articoli di giornale provenienti dalle principali testate generaliste quotidiane d'Italia, ognuno avente per oggetto uno dei tre filoni tematici inseriti nel progetto stesso: la discriminazione nei confronti delle donne nel mondo del lavoro, l'emarginazione delle persone con HIV/AIDS e il razzismo ai danni delle persone migranti.

L'intenzione di questo contributo è dunque quella di mostrare come, nell'unità di ricerca di Cagliari, è stato concepito e realizzato il corpus *Migrazioni*.<sup>1</sup> Prima di tutto, nel lavoro verranno resi chiari i criteri di selezione delle testate e di individuazione delle *query* di ricerca, che hanno permesso un primo isolamento degli eventi storici principali attorno ai quali si è accentrata la ricerca di articoli; verranno specificati anche i successivi momenti di raffinamento e ampliamento delle parole chiave. In secondo luogo, si esplicherà il lavoro di preparazione dei testi per il trattamento informatico, nonché di bilanciamento dei sottocorpora relativi alle varie testate giornalistiche scelte.<sup>2</sup> Infine, si proporranno alcune preliminari indagini linguistiche effettuate sul corpus stesso.

## 2. DESIGN DEL CORPUS

### 2.1 Individuazione delle testate e delle prime query di ricerca

Il primo passo per la costruzione del corpus *Migrazioni* è stato quello di individuare le testate da cui estrarre gli articoli d'interesse. L'attenzione si è focalizzata sulla stampa allo stesso tempo quotidiana e generalista, escludendo quindi sia i periodici sia le testate giornaliere con una diffusione territoriale troppo ristretta. Sono dunque entrati nel computo i seguenti quotidiani: *Corriere della Sera*, *La Repubblica*, *La Stampa*, *Il Messaggero*. Tutte queste testate hanno un orizzonte nazionale, hanno importanti tirature e possiedono archivi storici piuttosto accurati che hanno consentito il reperimento delle informazioni e la loro organizzazione attraverso filtri quali l'arco temporale o la parola chiave.

A essi si è aggiunto un quotidiano particolare come *l'Avanti!*, il quale, pur essendo l'organo ufficiale di stampa di uno specifico partito con una tiratura tutto sommato limitata, è stato inserito per la capillarità con cui ha seguito molti dei temi sociali che sono alla base del corpus *Migrazioni* stesso, e in generale del progetto LISDiGIO. Il giornale socialista si è infatti speso in prima linea con campagne stampa a favore della solidarietà e dell'accoglienza degli esuli cileni o dei profughi vietnamiti negli anni Settanta, senza dimenticare le infervorate prese di posizione che proprio dalle colonne dell'*Avanti!* si sono levate nel corso degli anni Ottanta in difesa delle persone migranti dall'Africa, brutalmente sfruttate nelle campagne italiane. Inoltre, il socialista Claudio Martelli, vicepresidente del consiglio in carica dal 1989 al 1992, è stato promotore di un decreto-legge, poi convertito, per la regolamentazione dei flussi

---

1 Non ci si soffermerà qui sulle prospettive teoriche alla base della costruzione dei corpora, per le quali si rinvia alla vasta e nota letteratura in merito. Sul punto cfr. almeno Atkins/Clear/Ostler (1992); Francis (1979); Leech (1991; 2007); Lenci/Montemagni/Pirrelli (2016); McEnergy/Wilson (2001); Rossini Favretti (2000); Sinclair (2004a; 2004b); Tognini Bonelli (2001).

2 Cfr. Biber (1993).

migratori in Italia, altro evento seguito da vicino dalla testata per via della comune appartenenza politica.<sup>3</sup> Per tutte queste ragioni, pur non raggiungendo la diffusione e l'importanza degli altri quotidiani selezionati, l'*Avanti!* è stato comunque inserito nel corpus.

Per quanto riguarda la prima selezione delle *query* di ricerca scelte per trovare gli articoli che componessero il corpus, va sottolineato che, in una fase embrionale, si è proceduto partendo dall'intuizione di chi scrive, cercando soluzioni di respiro più generale possibile. Come fa notare Orrù (2017: 19),

[l]a scelta dei termini di ricerca pone una quantità di problemi teorici e metodologici non indifferente [...]. Qualsiasi scelta fatta a monte di una ricerca scientifica è inevitabilmente influenzata dalla visione del mondo e dalle inclinazioni ideologiche del singolo ricercatore [...]. Tenere in considerazione tali limiti consente allo stesso tempo quella pratica autoriflessiva che può aiutare a tenere una maggiore distanza dal proprio oggetto di analisi e arricchire le proprie intuizioni.

Per questo motivo è stata compiuta una scelta generica ma di buon senso, che ha però comportato qualche doverosa esclusione. Un termine come *stranieri*, ad esempio, se pur in un momento di riflessione preliminare aveva trovato posto tra le possibili parole chiave, è stato eliminato, in quanto portatore di una semantica troppo vaga e spesso ambigua, non esclusivamente circoscrivibile alle migrazioni. Per la stessa ragione, nel restringere i campi di ricerca all'interno degli archivi digitali delle varie testate, sono state espunte alcune categorie tematiche: gli articoli di economia, sport e spettacoli sono stati esclusi in quanto, anche ove avessero presentato i termini target, non avrebbero quasi mai apportato argomenti che fossero di reale beneficio per la costruzione del corpus stesso.

Alla fine della selezione iniziale, dunque, le prime *query* di ordine generale che hanno permesso l'avvio della ricerca sono state: *immigrato, extracomunitario, profugo, rifugiato, clandestino*.

Le *query* sono state inserite negli archivi dei vari giornali con l'aggiunta della cosiddetta *wildcard* finale, ovvero un carattere speciale (in questo caso l'asterisco '\*') che sostituisce il morfema grammaticale di genere e/o numero. In questo modo, tutte le forme flesse di ogni singolo termine possono essere ricercate simultaneamente. Inoltre, all'interno degli archivi della singola testata le *query* sono state aggregate secondo la logica dell'operatore booleano OR. Ogni archivio digitale richiede una forma per attivare l'operatore OR che può essere quella della virgola o del punto e virgola: le parole, cioè, devono essere separate da uno di questi due segni di interpunzione.

---

3 La legge 28 febbraio 1990, n° 39, ampiamente nota come "legge Martelli", rappresenta la conversione in norma del decreto-legge 30 dicembre 1989, n° 416, volto a regolamentare l'ingresso in Italia, la permanenza e le modalità di richiesta dell'asilo politico e dei visti di lavoro da parte delle persone extracomunitarie, esuli o apolidi. Cfr. <<https://www.gazzettaufficiale.it/eli/id/1990/02/28/090G0075/sg>>, ultimo accesso: 20/05/2025.

ne per far sì che il sistema recepisca il comando di interrogare gli archivi restituendo tutti gli articoli in cui almeno una delle *query* è presente in una qualsiasi sua forma flessa per genere o numero.

Dal punto di vista tematico, oltre i già menzionati articoli di sport, economia e spettacoli, dopo una prima ricognizione sono stati esclusi tutti gli articoli che avessero per nucleo i seguenti argomenti:

- migrazioni in paesi terzi rispetto all'Italia, anche qualora esse coinvolgessero i popoli target migrati pure nel nostro Paese;
- eventi minuti di microcriminalità (risse, furti, etc.) a meno che non sfociassero apertamente in reazioni razziste di rilievo nazionale;
- opinioni e questioni a proposito di paesi terzi rispetto all'Italia.

In sostanza, tutti gli articoli selezionati riguardano la macroarea tematica delle migrazioni con focus sull'Italia come paese di arrivo e con motivazioni di partenza di tipo economico, sociale, di guerra o dittatura, di discriminazione etnica o religiosa.

Nei seguenti paragrafi si spiegherà come sono state approfondite le parole chiave per la ricerca e come sono stati poi bilanciati i vari sottocorpora corrispondenti alle diverse testate indagate.

## 2.2 Approfondimento delle *query* e raccolta dei testi

Dopo aver individuato le prime *query* di ricerca alquanto generiche, sono stati fatti dei primi tentativi di *information retrieval* sui portali delle testate.

La prima testata indagata in ordine cronologico è stata l'*Avanti!*. L'amplificatore delle battaglie culturali e sociali del PSI conobbe una stagione di rinnovato splendore tra gli anni Settanta e gli anni Ottanta: ciò fu dovuto anche alla vicinanza con lo stesso segretario del partito, Bettino Craxi, diretto contributore della testata sotto lo pseudonimo *Ghino di Tacco*. In quel periodo la tiratura arriverà a toccare e superare anche le 50.000 copie (al 1982).<sup>4</sup> Tale epoca fertile si interromperà bruscamente nel 1993 a causa della scomparsa del PSI in seguito alla dissoluzione della Prima repubblica.<sup>5</sup> L'*Avanti!* ha restituito 190 articoli sui temi delle migrazioni, distribuiti tra il 1967 e il 1993.

La seconda testata esplorata è stata *La Stampa*, che, per la vastità del suo archivio, ha permesso l'affinamento della metodologia di ricerca, l'isolamento dei principali eventi storici cardine attorno ai quali appuntare l'attenzione, nonché l'ampliamento massiccio delle *query*. *La Stampa* ha mantenuto, nel periodo compreso tra il 1975 e il

<sup>4</sup> Di qui in avanti, tutte le affermazioni sulle tirature delle singole testate avranno come fonte l'aDS (Accertamento Diffusione Stampa). I dati storici sono disponibili al seguente indirizzo: <https://www.adsnotizie.it/Dati/DatiStorici#>, ultimo accesso: 20/05/2025.

<sup>5</sup> Per maggiori dettagli sulla storia dell'*Avanti* cfr. Abruzzese (1994) e Intini (2012).

1995, una tiratura media attorno alle 400.000 copie<sup>6</sup>. La banca dati è particolarmente ricca di articoli – e il tema delle migrazioni non fa eccezione – in quanto, oltre all'edizione nazionale, il giornale possiede numerose edizioni locali sparse per il territorio del Piemonte, della Liguria e di quella parte della Lombardia che confina con le due regioni appena menzionate, ovvero la Lomellina e in generale la provincia di Pavia. Per le sue caratteristiche strutturali, *La Stampa* è stata la testata che ha permesso una mappatura a tappeto dei momenti storici salienti riguardanti la narrazione discriminatoria a danno delle persone migranti. Gli eventi fondamentali attorno ai quali si è addensata la ricerca, dunque, sono stati i seguenti:

- arrivo degli studenti greci in Italia e poi dei greci in fuga dal regime dei colonnelli (1967-1977);
- questione dei rifugiati politici cileni (1973-1975);
- salvataggio e accoglienza dei profughi vietnamiti (1978-1979);
- arrivo dei migranti economici dall'Africa, dall'Asia e dal Medioriente (anni Ottanta);
- arrivo dei rifugiati politici dai paesi del blocco sovietico (anni Ottanta);
- arrivo dei rifugiati politici dalla Jugoslavia in dissoluzione (dal 1991);
- esodo di massa dall'Albania (dal 1990 al 1994).

*La Stampa* ha restituito 1883 articoli compresi tra il 1960 e il 1994. Una volta individuati i momenti storici importanti, una nuova ricerca mirata è stata condotta a posteriori sull'*Avanti!*, ma senza produrre ulteriori frutti rispetto ai 190 testi già trovati. La ricerca su *La Stampa* ha consentito peraltro di ampliare moltissimo le *query*, inserendo anche quelle di orientamento più etnico che lavorassero in combinazione con quelle generiche per la raccolta di materiale negli archivi, e che sono state utilizzate anche nelle altre testate. A *immigrato, extracomunitario, profugo, rifugiato, clandestino* si sono quindi aggiunte almeno *greco, cileno, vietnamita, marocchino, slavo*,<sup>7</sup> *albanese*. Anch'esse sono state inserite negli archivi digitali aggregate con l'operatore OR e provviste di *wildcard* finale (“\*”).

La terza testata indagata è stata *La Repubblica*, uno dei principali quotidiani italiani, con una tiratura media che, nella seconda metà degli anni Ottanta e negli anni Novanta, si è aggirata attorno alle 600.000 copie.<sup>8</sup> Tale quotidiano ha, però, la caratteristica di restituire testi solo a partire dal 1976, anno della sua fondazione, molto più tardiva rispetto alle altre testate. Sono quindi esclusi tutti i temi compresi nell'arco

---

6 Cfr. dati dell'accertamento Diffusione Stampa, ultimo accesso: 20/05/2025..

7 *Marocchino* e *slavo* erano gli pseudo-etnonimi generici con i quali venivano rispettivamente individuate le persone migranti dal Medioriente e dall'Africa (quest'ultime talvolta definite anche con i termini spregiativi *nero/negro*) e quelle provenienti da paesi dell'ex blocco Sovietico e della ex Jugoslavia.

8 Cfr. dati dell'accertamento Diffusione Stampa, ultimo accesso: 20/05/2025.

temporale precedente. *La Repubblica* ha fornito, tra il 1976 e il 1994, 392 testi.

Il primo e più importante quotidiano italiano, ovvero il *Corriere della Sera*, fondato esattamente un secolo prima de *La Repubblica* (1876), tra il 1975 e il 1995 ha mantenuto una tiratura media oscillante tra le 600.000 e le 750.000 copie, confermandosi come il quotidiano più stampato e diffuso della penisola.<sup>9</sup> Dai suoi archivi, tra il 1967 e il 1994 sono stati estratti 454 testi, che sono frutto anche dell'esplorazione delle due principali cronache cittadine, ovvero quella di Roma e quella di Milano.

L'ultima testata indagata è stata *Il Messaggero*, quotidiano molto radicato nella città di Roma ma con una buona diffusione nazionale e con una tiratura media che, tra il 1975 e il 1995, ha oscillato tra le 250.000 e le 300.000 copie.<sup>10</sup> Dagli archivi de *Il Messaggero* è stato possibile estrarre 183 testi tra il 1970 e il 1994.

Nella tabella 1 viene mostrata, in sintesi, una panoramica degli articoli estratti per testata e per decennio.

Testata <sup>11</sup>	Anni '60	Anni '70	Anni '80	Anni '90	Totale
AV	2 (1%)	13 (7,3%)	125 (65,28)	50 (26,42%)	190
ST	10 (0,5%)	86 (4,5%)	223 (11,79%)	1564 (83,11%)	1883
RP	0 (0%)	39 (10,79%)	170 (42,2%)	183 (46,92%)	392
CS	19 (4,3%)	107 (23,51%)	148 (32,52%)	180 (39,56%)	454
MS	0 (0%)	28 (15,67%)	60 (32,43%)	95 (51,89%)	183

**Tab. 1.** Ripartizione degli articoli per testata e per decennio con relativa percentuale sul totale

### 2.3 Rappresentatività e bilanciamento dei sottocorpora e preparazione dei testi per il trattamento informatico

Finora abbiamo visto quanti testi sono stati reperiti dalle varie testate indagate, per così dire, “allo stato grezzo” della materia. Le intuizioni di chi scrive riguardo alle *query* di ricerca da scandagliare, più i suggerimenti che di volta in volta sono giunti dai testi stessi, permettendo di ampliare il parco di parole chiave usate, rappresentano una prospettiva sicuramente vasta ma del tutto soggettiva, e come tale limitata per sua stessa natura. Inoltre, anche alcune delle *query* più significative, che hanno permesso la restituzione di tanto materiale, hanno necessitato di rodaggio e pulizia di risultati talvolta ambigui o del tutto non pertinenti. Il termine *immigrato*, per esempio, consentirà di trovare molto materiale utile sugli archivi dei giornali a partire

<sup>9</sup> Cfr. dati dell'accertamento Diffusione Stampa, ultimo accesso: 20/05/2025.

<sup>10</sup> Cfr. dati dell'accertamento Diffusione Stampa, ultimo accesso: 20/05/2025.

<sup>11</sup> Uso qui, per la denominazione delle testate, il medesimo sistema di abbreviazioni scelto da Orrù nel suo lavoro all'interno di questo volume.

soltanto dagli anni Ottanta. Fino alla fine dei Settanta, infatti, specialmente in una testata come *La Stampa* nelle sue innumerevoli edizioni locali, esso era regolarmente impiegato per riferirsi al variegato mondo delle famiglie di origine meridionale stabilitesi a vivere al Nord Italia, in particolare nelle aree urbane di Piemonte e Liguria. Dunque un articolo dall'ipotetico titolo *Torino, rissa tra famiglie di immigrati* assume un significato completamente diverso se datato 1960 o 1985: ciò ha richiesto una rigida selezione manuale del materiale.

Nonostante queste premesse, come si è visto, il materiale reperito è stato piuttosto corposo, con più di 3000 testi scaricati e archiviati. Non sarà però sfuggita l'enorme sproporzione de *La Stampa* rispetto alle altre testate. Per tale motivo, per la costruzione finale del corpus è stato necessario un lavoro di bilanciamento del sottocorpus relativo a quella testata rispetto agli altri, per ricondurre il materiale testuale a una dimensione che fosse realisticamente rappresentativa del vero impatto de *La Stampa* nella narrazione discriminatoria italiana sulle persone migranti. Viceversa, il corpus *Migrazioni* sarebbe stato così tanto sbilanciato a favore de *La Stampa* stessa che sarebbe stato impossibile mappare correttamente i fenomeni discriminatori, portando a un'analisi, almeno in parte, viziata.

Come fa notare Sinclair (1991: 13), «[t]he beginning of any corpus study is the creation of the corpus itself. The decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently».

Volendo quindi assumere decisioni che fossero il più possibile realistiche e non condizionassero troppo lo sviluppo della ricerca, si è deciso di accogliere all'interno del corpus *Migrazioni* tutti i testi estratti da *Corriere della Sera*, *La Repubblica*, *Messaggero* e *Avanti!*, in quanto il loro numero totale rispecchia piuttosto fedelmente l'impatto che la singola testata ha avuto, nel periodo storico indagato, sulla stampa quotidiana nazionale italiana sia in termini di tiratura sia in termini di lettori giornalieri. Il materiale de *La Stampa*, invece, è stato ridotto riconducendo gli articoli estratti al rapporto con la sua tiratura. Il *Corriere della Sera*, in quanto quotidiano principale di riferimento, ha, come abbiamo visto, una tiratura sempre superiore alle 600.000 copie giornaliere, e ha prodotto poco più di 450 testi; il *Messaggero*, che può competere più o meno con *La Stampa* in termini di pregnanza locale sul territorio, si assesta all'incirca attorno alle 250.000 copie e ha prodotto 183 testi. Si è scelto di prendere questi due estremi di riferimento e di selezionare da *La Stampa*, con le sue circa 400.000 copie di tiratura, la cifra di 400 articoli. La quantità è sicuramente sbilanciata verso quella del *Corriere* a fronte di una tiratura visibilmente inferiore, ma la scelta è giustificata dal fatto che bisognava in qualche modo rappresentare in maniera credibile l'ingente apporto del sottocorpus *La Stampa* rispetto a quello dato da tutte le altre testate. All'interno dei 400 testi, si è scelto di rappresentare i decenni – ovvero anni Sessanta, Settanta, Ottanta e Novanta – con una riproduzione bilanciata del

loro impatto percentuale sul totale.<sup>12</sup> Una volta messi a punto i criteri di selezione, gli articoli sono stati estratti con una procedura di campionamento casuale stratificato, della quale si dà conto nelle tabelle 2 e 3.

La Stampa	Anni '60	Anni '70	Anni '80	Anni '90	Totale
Articoli	10	86	223	1564	1883
Percentuale	0,5%	4,5%	11,79%	83,11%	100%

**Tab.2.** *Corpus originale*

La Stampa - Bilanciato	Anni '60	Anni '70	Anni '80	Anni '90	Totale
Articoli	10	70	140	180	400
Percentuale	2,5%	17,5%	35%	45%	100%

**Tab.3.** *Corpus bilanciato*

Prima di costituire il corpus definitivo, i testi sono dovuti passare attraverso una fase di pulizia e preparazione per il trattamento informatico. Nel caso de *La Stampa* e *La Repubblica* (quest'ultima a partire dal 1984), i testi scaricati dagli archivi digitali erano già in formato .TXT, quindi il lavoro di pulizia ha riguardato solo la correttezza grammaticale, o talvolta semplicemente formale, del contenuto di ogni *file*. Gli articoli de *La Repubblica* dal 1976 al 1983, quelli del *Corriere della Sera*, de *Il Messaggero* e dell'*Avanti!*, invece, erano scaricabili o consultabili in diversi formati non testuali, talvolta reperibili soltanto in analogico.<sup>13</sup> La costruzione del corpus ha perciò richiesto un'intensa fase di lavoro per la trasformazione dei *file* con estensione .PNG o .JPG in documenti di testo attraverso l'uso del *software* OCR ABBYY.

La dimensione definitiva del corpus *Migrazioni* completo e interrogabile è perciò la seguente: 1619 testi per un totale di oltre 1.325.000 *token*.

<sup>12</sup> La riproduzione bilanciata è esclusa per il decennio 1960-1969, in quanto sono stati trovati solo dieci articoli relativi agli anni Sessanta, quindi essi sono stati inseriti nella loro totalità.

<sup>13</sup> È stato il caso degli articoli de *La Repubblica* usciti nel periodo compreso tra il 1976 e il 1983, i quali non erano stati digitalizzati in formato testo e neppure scansionati come immagini sull'archivio digitale del giornale. Essi sono stati consultati nel formato *microfilm* presso la Biblioteca civica Sormani di Milano. Si ringrazia di cuore Paolo Orrù per il prezioso aiuto nella digitalizzazione dei testi de *La Repubblica* dal 1976 al 1979.

### 3. PRIME INDAGINI LESSICALI SUL CORPUS *MIGRAZIONI*

Dopo aver costruito, bilanciato e pre-trattato il corpus, sono state condotte su di esso alcune iniziali analisi lessicali.

La prima indagine svolta è quella che ha riguardato l'estrazione delle parole chiave sia dell'intero corpus *Migrazioni* sia dei sottocorpora corrispondenti a ciascuna testata. Va sottolineato che l'estrazione delle parole chiave è stata operata scegliendo come corpus di controllo *laRepubblica*, una raccolta di testi giornalistici tratti dall'omonimo quotidiano. La raccolta è stata sviluppata all'università di Bologna e si compone di oltre 380 milioni di parole.<sup>14</sup>

Le *keyword* del corpus *Migrazioni* sono state poi selezionate escludendo nomi propri, parole grammaticali, parole flesse, numerali e non-parole, e sono state estratte con il tool *Wordsmith*. Sono stati invece conservati i toponimi in quanto particolarmente significativi per la narrazione delle migrazioni in Italia. L'approccio metodologico privilegiato per l'analisi delle *keywords* è stato il cosiddetto «approccio *exploratory*» basato sul concetto di salienza (*keyness*). La salienza delle parole chiave, come fa notare Orrù (2022: 176) «viene determinata attraverso un rapporto statistico tra la frequenza in un dato corpus e quella in un corpus di riferimento più generale della lingua». L'approccio esplorativo (*ibidem*) segue dei quesiti di ricerca generici per tracciare una strada che orienterà l'interpretazione dei primi dati rilevati e indirizzerà le fasi seguenti della ricerca. A partire da ciò, è stata quindi generata la lista delle parole chiave che, sottolinea ancora Orrù (*ibidem*), altro non è che «un'indicazione di *aboutness* [...] dei testi, per identificare, cioè, i temi, le idee o i particolari stilistici di un testo o un corpus».

Nella tabella 4 sono visibili le prime cinquanta *keywords* generali del corpus *Migrazioni*.

1	profughi	26	Brindisi
2	immigrati	27	nordafricani
3	albanesi	28	questura
4	stranieri	29	senegalesi
5	extracomunitari	30	tunisini
6	immigrazione	31	nave
7	clandestini	32	cileni
8	rifugiati	33	permesso
9	vietnamiti	34	ambulanti

<sup>14</sup> A tal proposito cfr. Baroni *et al.* (2004).

10	lavoro
11	soggiorno
12	razzismo
13	asilo
14	marocchini
15	sanatoria
16	Caritas
17	Albania
18	lavoratori
19	(vu) cumprà
20	paesi
21	accoglienza
22	esuli
23	africani
24	(di) colore
25	Tirana
35	Vietnam
36	comunità
37	assistenza
38	regolarizzazione
39	emigrazione
40	Polizia
41	ambasciata
42	bordo
43	fenomeno
44	cittadini
45	Durazzo
46	frontiere
47	porto
48	esodo
49	alloggio
50	sistemazione

**Tab. 4.** *Prime 50 parole chiave del corpus Migrazioni in generale ordinate per salienza*

Le parole chiave possono essere categorizzate in diversi sottogruppi, dei quali le prime cinquanta *keywords* rappresentano una panoramica piuttosto fedele:

- termini per descrivere gli immigrati dal punto di vista del loro status giuridico, del loro aspetto fisico o della loro appartenenza (pseudo)etnica (*profughi, immigrati, albanesi, stranieri, extracomunitari, clandestini, rifugiati, vietnamiti, marocchini, lavoratori, vu cumprà, esuli, africani, di colore, nordafricani, senegalesi, tunisini, cileni, ambulanti, cittadini*);
- termini per descrivere concetti o azioni (*immigrazione, lavoro, razzismo, emigrazione, fenomeno, frontiere, esodo*);
- termini giuridico/burocratici (*asilo, sanatoria, accoglienza, permesso, assistenza, regolarizzazione, sistemazione*);
- termini che indicano enti privati o apparati statali preposti alla gestione dei flussi migratori (*Caritas, questura, comunità, Polizia, ambasciata*);
- toponimi italiani o stranieri (*Albania, Tirana, Brindisi, Vietnam, Durazzo*);
- mezzi di trasporto o luoghi di stallo (*nave, bordo, porto, alloggio*).

La prima categorizzazione, in particolar modo, delinea un interessante spaccato delle narrazioni giornalistiche a proposito delle persone migranti negli ultimi trent'anni del XX secolo. Notiamo infatti che molti di questi termini rappresentano

le stesse *query* di ricerca inizialmente intuitive e poi formalizzate per la costruzione del corpus, come ampiamente descritto nei §§ 2.1 e 2.2.

In aggiunta, si notano innanzitutto degli etnonimi (*albanesi, vietnamiti, marocchini, africani, nordafricani, senegalesi, tunisini, cileni*) a diverso grado di realismo e di approfondimento. Infatti, un etnonimo come *cileno* o *vietnamita* indica sicuramente un popolo ben individuabile e la cui circoscrivibilità è universalmente nota sia a chi scrive gli articoli sia al pubblico leggente. Un termine come *nordafricano* o *africano*, invece, è ambiguo, in quanto non delimita una specifica nazionalità – anzi, *africano* si estende addirittura a un intero continente – contribuendo piuttosto ad alimentare una narrazione eurocentrica e razzializzante per cui gli “africani” sono un unico popolo all’interno del quale le singole nazionalità sono indistinguibili. Vanno pertanto presi con un certo grado di astrazione anche etnonimi che mostrano un apparentemente maggiore tasso di specificità, come *marocchini, senegalesi e tunisini*: infatti, anch’essi possono rientrare a pieno titolo negli pseudo-etnonimi, ovvero quei termini di nazionalità sovraestesi a chiunque abbia delle caratteristiche fisiche assimilabili a quelle che, stereotipicamente, sono attribuite all’etnonimo stesso. Non vi è, però, alcuna garanzia che essi, quando impiegati, siano realmente rappresentativi della popolazione che pretendono di individuare. Non è infrequente, infatti, l’impiego di *marocchino* o *tunisino* per individuare qualsiasi persona dalla pelle più o meno scura proveniente da una zona compresa tra il Nord Africa, il Corno d’Africa e il Medio Oriente, senza pressoché alcuna certezza dell’eventuale legame degli individui target con paesi quali la Tunisia o il Marocco.

Si vuole inoltre attirare l’attenzione su almeno un’altra tipologia dei termini descrittivi, ovvero quella che individua le eventuali professioni delle persone migranti. Si noti che al genericismo *lavoratori* si accompagnano il più specifico *ambulanti* e l’evocativo *vu cumprà*, termine assai dispregiativo stante a indicare in particolar modo i venditori abusivi con l’abitudine di mercanteggiare piccoli oggetti di poco valore sulle spiagge o nei centri storici a forte vocazione turistica. Secondo Faloppa (2011: 35) la prima attestazione di *vu’ cumprà* sulla stampa italiana si ha nel 1986,<sup>15</sup> anche se nel nostro corpus non appare prima del 1988. Precisa Faloppa (*ibidem*) che i *vu cumprà* «venivano presentati come icone dell’immigrazione *tout court*: quasi sempre ‘neri’, appunto, spesso ‘clandestini’, di solito marginali [...]. E che – a leggere i quotidiani – avevano invaso, o stavano per invadere, l’intera penisola».

Oltre all’estrazione delle parole chiave generali dell’intero corpus, si è poi proceduto con la stessa metodologia descritta all’estrazione mirata per ogni sottocorpus, cioè per ogni testata analizzata. I risultati delle prime venticinque *keywords* per testata disposte in ordine di salienza sono visibili nella tabella 5.

---

15 Faloppa riprende la datazione proposta dal GRADIT (s. v.).

N.	AV	ST	RP	CS	MS
1	immigrati	profughi	immigrati	profughi	profughi
2	immigrazione	extracomunitari	profughi	immigrati	albanesi
3	stranieri	immigrati	immigrazione	stranieri	immigrati
4	profughi	albanesi	clandestini	albanesi	clandestini
5	rifugiati	stranieri	extracomunitari	vietnamiti	Albania
6	lavoratori	vietnamiti	albanesi	clandestini	Tirana
7	extracomunitari	Caritas	razzismo	extracomunitari	rifugiati
8	paesi	marocchini	rifugiati	(vu) cumprà	stranieri
9	albanesi	lavoro	stranieri	marocchini	cileni
10	Italia	soggiorno	sanatoria	rifugiati	esuli
11	sanatoria	clandestini	asilo	asilo	Brindisi
12	legge	immigrazione	Villa Literno	soggiorno	immigrazione
13	razzismo	nordafricani	lavoro	lavoro	nave
14	lavoro	accoglienza	soggiorno	africani	Villa Literno
15	asilo	Asti	vietnamiti	questura	bordo
16	fenomeno	senegalesi	paesi	sanatoria	Puglia
17	emigrazione	(vu) cumprà	polacchi	Albania	Durazzo
18	regolarizzazione	asilo	marocchini	razzismo	porto
19	cittadini	tunisini	nave	Brindisi	lavoro
20	clandestini	rifugiati	ambasciata	zingari	soggiorno
21	problema	Novara	Albania	(di) colore	Caritas
22	decreto	ambulanti	ambulanti	Vietnam	Otranto
23	diritti	(di) colore	accoglienza	esuli	Cile
24	accoglienza	questura	frontiere	Tirana	mare
25	comunità	africani	(di) colore	senegalesi	vietnamiti

**Tab. 5.** *Comparazione orizzontale delle prime 25 parole chiave di ogni sottocorpus ordinate per salienza*

Per quanto riguarda le parole chiave comparate delle cinque testate, i dati che deduciamo dalla tabella 5 sono interessanti per diverse ragioni. Se, infatti, vediamo molte delle *query* di ricerca ripetersi, ed essere sempre presenti più o meno ai primi posti della lista, vediamo anche che le diverse testate presentano alcune peculiarità proprie.

L'*Avanti!*, per esempio, restituisce talune *keywords* uniche, che non si ritrovano nelle altre testate, come *lavoratori*, *asilo*, *regolarizzazione*, *diritti*. Come già precedentemente sottolineato nel § 2.2, l'*Avanti!* è un quotidiano dalla spiccata coloritura politica, che ha seguito da vicino tutte le fasi della cosiddetta legge-Martelli e che ha sempre avuto una propensione per gli argomenti di solidarietà sociale, fatto che puntualmente si riflette nella sua lista di *keywords*.

È da notare anche la presenza dei toponimi tra le parole chiave delle varie testate. Diverse di esse (*Corriere della Sera*, *La Repubblica*, *Il Messaggero*) hanno, tra le prime venticinque *keywords*, toponimi che rimandano a casi assurdi agli onori delle cronache nazionali, quali *Albania*, *Tirana*, *Durazzo*, *Otranto* o *Villa Literno*.<sup>16</sup> Non così, invece, per *La Stampa*, che presenta toponimi molto specifici e legati a un contesto territoriale più ristretto (*Asti*, *Novara*). In larga parte, ciò dipende dal forte legame della testata con il territorio del Piemonte e dalla sua capacità di coprire con capillarità eventi anche minuti di tipo locale.

Una seconda indagine lessicale svolta sui primi dati del corpus *Migrazioni* ha riguardato i collocati al livello di corpus generale, per meglio comprendere i contesti d'uso di alcuni dei principali termini attorno ai quali si addensa la narrazione discriminatoria delle persone migranti sui giornali italiani. I collocati vengono qui considerati a partire dall'approccio metodologico cosiddetto «*collocation window*», che «esamina quali parole ricorrano frequentemente in un dato spazio di testo (tipicamente 5 parole a destra e a sinistra) rispetto a un dato termine. È possibile identificare in questo modo quali connotazioni vengano a esso attribuite; quali verbi possano essere coinvolti; in quali strutture grammaticali entri» (Orrù 2022: 180-81).

L'algoritmo qui utilizzato è quello che predilige le parole lessicali (LogDice): con questo metodo, si confronta la frequenza della collocazione rispetto a quella delle singole parole che la compongono e si aggiunge un valore standard; tipicamente il valore finale non supera il 10.

---

16 Albania, Tirana, Durazzo e Otranto sono quattro toponimi noti alle cronache nei primi anni Novanta, quando si verificò il cosiddetto “esodo degli Albanesi”, ovvero una migrazione massiccia dalle coste dell'Albania in direzione di quelle pugliesi. Durazzo era il principale porto di partenza, assieme a Valona; Otranto fu, con Brindisi, il centro nevralgico degli arrivi in Italia, corredati da non poche tensioni sociali.

Villa Literno, un piccolo comune rurale in provincia di Caserta, fu invece teatro, nell'agosto 1989, di una brutale aggressione ai danni di alcuni migranti che ivi si trovavano per la stagione della raccolta dei pomodori. Un gruppo di braccianti africani di varia nazionalità, infatti, dormiva in alcune cascine dismesse della campagna casertana, ove una banda di ragazzi italiani armati di mazze e spranghe fece irruzione a scopo di rapina e pestaggio a sfondo razzista. In questo *raid* perse la vita un bracciante, rifugiato politico e studente sudafricano, Jerry Essan Masslo. Il delitto di Masslo rese palesi le condizioni del bracciantato agricolo nelle campagne italiane e scosse l'opinione pubblica a tal punto da provocare la richiesta – e ottenere lo svolgimento – dei funerali di Stato per il giovane sudafricano. Le esequie pubbliche si tennero il 28 agosto 1989 a Roma alla presenza di diverse cariche istituzionali, tra cui l'allora Ministro per gli Affari Esteri Gianni De Michelis.

I termini target scelti rispetto ai quali ricercare i collocati sono stati *immigrato*, *extracomunitario*, *profugo*, *rifugiato*. Per motivi di comparazione orizzontale, i termini target sono stati analizzati solo nelle loro occorrenze come nomi, escludendo quindi i casi in cui occorrono come aggettivi.

Per quanto riguarda *immigrato*, tra i collocati più tipici troviamo *colore*, *extracomunitario*, *clandestino*, *Terzo Mondo*, *straniero*, *Italia*, *numero*, *vivere*, *diritto*, *sanatoria*. Notiamo ai primi posti parole quali *colore*, *extracomunitario* e *clandestino*, la cui preponderante co-occorrenza con *immigrato* si spiega sulla base di una riflessione già precedentemente esperita nel § 2.3. Il termine *immigrato*, infatti, ha per lungo tempo individuato, specie nelle testate giornalistiche aventi come base una città del Nord, le persone provenienti dalle regioni del Sud Italia. Ecco perché, allorquando le famiglie meridionali al Nord sono diventate completamente integrate nel tessuto sociale e il termine ha spostato il suo focus per indicare persone provenienti da altri paesi, è divenuto necessario precisare che non si tratta più di *immigrati* e basta, ma di una categoria particolare di immigrati: quelli *di colore*, *stranieri* o, per meglio dire, *extracomunitari*, cioè letteralmente provenienti da paesi che non fanno parte della Comunità Europea. Tra gli altri collocati, vediamo che molti hanno una semantica neutra (*straniero*, *numero*, *vivere*) se non dichiaratamente positiva, come *diritto* o *sanatoria*.

I collocati più tipici di *extracomunitario* sono *ingresso*, *presenza*, *vivere*, *accoglienza*, *presente*, *comune*, *mila*, *regola*, *numero*, *ghetto*. Notiamo che nel caso di *extracomunitario* non c'è co-occorrenza tipica con gli altri termini target qui analizzati, che quindi non diventano suoi modificatori, come invece talvolta accade all'inverso. Vi è, però, una forte insistenza su temi burocratico-amministrativi quali la conta delle persone (*numero*, *mila*), la regolarizzazione dei flussi migratori (*ingresso*, *presenza*, *Comune*, *regola*) e la necessità di far integrare correttamente queste persone con il tessuto sociale anziché lasciare che vivano in situazioni di marginalità (*accoglienza*, *ghetto*). Notiamo che *extracomunitario* è, tra i termini analizzati, quello i cui collocati insistono di più, semanticamente, sul numero delle persone coinvolte nei flussi di arrivo: i collocati sembrano dirci che serve sapere il loro numero – quanti *mila* sono, appunto – e attrezzarsi di conseguenza, trasmettendo un senso di urgenza.

Giungiamo ora a esaminare la situazione di *profugo*. I collocati sono del tutto diversi rispetto a quelli che si riscontrano per *immigrato* e per *extracomunitario*, in quanto in questo caso sono molto focalizzati sulla nazionalità/etnia: i due più tipici sono infatti *albanese* e *vietnamita*, seguiti da *Vietnam*, *politico*, *Italia*, *accogliere*, *problema*, *ospitare*, *arrivo*, *gruppo*. Notiamo che tra questi collocati gli unici verbi che ci sono, ovvero *accogliere* e *ospitare*, hanno una semantica decisamente positiva. L'unico termine dalla semantica negativa è *problema*, che rimane comunque in minoranza.

Infine, possiamo analizzare i collocati di *rifugiato*. Naturalmente, il modificatore più tipico è *politico*, seguito da *status*, *riconoscimento*, *ONU*, *Nazioni Unite*, *commissariato*, *ambasciata*, *riconosciuto*, *asilo*, *considerato*. Notiamo che qui la semantica di tutti i collocati più tipici ha del tutto a che fare con questioni di diritto sovranazionale

e di protezione internazionale dei rifugiati, nonché con la necessità di accoglierli e di fornire loro strumenti di vita e dignità.

Sulla base di questi dati, possiamo condurre alcune riflessioni.

Innanzitutto, possiamo dire che questi quattro termini fondamentali quando si parla di migrazioni, talvolta erroneamente mescolati o ritenuti sinonimi, in realtà sono dotati di una semantica autonoma che si dispone lungo una scala a decrescente tasso di accettabilità sociale. Al gradino più alto troviamo *rifugiato*. Quello di rifugiato politico è un vero e proprio *status* giuridico riconosciuto a livello globale, per cui vi sono sin dal Secondo dopoguerra delle regole che normano la richiesta di asilo e protezione.<sup>17</sup> Nei confronti delle persone rifugiate vi è pertanto una particolare attenzione istituzionale e una spiccata sensibilità sociale, riscontrabile anche nella semantica dei collocati precedentemente menzionati; essi sembrano voler sottolineare che vi sono dei popoli per i quali, a ogni arrivo di nuove persone, bisognerà provvedere necessariamente alla loro accoglienza e stabilizzazione per consentire loro un fruttuoso inizio di una nuova vita.

Un ragionamento molto simile può essere fatto per *profugo*, che si situa più o meno allo stesso livello di accettabilità di *rifugiato*. La condizione di profugo è prodromica a quella di rifugiato nella misura in cui tutte le persone che scappano da guerre, dittature, colpi di stato o persecuzioni di vario genere sono profughi, che poi si trasformano in richiedenti asilo nei paesi di arrivo e, se la loro domanda viene accolta, diventano ufficialmente rifugiati politici. Le collocazioni di *profugo* esplo- rate ci suggeriscono l'insistenza sugli etnonimi riconoscibili, così come i verbi dalla semantica tutto sommato positiva ci restituiscono un atteggiamento linguistico – e di conseguenza sociale – di apertura nei confronti delle persone individuate da questo termine. A metà strada nella scala di accettabilità delle persone straniere e migranti si situa il termine *immigrato*, forse il più generico e universalmente applicabile dei quattro qui presi in esame. Le collocazioni più tipiche ovvero *di colore*, *extracomunitario*, *clandestino* ci suggeriscono già una presa di posizione nei confronti delle persone individuate dallo stesso termine *immigrato*: le narrazioni sono in questo caso maggiormente improntate al rafforzamento di un *ingroup* e all'esclusione di un *outgroup*, quest'ultimo rappresentato da persone che, per caratteristiche fisiche, di provenienza geografica, sociali o culturali sono evidentemente diverse dal 'noi', e che hanno quindi bisogno di essere circoscritte da termini che sottolineano l'alterità.

---

17 Va sottolineato però che fino al 1990, cioè fino all'entrata in vigore della legge Martelli (vd. *supra*), l'Italia manteneva ancora attiva una clausola di limitazione geografica che aveva conservato all'atto di ratifica della Convenzione di Ginevra: potevano quindi, almeno formalmente, chiedere asilo in Italia solo quei rifugiati provenienti da paesi europei. In realtà, nella prassi quotidiana, anche ben prima della legge Martelli erano state fatte delle eccezioni in favore di gruppi abbastanza contenuti di rifugiati provenienti da altre parti del mondo, come per esempio i cileni del post-golpe di Pinochet e i vietnamiti in fuga da un regime di terrore.

Al gradino più basso della piramide della narrazione giornalistica si trova *extracomunitario*, il termine dal significato più connotato e più semanticamente orientato verso il polo negativo tra i quattro presi in esame. Prima di tutto, come già detto, notiamo che *extracomunitario* non co-occorre con gli altri termini target analizzati, quindi, se esso viene usato come N, non accade che *immigrato*, *profugo* o *rifugiato* siano suoi modificatori. Ciò ci testimonia almeno due tendenze linguistiche. Prima di tutto, negli articoli di giornale, noti per lo stile brillante ma sintetico, si tende a condensare molte informazioni semantiche in poche parole. Un termine come *extracomunitario* funziona dunque perfettamente come incapsulatore semantico di concetti che hanno a che fare con le migrazioni, ma in particolare con il lato delle migrazioni più grigio, svantaggioso e squalificante, e non necessita di ulteriori specificazioni al suo fianco. In secondo luogo, appellare qualcuno *extracomunitario* lo pone immediatamente in una posizione meno vantaggiosa di quella delineata per un *profugo* o un *rifugiato*, che quindi sarebbero eventualmente dei collocati ridondanti.

A partire da tutto ciò, possiamo condurre anche una seconda riflessione. Prendendo qui il modello teorico proposto da Bianchi (2017; 2021), si parlerà della cosiddetta ‘ingiustizia discorsiva’.<sup>18</sup> Secondo Bianchi (2021: 19) «certi individui subiscono forme di oppressione a causa di pregiudizi legati alla loro identità sociale», e in particolare «il fenomeno dell’ingiustizia discorsiva tocca i soggetti nella loro capacità di agire sul mondo con le parole». Nel caso di un gruppo socialmente minoritario o oppresso, come può essere quello rappresentato dalle persone migranti, continua Bianchi (*ivi*: 20 e ss.) si può verificare un fenomeno di ‘riduzione al silenzio’, ovvero incapacità totale dei membri del gruppo oppresso di agire sul mondo con le proprie pratiche discorsive. Se le persone migranti sono disposte lungo un tracciato discendente che tende a separare quelle apertamente accolte (*rifugiati*) da quelle meglio tollerate (*profughi*, alcune categorie di *immigrati*) e soprattutto da quelle poco o per nulla accettate (gli *extracomunitari* e, in una certa misura, anche le tipologie meno integrate di *immigrati*), è chiaro che, a seconda del gradino della scala ove una persona si colloca, avrà più o meno potere di agire sul mondo con le proprie parole, di farsi sentire e di rivendicare condizioni migliori per sé e per altre persone nella stessa situazione.

Insomma, secondo i primi dati raccolti dal corpus *Migrazioni* sembrano emergere delle tendenze linguistiche e discorsive ben precise, che avranno modo di essere approfondite nel prosieguo della ricerca. Per ora, sembra di poter affermare che le narrazioni disponibili sulla stampa italiana a proposito delle persone che per qualsiasi motivo sono migrate nel nostro Paese nell’ultimo trentennio del XX secolo, anche quando non costruite con intenti palesemente razzisti e presenti su testate meno politicamente connotate, sono state spesso intrise di paternalismo e razzializzazione, e

---

18 Per formulare questa teoria, Bianchi si rifà a sua volta a quella delineata nel 2007 da Miranda Fricker a proposito della *ingiustizia epistemica* (cfr. Fricker 2007).

hanno comunque contribuito a mantenere lontano dai soggetti interessati un potere linguistico, e di conseguenza sociale, che essi sono riusciti a conquistare soltanto molto recentemente, nel XXI secolo, e grazie anche a forme di comunicazione alternativa e autogestita, come i *social media*.

#### 4. CONCLUSIONI

Il presente contributo ha illustrato tutte le fasi di creazione del corpus *Migrazioni* all'interno del progetto PRIN *LiSDiGio*, a partire dalla concezione delle *query* di ricerca, passando attraverso il design concreto del corpus stesso e il bilanciamento dei suoi sottocorpora per arrivare al prodotto finale.

In secondo luogo, abbiamo illustrato alcune prime indagini lessicali su questo corpus, mostrando come possono essere utilizzati i dati e quali strade si sono intraprese nel contesto di questo progetto. Possiamo dire che la creazione di questo corpus si pone in continuità con il lavoro avviato da Orrù (2017), che ha inteso mappare la discriminazione linguistica nei confronti delle persone migranti esperita sulla stampa quotidiana nel nuovo millennio, con particolare focus sul decennio 2000-2010. Nulla, invece, era disponibile sugli ultimi trent'anni del Novecento, e si può affermare che questa ricerca sia stata in grado di colmare, almeno dal punto di vista della raccolta quantitativa, tale vuoto.

Questo contributo intende fornire un primo ragguaglio di una ricerca ancora in itinere. I futuri passi del progetto saranno il caricamento di una quota selezionata di articoli su un'interfaccia dedicata, la loro taggatura per fenomeni lessicali e fenomeni discorsivi e la loro pubblicizzazione al fine di incoraggiare altri gruppi di ricerca o altre singole studiose e singoli studiosi all'indagine delle discriminazioni linguistiche nei confronti di persone migranti, donne e soggetti con HIV/AIDS.

#### BIBLIOGRAFIA

- Abruzzese 1994 = Alberto Abruzzese, *La stampa italiana nell'età della TV: 1975-1994*, in Valerio Castronovo / Nicola Tranfaglia (a cura di), *Storia della stampa italiana*, 7 voll. (1976-1994), Roma-Bari, Laterza.
- Atkins/Clear/Ostler 1992 = Sue Atkins / Jeremy Clear / Nicholas Ostler, *Corpus Design Criteria*, in «Literary and Linguistic Computing», 7 (1), pp. 1-16.
- Baroni *et al.* 2004 = Marco Baroni / Silvia Bernardini / Federica Comastri / Lorenzo Piccioni / Alessandra Volpi / Guy Aston / Marco Mazzoleni, *Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian*, in Proceedings of the Fourth International Conference on Language Resources and evaluation (LREC).
- Bianchi 2017 = Claudia Bianchi, *Linguaggio d'odio, autorità e ingiustizia discorsiva*, in «Rivista di

- Estetica», 64, pp. 18-34.  
 (<<https://journals.openedition.org/estetica/2059>>)
- Bianchi 2021 = Claudia Bianchi, *Hate speech. Il lato oscuro del linguaggio*, Roma-Bari, Laterza.
- Biber 1993 = Douglas Biber, *Representativeness in corpus design*, in «Literary and Linguistic Computing», 8 (4), pp. 243-257.
- Faloppa 2011 = Federico Faloppa, *Razzisti a parole (per tacer dei fatti)*, Roma-Bari, Laterza.
- Francis 1979 = Nelson Francis, *Problems of assembling and computerizing large corpora*, in Henning Bergenholtz / Burkhard Schaefer (eds.), *Empirische Textwissenschaft: Ausbau und Auswertung von Text-Corpora*, Königstein, Scriptor, pp. 110-123.
- Fricker 2007 = Miranda Fricker, *epistemic Injustice. Power and the Ethics of Knowing*, Oxford, Oxford University Press.
- Intini 2012 = Ugo Intini, *Avanti!: un giornale, un'epoca*, Roma, PonteSisto.
- Leech 1991 = Geoffrey Leech, *The state of the art in corpus linguistics*, in Karin Aijmer / Bengt Altenberg (eds.), *English corpus linguistics: studies in honor of Jan Svartvik*, London, Routledge, pp. 8-29.
- Leech 2007 = Geoffrey Leech, *New resources, or just better old ones? The Holy Grail of representativeness*, in Marianne Hundt / Nadja Nesselhauf / Carolin Biewer (eds.), *Corpus Linguistics and the Web*, Amsterdam, Rodopi, pp. 133-149.
- Lenci/Montemagni/Pirrelli 2016 = Alessandro Lenci / Simonetta Montemagni / Vito Pirrelli, *Testo e computer. Elementi di linguistica computazionale*, Roma, Carocci.
- McEnery/Wilson 2001 = Tony McEnery / Andrew Wilson, *Corpus Linguistics*, Edinburgh, University Press.
- Orrù 2017 = Paolo Orrù, *Il discorso sulle migrazioni nell'Italia contemporanea. Un'analisi linguistico-discorsiva sulla stampa (2000-2010)*, Milano, FrancoAngeli.
- Orrù 2022 = Paolo Orrù, *Linguistica dei corpora e analisi del discorso: tecniche per l'analisi della stampa con un caso di studio sulla rappresentazione del Sud*, in Paola Carlucci / Eugenio Salvatore (a cura di), *Giornali italiani dal 1950 a oggi: rappresentazione della realtà e della lingua*, Siena, Edizioni Università per stranieri di Siena, pp. 167-188.
- Rossini Favretti 2000 = Rema Rossini Favretti, *Progettazione e costruzione di un corpus di italiano scritto – CORIS/CODIS*, in Rema Rossini Favretti (a cura di), *Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento*, Roma, Bulzoni, pp. 57-73.
- Sinclair 1991 = John Sinclair, *Corpus, Concordance and Collocations*, Oxford, University Press.
- Sinclair 2004a = John Sinclair, *Corpus and Text — Basic Principles*, in Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, Oxbrow, pp. 17-29.
- Sinclair 2004b = John Sinclair, *Current issues in corpus linguistics*, in John Sinclair, *Trust the Text: Language, Corpus and Discourse*, London, Routledge.
- Tognini-Bonelli 2001 = Elena Tognini Bonelli, *Corpus Linguistics at Work*, Amsterdam, John Benjamins.