PAOLO ORRÙ

LINGUISTICA DEI CORPORA E ANALISI DEL DISCORSO: TECNICHE PER L'ANALISI DELLA STAMPA, CON UN CASO DI STUDIO SULLA RAPPRESENTAZIONE DEL SUD

1. Stampa, discorso e società

Nonostante il mercato editoriale odierno sia sempre più caratterizzato dalla disintermediazione digitale e lo sviluppo di Internet e dei social media abbia fatto aumentare enormemente il numero di fonti consultabili online, il giornalismo ha ancora un ruolo importante nel raccogliere, interpretare e dare senso alla miriade di notizie che ogni giorno attraversano i nostri schermi. Molte delle nuove proposte in questo campo sono attive solamente in rete (*Il Post*, *Fanpage*, *Open*, solo per citare alcune tra le più note); ciononostante, i siti di informazione più visitati ogni giorno rimangono quelli delle testate più tradizionali: *la Repubblica*, *Corriere della Sera*, *Il Messaggero*. I quotidiani nazionali agiscono, quindi, ancora in modo importante come filtro tra la realtà politico-sociale e i cittadini/lettori, e il giornalismo più in generale svolge un'azione capillare nella costruzione e nella circolazione dei discorsi, intesi come stratificazioni di produzioni semiotiche che strutturano un campo dell'esperienza sociale. La lingua non è, infatti, solo rappresentazione, ma è anche creazione, nel senso in cui opera per attribuire significati ai fenomeni sociali di cui facciamo esperienza quotidianamente. I discorsi non vanno trattati «come degli insiemi di segni (di elementi

significanti che rimandino a contenuti o a rappresentazioni), ma come delle pratiche che formano sistematicamente gli oggetti di cui parlano» (Foucault 2009: 45). La stampa è una tra queste pratiche: contribuisce a definire i limiti interpretativi dei fenomeni del reale, dà voce e forma allo *status quo*, è una forma di esercizio di potere. Come sostiene Fairclough:

> Discourse is socially constitutive as well as socially shaped: it constitutes situations, objects of knowledge, and the social identities of and relationships between people and groups of people. It is constitutive both in the sense that it helps to sustain and reproduce the social status quo, and in the sense that it contributes to transforming it. Since discourse is so socially influential, it gives rise to important issues of power (Fairclough/Wodak 1997: 258).

Il discorso, insomma, non riflette una realtà che esiste fuori o prima del sociale e del politico: è sempre parte attiva di ciò che è sociale e politico poiché attraverso la lingua formuliamo le nostre idee del mondo. Il compito dell'analisi del discorso è, allora, di mostrare come le rappresentazioni delle nostre esperienze non vadano interpretate nel segno di ciò che è vero o falso, ma vadano piuttosto pensate come produzioni discorsive contingenti, formulate in larga parte da chi è in grado di detenere il controllo sul discorso in un certo periodo storico. Ciò ovviamente non esclude la possibilità di riconoscere e contrastare tali sistemi attraverso discorsi alternativi e per ciò che è alternativo di diventare un giorno il nuovo *status quo*.

Per tutti questi motivi, un'analisi linguistico-discorsiva delle rappresentazioni sociali veicolate dalla stampa, ossia un'analisi che attraverso l'indagine delle forme linguistiche e delle loro funzioni possa rivelarne i discorsi soggiacenti, è quanto mai pertinente e, anzi, necessaria.

2. Analisi del discorso assistita dai corpora

Un altro concetto fondamentale da tenere in considerazione è quello che, insieme a Fairclough, possiamo definire l'aspetto *cumulativo* del discorso mediatico:

> The hidden power of media discourse and the capacity of […] power-holders to exercise this power depend on systematic tendencies in news reporting and other media activities. A single text on its own is quite insignificant: the effects of media power are cumulative, working through the repetition of particular ways of handling causality and agency, particular ways of positioning the reader, for instance, media discourse is able to exercise a pervasive and powerful influence in social reproduction because of the very scale of the modern mass media and the extremely high level of exposure of whole populations to a relatively homogeneous out-put (Fairclough 1989: 54).

La ripetizione nel tempo attraverso i mass media di forme linguistiche, narrative, testuali è ciò che permette a un discorso di diventare senso comune, di penetrare nell'immaginario collettivo, di farsi stereotipo e chiave interpretativa immediata. Per questo motivo, l'uso delle analisi quantitative dei testi è da ritenersi estremamente utile per osservare le tendenze di medio-lungo termine. Proprio in Italia (Partington

2004), a tal proposito è stata coniata l'etichetta di *Corpus Assisted Discourse Studies* (CADS): «The aim of the CADS approach is the uncovering, in the discourse type under study, of what we might call non-obvious meaning, that is, meaning which might not be readily available to naked-eye perusal» (Partington *et al*. 2013: 11). È questo il campo in cui ci muoviamo in questo contributo.

Può essere facile confondere gli strumenti della linguistica dei corpora solamente con la ricerca del riscontro numerico su grandi insiemi di dati, ovverosia di associare tali metodi puramente ad analisi di tipo quantitativo in opposizione a più classiche letture qualitative. In realtà, non sono altro che strumenti informatici utili per evidenziare o ricercare nei testi parole, forme, costruzioni in modo automatico e agile, non per forza devono essere impiegati su corpus di grandi dimensioni, anche se in prevalenza è ciò che viene fatto. Spesso, inoltre, il fattore numerico e statistico può indurre l'errata convinzione di offrire un'interpretazione oggettiva dei fatti rispetto a metodi più tradizionali; si tratta anche in questo caso di una scorciatoia da respingere. I dati possono facilmente alimentare il proprio *confirmation bias* (la tendenza a cercare conferma di quanto già crediamo); i metodi quantitativi, al contrario, possono e devono avere un ruolo importante nel processo autoriflessivo di verifica, ed eventuale rigetto, delle proprie ipotesi iniziali.

I testi giornalistici hanno rappresentato una delle fonti privilegiate per la costruzione di basi di dati delle grandi lingue di cultura fin dai primordi della disciplina: sono utili non solo per gli studi sulle tendenze linguistiche in sincronia e diacronia, o miniera inesauribile di neologismi per i lessicografi; giornali e riviste sono anche una fonte preziosa per gli studi interdisciplinari che vogliano indagare con puntualità i fatti sociali e le loro manifestazioni discorsive. Tra i più noti[1] corpus dell'italiano vi è sicuramente il corpus *la Repubblica*[2] (Baroni *et al*. 2004), sviluppato dall'Università di Bologna, raccoglie testi del quotidiano romano pubblicati tra il 1985 e il 2000, per un totale di circa 380 milioni di token.[3] Il corpus di riferimento attualmente più vasto e aggiornato per l'italiano giornalistico è però il Timestamped JSI; esso fa parte di una famiglia di corpora aggiornati quotidianamente estraendo articoli (tra i 100 e i 150 mila al giorno) da una lista di 75 mila fonti web in 18 lingue (tra cui arabo, inglese, spagnolo, catalano, ungherese, francese, russo, coreano ecc.).[4] Oltre all'immenso la-

1 Per una rassegna aggiornata su alcuni dei principali corpora in varie lingue si veda il sito dell'iniziativa CLARIN (Common Language Resources and Technology Infrastructure) https://www.clarin.eu/resource-families/newspaper-corpora (ultimo accesso: 15/6/2021).

2 Il corpus è indagabile attraverso la piattaforma gratuita NoSketchEngine: https://corpora.dipintra.it/public/run.cgi/first_form (ultimo accesso: 15/6/2021).

3 Con *token* si intende l'unità minima di cui è composto un corpus: parole, segni interpuntivi, cifre, sigle, qualsiasi elemento testuale tra due spazi bianchi. La *parola* è invece un *token* che inizia con una lettera dell'alfabeto. La quantità di token è sempre più elevata del numero di parole.

4 Per una descrizione sintetica si veda la pagina https://www.sketchengine.eu/jozef-ste-

voro di raccolta e archiviazione di questa ingente mole di testi e dati, la peculiarità del progetto risiede nella ricchezza di metadati con cui è annotato il corpus. L'annotazione temporale dei testi permette un'accurata ricerca diacronica, che, seppur circoscritta agli anni dal 2014 in poi, consente sicuramente di effettuare precise analisi su fenomeni sociali recenti. La base di dati italiana conta nel momento in cui scriviamo oltre 7,5 miliardi di token. Ogni articolo è annotato non solo per anno, ma anche per quadrimestre, mese e data specifica. Oltre ai dati cronologici, poi, i testi sono etichettati attraverso un ricco sistema di categorie tematiche; è possibile insomma svolgere indagini molto specifiche, su una singola testata, su un dato periodo, su una categoria di notizie.

Sono due i maggiori problemi che limitano il pieno sfruttamento e l'attendibilità dei dati ottenuti con questi corpus: la duplicazione dei testi e i procedimenti di assegnazione automatica delle parti del discorso. Molto spesso l'attività di *webcrawling* (l'estrapolazione automatica degli articoli da un set di fonti) produce un numero indefinibile con precisione di duplicati dello stesso articolo; ciò comporta una corruzione dei risultati in una misura non sempre trascurabile. Le procedure per l'eliminazione automatica dei doppioni dei testi non sembrano essere totalmente risolutive. Per quanto riguarda la seconda questione, lo sviluppo di tecniche di assegnazione automatica dei *tag* per consentire l'analisi sintattica dei testi, pur avendo fatto grandi passi avanti, non risulta soddisfacente. Per una lingua molto ricca morfologicamente come l'italiano, non sembra si tratti di un problema facilmente risolvibile allo stato attuale. Anche se, bisogna dirlo, non sempre si tratta di errori, del tutto comprensibili, dovuti all'omografia tra forme.

3. Strumenti e tecniche per l'analisi del discorso giornalistico: un caso di studio

Per meglio rispondere al tema del volume, ci sembra importante offrire un piccolo caso di studio attraverso cui mostrare il metodo CADS. Tratteremo, dunque, di alcuni aspetti del dualismo tra Nord e Sud Italia nella stampa quotidiana online. È un soggetto ancora inesplorato dal punto di vista linguistico, se non per alcune prime e interessanti considerazioni di Fabio Rossi (2015: 183-9), che, all'interno di un più ampio discorso sull'uso della lingua a fini discriminatori, si è soffermato sul diverso approccio riservato a settentrionali e meridionali in un piccolo campione di articoli del *Corriere della Sera* e di *Repubblica*. Il sondaggio di Rossi mostrava come la provenienza meridionale negli articoli di cronaca venisse messa in rilievo anche laddove non fosse per niente informativa: etnonimi come *calabrese*, *siciliano*, *napoletano* comparivano in contesti negativi con più frequenza rispetto a *lombardo*, *piemontese*,

fan-institute-newsfeed-corpus/#toggle-id-1; per una disamina più completa invece Bušta/Herman (2017).

milanese. Studi di ambito sociologico (Cremonesini 2015) condotti sulla rappresentazione giornalistica del Sud sui due maggiori quotidiani italiani (*Corriere* e *Repubblica*) dal 1984 al 2010 hanno mostrato come le notizie relative al Meridione siano drasticamente diminuite dal 2000 in poi, sancendo una minore attenzione per l'area; gli articoli si concentravano, inoltre, su quattro ambiti tematici: criminalità, cronaca, politica, welfare. Quasi una notizia su due trattava di criminalità e mafia. All'interno dello stesso progetto è stato condotto anche un analogo scrutinio (Cristante 2015) sui servizi del Tg1 dedicati al Sud che ha mostrato tendenze molto simili; i servizi sul resto del Paese vedevano, invece, prevalere notizie di politica, cronaca, economia e cultura. Tra gli anni Ottanta e Novanta «il Meridione è in genere raccontato come un territorio pericoloso e contraddistinto da degrado morale, come un luogo che deve ancora intraprendere la via dello sviluppo, come un luogo diverso dal resto d'Italia» (Cremonesini 2015: 190). Nell'ampio periodo considerato, il Sud era ridotto a due sole regioni, Campania (e in particolare Napoli) e Sicilia, e la sua narrazione quasi esclusivamente incentrata sulla criminalità organizzata e in seconda battuta sulle questioni politiche, relative all'arretratezza del Mezzogiorno nel suo complesso.

È necessario ricordare che la maggior parte dei quotidiani risiede al Nord: tra nazionali e locali sono 32; 12 al centro e 10 al Sud. Ciò va sicuramente a influire sulla quantità di articoli e probabilmente anche sulla qualità degli stessi: senza scomodare a tutti i costi il pregiudizio antimeridionale, un articolo redatto senza un'esperienza diretta o una conoscenza approfondita del territorio non può che risentire di filtri interpretativi formati su concetti e categorie precostituite. Inoltre

> Le analisi più serie e circostanziate hanno difficilmente accesso ai media nazionali. Spesso a parlare di Mezzogiorno sono intellettuali o giornalisti che da anni ne stanno fuori e guardano alla loro terra natia, usando talora gli occhi della nostalgia, talaltra gli accenti della deprecazione, senza misurarsi con la complessità del mondo reale (Gribaudi 2010: 117-118).

Tenteremo, quindi, di offrire alcuni ulteriori spunti sulla questione, espandendo e aggiornando la base di dati rispetto a quanto fatto in precedenza. Gli studi CADS normalmente prevedono la costruzione di un corpus ad hoc per rispondere alla specifica domanda di ricerca; nella prossima sezione, pertanto, descriveremo la procedura usata per costruire la base di dati che sarà poi oggetto di analisi. Alcune tecniche sono tipiche e si procede secondo una sorta di routine consolidata. Una volta raccolto, preparato e trattato il corpus (e un eventuale secondo corpus di controllo o riferimento), viene elaborata una lista di frequenza delle parole e calcolata una lista delle parole chiave. Dopo l'analisi e la categorizzazione delle parole chiave per area semantica, si procede alla lettura delle linee di concordanza (stringhe di testo ricavate dalla ricerca di una forma linguistica) di quelle che appaiono come le voci più interessanti da indagare; in seguito, sulla scorta di questi primi sondaggi vengono applicate altre tecniche, come l'analisi delle collocazioni, dei cluster linguistici (o *n-grams*) o l'analisi dei *dispersion plot* (la concentrazione di certi vocaboli in parti specifiche del corpus).

Nel nostro caso ci limiteremo al calcolo delle parole chiave e all'analisi di alcune collocazioni; le linee di concordanza (cfr. fig. 1) sono uno strumento utilizzato in ogni momento per la verifica delle ipotesi e per la lettura del contesto discorsivo e quindi per ricavare gli esempi concreti di lingua da analizzare.

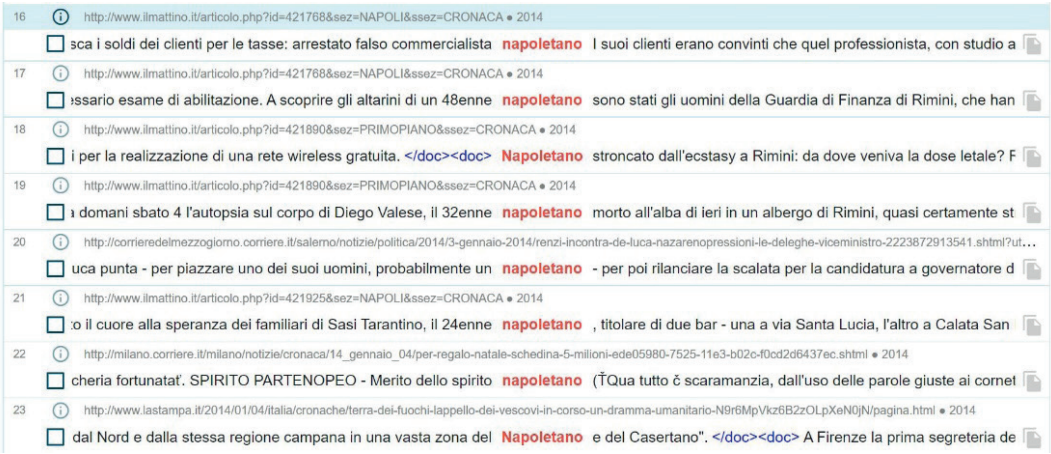

fig. 1. *Esempio di linee di concordanza*

3.1 Corpus

Per la nostra analisi abbiamo utilizzato il corpus Timestamped Jsi,[5] selezionando un numero limitato di fonti tra le tante disponibili. Tra queste abbiamo optato per i maggiori quotidiani nazionali,[6] poiché si propongono come testate di ampio respiro contenutistico; si rivolgono indifferentemente a tutta la popolazione; hanno redazioni locali o i mezzi per svolgere inchieste e servizi sull'intero territorio nazionale; hanno una circolazione anche su altri media come radio e televisione. I quotidiani oggetto di analisi sono dunque: *Corriere della Sera* (CS),[7] *la Repubblica* (RP), *La Stampa* (ST), *Il Giornale* (GN), *Il Messaggero* (MS), *Il Mattino* (MaT) e *Il Fatto quotidiano* (FT). Tre del Nord (Milano e Torino), tre del Centro (Roma), uno del Sud (Napoli). Si potrebbe obiettare che per bilanciare la selezione delle fonti avremmo potuto includere altri quotidiani locali di buona diffusione o diffusione sovraregionale; ma il nostro intento è di rispecchiare la reale situazione della proposta editoriale nazionale.

5 Il corpus è consultabile attraverso la piattaforma Sketch Engine, Kilgarriff *et al.* (2014).

6 I dati sulla diffusione sono raccolti mensilmente da Audiweb, che misura gli accessi unici ai siti delle testate, e da Ads (Accertamenti Diffusione Stampa), che misura il totale di vendite cartacee e abbonamenti digitali, consultabili sul sito https://www.adsnotizie.it/_dati_DMS.asp (ultimo accesso: 10/7/2021).

7 Il sistema di sigle qui indicato verrà utilizzato più avanti come forma di notazione degli esempi, che verrà riportata tra parentesi nella forma (SIGLA data dell'articolo).

Abbiamo costruito due corpus all'interno della base di riferimento, partendo da due serie di termini di ricerca espressi nella sintassi Corpus Query Language (CQL):[8]

Corpus-Sud: [lc="napoletano" | lc="napoletana" | lc="napoletani" | lc="napoletane" | lc="calabrese" | lc="calabresi" | lc="siciliano" | lc="siciliani" | lc="siciliana" | lc="siciliane" | lc="pugliese" | lc="pugliesi" | lc="meridionale" | lc="meridionali"]
Corpus-Nord: [lc="milanese" | lc="milanesi" | lc="settentrionale" | lc="settentrionali" | word="ligure" | lc="liguri"| lc="piemontese" | lc="piemontesi"| word="veneto" | lc="veneti" | lc="veneta"| lc="venete"]

Abbiamo optato per selezionare solo alcuni aggettivi che possano richiamare sia protagonisti sia luoghi relativi alle principali regioni del Sud e del Nord. Si noterà che nella maggior parte dei casi si tratta di espressioni relative alla regionalità, in altri due abbiamo invece ritenuto opportuno far riferimento ai due maggiori centri (Napoli e Milano), due, infine, sono di portata sovraregionale (*meridionale*/*i*, *settentrionale*/*i*), così da esplorare i diversi livelli territoriali. I vocaboli sono stati cercati indipendentemente dalla forma maiuscola/minuscola, ad eccezione di due: *veneto* è stata ricercata solamente nella sua versione minuscola, così da eliminare l'omografia con il nome della regione; per motivi simili abbiamo limitato la ricerca di *ligure* per evitare la coincidenza con vari toponimi.

Per quanto riguarda l'orizzonte temporale, ci siamo limitati al periodo 2014-2019, preferendo escludere dalla ricerca il biennio 2020-2021: la pandemia causata dal nuovo Coronavirus ha letteralmente stravolto i normali ritmi giornalistici. Per quanto possa trattarsi di un tema assolutamente degno di interesse per lo studio del dualismo tra Nord e Sud Italia, l'emergenza sanitaria ha comportato la produzione di una quantità spropositata di testi – la cosiddetta *infodemia* –,[9] la cui inclusione nel corpus influenzerebbe qualsiasi calcolo statistico e non farebbe emergere la normale copertura del fenomeno sociale di nostro interesse. La fig. 2 mostra la distribuzione dei testi per anno nel corpus generale, da cui emerge chiaramente che quasi il 40% dell'intera base di dati è relativa al solo anno 2020.

8 Nella sintassi CQL di Sketch Engine il comando 'lc' (*lowercase*) è utilizzato a tale scopo, mentre il comando 'word' cerca l'esatta forma digitata. La barra verticale '|' indica invece l'operatore booleano 'OR', serve quindi a ottenere qualsiasi testo in cui sia presente almeno una volta almeno una delle parole indicate.

9 Vd. https://www.treccani.it/vocabolario/infodemia_%28Neologismi%29/ (ultimo accesso 25/8/2021).

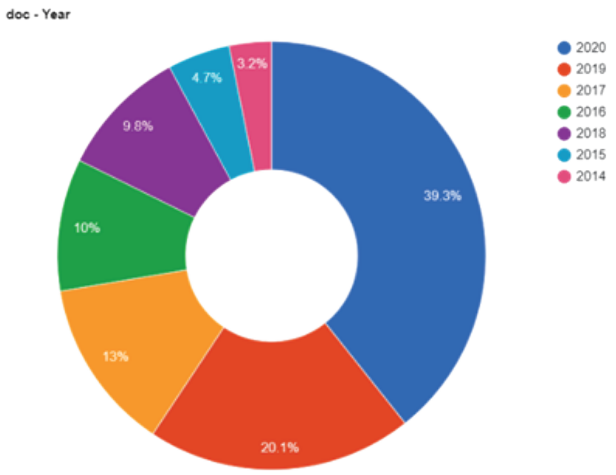

fig. 2. *Composizione corpus Timestamped per anni*

Il corpus Timestamped, come si diceva, è ampiamente annotato con una grande varietà di metadati, non solo temporali, ma anche qualitativi (tipologia di articoli, area geografica, tema). Abbiamo scelto di effettuare la ricerca su tutte le categorie di notizie, così da osservare il quadro più ampio degli articoli su Meridione e Settentrione.

Nella tab. (1) possiamo osservare la consistenza dei due corpora. È da notare che vi sarà una certa sovrapponibilità, alcuni testi possono ovviamente comprendere termini che riguardano sia il Sud sia il Nord.

	Tokens	Words	Testi
Corpus-Nord	43.808.727	37.040.035	79.350
Corpus-Sud	45.938.298	38.840.576	85.828

tab. 1. *Consistenza corpus e subcorpus*

Grazie alla ricchezza di metadati di Timestamped è stato possibile selezionare i siti generali delle testate in modo che comprendessero anche i sotto-siti tematici, ciò per poter attingere automaticamente anche alle edizioni locali (ad esempio, al *Corriere del Mezzogiorno* e al *Corriere del Veneto*, e alle tante edizioni di *Repubblica*), ai blog e alle varie sezioni. Mediante la sintassi CQL abbiamo escluso, però, le pagine di annunci del sito annunci.repubblica.it poiché avrebbero prodotto migliaia di risultati irrilevanti ai nostri scopi.[10]

10 Riportiamo il comando specifico: [termini di ricerca] ! within <doc urldomain="annunci\.repubblica\.it" />, l'operatore "!" nel linguaggio CQL ha il valore di NOT, in combinazione con "within" e il comando tra le parentesi uncinate indica quale specifico dominio web vada ignorato nella ricerca.

Testata	Corpus-Nord	Corpus-Sud
corriere.it	23.927	20.903
ilfattoquotidiano.it	2.934	3.519
ilgiornale.it	10.421	8.408
ilmattino.it	1.723	13.326
ilmessaggero.it	7.168	6.096
lastampa.it	8.252	4.360
repubblica.it	24.905	29.216

tab. 2. *Corpus, articoli per testata*

Le due testate più importanti (*Corriere* e *Repubblica*) mostrano una copertura ampia su entrambe le aree, grazie anche alle sezioni locali e ai siti loro dedicati. Anche *Il Messaggero* e *Il Fatto* hanno un rapporto bilanciato tra gli articoli dedicati al Nord e al Sud, mentre *Il Mattino* è fortemente specializzato sul Sud e la *La Stampa* sul Nord.

	CS	FQ	GN	MaT	MeS	RP	ST	Totale
napoletan*	10.235	1.170	3.190	15.512	2.556	12.863	1.203	46.729
sicilian*	6.012	1.959	3.845	691	2.242	15.565	1.704	32.018
puglies*	6.899	770	1.731	514	1.055	8.131	698	19.798
calabres*	2.516	897	1548	693	1.068	4.880	933	12.535
milanes*	17.016	1.949	8.999	828	4.968	14.262	3.060	51.082
venet*	11.428	772	2.083	324	1.503	4.509	1.044	21.483
ligur*	1.694	566	1.424	289	874	8.298	2.012	15.157
piemontes*	2.064	434	1.076	212	616	5.709	4.204	14.315
meridional*	3.508	785	1.706	1.231	1.482	5.286	1.473	15.471
settentrional*	1.422	363	870	439	932	2.944	928	7.905

tab. 3. *Occorrenze lemmi per testata*

I due lemmi più importanti risultano, com'era lecito attendersi, quelli relativi alle due maggiori città; ciò conferma, inoltre, la bontà della scelta iniziale poiché lemmi come *torines**, *lombard**, *campan** o *palermitan** registrano tutti frequenze significativamente inferiori rispetto ai termini di ricerca. Dalla tabella si evince come la maggiore dimensione della componente di *Repubblica* si rifletta anche nella distribuzione delle occorrenze: il quotidiano romano presenta quasi sempre il maggior numero di occorrenze per ogni lemma; le eccezioni sono rappresentate da due lemmi settentrionali (*milanese* e *veneto*) maggiormente rappresentati sul *Corriere* e dal lemma *napoletan** più utilizzato sul *Mattino*, che, allo stesso tempo, registra un minore interesse

per tutti gli altri lemmi.

3.2 Parole chiave

Il primo passo nella nostra indagine è il calcolo delle parole chiave. I concetti di *keyword* e *keyness* sono stati introdotti a metà degli anni Novanta da Mike Scott (1997). Le keyword sono parole identificate non per la loro frequenza grezza, ma per la loro salienza; essa viene determinata attraverso un rapporto statistico[11] tra la frequenza in un dato corpus e quella in un corpus di riferimento più generale della lingua. La lista delle parole chiave serve come un'indicazione di *aboutness* (Phillips 1989) dei testi, per identificare, cioè, i temi, le idee o i particolari stilistici di un testo o un corpus. Com'è facile intuire si tratta di uno strumento assai proficuo per individuare gli argomenti più frequentemente trattati in un insieme di articoli di giornale: scorrendo la lista di parole chiave si possono derivare, infatti, gruppi di parole vicine tra loro semanticamente; oppure, se il corpus è composto da diverse testate, è possibile evidenziare le specificità di ognuna di esse calcolando i rispettivi elenchi; o ancora, se il corpus presenta un'annotazione diacronica, si possono verificare cambiamenti nella salienza di un tema o nella sua trattazione nel tempo. Sono due gli approcci prevalenti per l'analisi: *focused* and *exploratory*. Nel primo, una serie di parole o forme specifiche scelte dal ricercatore viene confrontata in due corpus differenti; il secondo invece compara le frequenze di tutte le parole di un corpus con quelle di un corpus di riferimento più generale, per far emergere quelle che ricorrono in maniera inusuale. Un approccio *focused* parte quindi da domande di ricerca ben delineate; uno *exploratory* invece si fonda su quesiti più generali e guarda ai dati per guidare l'interpretazione e le fasi successive della ricerca. I due approcci non sono mutualmente esclusivi: benché nell'analisi del discorso prevalga soprattutto il secondo metodo, nella verifica delle ipotesi subentra frequentemente anche un confronto di tipo *focused*.

Nel calcolare le parole chiave dei nostri due corpus ci siamo basati su un approccio esplorativo: il calcolo avviene raffrontando la frequenza relativa di ogni parola nel corpus con quella della stessa parola nel corpus di riferimento, a ogni parola viene aggiunto un valore numerico standard,[12] da questo rapporto si ottiene un punteggio di *keyness*. La piattaforma Sketch Engine consente di scegliere se nel calcolo delle parole chiave si vogliano prediligere parole rare o comuni; in questo caso si è deciso di cercare una via di mezzo, preferendo parole non troppo comuni, ma non rare. Ciò deriva dalla volontà di far emergere dagli articoli temi sì peculiari, ma allo stesso

11 Fare riferimento solamente alla frequenza delle parole è in realtà molto riduttivo; le metriche per il calcolo delle parole chiave includono diversi possibili fattori, tra cui calcoli probabilistici e/o l'effetto della dimensione del corpus sulla frequenza attesa di una parola. Sono varie le misure statistiche per determinare la *keyness* di una parola, rimandiamo a Gabrielatos (2018) per una trattazione esaustiva.

12 Per una disamina più completa, ma allo stesso tempo chiara, rimandiamo a Kilgarriff (2009).

tempo piuttosto ordinari.

Dai risultati abbiamo eliminato tutti i nomi propri di persona e i toponimi, poiché sono tra le parole che più facilmente possono risaltare;[13] abbiamo espunto dalle liste anche i nostri termini di ricerca, in quanto tali emergono con una frequenza ovviamente più alta rispetto al normale.

Per quanto riguarda il corpus-Sud, nel raffronto con un corpus più generale del web (il corpus itTenTen16)[14] 31 tra le 50 parole più rilevanti riguardano il tema della criminalità, della mafia e della sicurezza: *ndrangheta*, *procura*, *clan*, *inchiesta*, *antimafia*, *governatore*, *boss*, *mafia*, *indagati*, *arrestato*, *pm*, *camorra*, *procuratore*, *gip*, *carabinieri*, *inquirenti*, *investigatori*, *indagato*, *arresti*, *omicidio*, *indagini*, *accusa*, *domiciliari*, *mafiosa*, *arrestati*, *mafioso*, *sequestro*, *droga*, *carcere*, *magistrati*, *cocaina*. La differenza con il corpus-Nord è piuttosto marcata: in esso solo dieci parole possono richiamare il tema della sicurezza. Nel corpus-Nord 14 parole riguardano l'economia (solo 1 nel Sud); 8 la politica (7 nel Sud); 3 la cronaca; le restanti sono relative allo sport o al meteo.

Se volessimo invece calcolare le parole chiave usando come riferimento la restante parte del corpus Timestamped, la differenza sarebbe ancora più marcata: in questo caso, infatti, nel corpus-Nord vi sono solo due vocaboli legati a criminalità e sicurezza (*vittore*, il noto carcere milanese di San Vittore, e *pm*), mentre sono 18 su 50 nel corpus-Sud, di cui molti legati alle grandi organizzazioni di carattere mafioso.

13 Abbiamo conservato solamente nomi propri che rimandano a fatti di cronaca specifici, come *morandi*, che si riferisce ovviamente al crollo del ponte sul Polcevera, e i nomi dei quartieri, poiché possono essere legati a specifici temi o casi di cronaca.

14 Si tratta di un vasto corpus di riferimento (4,9 miliardi di parole) dell'italiano del web, compilato nel 2010 e poi aggiornato nel 2016, vd. Jakubíček *et al.* (2013).

Rank	Nord	Sud
1	enne	enne
2	ftse	ndrangheta
3	allerta	procura
4	carige	clan
5	centrodestra	migranti
6	teleborsa	inchiesta
7	leghista	antimafia
8	procura	governatore
9	inchiesta	boss
10	governatore	mafia
11	lega	indagati
12	rialzo	arrestato
13	pm	pm
14	maltempo	centrodestra
15	carroccio	camorra
16	indagati	procuratore
17	capoluogo	gip
18	listino	carabinieri
19	piogge	inquirenti
20	migranti	investigatori
21	expo	pd
22	calo	allerta
23	gip	premier
24	tav	indagato
25	investigatori	arresti
26	btp	ilva
27	mib	dem
28	pd	maltempo
29	premier	omicidio
30	temporali	indagini
31	sindaca	imprenditore
32	vigilia	accusa
33	atm	domiciliari
34	spread	tifosi
35	inquirenti	ars
36	banca	mafiosa
37	accusa	mezzogiorno
38	cda	consip
39	mps	ultrà
40	atp	grillini
41	procuratore	arrestati
42	perturbazione	mafioso
43	inter	sequestro
44	derby	droga
45	siro	carcere
46	ex	ex
47	arrestato	deputato
48	imprenditore	magistrati
49	unicredit	juve
50	precipitazioni	cocaina

tab. 3. *Parole chiave calcolate sul corpus ItTenTen*

Rank	Nord	Sud
1	enne	enne
2	carige	ndrangheta
3	tav	clan
4	leghista	camorra
5	atm	pizza
6	ftse	mafia
7	listino	consip
8	sweet	boss
9	crude	ars
10	lombardo	antimafia
11	vicepremier	mezzogiorno
12	carroccio	scampia
13	autostrade	mafioso
14	torinese	mafiosa
15	navigli	governatore
16	brera	mafiosi
17	morandi	tap
18	duomo	anm
19	capoluogo	ultrà
20	ribassi	pm
21	tallio	palermitano
22	linate	partenopea
23	centrodestra	mafie
24	light	campani
25	hinterland	partenopeo
26	frazionale	poggioreale
27	oil	sud
28	leghisti	barese
29	chef	magistrati
30	torinesi	grillini
31	btp	campano
32	lombarda	teatro
33	spread	cosche
34	governatore	ilva
35	panettone	criminalità
36	autonomia	bagnoli
37	atp	gomorra
38	pm	dda
39	politecnico	criminale
40	pianura	dialetto
41	lega	mdp
42	pfas	deputato
43	all-share	ong
44	guadagno	dem
45	vittore	centrodestra
46	oncia	scrittore
47	mib	magistrato
48	malpensa	viminale
49	nevicate	rione
50	comparti	imprenditore

tab. 4. *Parole chiave rispetto al corpus Timestamped*

Rispetto quindi al complesso del panorama giornalistico online, la frequenza relativa dei termini associati alla sicurezza diminuisce, poiché tale tipologia di articoli è ritenuta tra le più notiziabili. Tuttavia, nel corpus-Sud la quantità di parole legate alla criminalità rimane cospicua, si tratta di oltre un terzo del totale, mentre nel corpus-Nord si ferma al 4%. Non mancano, però, anche vocaboli appartenenti a fatti di attualità importanti come *tav*, *morandi*, *autostrade* e *pfas* al Nord e *consip*, *tap* e *ilva* per il Sud. Tali parole assumono un alto punteggio di *keyness* anche per la loro struttura (si tratta soprattutto di sigle), che le rende meno frequenti in un corpus di riferimento.

Le dimensioni notevoli dei due corpus consentirebbero di estendere l'analisi ad almeno le prime 100 parole chiave, ma per motivi di brevità non sarà possibile farlo in questa sede; ad ogni modo, dalla lettura dei due elenchi emerge chiaramente una distribuzione tematica che vede gli articoli sul Nord incentrati soprattutto su aspetti economico-finanziari e politici e quelli sul Sud dove l'elemento della mafia e della criminalità è prevalente o comunque fortemente caratteristico. Questa prima lettura conferma precedenti studi sulla rappresentazione del Sud. Non si può (e non si vuole) sottovalutare l'importanza e la salienza della criminalità organizzata nelle realtà quotidiana di alcune regioni del Mezzogiorno; tuttavia, è bene rimarcare che una narrazione così fortemente sbilanciata continua in parte a oscurare altri lati positivi degli stessi territori e a confermare stereotipi e interpretazioni fortemente radicate nell'immaginario collettivo attraverso gli altri media.[15]

Vale la pena, infine, notare un ultimo caso: tutti e quattro gli elenchi riportati nelle tabelle sopra hanno la stessa parola chiave come più saliente: *enne*. I processi di separazione delle parole dei testi (*tokenization*) e di assegnazione di una parola a un lemma (*lemmatization*) possono produrre risultati in apparenza erronei o curiosi. Si tratta in questo caso del suffisso per la creazione di sostantivi e aggettivi numerativi, trattato come parola a tutti gli effetti, poiché isolata da due spazi. Si potrebbe quindi tralasciare, non foss'altro poiché è tipico degli articoli giornalistici e di cronaca riportare l'età dei protagonisti delle vicende, vedremo però in seguito perché tale forma possa essere interessante per la nostra disamina.

3.3 Collocazioni

Se il concetto di collocazione come combinazione preferenziale tra due parole è di certo trasparente per i linguisti, nello specifico campo in cui ci muoviamo necessita di alcune precisazioni; essa va, infatti, intesa in senso più estensivo: non si tratta solamente di combinazioni fraseologiche cristallizzate e tipiche di una lingua, ma anche di associazioni statisticamente significative tra parole. Sono due le tipologie di approcci in questo senso: il *collocation window approach* esamina quali parole ri-

15 Rimandiamo ancora al volume di Cristante/Cremonesini (2015) in questo senso.

corrano 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. Si ricorre spesso in questo senso anche alla definizione di prosodia semantica (o discorsiva): «The consistent aura of meaning with which a form is imbued by its collocates» (Louw 1993: 157). L'altro approccio, definito *n-gram approach*, è basato sul calcolo di combinazioni di parole adiacenti (bigrammi, trigrammi, ecc.) e mira quindi a evidenziare strutture ricorrenti.

La nostra analisi si basa sul primo approccio. Sono vari gli algoritmi[16] utilizzabili a tal fine, ognuno tende a prediligere una classe di parole sulle altre. Alcuni fanno emergere rapporti molto stretti tra le parole, ma più rari nelle frequenze (Mutual information); altri privilegiano le parole grammaticali (Log-likelihood), quelle lessicali (LogDice) o un mix tra queste (T-Score). Nell'analisi del discorso giornalistico i più utilizzati sono il T-Score e il LogDice. Il T-score è una misura che vuole indicare il grado di certezza con cui si può sostenere che la co-occorrenza di due parole non sia casuale, per cui combinazioni molto frequenti ottengono un punteggio elevato pur non essendo particolarmente significative dal punto di vista del significato. Il T-score assegna un punteggio che si basa sulla dimensione del corpus e non ha una scala definita da un minimo e un massimo; non è quindi possibile comparare i punteggi ottenuti in insiemi di dati di misura differente tra loro. Il LogDice viene calcolato misurando il rapporto tra la frequenza della collocazione e delle singole parole che la compongono, a cui viene aggiunto un valore standard; offre un indice scalare indipendente dalle dimensioni del corpus, permette così di comparare insiemi di dati differenti. Il valore massimo non può essere superiore a 14, il che si verifica quando la parola X e la parola Y occorrono sempre insieme in un corpus; tipicamente, il valore è inferiore a 10. Il punteggio può assumere anche valori sotto lo zero, il che significa che la collocazione non ha alcuna valenza statistica.

Il T-score ha un rapporto molto stretto con la frequenza grezza delle co-occorrenze, il che può essere sicuramente utile in certi casi; il LogDice favorisce invece un bilanciamento tra frequenza ed esclusività delle combinazioni. La scelta della misura è in larga parte dipendente dalla domanda di ricerca: nel nostro caso abbiamo optato per l'uso del secondo algoritmo, preferendo una più forte significatività semantica rispetto alla sola frequenza. Partiremo da una visione generale delle collocazioni dei nostri termini di ricerca, riportandone dunque solo alcune e soffermandoci sulle categorie semantiche più rappresentate, senza riprodurre per brevità gli interi elenchi.

In via preliminare, è necessario puntualizzare che i termini *meridionale/i* e *settentrionale/i* registrano solo collocazioni di scarso interesse per i nostri scopi: sono

16 Per una descrizione più completa rinviamo a Baker (2006) e Glabasova *et al*. (2017) e nello specifico sul LogDice, Rychlý (2008).

infatti tutte relative a fatti di natura geopolitica o meteorologica.

Milanese è associata a una nutrita gamma di vocaboli legati all'economia e alla finanza: *listino*, *azienza*, *società*, *comparti*, *borsa*, *guadagno*, *seduta*, *mib*, *azioni*, *ftse*, *istituto* (spesso riferito a istituti di credito), ma sono presenti anche diverse parole legate al tema della criminalità: *procura*, *carcere*, *inchiesta*, *processo*, *dda* (direzione distrettuale antimafia).

Per *veneto* troviamo un'ampia tipologia di termini dell'economia (*imprenditore*, *istituto*, *produttivo*, *sistema*, *credito*, *bancario*, *colosso*), della politica (*governatore*, *consigliere*, *segretario*, *presidente*, *referendum*, *deputato*), dello sport (*derby*, *club*, *ciclista*, *nuotatore*, *scalatore*, *portiere*) e, in una posizione molto significativa, la parola *accento*, su cui torneremo più avanti.

Per quanto concerne *ligure*, le collocazioni sono per lo più relative all'accezione geografica dell'aggettivo (*riviera*, *entroterra, cittadina*, *territorio*, *appennino*, *borgo*, *località*); altre ancora riguardano l'economia (*banca*, *rete*, *azienda*) e altre sono di varia natura semantica dalle quali non emerge una caratterizzazione precisa.

Tra le collocazioni di *piemontese* abbondano ancora una volta vocaboli geografici; sono tante le parole attinenti al cibo: *razza*, *tradizione*, *vino*, *carne*, *cucina*, *fassona*. Rimandano alle persone solo *origine*, *giovane*, *tecnico* e una serie di ruoli politici (*governatore*, *deputato*, *assessore*, *segretario*).

Considerando la lista delle parole chiave ricavata dal corpus sul Sud, ci si potrebbe invece attendere varie collocazioni relative alla criminalità e al tema della sicurezza. L'impressione è confermata solo per due termini: *napoletano* e *calabrese*. Riguardo al primo, sono nove tra le prime 50 le collocazioni che riguardano la sicurezza: *arrestato*, *carcere*, *poggioreale*, *magistrato*, *morto*, *pm*, *pregiudicato*, *ucciso*. Sono molte, però, anche le parole che richiamano contesti positivi o neutri (*imprenditore*, *tifoso*, *cantautore*, *regista*, *giovane*, *attore*, *scrittore*, *cantante*) o inerenti aspetti culturali (*dialetto*, *presepe*, *arte*, *teatro*). Per *calabrese* l'elenco invece è più nutrito: 9 vocaboli tra i primi 50 sono connessi all'organizzazione 'ndranghetista (*ndrangheta*, *mafia*, *malavita*), alla sua struttura (*cosca*, *clan*, *organizzazione*), ai suoi affiliati (*boss*, *esponenti*, *esponente*); sono presenti poi alcune entrate più generiche (*organizzata*, *criminale*, *pregiudicato*). Tali associazioni sono meno presenti per *siciliano* e *pugliese*, per i quali troviamo invece collocazioni come *scrittore*, *artista*, *cantautore*, *imprenditore*, *campione*, *regista* o ancora *accento* e *cannolo* per il primo; per il secondo inoltre abbondano collocazioni che fanno riferimento a *pugliese* come specificazione geografica: *costa*, *società*, *turismo*, *comunità*, *eccellenza*, *stabilimento*.

Anche una collocazione, in apparenza, neutra come *imprenditore*, presente in tutti gli elenchi, rivela tra le linee di concordanza articoli di cronaca, corruzione e altri illeciti. Seguendo solo la catena di collocazioni, avremmo risultati numerici non troppo consistenti, e leggere tutte le linee necessiterebbe di molto tempo e restituirebbe solo un quadro parziale. Il corpus Timestamped, in questo caso, offre un altro utile strumento. Tutti gli articoli vengono taggati automaticamente dal sistema attingen-

do a un'amplissima serie di categorie tematiche (ogni articolo può contenere più di un tag): è possibile, dunque, verificare a quale genere di notizie appartengano i testi contenenti i sintagmi che ci interessano. Per quanto riguarda il sintagma con più frequenza, *imprenditore napoletano*, la categoria 'crime and justice' è la 5ª più applicata agli articoli, e sono 13 le categorie di tag relative al crimine assegnate agli articoli (su 236 complessive). In tutto, la frequenza dei tag per questa forma è di 416 su 2051 (20,3%) su quelli applicati. Gli articoli con queste etichette interessano oltre un terzo delle occorrenze del sintagma (158 su 443). Lo stesso calcolo sulla frequenza dei tag per gli altri termini di ricerca dà i seguenti risultati: per *i. pugliese* si tratta di 16 tag su 421; per *i. siciliano* di 59 su 422; per *i. calabrese* di 122 su 613; per *i. veneto* di 3 su 609; per *i. ligure* di 3 su 208; per *i. piemontese* di 8 su 385; per *i. milanese* di 50 su 1022. Nonostante possano esserci errori nell'attribuzione automatica del tag, com'è facile intuire dalle cifre, si tratta di percentuali assai ridotte per quanto riguarda tutti gli etnonimi del Nord, mentre le parole sul Sud, a eccezione di *pugliese*, sono fortemente legate al tema della criminalità e della sicurezza. Per completezza va notato, però, che una buona parte delle occorrenze di *imprenditore napoletano* appaiono nell'anno 2017 e sono relative a un singolo caso di cronaca, ciò influenza in modo inevitabile i risultati.

Si è ricordato in precedenza lo studio di Rossi (2015), nel quale si osservava l'accento meridionale come elemento ricorrente in articoli (non solamente) di cronaca come forma di identificazione e caratterizzazione dei protagonisti. In effetti, *accento* figura come collocazione significativa di alcuni dei nostri termini di ricerca: *napoletano*, *siciliano*, *calabrese* e *veneto*; è del tutto assente per quanto riguarda *ligure* e poco significativa per *milanese*, *piemontese* e *pugliese*. Nel caso di *veneto*, in cui ha l'indice di associazione più alto (8,15), l'accento è *inconfondibile*, *spiccato*, *marcato,* assume un carattere di connotazione personale dei protagonisti degli articoli, soprattutto attori, scrittori, sportivi, come la pallavolista Paola Egonu, della quale viene sempre rimarcata l'origine straniera, e un valore identitario:

(1) Paola Egonu, afro-azzurrina oro della pallavolo Paola Egonu, genitori nigeriani ma **accento veneto**, è la schiacciatrice della Nazionale Under 18 che è tornata dai Mondiali in Perù con la medaglia d'oro al collo (CS 18/8/2015).

Sono presenti, tuttavia, due casi di cronaca:

(2) "La scorsa Pasquetta ci hanno rubato 4mila euro di Gratta e vinci di notte. Stavolta erano italiani, **l'accento veneto è inconfondibile**". I banditi per arrivare dietro al bancone hanno rovesciato tutti gli articoli in esposizione e hanno puntato la pistola contro il marito della proprietaria, prima di rubare e scappare (CS 9/8/2017).

(3) I malviventi hanno colpito in maniera fulminea. In pochi minuti sono entrati, hanno puntato la pistola contro il marito della proprietaria, si sono fatti consegnare il denaro dalla cassa (circa 600 euro) e sono scappati. Entrambi erano a volto coperto, **con accento veneto** ed emanavano un forte

odore di alcol (CS 14/12/2017).

Anche per *siciliano* la situazione è simile, ma più marcata sul versante cronachistico. Decisamente più problematica è l'associazione nel caso di *napoletano*: sono 71 le co-occorrenze totali, ben 40 tra queste riguardano articoli su rapine, furti, violenze.

(4) Rapina in banca a Pontecorvo, due giovani "traditi" **dall'accento napoletano** I due sono stati arrestati dagli agenti del Commissariato di polizia di Giugliano, diretti dal primo dirigente Pasquale Trocino (MaT 27/5/2014).

(5) Dall'abitazione sono stati portati via 3 mila euro e alcuni liquori, ma forse i malviventi - che **parlavano con accento napoletano** - cercavano altro, visto che hanno tagliato i materassi e messo a soqquadro la casa (CS 27/1/2016).

(6) Valmontone, i carabinieri sventano rapina in banca, catturati tre banditi – Erano le 14.30 di ieri pomeriggio quando tre persone **dall'accento napoletano** con il volto coperto da passamontagna si sono presentati nella Banca dei Monti di Paschi di Siena a Valmontone (MS 20/3/2015).

Nell'esempio (7) è interessante notare l'uso delle due parentetiche per specificare in successione l'origine dei rapinatori.

(7) I cinque rapinatori, pare tutti italiani, **uno con accento napoletano**, sono sbucati poco prima dell'orario di apertura nel seminterrato della filiale della Banca popolare di Novara di piazza Otto Novembre (GN 26/8/2017).

Il primo inciso viene in parte modulato dal verbo *pare*, che introduce un'informazione non certa sull'italianità dei rapinatori; il secondo specifica però la provenienza di uno di essi. Si vuole forse dare uno dei pochi dettagli certi per identificare i rapinatori, ma ci si potrebbe chiedere in che modo renda il testo più informativo. Quella di marcare la provenienza attraverso l'accento delle persone coinvolte è una strategia discorsiva già studiata nel nesso tra criminalità e immigrazione nella stampa;[17] che possa trattarsi della stessa soluzione è palesato in modo quasi inequivocabile, e a dire il vero un po' goffo, nell'articolo dell'esempio (8), tratto però da *Il Mattino*:

(8) Un particolare, questo, da tenere in considerazione: i banditi non hanno aperto bocca. Perché? Per timore di farsi riconoscere? **Perché non volevano rivelare l'accento napoletano**, **o magari straniero**? (MaT 4/10/2018).

Se è vero, come abbiamo visto, che questa scelta discorsiva non è riservata esclusivamente ai meridionali, lo è però nella larghissima parte e non possono essere una manciata di occorrenze con un etnonimo settentrionale a bilanciare il quadro. Il fatto, poi, che sia una forma condivisa con articoli su altre persone definite per la loro alterità (gli immigrati) permette forse di intravedere il filo comune del pregiudizio.

17 Cfr. Orrù (2017: 139-141).

Veniamo ora alla parola *enne*, a cui abbiamo accennato sopra per il suo valore di *keyness*. Le collocazioni della forma illustrano come essa sia a sua volta associata nel corpus-Sud a parole come *arrestato*, *residente*, *denunciato*, *incensurato*, *vittima*, *pregiudicato*, *morto*, *muore*, *ferito*, *carabinieri*, *bloccato*, *accoltellato*, *ucciso*, *manette*. La situazione non è molto diversa nel corpus-Nord, anche se vi si possono trovare diversi etnonimi stranieri, come *marocchino*, *senegalese*, *albanese*, *egiziano*.

Nel corpus-Sud le prime due collocazioni sono proprio *arrestato* e *napoletano* (613 e 1019 occorrenze, entrambe con un indice LogDice di oltre 10 su 14); nel corpus-Nord, invece, sono 135 le co-occorrenze con *arrestato* e 435 con *milanese* (LogDice 8,89 e 8,42). Continuando a catena tra le associazioni, sono solo 20 i casi di *enne arrestato* con *milanese* nell'area delle 5 parole a destra e a sinistra, mentre sono 247 con *napoletano*. I freddi numeri, insomma, sembrano dirci che è molto più frequente indicare la provenienza napoletana rispetto a quella milanese. Va in realtà precisato che la gran parte di questi casi è legato proprio all'unico quotidiano partenopeo. Effettuando le stesse ricerche senza gli articoli del *Mattino*, *napoletano* perderebbe in salienza e si otterrebbero solo 34 articoli. Scorrendo i testi non si trovano forti differenze di trattamento tra Nord e Sud: gli articoli, ad esempio, non legano i reati commessi alla componente etnico-geografica. Una differenza è però ravvisabile e sostanziale e potrebbe essere indicativa per ulteriori scrutini: sui 20 casi estratti dal corpus-Nord, solo 6 riguardano l'effettivo etnonimo *milanese*, tutti gli altri sono invece toponimi; la proporzione è sostanzialmente ribaltata al Sud in cui l'etnonimo è usato 25 volte su 34. Si può quindi ritenere che la provenienza sia maggiormente enfatizzata per i napoletani.

(9) Napoli, arrestato latitante: aveva prenotato i biglietti per la partita **Un 33 enne latitante napoletano** è stato arrestato dai carabinieri dopo essere stato tradito dalla sua irrefrenabile passione per... il Napoli calcio! (GN 12/10/2017).

(10) "Deve pagare, altrimenti arrestano suo figlio": l'ennesima truffa Ma in cella finisce un finto avvocato I carabinieri della Compagnia di Civitavecchia hanno arrestato **un 47 enne napoletano**, già conosciuto alle forze dell'ordine, con l'accusa di truffa aggravata. L'imbroglione è stato scoperto grazie ad alcune segnalazioni, giunte ai militari nelle prime ore della mattinata di ieri (MS 24/1/2017).

(11) Napoli, rapinava orologi preziosi in Costa Azzurra: arrestato dalla polizia **Il 26 enne napoletano** è stato sorpreso mentre era in vacanza con i familiari a Minturno (CS 10/8/2018).

Per completezza diamo di seguito alcuni esempi con *milanese*:

(12) Milano, rapina market e fugge in scooter. Ma viene intercettato e arrestato In carcere **un 22 enne milanese**. Si era fatto consegnare dalle commesse del Carrefour 4mila euro (CS 17/9/2017).

(13) Vendono casa, ma è di un altro **Un milanese 50 enne arrestato**, in fuga la complice finta proprietaria (GN 25/4/2018).

Nell'estratto seguente si può notare, invece, l'uso prevalente di *milanese* come spe-

cificazione toponimica:

(14) Litiga col vicino di casa per i rumori e lo aggredisce per strada con un taglierino: arrestato **62 enne nel milanese** […] A Nova Milanese, ieri mattina alle 11.30 in via Garibaldi, un uomo italiano nato nel 1956 ha aggredito per strada un suo vicino di casa, anche lui italiano, nato nel 1960 (RP 24/2/2019).

Nel titolo viene fornita l'informazione sulla localizzazione degli avvenimenti; nell'apertura dell'articolo si precisa la nazionalità delle persone coinvolte, forse per escludere nel lettore la chiave interpretativa dell'immigrazione, ma non si sa o non si vuole dire l'esatta provenienza. L'effetto di sostituire l'etnonimo con la semplice posizione geografica del reato potrebbe suggerire l'attivazione di una sorta di implicito che agisce per metonimia: vicinanza del luogo = provenienza dell'autore del reato. È un accostamento che può risultare quasi del tutto automatico per il lettore non troppo attento, ma che lascia comunque spazio ad altre possibili interpretazioni, soprattutto quando dal testo vengono espunte le generalità o altri riferimenti espliciti. Rimane poi da considerare che nei casi in cui il protagonista sia un napoletano, il dettaglio è reso esplicitamente più del doppio delle volte.

4. Conclusioni

L'obiettivo del presente contributo era quello di delineare un metodo per costruire un corpus di articoli su un aspetto della realtà sociale e illustrare alcune tra le principali tecniche per la sua analisi. Non siamo, quindi, andati nel dettaglio dell'analisi puntuale degli esempi testuali per privilegiare l'aspetto più prettamente metodologico. Abbiamo comunque voluto indicare alcuni spunti da approfondire in futuro per l'indagine della rappresentazione del dualismo tra Nord e Sud sulla stampa quotidiana.

Una delle questioni aperte relative all'uso dei mezzi quantitativi è che rende semplice identificare cosa succeda nelle vicinanze di una data parola, ma non fenomeni più distanti o complessi (ad esempio i rimandi anaforici) dal termine ricercato; i dati vanno, insomma, costantemente verificati e controllati nel contesto di discorso più ampio, fermarsi solamente alle prime evidenze numeriche, che pure possono sembrare illuminanti, rischia di offrire un'immagine distorta di ciò che realmente può trovarsi nei testi. L'analisi quantitativa, quindi, va sempre raffinata attraverso quella qualitativa. Se da un lato l'uso delle tecniche statistiche e informatiche consente di verificare e formulare ipotesi più rapidamente e di trovare il conforto dei numeri e della rilevanza statistica, allo stesso tempo è necessario investire tempo nell'attento scrutinio dei testi. Il punto forte può essere, insomma, anche un punto debole del metodo, di fatto risulta dispendioso se non impossibile verificare decine di migliaia di esempi. Il corpus da noi utilizzato presenta poi importanti problemi dal punto di vista dell'annotazione automatica delle parti del discorso; ciò impedisce una più completa analisi sintattica. Tuttavia, il corpus Timestamped attraverso la piattaforma

Sketch Engine si dimostra un mezzo estremamente utile, semplice e flessibile per l'analisi dei fenomeni sociali attraverso la stampa quotidiana.

I pochi esempi discussi non hanno certo alcuna pretesa di esaustività o di generalizzazione. Al contrario, il saggio offerto ci mostra come gli strumenti informatici possano allo stesso tempo indicarci vie di ricerca inaspettate (consultando ad esempio la forma *enne*), confermare ipotesi formulate attraverso la pura intuizione personale (è il caso dell'*accento* o delle differenze tematiche tra Nord e Sud) oppure aiutarci a rigettare e rifinirle (il fatto che il Sud sia raccontato soprattutto per aspetti criminali). È indubbio che le "brutte notizie" siano da anni ormai ritenute dalle redazioni più seducenti per il lettore; non deve, quindi, sorprendere che dai dati del corpus emergano costantemente articoli di questo genere.[18] Appare altrettanto evidente da queste prime cursorie esplorazioni del corpus come tale tendenza generale vada a concentrarsi forse più del dovuto su un'area specifica del Paese. Per converso, l'idea che la stampa nazionale sia focalizzata solo sugli aspetti negativi del Meridione è forse da rivalutare.

BIBLIOGRAFIA

Baker 2006 = Paul Baker, *Using Corpora in Discourse Analysis*, London, Continuum.

Baker *et al.* 2008 = Paul Baker *et alii*, *A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK Press*, in «Discourse and Society», 19, 3, pp. 273-306.

Baroni *et al.* 2004 = Marco Baroni *et alii*, *Introducing the "la Repubblica" corpus*: *A large, annotated, TEI(XML)-compliant corpus of newspaper Italian*, in Proceedings of LREC 2004.

Bušta/Herman 2017 = Jan Bušta / Ondřej Herman, *JSI Newsfeed Corpus*, in *The 9th International Corpus Linguistics Conference*. *Corpus Linguistics 2017 Conference*, University of Birmingham, 25-28 July 2017, https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper382.pdf.

Cremonesini 2015 = Valentina Cremonesini, *Il Sud nei giornali*: *La Repubblica e il Corriere della Sera (dal 1980 al 2010)*, in Stefano Cristante / Id. (a cura di), *La parte cattiva dell'Italia*: *Sud*, *media e immaginario collettivo*, Milano, Mimesis, pp. 177-282.

Cristante 2015 = Stefano Cristante, *Cosa dice il Tg1 del Sud? Parole, immagini e cornici cognitive dal telegiornale più visto in Italia (1980-2010)*, in Id. / Valentina Cremonesi (a cura di), *La parte cattiva dell'Italia*: *Sud*, *media e immaginario collettivo*, Milano, Mimesis, pp. 112-185.

Fairclough 1989 = Norman Fairclough, *Language and Power*, New York, Longman.

Fairclough / Wodak 1997 = Norman Fairclough / Ruth Wodak, *Critical Discourse Analysis*, in Teun Adrianus van Dijk (ed.), Discourse Studies: A Multidisciplinary Introduction. Vol. 2

18 Anche se andrebbe ricordato che quasi tutti i tipi di reati sono in diminuzione costante da anni

Discourse as Social Interaction, London, Sage, pp. 258-284.
Foucault 2009 = Michel Foucault, *L'archeologia del sapere*, Milano, Bur (5a ediz.).
Gablasova *et al.* 2017 = Dana Gablasova *et alii*, *Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence*, «Language Learning», 67, S1, pp. 130-154.
Gabrielatos 2018 = Costas Gabrielatos, *Keyness Analysis*: *Nature*, *Metrics and Techniques*, in Charlotte Taylor / Anna Marchi (eds), *Corpus Approaches To Discourse*: *A Critical Review*, Oxford, Routledge, pp. 225-258.
Gribaudi 2010 = Gabriella Gribaudi, *Nord e Sud*: *una geografia simbolica*, in «Contemporanea», XIII (1), pp. 105-118.
Jakubíček *et al.* 2013 = Miloš Jakubíček *et alii*, *The TenTen Corpus Family*, in *7th International Corpus Linguistics Conference CL*, pp. 125-127, https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf.
Kilgarriff 2009 = Adam Kilgarriff, *Simple maths for keywords*, in Michaela Mahlberg *et alii* (eds.), *Proceedings of Corpus Linguistics Conference CL2009*, University of Liverpool, https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf.
Kilgarrif *et al*. 2014 = Adam Kilgarriff *et alii*, *The Sketch Engine*: *Ten Years on*, in «Lexicography», 1, pp. 7-36.
Louw 1993 = Bill Louw, *Irony in the text or insincerity in the writer*? *The diagnostic potential of semantic prosodies*, in Mona Baker / Gill Francis / Elena Tognini-Bonelli (eds.), *Text and technology: In Honour of John Sinclair* Amsterdam, John Benjamins, pp. 157-175.
Partington 2004 = Alan Partington, *Corpora and discourse*: *A most congruous beast*, in Id. / John Morley / Louann Haarman (eds), *Corpora and Discourse*, Bern, Peter Lang, pp. 11-20.
Partington *et al.* 2013 = Alan Partington *et alii*, *Introduction*, in Idd., *Patterns and Meanings in Discourse*. *Theory and Practice in Corpus-assisted discourse studies*, Amsterdam, John Benjamins, pp. 1-24.
Phillips 1989 = Martin Phillips, *Lexical structure of texts*, Birmingham, English Language Research, University of Birmingham.
Orrù 2017 = Paolo Orrù, *Il discorso sulle migrazioni nell'Italia contemporanea*: *un'analisi linguistico-discorsiva sulla stampa (2000-2010)*, Milano, FrancoAngeli.
Rossi 2015 = Fabio Rossi, *Dalla questione della lingua all'aggressione linguistica: le idee sulla lingua nei giornali italiani dell'ultimo decennio*, in «Circula, Revue d'idéologies linguistiques», 1, pp. 173-195.
Rychlý 2008 = Pavel Rychlý, *A Lexicographer-Friendly Association Score*, in *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing*, RASLAN, 2, pp. 6-9.
Scott 1997 = Mike Scott, *PC analysis of key words - And key key words*, in «System», 25, 2, pp. 233-245.