4 Metodologia e *corpus* di analisi

Sommario 4.1 Domande e metodologia di ricerca. – 4.2 Modelli teorici e strumenti di analisi. – 4.3 *Corpus* di analisi.

4.1 Domande e metodologia di ricerca

Come anticipato nell'Introduzione, questo volume presenta una ricerca sulle rappresentazioni di genere nel linguaggio dei TG italiani che si propone di essere interdisciplinare, linguistica e mediale, partendo dal presupposto che i media contribuiscono alla costruzione della realtà attraverso rappresentazioni sociali, incluse quelle di genere, codificate, decodificate e trasmesse (anche) dall'uso della lingua (Moscovici 1969; Losito 1998). Per questo si è scelto di condurre l'analisi di un *corpus* linguistico che raccoglie le trascrizioni di un campione di TG trasmessi nel triennio 2018-20, e che è descritto dettagliatamente nel paragrafo 4.3. Il *corpus* è stato esplorato attraverso due set di domande funzionali a orientare l'analisi in una prospettiva di genere, binaria e comparativa (cf. capitolo 1). La prima serie di domande indaga il parlato delle fonti giornalistiche, ovvero il parlato delle persone intervistate (o di cui viene trasmessa una dichiarazione in voce) ed è articolata nei seguenti quesiti: quali sono le

DOI 10.30687/978-88-6969-728-9/004

rappresentazioni delle donne e degli uomini come fonte d'informazione nei TG? Ci sono somiglianze o differenze fra gli anni? Ci sono somiglianze o differenze fra le testate giornalistiche? La seconda serie di domande indaga la tematizzazione di donne e uomini, assumendo i lemmi *donna* e *uomo* come parole *target* del discorso giornalistico, ed è articolata come segue: quali sono le rappresentazioni di genere di donne e uomini come argomento d'informazione nei TG? Ci sono somiglianze o differenze fra gli anni? Ci sono somiglianze o differenze fra le testate giornalistiche?

Per rispondere a queste domande, si è optato per una metodologia di analisi del contenuto semiautomatica, che ha una consolidata tradizione di ricerca nelle scienze umane e sociali (Bolasco 2013; Pandolfini 2017), dove viene tradizionalmente utilizzata per analizzare dati testuali, al fine di rilevarne contenuti latenti. Nell'ambito dei *media studies* è considerata un metodo di analisi del contenuto che, a differenza di altri, ha il vantaggio di superare la 'storica' distinzione fra analisi qualitativa e quantitativa, consentendo un approccio funzionale a integrare fra loro non solo aspetti qualitativi e quantitativi, ma anche linguistici e mediali.

L'analisi del contenuto, con riferimento a contenuti mediali, può essere definita come un insieme di metodi finalizzati a rispondere a domande di ricerca, o verificare ipotesi, su

> fatti di comunicazione (emittenti, messaggi, destinatari e loro relazioni) e che a tale scopo utilizzano procedure di scomposizione analitica e di classificazione, normalmente a destinazione statistica, di testi e di altri insiemi simbolici. (Rositi 1988, 66)

Tradizionalmente, questo insieme di metodi viene distinto in due grandi categorie: l'analisi del contenuto qualitativa e l'analisi del contenuto quantitativa (Losito 1996; Tuzzi 2003; Pandolfini 2017). I criteri su cui si basa questa distinzione sono principalmente due: il modo in cui i contenuti vengono scomposti e classificati, e il modo in cui essi vengono analizzati (Pandolfini 2017). L'analisi del contenuto qualitativa procede per scomposizione e classificazione, mediante individuazione di temi o concetti per deduzione o per induzione, in ogni caso secondo una modalità *open-coding*: le categorie di analisi vengono definite e ridefinite in corso d'opera. Nel processo per deduzione, i contenuti vengono scomposti e classificati con categorie identificate *a priori* sulla base di un quadro teorico di riferimento e un'ipotesi di ricerca, sviluppati poi *in itinere*, durante l'analisi, la quale si caratterizza per una progressiva revisione delle categorie classificatorie. Nel processo per induzione le categorie sono invece individuate e definite in fase di analisi, mediante un processo di progressiva creazione e revisione delle categorie stesse. Indipendentemente dall'adozione di un approccio di tipo deduttivo o induttivo,

l'analisi del contenuto qualitativa procede per interpretazione argomentativa dei risultati, che vengono confrontati con l'ipotesi di ricerca e la cornice teorica di riferimento, in genere senza il supporto di processi di automazione computerizzata né statistiche, tradizionalmente impiegate invece nell'analisi del contenuto quantitativa (Pandolfini 2017).

L'analisi del contenuto quantitativa procede per scomposizione e classificazione dei contenuti tramite categorie *a priori* che, una volta individuate sulla base delle domande di ricerca e di un *framework* teorico di riferimento, rimangono invariate e vengono trattate come variabili categoriali. Il processo di analisi si avvale di misurazioni statistiche che possono essere di vario tipo e possono riguardare le unità di analisi nella loro interezza, oppure elementi che costituiscono le unità di analisi. In quest'ultimo caso, le unità di analisi vengono assunte come unità di contesto, mentre gli elementi analizzati diventano le unità di analisi nel contesto.

L'analisi quantitativa può essere svolta in modo automatico o semiautomatico, in entrambi i casi con il supporto di strumenti informatici. L'analisi automatica prevede l'analisi di testi non pre-processati, ovvero non preliminarmente scomposti e classificati. L'analisi viene svolta automaticamente da un programma informatico che, al più, richiede la configurazione di parametri per l'esecuzione di operazioni specifiche. Occorre tuttavia precisare che, anche se tutti i processi sono automatizzati, l'analisi automatica non può prescindere, sia nella fase di disegno della ricerca sia nella fase di interpretazione dei risultati, da una discussione che consideri aspetti qualitativi, come lo stato dell'arte, per esempio, necessario per definire la cornice teorica e contestuale della ricerca e formulare domande di indagine. Valutazioni di ordine qualitativo coinvolgono anche la scelta del software, poiché ogni programma ha caratteristiche proprie e si basa su algoritmi e operazioni, che implementano un metodo riferito a un modello teorico piuttosto che a un altro (Chartier, Meunier 2011; Lahlou 2012; Bolasco 2013).

L'analisi semiautomatica introduce elementi qualitativi non solo nelle fasi preliminari e finali della ricerca, ma anche nella fase empirica, che richiede un pre-processamento manuale dei contenuti, preliminarmente scomposti e classificati in unità di analisi distinte e annotate con variabili categoriali, le unità di contesto iniziali (UCI). Questo pre-trattamento dei dati consente di combinare l'approccio qualitativo con quello quantitativo direttamente nella fase sperimentale della ricerca, perché le variabili categoriali introducono informazioni qualitative, non numeriche, che possono però essere assunte come oggetto di misurazioni statistiche. Per esempio, una preliminare scomposizione del testo trascritto di una serie TV in dialoghi classificati per genere del/la parlante, rende disponibile, per una fase di analisi successiva, la quantificazione dello spazio di

parola dei personaggi femminili vs. quelli maschili, restituendo un risultato quali-quantitativo. Sebbene sia un approccio che implica un impiego di risorse piuttosto oneroso, impraticabile nell'analisi di *big data*, l'analisi semiautomatica ha il vantaggio di evitare la perdita di informazioni spesso difficili da recuperare in un processo puramente automatico. Tornando all'esempio delle serie TV, le informazioni sul genere, o altre caratteristiche dei personaggi, sono ricavabili attraverso una visione audiovisiva integrale delle serie TV, sulla base della quale i contenuti possono essere scomposti per dialoghi e classificati per caratteristiche dei personaggi.[1]

Scegliere fra un approccio completamente automatico o semiautomatico può non sempre essere facile, occorre valutare una serie di elementi, quali l'obiettivo dell'analisi e le caratteristiche del campione. Dato che l'obiettivo della presente ricerca è quello di rilevare le rappresentazioni di genere nel linguaggio dei TG, partendo dal presupposto teorico che le rappresentazioni di donne e uomini veicolati dall'informazione si codificano, decodificano e trasmettono anche attraverso l'uso della lingua, l'analisi semiautomatica di un *corpus* testuale si è configurata come la scelta più adeguata. La parte 'semi', ovvero il pre-processamento manuale del *corpus*, consente infatti di inserire variabili categoriali che tengono conto di elementi mediali, quali la struttura del TG e le caratteristiche delle fonti giornalistiche, in modo tale che la parte 'automatica', ovvero l'analisi computerizzata, permetta poi di fare emergere elementi linguistici codificati nel *corpus*, anche in relazioni alla struttura del TG e alle caratteristiche delle fonti.

Inoltre, generalmente, l'approccio automatico risulta la scelta migliore, talvolta obbligata, per l'analisi di *big corpora*, perché il pretrattamento di *big data* sarebbe troppo oneroso. L'approccio semiautomatico può invece essere una buona scelta per l'analisi di piccoli *corpora*, perché la loro dimensione contenuta comporta tempi di preprocessamento manuale 'ragionevoli'. Il *corpus* analizzato, come vedremo dettagliatamente nell'ultimo paragrafo di questo capitolo, è piccolo, sia che si considerino i parametri dimensionali tradizionalmente utilizzati nella linguistica dei *corpora* (Davies 2015, 11), come per esempio il numero di *token* che sono poco più di un milione (N=1.434.733), sia che si consideri la dimensione del contenuto da visionare, in questo caso circa 139,5 ore di trasmesso televisivo, una durata ragionevolmente visionabile, per il pre-processamento manuale dei dati.

1 Negli anni più recenti, l'Intelligenza Artificiale ha sviluppato algoritmi sempre più affidabili nel riconoscimento e nella trascrizione del parlato in contenuti audiovisivi che potrebbero eseguire automaticamente questo compito (cf. Potamianos et al. 2012).

4.2 Modelli teorici e strumenti di analisi

Per la parte automatica dell'analisi del *corpus* è stato utilizzato il software IRaMuTeQ (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires). Il programma è scritto in Python, basato su pacchetti di analisi statistica di R, distribuito sotto licenza software libero GNU GPL (v2), ed è stato sviluppato da Pierre Ratinaud nell'ambito del LERASS (*Laboratoire d'Études et de Recherches Appliquées en Science Sociales*) dell'Università di Tolosa 3, originariamente in francese e per l'analisi della lingua francese (Baril, Garnier 2015). Ora è disponibile anche in lingua inglese e supporta l'analisi di undici lingue diverse, incluso l'italiano.[2] La versione utilizzata è la numero 0.7 alpha 2 (2020), basata sul pacchetto R 4.0.3.[3]

La scelta di IRaMuTeQ, fra i diversi programmi di analisi testuale disponibili, è stata orientata da tre motivi: il primo è che si tratta di una *open source*, essendo distribuita sotto licenza di software libero, come già scritto; il secondo è che, come vedremo di seguito, dispone di un'ampia varietà di tecniche di analisi; il terzo riguarda i due modelli teorici cui il software fa riferimento, ovvero il modello della semantica distribuzionale e il modello dei 'mondi lessicali'. Il primo si basa sull'ipotesi distribuzionale originariamente intuita da Wittgenstein (1953) e Firth (1957) e sviluppata in modo più articolato da Harris (1954)), che non si limita a osservare come «the meaning of words lies in their use» (Wittgenstein 1953, 80) o «you shall know a word by the company it keeps» (Firth 1957, 11), ma afferma che «difference of meaning correlates with difference of distribution» (Harris 1954, 156). Questo modello è ripreso anche da Benzécri (1973; 1980) nell'ambito della scuola francese dell'*analyse des données* per sviluppare una metodologia di analisi che si focalizza sulla distribuzione delle parole in un *corpus*, o in *corpora* diversi a confronto, e che è alla base dello sviluppo di IRaMuTeQ.[4]

Il secondo modello implementato dal software francese si propone come revisione del modello di Benzécri (1973; 1980), nel senso di un'analisi di dati linguistici focalizzata sulla distribuzione delle parole non nel *corpus* integralmente considerato, bensì nel *corpus* suddiviso in segmenti di testo assunti come unità di classificazione, e sulla distribuzione e l'interrelazione di questi segmenti nel *corpus* la cui struttura si suppone 'memorizzi' le condizioni di produzione del testo (Reinert 1993, 9). Reinert (1990) parte dal presupposto che la

2 Inglese, francese, galiziano, greco, italiano, norvegese, olandese, portoghese, spagnolo, svedese, tedesco.

3 Versione rilasciata il 25/11/2020.

4 Per un approfondimento sui modelli di semantica distribuzionale cf. Lenci (2010).

semantica di un enunciato si differenzia dalla semantica di una parola, poiché contiene 'l'impronta' di un soggetto psichico, vale a dire 'memorizza' la codifica e decodifica della realtà da parte di un soggetto. Un enunciato semplice è la parte più piccola di un discorso in cui un soggetto esprime la propria rappresentazione del mondo, rappresentazione mentale che lo mette in relazione con la realtà esterna. Un *corpus* può essere considerato, oltre che un insieme di parole, come un insieme di enunciati semplici, che sono rappresentazioni elementari della realtà. Studiando somiglianze e differenze del vocabolario fra i segmenti di testo che compongono un *corpus*, possiamo cogliere somiglianze o differenze fra le forme di relazione con il mondo, forme che, sempre secondo Reinert (1990), ci sono accessibili solo come rappresentazioni della realtà, che Reinert chiama 'mondi lessicali' o 'tipi di mondo': «les types de mondes référentiels les plus sollicités par un sujet psychique, lors de l'élaboration du *corpus*» (Reinert 1990, 21-2), formulando così una nozione congruente con la teoria delle rappresentazioni sociali di Moscovici, su cui si basa la ricerca presentata in questo volume, come ampiamente discusso nei precedenti capitoli.

Partendo da questa ipotesi, il linguista francese mette a punto il metodo della Classificazione Gerarchica Discendente (CGD) in grado di suddividere un *corpus* in 'classi' caratterizzate, sul piano lessicale, da omogeneità interna e diversità esterna, rispetto alle altre classi e, sul piano semantico, da 'mondi lessicali', ovvero rappresentazioni della realtà latenti nella struttura di un *corpus*. In questo contesto, il termine 'mondo' non va ovviamente inteso in senso realistico, ma cognitivo: «un monde apparait, au niveau cognitif, à travers un ensemble plus ou moins organisé de signes relatifs à des objets, des actes, des jugements, ecc.» (Reinert 1993, 13), e la CGD come un approccio euristico per rilevare non il mondo 'reale', ma il suo substrato simbolico costruito mediante l'uso della lingua. Questa particolarità della CGD è potenzialmente interessante per l'analisi di un *corpus* di telegiornali, focalizzata sulle rappresentazioni di donne e uomini. Le diverse testate giornalistiche possono infatti essere incluse fra quei 'soggetti collettivi' di cui scrive Reinert (1993, 12), i cui 'mondi lessicali' si caratterizzano come 'luoghi comuni' collocati in uno spazio intermedio fra le rappresentazioni individuali (il punto di vista sul mondo del singolo soggetto) e i pre-costrutti culturali (il punto di vista sul mondo condiviso da una comunità). Questi ultimi si possono 'imporre' all'enunciatore/trice più di quanto l'enunciatore/trice non li scelga, anche se li ricostruisce conferendo loro una particolare 'colorazione', cioè un senso entro un universo (simbolico) socialmente condiviso. Una considerazione, quest'ultima, che appare particolarmente rilevante se si considera che il *corpus* analizzato è una raccolta di testi informativi e l'informazione televisiva si contraddistingue, come ampiamente discusso nei capitoli 2 e 3,

per rappresentazioni di genere ambivalenti fra stereotipi e innovazioni. Entro questa ambivalenza giocano un ruolo fondamentale sia la cultura personale, espressa da giornaliste/i e fonti dell'informazione, sia la cultura del paese, di cui entrambi sono portatori, sia la cultura giornalistica e/o redazionale, condivisa da chi scrive e produce le notizie. E ovviamente l'uso della lingua, attraverso la quale le rappresentazioni sociali si codificano, decodificano e trasmettono (Moscovici 1969).

Venendo ora alle caratteristiche di IRaMuTeQ, questo software consente di analizzare dati linguistici raccolti sotto forma di *corpus* di tipo specialistico, ovvero circoscritto a un determinato genere testuale e/o argomento, anche pre-processato manualmente, attraverso uno standard di annotazione che utilizza il simbolo dell'asterisco per suddividere il *corpus* in UCI, ovvero segmenti di testo marcati con le variabili categoriali precedentemente selezionate, per esempio il genere del/la parlante. Le tipologie di analisi eseguibili sono cinque: l'analisi lessicometrica; l'analisi delle specificità e delle corrispondenze lessicali; la classificazione gerarchica discendente; l'analisi delle somiglianze; la 'nuvola di parole' (*word cloud*).

L'analisi lessicometrica identifica e riformatta il *corpus*, o le UCI nel caso in cui il testo sia stato segmentato in una fase di pre-processamento dei dati, in unità di classificazione elementari (UCE), sulla base di un algoritmo originariamente messo a punto per il software francese Alceste (Reinert 1983). Dopo aver effettuato questa operazione, l'analisi calcola la frequenza delle UCE, il numero di *token*, ovvero il numero delle parole e di tutte le altre unità linguistiche, come per esempio numeri, sigle, segni di punteggiatura, e così via, occorrenti nel *corpus*, il numero di *type*, ovvero delle forme flesse delle parole, il numero di lemmi, se si è optato per una lemmatizzazione del *corpus* (e in questo caso *type* e lemmi coincidono) e il numero degli *hapax*,[5] il rapporto fra numero di *hapax* e numero di *token*, fra numero di *hapax* e numero di *type*, e, infine, calcola il rapporto fra il rango di frequenza e la frequenza dei *token* attestata nel *corpus*, verificando la legge di Zipf (1949), in base alla quale il rango di una parola in un *corpus* cresce al decrescere della sua frequenza nel *corpus* stesso. Con IRaMuTeQ la frequenza dei *type* può essere calcolata secondo due criteri diversi: lemmatizzando o non lemmatizzando il *corpus*. Come vedremo meglio di seguito, il *corpus* analizzato è stato lemmatizzato.

IRaMuTeQ esegue la lemmatizzazione sulla base di due dizionari: quello delle parole e quello delle espressioni. Il dizionario delle parole è organizzato in una tabella in cui a ogni parola corrisponde un lemma e la rispettiva annotazione grammaticale, che, per la

5 Gli *hapax* sono parole con occorrenza N=1.

lingua italiana, si limita a un livello basilare di PoS *tagging* (Part of Speech), riconoscendo aggettivi, articoli definiti e indefiniti, avverbi, congiunzioni, nomi, numeri e cifre, preposizioni, pronomi, verbi. Una volta annotate, le parole vengono anche distinte in lessicali e grammaticali.[6] Il software propone una classificazione che considera lessicali gli aggettivi, i nomi comuni, e i verbi, e grammaticali gli articoli, gli avverbi, le preposizioni, i pronomi, i nomi propri e tutte le altre unità linguistiche non riconosciute dal PoS *tagging*.[7] Questa tassonomia può essere variata dall'utente. Nell'analisi qui presentata si è optato per una modifica che include i nomi propri fra le parole lessicali. Il dizionario delle espressioni raccoglie «multi-word expressions» (Masini 2019, 1), ovvero associazioni fra due o più parole che si comportano come un'unità grammaticale e lessicale. Sulla base di questi dizionari e delle UCE, IRaMuTeQ individua le parti del discorso di ogni UCE, recuperando la funzione grammaticale delle parole nel *corpus*, secondo il contesto, e disambiguandone la funzione in caso di omonimia.

L'analisi delle specificità e delle corrispondenze lessicali è un'analisi di tipo comparativo che permette di analizzare il *corpus* sulla base delle variabili categoriali predefinite e utilizzate per annotare il testo, confrontando fra loro le diverse modalità della variabile, che devono essere almeno tre. Per esempio, se si sceglie di annotare il *corpus* sulla base dell'anno di produzione del testo (es. 2018, 2019, 2020), l'analisi delle specificità consente di confrontare fra loro i testi del 2018, 2019, 2020.[8]

I risultati dell'analisi delle specificità possono essere utilizzati anche per la costruzione di un piano fattoriale, che rappresenta graficamente la vicinanza e la lontananza di linguaggio tra le diverse modalità delle variabili. La costruzione del piano fattoriale viene effettuata tramite un'Analisi delle Corrispondenze Lessicali (ACL), un'applicazione dell'Analisi delle Corrispondenze Multiple (ACM) a dati testuali, tipicamente usata nell'analisi automatica dei testi. La sua funzione è quella di fare emergere relazioni latenti tra un alto numero di variabili interdipendenti, la cui relazione viene ricondotta a poche variabili sintetiche, definite 'fattori' o 'dimensioni latenti'. L'applicazione

6 IRaMuTeQ utilizza l'etichetta 'attive' per definire le parole lessicali e l'etichetta 'supplementare' per definire le parole grammaticali (Baril, Garnier 2015, 8).

7 Il PoS *tagging* di IRaMuTeQ utilizza l'etichetta 'nr' per tutti i *token* che non sono presenti nel dizionario delle parole e dunque non riconosce come aggettivi, articoli definiti e indefiniti, avverbi, congiunzioni, nomi, numeri e cifre, preposizioni, pronomi, verbi.

8 Nel caso in cui una variabile presenti soltanto due modalità (per esempio il genere del/la parlante uomo o donna), e non è dunque possibile eseguire l'analisi delle specificità, è possibile svolgere un'analisi contrastiva costruendo due *sub-corpora* e confrontandoli a tutti i livelli di analisi.

dell'ACM a dati linguistici assume il testo come variabile e le forme lessicali (o i lemmi) occorrenti nel testo come modalità della variabile 'testo' e misura il variare della variabile 'testo' in relazione alle variazioni del 'lessico'. In pratica, l'ACL si basa sulla costruzione di una tabella con in riga segmenti di testo (ogni riga un segmento) e in colonna forme lessicali (ogni colonna una forma). A partire da questa matrice, attesta associazioni significative tra forme lessicali, in relazione al loro profilo di ripartizione tra i segmenti di testo; tra segmenti, in relazione alla loro somiglianza sotto il profilo lessicale; fra forme lessicali e segmenti che compongono un testo. I segmenti di testo possono essere pre-selezionati sulla base di variabili categoriali con cui il *corpus* è stato pre-processato.

La classificazione gerarchica discendente (CGD) è stata messa a punto da Reinert (1983; 1990) e originariamente implementata con Alceste, il software da cui IRaMuTeQ eredita alcuni modelli e tecniche di analisi. La CGD procede per progressiva suddivisione del *corpus* in 'classi lessicali', raggruppate secondo i rispettivi vocabolari costruiti sulla base della frequenza delle parole, o dei lemmi, nei segmenti di testo. Il *corpus* viene segmentato dapprima in due classi e poi successivamente per un numero *n* di classi, variabile anche in relazione alla dimensione e alla ricchezza di vocabolario del *corpus*. Poiché la CGD procede per individuazione di classi lessicali che hanno contestualmente un vocabolario simile al loro interno e un vocabolario diverso da altre classi, tanto più il vocabolario di un *corpus* è ricco, tanto più elevato sarà il numero di classi ottenute. IRaMuTeQ consente di sottoporre i risultati della CGD anche a un'ACL che proietta le diverse classi lessicali, ed eventuali variabili categoriali associate, su un piano cartesiano (proiezione fattoriale), dove l'asse delle ascisse e l'asse delle ordinate rappresentano i due principali fattori latenti che spiegano la varianza e la cui natura deve essere interpretata. L'interfaccia consente di recuperare nel *corpus* originale i segmenti di testo associati a ogni classe, permettendo un'analisi più qualitativa dei dati, utile anche a interpretare le dimensioni latenti della proiezione fattoriale.

L'analisi delle somiglianze visualizza graficamente i rapporti di prossimità fra le parole del *corpus*, sulla base di indicatori standard, ovvero misure di tipo distribuzionale, disponibili nella libreria proxy di R, implementata dal software francese.[9] L'indicatore proposto di default da IRaMuTeQ, per cui si è optato in tutte le analisi effettuate, è quello delle coricorrenze testuali assolute, che misura quante volte una parola occorre insieme a un'altra nel *corpus*. Le rappresentazioni grafiche possibili sono più di una, ma hanno tutte in comune l'obiettivo di visualizzare la struttura del *corpus*. Che abbiano una

9 `https://cran.r-project.org/web/packages/proxy/index.html`.

forma simile alle ramificazioni di un albero o di una ragnatela, esse si caratterizzano per rispettare, nella distanza fra parole, le proporzioni reali di vicinanza e lontananza delle stesse nel *corpus* e, nello spessore dei tratti che le uniscono, la misura della relazione effettiva delle parole nel *corpus*, in entrambi i casi sulla base dell'indicatore prescelto.

La *word cloud*, che non è stata utilizzata, è un'altra analisi di tipo grafico che raggruppa e organizza le parole o i lemmi di un *corpus* sulla base della loro frequenza, rappresentandole nella classica forma a nuvola che le visualizza l'una in prossimità all'altra, con dimensioni diverse a seconda del numero di occorrenze e con al centro le parole più frequenti.

4.3 *Corpus* di analisi

Secondo Lenci, Montemagni e Pirrelli (2016, 26), un *corpus* può essere definito come: «una collezione di testi selezionati e organizzati in maniera tale da soddisfare specifici criteri che li rendono funzionali per le analisi linguistiche». A partire da questa definizione, vedremo di seguito quali sono, in generale, i criteri di organizzazione che identificano una raccolta di testi come *corpus* e, in particolare, i criteri con cui è stato costruito e organizzato il *corpus* esplorato in questa ricerca. Tale *corpus* è stato denominato *corpus* TG e raccoglie le trascrizioni integrali dei telegiornali trasmessi in fascia *prime time* da Rai 1, Rai 2 e Canale 5 – TG1 ore 20:00, TG2 ore 20:30 e TG5 ore 20:00 – trasmessi nei mesi di gennaio 2018-2020, coprendo 279 edizioni e circa 139,5 ore di trasmesso, se si considera che ogni edizione dura circa 30 minuti.

Un primo criterio di organizzazione di un *corpus* riguarda il canale dei testi raccolti, che possono essere originariamente scritti o orali; in quest'ultimo caso, vengono trascritti e, in alcuni casi, la trascrizione è associata alla registrazione audio o audiovisiva (*corpora* multimediali). Lo sviluppo delle tecnologie dell'informazione e della comunicazione ha reso disponibili nei tempi più recenti anche i cosiddetti web-*corpora* o *corpora* basati sul web: raccolte liberamente accessibili, scaricabili e in alcuni casi analizzabili direttamente online (Davies 2015, 11). Un secondo criterio riguarda la loro rappresentatività. A differenza di altre tipologie di raccolte, come per esempio le antologie, i *corpora* sono raccolte di testi esemplificativi del linguaggio naturale, che ambiscono a essere rappresentativi di una lingua, una sua varietà o dominio (Leech 1991, 28). A seconda dei casi possono essere *corpora* generali o di riferimento, rappresentativi di una lingua, più lingue, o varietà linguistica, oppure *corpora* verticali o specialistici, rappresentativi di un dominio particolare (Lenci, Montemagni, Pirrelli 2016, 29). Un terzo criterio,

strettamente legato alla rappresentatività, riguarda il bilanciamento: un *corpus*, per essere rappresentativo di una lingua, una varietà o anche un dominio specifico, deve essere esemplificativo della variabilità dei tratti linguistici sia della popolazione di riferimento sia dei testi raccolti. Questa variabilità può essere più o meno ampia a seconda che un *corpus* sia generalista o di riferimento, e anche a seconda della sua dimensione; in ogni caso va rappresentata con adeguati criteri di bilanciamento.

Altri criteri di organizzazione dei *corpora* riguardano la cronologia, la lingua, l'integrità e la codifica digitale (Lenci, Montemagni, Pirrelli 2016, 27-34). Con riferimento alla cronologia, un *corpus* può essere sincronico o diacronico, a seconda che comprenda testi prodotti in un periodo di tempo particolare, per esempio un anno, oppure testi di periodi diversi. Quanto alla lingua, un *corpus* può contenere testi prodotti in una sola lingua o in due o più lingue (*corpora* bilingue e multilingue). Il criterio dell'integrità distingue fra raccolte di testi integrali e raccolte di porzioni di testi di lunghezza fissa, che talvolta vengono ritenute più adeguate al bilanciamento di un *corpus*. Testi di lunghezza molto diversa fra loro possono infatti portare a una collezione sbilanciata a favore dei testi più lunghi che condizionano l'intero *corpus*, distorcendo il campione, per questo in alcuni casi si preferisce optare per una selezione di testi della stessa dimensione. Infine, per quanto riguarda la codifica digitale, i *corpora* possono essere annotati, a vari livelli, con etichette che ne descrivono le caratteristiche linguistiche (per esempio morfologiche, sintattiche, semantiche, pragmatiche) o etichette che introducono variabili categoriali in fase di pre-processamento dei dati.

Il *corpus* TG raccoglie trascrizioni di telegiornali, dunque testi originariamente orali, di tipo semispecialistico, includendo sia trascrizioni delle notizie lette in studio, dei servizi e di eventuali collegamenti in diretta, dunque testi orali basati su uno scritto o semiscritto redazionale, parlato da giornaliste e giornalisti, sia trascrizioni del parlato delle fonti, ovvero persone intervistate o di cui viene trasmesso un discorso o una dichiarazione rilasciata in un altro contesto mediale (per esempio un programma radiofonico o televisivo), oppure extra-mediale (per esempio un convegno). Riguardo alla cronologia, il *corpus* TG è di tipo diacronico, perché raccoglie testi prodotti in tre anni diversi ed è stato organizzato in modo tale da poter essere analizzato su scala longitudinale per una comparazione annuale, ancorché limitata a un periodo breve, qual è un triennio. Si tratta, poi, di un *corpus* in lingua italiana, che può comprendere prestiti da altre lingue, attestandone, nel caso, l'uso nel linguaggio giornalistico. È una raccolta di testi integrali annotata, come vedremo meglio di seguito, in due fasi e con due diverse procedure: manualmente è stata eseguita una annotazione per l'introduzione di variabili categoriali, attraverso etichette contenenti informazioni

riguardanti la struttura del TG e alcune caratteristiche delle/dei parlanti; automaticamente è stata eseguita invece un'annotazione grammaticale.

Le trascrizioni integrali dei TG sono state fornite dall'Osservatorio di Pavia,[10] che dal 2018 utilizza l'API Google Cloud Text-to-Speech[11] per la trascrizione automatica dei telegiornali trasmessi dai canali generalisti di Rai, Mediaset e La7 in fascia *prime time*. La selezione del campione di notiziari si è basata su criteri di *audience*, rappresentatività, bilanciamento e comparabilità. Il TG1 e il TG5 delle ore 20:00 sono i telegiornali più seguiti in Italia: nel 2018 hanno avuto rispettivamente una media di 4.916.212 e 2.775.011 spettatrici/ori,[12] per uno *share* medio rispettivamente del 23,6% e 18,6%; nel 2019 un'*audience* di 5.636.000 e 4.523.000, per uno *share* medio rispettivamente del 24,1% e 19,1%;[13] nel 2020 un'audience di 6.219.000 e 5.230.000 spettatrici/ori e uno *share* medio del 24,1% e 19,9%.[14] Il TG1 è trasmesso da Rai 1, prima rete della concessionaria del servizio pubblico radio-televisivo italiano, il TG5 da Canale 5, la prima rete di Mediaset, che è la seconda *media company* privata nazionale, storica concorrente della Rai, con cui continua a condividere la *leadership* in termini di offerta e ascolto dei notiziari (AGCOM 2020). Il TG2 delle 20:30 è stato scelto perché nel 2018 era diretto da Ida Colucci, una giornalista dichiaratamente impegnata a favore di un linguaggio paritario, inclusivo e non stereotipato.[15] La sua inclusione nel *corpus* TG consente così di verificare similarità o differenze fra linee editoriali diversamente impegnate sul fronte di un linguaggio giornalistico paritario e non stereotipato. Inoltre, è anch'esso fra i TG nazionali più seguiti: nel 2018 ha registrato una media di 1.749.859 spettatrici/ori e uno *share* medio del 7,6%;[16]

10 Per una presentazione dell'Osservatorio di Pavia: https://www.osservatorio.it.

11 https://cloud.google.com/speech-to-text/?hl=it&utm_source=google&utm_medium=cpc&utm_campaign=emea-it-all-it-dr-bkws-all-all-trial-e-gcp-1010042&utm_content=text-ad-none-any-DEV_c-CRE_170511603295-ADGP_Hybrid%20%7C%20BKWS%20-%20EXA%20%7C%20Txt%20~%20AI%20%.

12 https://www.affaritaliani.it/blog/prima-serata/ascolti-tv-auditel-tg1-resta-leader-mentana-579773.html.

13 https://www.adginforma.it/ecco-la-top-ten-dei-tg-nazionali-nel-2019-in-base-allaudience.

14 https://bubinoblog.altervista.org/analisi-auditel-chi-segue-i-telegiornali-delle-ammiraglie/.

15 Cf. «Ida Colucci, la direttrice che ha cambiato linguaggio al Tg2», GiULia Giornaliste, 19 agosto 2020. https://giulia.globalist.it/attualita/2019/08/19/ida-colucci-la-direttrice-che-ha-cambiato-linguaggio-al-tg2/.

16 https://www.affaritaliani.it/blog/prima-serata/ascolti-tv-auditel-tg1-resta-leader-mentana-579773.html.

nel 2019, 1.811.000 spettatrici/ori, per uno *share* del 7,2%;[17] nel 2020, un'audience di 1.791.000 e *share* del 6,69%.[18]

La scelta di limitare la costruzione del *corpus* a testi di TG di gennaio è dipesa dalla disponibilità dell'Osservatorio di Pavia a fornire gratuitamente le trascrizioni dei notiziari di un solo mese nel corso dell'anno;[19] si è quindi optato per gennaio, in modo da coprire due diverse legislature (17esima nel gennaio 2018 e 18esima nel gennaio 2019 e 2020), evitando al contempo i giorni della campagna elettorale per le politiche del 2018 (iniziata a febbraio), in cui la visibilità televisiva, in particolare femminile (Azzalini 2009), è tradizionalmente monopolizzata da *leader* di partito (Mazzoleni 2012), attualmente tutti uomini, con la sola eccezione di Giorgia Meloni. L'inclusione nel *corpus* TG di testi che coprono due legislature diverse è stata motivata dall'esigenza di verificare eventuali interrelazioni fra i risultati dell'analisi e il contesto politico, in particolare in relazione al diverso impegno delle istituzioni nella promozione delle pari opportunità di un linguaggio *gender-fair* nei due diversi periodi legislativi. Il 23 marzo 2018 si è infatti conclusa una legislatura distinta dall'impegno della presidente della Camera Laura Boldrini per un uso non sessista della lingua a partire dall'istituzione della Commissione Jo Cox sui fenomeni d'odio, intolleranza, xenofobia e razzismo,[20] volta a contrastare anche gli stereotipi diffusi a livello linguistico, fino alla promozione dell'uso dei femminili per nominare le cariche politiche di cui si è scritto nel capitolo 2, e si è aperta una legislatura meno impegnata su questo fronte.

Le trascrizioni fornite dall'Osservatorio di Pavia in singoli file di testo (uno per ogni edizione quotidiana di ciascuna testata giornalistica) sono state controllate, corrette e unificate in un documento elettronico in formato testo (txt). La costruzione del *corpus* TG ha avuto inizio da questo *file* che è stato normalizzato e annotato, a un primo livello, manualmente. In fase di normalizzazione, il *corpus* è stato uniformato in relazione all'uso di maiuscole e minuscole, all'ortografia di acronimi (con o senza punti), alla scrittura di parole composte (con o senza trattino) o espressioni multi-parola. Non è stato necessario alcun controllo della punteggiatura, perché le trascrizioni originali non ne contenevano e non si è ritenuto necessario introdurla. In fase di annotazione manuale, il *corpus* è stato scomposto in UCI, ciascuna

17 https://www.adginforma.it/ecco-la-top-ten-dei-tg-nazionali-nel-2019-in-base-allaudience/.

18 https://bubinoblog.altervista.org/analisi-auditel-chi-segue-i-telegiornali-delle-ammiraglie/.

19 È in fase di sviluppo un'interfaccia di consultazione del *corpus* liberamente accessibile online dal sito dell'Osservatorio di Pavia.

20 https://www.camera.it/leg17/1264m.

preceduta da una stringa di testo contenente tutte le informazioni utili a classificarla sulla base di una serie di variabili categoriali pertinenti ad alcune caratteristiche della struttura del TG e delle/dei parlanti (giornaliste/i e fonti). Seguendo lo standard di IRaMuTeQ descritto più sopra, ogni stringa di testo è stata introdotta da quattro asterischi che indicano al software che si tratta di una stringa di istruzione e non di una parte del *corpus*, e a ciascuna variabile è stata assegnata una lettera identificativa, preceduta da un asterisco, seguita da un *underscore* e poi dal nome della modalità della variabile adeguata alla classificazione di ogni segmento di testo. Per esempio l'annotazione **** *A_2018 *G_06-01 *T_TG2 *N_Intervista *P_Fonte *S_Uomo identifica il segmento di testo di un'intervista a una fonte giornalistica di genere maschile, trasmessa dal TG2 il 6 gennaio 2018.

Operativamente questa annotazione manuale è stata eseguita leggendo i testi trascritti e organizzati per data e testata giornalistica, visionando contestualmente le registrazioni audiovisive corrispondenti. La lettura del *corpus*, accompagnata dall'ascolto del parlato e dalla visione delle immagini, ne ha guidato la suddivisione in segmenti di testo, classificati sulla base delle seguenti variabili categoriali: Anno (*A), Giorno (*G), Testata giornalistica (*T), tipologia della Notizia (*N), tipologia dei/lle Parlanti (*P), genere sociale, variabile quest'ultima che è stata chiamata Sesso (*S) per motivi meramente pratici, cioè per evitare una possibile confusione con la *G della variabile Giorno. L'esempio 1 riporta un segmento di testo annotato, l'elenco 1 riporta, per tutte le variabili, le relative modalità.

Dopo il pre-trattamento manuale, l'annotazione grammaticale del *corpus* è stata eseguita automaticamente, tramite PoS *tagging* di IRaMuTeQ (si veda più sopra, § 2).

La dimensione complessiva del *corpus* TG è stata calcolata attraverso l'analisi lessicometrica di IRaMuTeQ, eseguita optando per la lemmatizzazione automatica. Quest'ultima operazione pone alcuni problemi in termini di controllo efficace del processo e di risultati non sempre corretti, soprattutto nei casi di omografia, ma è ritenuta più funzionale a un'analisi prevalentemente di natura testuale (Bolasco 2013, 81). Al fine di ridurre quanto più possibile eventuali errori dovuti alla lemmatizzazione, è stata apportata qualche modifica al dizionario italiano fornito dalla libreria di R utilizzata da IRaMuTeQ, così come è stato ampliato il dizionario delle espressioni multiparola, aggiungendo alla lista nomi di partito, molto frequenti nei TG che danno sempre ampio spazio ai soggetti politici, anche collettivi.[21]

21 https://www.agcom.it/pluralismo-politico-sociale-in-televisione.

Esempio 1 Un segmento del *corpus* annotato

**** *A_2018 *G_06-01 *T_TG2 *N_Intervista *P_Fonte *S_Uomo
lei ha rappresentato un passaggio d epoca ed è stata capace di chiudere l epoca in cui le donne erano fatali e aprire l epoca in cui le donne conquistavano il mondo

Elenco 1 Elenco delle variabili con cui è stato annotato manualmente il *corpus* TG

La prima riga riporta i nomi delle variabili. Le modalità di ciascuna variabile sono indicate, colonna per colonna, nelle righe successive.

Anno (*A)	Giorno *(G)	TG (*T)	Notizia (*N)	Parlante (*P)	Sesso (*S)
2018	1-01	TG1	Introduzione (saluti a inizio TG)	Giornalista	Donna
2019	2-01	TG2	Lancio (annuncio di un servizio)	Fonte	Uomo
2020	3-01	TG5	Notizia da studio (notizia letta in studio)		
	4-01		Servizio (servizio)		
	5-01		Collegamento (collegamento in diretta)		
	6-01		Commento (commento giornalistico)		
	…		Chiusura (saluti a chiusura TG)		
	…		Messaggio (messaggio rilasciato dalla fonte in contesto extra-TG e trasmesso dal notiziario)		
	31-01		Intervista (intervista a/di fonte)		

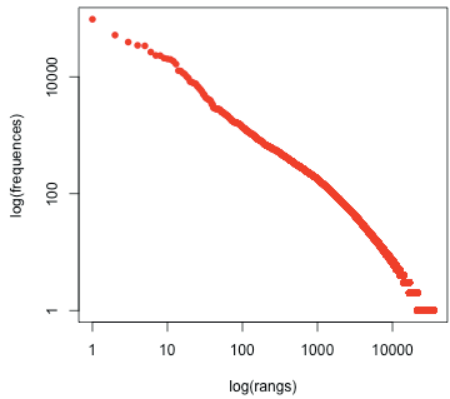

Grafico 1 *Corpus* TG: profilo rango/ frequenza (n. di *token*)

La dimensione complessiva del *corpus* è risultata pari a N=1.434.733 *token*, N=35.447 *type* (lemmi) e N=25.929 segmenti di testo. La *ratio* (*type*/*token*), misura indicativa della ricchezza del vocabolario, è pari a 0,025, un valore piuttosto basso, indicatore di una certa povertà lessicale, e tuttavia atteso, in considerazione del carattere specialistico del *corpus*. La curva disegnata dal rapporto fra il rango di frequenza e la frequenza delle parole empiricamente attestata nel *corpus* TG **[graf. 1]** mostra un andamento in linea con la legge di Zipf (1949), per cui la frequenza di una parola in un testo è inversamente proporzionale al suo rango, secondo la formula:

$$f(z) = C/Z^{a}$$

dove $f(z)$ è la frequenza di una parola di rango z, C è la frequenza della parola più frequente e a un indice inversamente proporzionale alla ricchezza del *corpus*.