DOI: 10.17469/O2109SLI000017

ARIANNA BIENATI, PAOLO BRASOLIN, FEDERICA COMINETTI

Il potere in parole povere: un'analisi diacronica corpus-based sulla semplificazione dell'italiano politico repubblicano[1]

Questo lavoro esplora l'ipotesi che l'italiano politico abbia subito una semplificazione nel corso della storia repubblicana. Tale ipotesi, per quanto molto diffusa, non è finora stata suffragata da metodi statistici applicati a un corpus rappresentativo. Il lavoro esamina diverse misure di complessità lessicale nel corpus di italiano politico IMPAQTS (1500 discorsi, 1946-2023) con lo scopo di verificare se il discorso politico si è semplificato dal punto di vista della ricchezza e varietà del lessico. Mette inoltre a confronto i discorsi delle "tre Repubbliche" (1946-1994, 1994-2013, 2013-2023) per individuare eventuali variazioni significative in termini di complessità lessicale. I risultati mostrano che le misure confermano l'ipotesi della semplificazione e presentano un'analisi qualitativa dei testi più rappresentativi dello stile lessicale dell'italiano politico nelle tre Repubbliche.

Parole chiave: semplificazione, discorso politico, complessità lessicale, Teoria dell'Informazione.

1. *Introduzione*

In questo contributo si illustrano i presupposti e i primi risultati di un'analisi che esplora con tecniche computazionali, statistiche e linguistiche l'ipotesi che il linguaggio dei politici della Repubblica Italiana si sia semplificato nel corso del tempo. Le domande di ricerca derivano da un assunto spesso dato per scontato sia nell'opinione pubblica sia nel discorso scientifico, e tuttavia mai verificato con tecniche adeguate su un corpus rappresentativo. In particolare, in questo lavoro approcciamo il tema della semplificazione linguistica conside-

[1] L'articolo è frutto di riflessioni condivise. Per scopi accademici, Arianna Bienati è responsabile delle sezioni 3.4, 4, 4.1, 4.2, 5, 6. Paolo Brasolin è responsabile delle sezioni 3, 3.2, 3.3, Federica Cominetti delle sezioni 1, 2, 3.1, 5.1, 5.2.

rando un aspetto della componente lessicale: quello legato alla varietà e ricchezza del vocabolario.

Il contributo è organizzato come segue: in § 2 si presentano l'ipotesi di partenza e le domande di ricerca; in § 3 si presentano il corpus utilizzato e le misure di complessità linguistica e i modelli statistici adottati; in § 4 sono illustrati i risultati delle analisi statistiche e in § 5 quelli delle analisi linguistiche qualitative; in § 6 si propongono delle brevi conclusioni.

2. *L'ipotesi della semplificazione*

"Con l'avvento della «seconda Repubblica», prese piede il cosiddetto «gentese», il discorso che si voleva chiaro, diretto, esplicito, non «difficile», che doveva parlare alla gente comune." (Gian Luigi Beccaria, La Stampa, 27 novembre 2006, p. 34). Sarebbe così stato abbandonato il *politichese*, il linguaggio "complicato, inintelligibile e pomposo" della Prima Repubblica (Dell'Anna 2010).

L'idea che il discorso politico si sia semplificato nella storia repubblicana è ripresa da molti osservatori, dai quali è normalmente data per scontata. Rispetto a quelli della Prima Repubblica, i politici di oggi premerebbero sul "pedale della comprensibilità" (Gualdo 2004), cercherebbero un "vocabolario comprensibile" (Bolasco *et al*. 2006), avrebbero "cura della semplicità" (Lotti 2015). Dando per scontato che la semplificazione si sia verificata, Antonelli (2007) ne offre una spiegazione: la semplificazione del discorso politico repubblicano sarebbe legata al passaggio dal "paradigma della superiorità" a quello del "rispecchiamento", con il livello stilistico della classe politica che si abbassa per riflettere quello dell'elettorato.

Tutte le considerazioni appena fatte sono spesso proposte come generalizzazioni, ma sono in realtà normalmente basate sull'analisi di piccolissimi campioni di discorsi o delle varietà di singoli politici. Lo stesso *politichese* è in effetti prevalentemente modellato sullo stile comunicativo di Aldo Moro (Desideri 1998), mentre il *gentese* si basa sullo stile comunicativo "pubblicitario" di Silvio Berlusconi (Gualdo 2004). La "cura della semplicità" è stata osservata nei discorsi di Beppe Grillo (Lotti 2015).

Ricerche condotte con strumenti automatici su campioni di discorso politico con estensione diacronica non hanno finora confema-

to l'ipotesi della semplificazione (cfr. Giuliano 2016; Ondelli 2020). Inoltre, studi sul discorso politico sui social media – necessariamente basati su produzioni molto recenti – ne mettono in luce la densità e varietà lessicale (ad es. Spina 2012, Tavosanis 2016). Alla prova dei dati, l'ipotesi della semplificazione sembra insomma meno solida di quanto la letteratura non quantitativa lasci intendere.

Da tali considerazioni emergono le tre domande di ricerca (DR) che sostanziano questo studio. Applicando tecniche statistiche a un corpus rappresentativo, intendiamo:

> DR1: verificare se l'italiano politico repubblicano si sia semplificato nel corso del tempo;
> DR2: verificare se si sia semplificato nel passaggio dalla Prima alla Seconda Repubblica (1994) e nel passaggio dalla Seconda alla Terza Repubblica (2013).

Infine, intendiamo

> DR3: trarre considerazioni linguistiche che permettano di descrivere come si configura la eventuale semplificazione.

Come si deduce da DR2, in questa ricerca adottiamo l'ipotesi, non condivisa da tutti gli osservatori, che le elezioni politiche del 2013 e il conseguente avvio della XVII Legislatura abbiano segnato una svolta nella storia repubblicana, legata alla fine del bipolarismo. Tale rottura sarebbe interpretabile come il passaggio dalla Seconda alla Terza Repubblica e sarebbe apprezzabile anche sul piano linguistico (Lotti 2015).

3. *Metodologia*

La metodologia è organizzata in fasi distinte poste in sequenza: dapprima è stato individuato un corpus idoneo all'esplorazione diacronica, poi i testi sono stati pre-processati e si è proceduto con il calcolo delle misure di complessità lessicale, infine sono state condotte le analisi statistiche e linguistiche. In quel che segue illustreremo ogni fase nel dettaglio.

Ciascuna fase è implementata o supportata da codice, e per garantire la massima riproducibilità abbiamo reso disponibile l'intero flusso di lavoro in repository pubblici. La pipeline linguistica destinata alla preparazione dei testi ed al calcolo delle misure è disponibile in

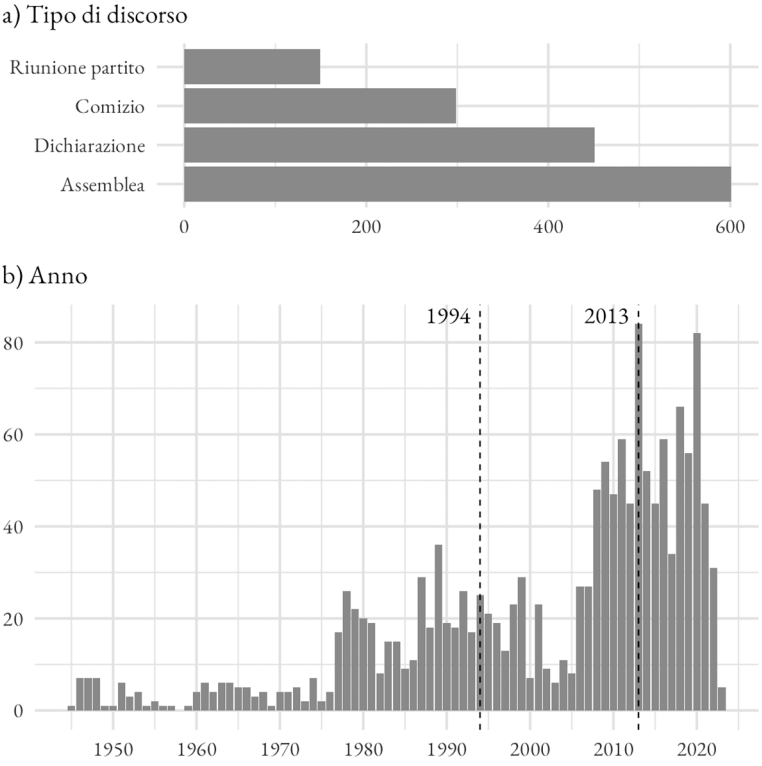
Brasolin *et al*. (2025a). Essa non è immediatamente eseguibile poiché il corpus individuato per le analisi è pubblicamente interrogabile ma non scaricabile integralmente; abbiamo pertanto reso disponibile il dataset di output delle misure in Brasolin *et al*. (2025b). Quest'ultimo supporta completamente l'esecuzione dei quaderni di analisi consultabili in Bienati *et al*. (2025).

Gli autori auspicano che questa trasparenza metodologica non solo permetta la verifica indipendente dei risultati, ma possa anche ispirare e facilitare ricerche future.

3.1 Il corpus IMPAQTS

Questo studio si avvale di IMPAQTS (Cominetti *et al*. 2022; 2024), un corpus di 1500 discorsi politici italiani pronunciati nella storia della Repubblica fino alla primavera del 2023, corrispondenti a 2.65 milioni di token di parlato trascritto. Il corpus è bilanciato per parlante, includendo 10 discorsi pronunciati da ciascuna delle 150 personalità politiche più importanti della storia repubblicana. È inoltre bilanciato per i principali tipi di monologo politico, includendo per ogni parlante quattro discorsi d'aula, due comizi, un discorso in assemblea di partito e tre dichiarazioni (Figura 1a). L'importanza di questo parametro per la verifica dell'ipotesi della semplificazione è stata messa in luce ad es. da Ondelli (2020). In diacronia, il corpus presenta una numerosità crescente rispetto alla progressione temporale (Figura 1b); comprende infatti 452 discorsi pronunciati nei quasi cinquant'anni della Prima Repubblica (1946-1994), 506 nei quasi vent'anni della Seconda e 542 nei dieci anni della Terza (dal 15 marzo 2013 alla primavera 2023).

Figura 1 - *a) Numero di testi nel corpus per tipologia di monologo politico; b) Numero di testi nel corpus per anno, in alto gli anni di passaggio tra le Repubbliche*



3.2 Preparazione del corpus

Per la preparazione dei testi è stata implementata una pipeline basata su Stanza v1.7.0 (Qi *et al*. 2020). Essa ha tokenizzato (scomponendo i token multi-parola) e lemmatizzato i testi, producendo la rappresentazione CoNLL-U utilizzata per il calcolo delle misure.

3.3 Misure di complessità linguistica

Questo studio si avvale di un insieme di misure che spaziano da indicatori classici di diversità lessicale a metriche mutuate dalla Teoria dell'Informazione. Queste ultime permettono di andare oltre il semplice approccio *bag-of-words* e di integrare alla varietà del lessico anche aspetti della sua organizzazione sequenziale.

Prima di esaminare nel dettaglio ognuna delle metriche scelte, elenchiamo le convenzioni notazionali che adotteremo nel seguito. Dato un testo,

- L è il numero di token (ovvero la lunghezza del testo);
- N è il numero di tipi di token (ovvero la dimensione del vocabolario usato nel testo);
- i tipi di token sono etichettati con numeri da *1* ad N;
- n_i è la frequenza assoluta del tipo di token i;
- $p_i = n_i / L$ è la frequenza relativa del token di tipo i;
- n_{ij} è la frequenza assoluta del bigramma (i, j);
- $p_{ij} = n_{ij} / (L - 1)$ è la frequenza relativa del bigramma (i, j).

3.3.1 Type-token ratio (TTR)

Il *type-token ratio* (TTR) è definito come il rapporto tra il numero di tipi lessicali ed il numero totale di token presenti nel testo:

$$N/L$$

Si tratta della misura di varietà lessicale più comunemente utilizzata, ma presenta il limite intrinseco di essere fortemente influenzata dalla lunghezza del testo. La natura di questo limite può essere compresa nel contesto della legge empirica di Herdan-Heaps (Corral & Font-Clos 2017; Serra-Peralta *et al*. 2021), secondo cui N ed L sono correlati da una legge di potenza, cioè $N \simeq \beta L^{\alpha}$ per opportune costanti $0 < \beta$ e $0 < \alpha \leq 1$. La definizione di TTR, infatti, confonde la dimensione del vocabolario con la varietà lessicale, ignorando che il soddisfacimento della legge di Herdan-Heaps implica che la dimensione del vocabolario cresca in maniera non lineare al crescere della lunghezza del testo. In pratica, testi più lunghi tendono ad avere un TTR più basso, indipendentemente dalla reale varietà lessicale. Per questo motivo, il TTR viene utilizzato principalmente come punto di partenza, richiedendo poi ulteriori aggiustamenti per consentire confronti equi tra testi di lunghezze differenti.

3.3.2 Indice di Guiraud (RTTR)

L'*indice di Guiraud* (RTTR) è una correzione empirica alla dipendenza del *type-token ratio* dalla lunghezza del testo. Esso si basa su uno studio di corpus (Guiraud 1954) ed è definito come

$$N/\sqrt{L}$$

Nel contesto della legge empirica di Herdan-Heaps si vede che esso è equivalente al parametro β nell'assunzione che $\alpha = 1/2$. Da questo si conclude che questa misura, al contrario del TTR, riconosce la dipendenza non lineare tra le grandezze coinvolte, ma presenta comunque una criticità, ovvero l'implicita assunzione dell'universalità di un valore specifico di α.

3.3.3 Moving average type-token ratio (MATTR)

Il *moving average type-token ratio* (MATTR, Covington & McFall 2010) è una correzione statistica alla dipendenza del *type-token ratio* dalla lunghezza del testo. Con questo approccio, il TTR viene calcolato su finestre mobili di lunghezza prefissata lungo il testo, e la media dei valori ottenuti viene poi considerata come indicatore della diversità complessiva.

L'uso di finestre di larghezza fissa, pur limitando la scala su cui la varietà viene misurata, offre una stima robusta che permette di comparare testi con lunghezze eterogenee. Pur evitando i problemi legati alla lunghezza del testo che il TTR presenta, MATTR rende necessario scegliere una dimensione per la finestra, decisione che non risulta immediata e incide significativamente sui risultati. Abbiamo utilizzato una finestra di 128 token per assicurarci che la metrica fosse calcolabile su quanti più testi possibile (solo due outliers più brevi di 150 token risultano esclusi).

3.3.4 Entropia

La Teoria dell'Informazione si fonda sul concetto di informazione come misura dell'incertezza. L'idea chiave è che il contenuto di informazione associato ad un evento E (ad esempio, il lancio di una moneta che risulta in testa) è inversamente proporzionale alla sua probabilità di accadere p_E, ed è quantificato come $I_E = -log_2 p_E$. A causa delle proprietà del logaritmo, I_E è zero se $p_E = 1$ e tende all'infinito quando p_E si avvicina a zero; in parole semplici, questo significa che il contenuto di informazione è minimo per eventi certi e aumenta rapidamente per eventi più rari, spiegando così perché *sorpresa* è comunemente utilizzato come sinonimo di contenuto di informazione. L'unità di misura dell'informazione è il bit (ad esempio, 1 bit corrisponde esattamente al contenuto di informazione di un lancio equo di una moneta, poiché $-log_2 (1/2) = 1$).

Considerando tutti i possibili esiti Ω di un esperimento (ad esempio, un lancio di moneta ha Ω = *{testa, croce}*), una grandezza interessante è il contenuto di informazione medio trasmesso dall'osservazione dell'esito di un esperimento. L'*average information content* (AIC) è calcolato come $H = \sum_{E \in \Omega} p_E I_E$ ed è più comunemente noto col nome di *entropia*.

Pensando ai tipi di token come agli esiti dell'osservazione di un token in un testo, tutte le nozioni sopra si applicano immediatamente anche ai testi. Il contenuto di informazione medio – ovvero, la sorpresa media – associata all'osservazione di un token del testo è

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i$$

L'entropia è frequentemente richiamata come una misura di *disordine*, e vale la pena di chiarire la valenza che assume nel contesto della linguistica quantificando la quantità media di incertezza (o sorpresa) associata all'osservazione di un token in un testo. Un'alta entropia indica che i token si distribuiscono in maniera più uniforme e imprevedibile, suggerendo una maggiore varietà lessicale. Al contrario, una bassa entropia denota una maggiore disomogeneità di frequenze, riflettendo una maggiore prevedibilità e, in un certo senso, un *ordine* più marcato.

L'entropia è una quantità non negativa. Inoltre, è possibile dimostrare che assume il suo valore minimo possibile 0 per una distribuzione triviale (quando un token type ha probabilità certa) ed il suo valore massimo possibile $\log_2 N$ per una distribuzione uniforme ($p_i = 1/N$). Notiamo quindi che la gamma di valori che può assumere non dipende dalla lunghezza del testo, ma dipende dalla dimensione del vocabolario usato nel testo.

3.3.5 Evenness

La misura nota come *evenness* (uniformità) è definita come la normalizzazione dell'entropia per il suo valore massimo possibile:

$$E = H / \log_2 N$$

L'*evenness* è pertanto una misura che assume valori tra 0 ed 1. Essa quantifica quanto la distribuzione delle frequenze dei token si avvicini a una distribuzione uniforme. In altre parole, l'*evenness* valuta l'equilibrio nella distribuzione delle probabilità dei token: un valore di

evenness pari a 1 indica che tutti i token hanno la stessa probabilità di occorrenza (massima uniformità), mentre un valore prossimo a 0 suggerisce una distribuzione altamente squilibrata, con pochi token dominanti e molti token rari.

3.3.6 Information Fluctuation Complexity (IFC)

L'entropia si basa su un modello *bag-of-words* e non può tenere conto dell'ordine delle parole in un testo. Per catturare gli aspetti dell'informazione sequenziale nel linguaggio, è necessario considerare misure più sofisticate.

La quantità teorica successiva più semplice che si può considerare per affrontare questa limitazione è il *net information gain* associato all'osservazione dell'evento *F* dopo l'evento *E*, definito semplicemente come la differenza tra il loro contenuto di informazione:

$$\Gamma_{EF} = I_F - I_E$$

Il modo più immediato di applicare questa nozione ai testi è quello di calcolare il guadagno netto medio di informazione sulle coppie di token

$$\Gamma = \sum_{i,j=1}^{N} p_{ij} \Gamma_{ij}$$

e la relativa deviazione quadratica media dalla media

$$\sigma_{\Gamma^2} = \sum_{i,j=1}^{N} p_{ij} (\Gamma_{ij} - \Gamma)^2$$

Si può dimostrare che Γ non è d'interesse poiché è sempre zero. Dimostrare questo fatto richiede una teoria dei sistemi dinamici leggermente sofisticata, che non approfondiremo qui; si raccomanda Bates (2020) o Bates & Shepard (1993) per una guida accessibile e Wackerbauer *et al*. (1994) per una rassegna completa sul tema.

La radice quadrata di σ_{Γ^2} non è in generale zero, ed è chiamata *information fluctuation complexity* (IFC):

$$\sigma_\Gamma = \sqrt{\sum_{i,j=1}^{N} p_{ij} (\Gamma_{ij} - \Gamma)^2}$$

Poiché σ_I è la deviazione standard di I, IFC è immediatamente interpretabile come la volatilità della variazione della sorpresa tra i token dei bigrammi, ovvero la volatilità del guadagno netto di informazione tra token.

3.3.7 Moving Average IFC (MAIFC)

L'IFC è una quantità non negativa. Tuttavia, al contrario dell'entropia, individuare il suo massimo valore possibile è un problema complesso. Nell'impossibilità di derivare da essa una quantità normalizzata che elimini effetti di dipendenza dalla dimensione del vocabolario usato nel testo come fatto con l'*evenness*, abbiamo deciso invece di calcolare la *moving average IFC* su una finestra di 128 token.

3.4 Modelli statistici

In una prima fase esplorativa, sei modelli di regressione lineare a effetti misti, uno per misura di complessità escluso il TTR semplice, sono stati costruiti per determinare la variazione delle misure di complessità in funzione dell'anno (DR1) o della Repubblica (DR2). Sono stati scelti modelli di regressione ad effetti misti poiché IMPAQTS presenta interdipendenza tra i dati. Il corpus è infatti composto da 10 discorsi pronunciati da 150 personalità politiche. La scelta di questi modelli limita l'occorrenza degli errori di tipo I (errori derivanti dal rifiuto dell'ipotesi nulla quando questa è effettivamente vera), che si verificano nei modelli lineari semplici quando l'assunzione di indipendenza non è rispettata. L'effetto random (*random effect*) inserito nei modelli è relativo al codice identificativo della personalità politica che ha pronunciato i discorsi (*speaker_id*). Di seguito, riportiamo per chiarezza la struttura dei modelli utilizzati, con la sintassi della libreria R *lme4* 1.1-35.5 (Bates *et al*. 2015). Il predittore relativo all'anno di enunciazione del discorso è stato centrato (*year_centered*) per garantire più interpretabilità ai risultati. Il predittore relativo alla Repubblica (*republic_id*), invece, è un fattore di tipo categoriale con tre livelli: I, II, III. Il livello di riferimento è stato impostato sulla Seconda Repubblica.

DR1. `lmer(value ~ year_centered + (1 | speaker_id), data = x)`

DR2. `lmer(value ~ republic_id + (1 | speaker_id), data = x)`

La significatività del predittore è stata determinata utilizzando due metodi complementari[2]:

a. Il metodo Satterthwaite per la determinazione dei gradi di libertà, così come è implementato nella libreria *lmerTest* 3.1-3 (Kuznetsova *et al*. 2017)
b. *likelihood-ratio test* (Winter 2019) tra il modello specificato e il modello nullo (ovvero senza predittore), dove il modello nullo è `lmer(value ~ 1 + (1 | speaker_id), data = x.`

La soglia della significatività (*alpha level*) è stata stabilita utilizzando la correzione di Bonferroni ed è stata determinata dividendo il livello *alpha* classico (0.05) per il numero totale di modelli eseguiti (6) = 0.008.

In una seconda fase, i modelli relativi alle misure la cui varianza spiegata dai predittori raggiungeva i livelli maggiori sono stati analizzati e interpretati più in dettaglio.

4. *Risultati delle analisi statistiche*

In quel che segue presentiamo i risultati delle analisi statistiche che abbiamo svolto per rispondere alle prime due domande di ricerca: se l'italiano politico repubblicano si sia semplificato nel corso del tempo (DR1) e tra Repubbliche (DR2).

4.1 Risultati DR1: L'italiano politico repubblicano si è semplificato nel corso del tempo?

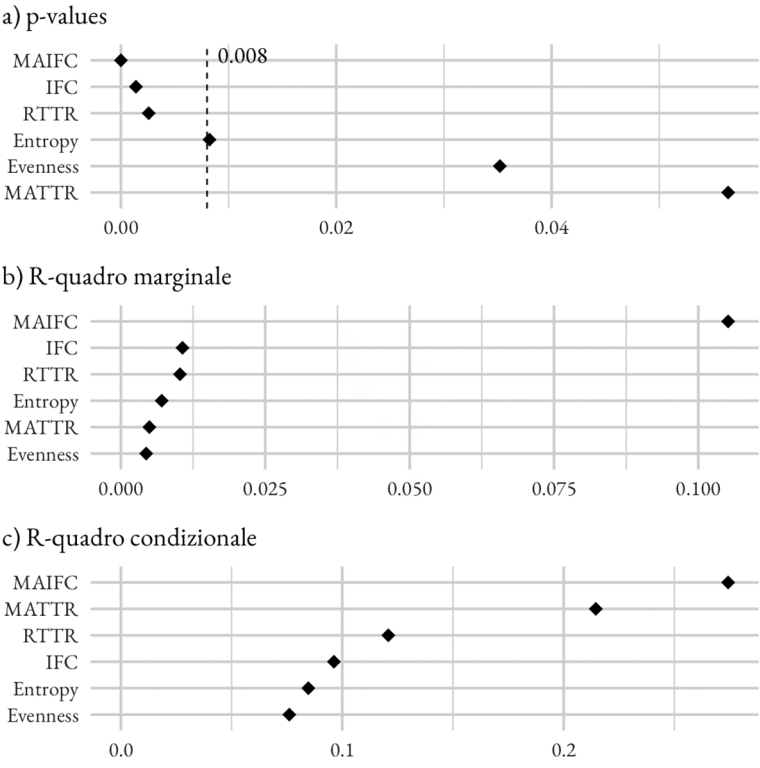
L'anno di enunciazione del discorso, utilizzato per modellare la componente diacronica del corpus, ha un effetto statisticamente significativo sui valori di RTTR (p-value a = 0.003; p-value b = 0.003), IFC (p-value a = 0.001; p-value b = 0.001) e MAIFC (p-value a < 0.001; p-value b < 0.001) (Figura 2a). Per tutte queste misure il trend è negativo. L'entropia, invece, sulla soglia di significatività (p-value a = 0.008; p-value b = 0.008), mostra un trend positivo. Tuttavia, esplo-

[2] Nel riportare i risultati dei test statistici usiamo questa convenzione: p-value a: p-value relativo al predittore *year_centered* calcolato con il metodo Satterthwaite; p-value b: p-value relativo a il *likelihood-ratio test*. Essendoci solo un predittore, è normale e auspicabile che i risultati dei due test coincidano.

rando i residui con istogrammi e *qq-plots*, si può concludere che le assunzioni del modello non sono rispettate: i residui non sono normalmente distribuiti e la loro varianza cresce al diminuire dei valori di entropia. Si è deciso, dunque, di non considerare questo modello come attendibile e di considerare nella trattazione e nell'interpretazione dei risultati solo i modelli relativi a RTTR, IFC e MAIFC.

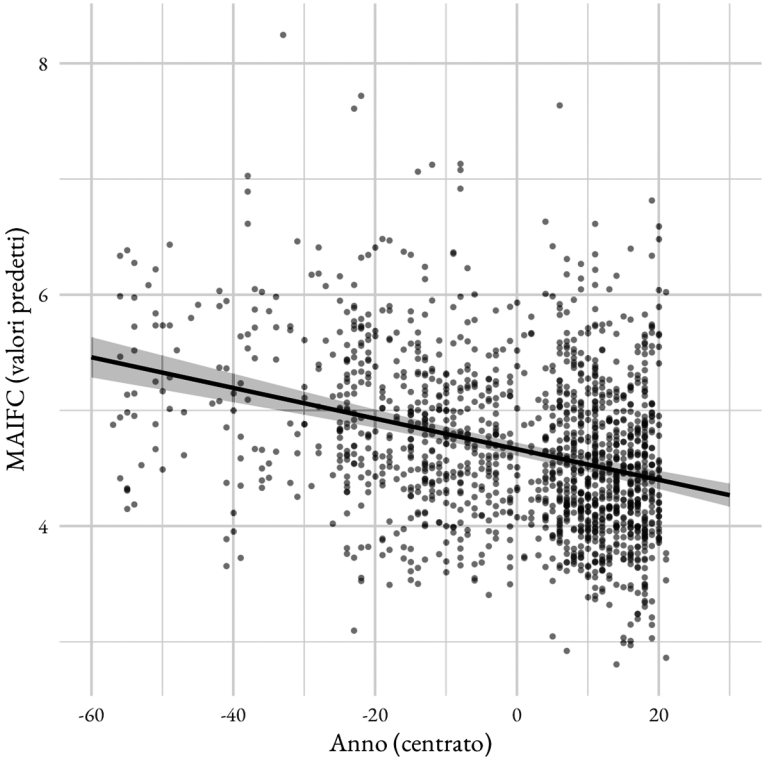
La grandezza dell'effetto (*effect size*) del predittore in questi tre modelli è in generale piccola (R^2 marginale: 0.010-0.105), con MAIFC che emerge come misura di complessità la cui variazione viene meglio descritta dall'anno di enunciazione (Figura 2b). Complessivamente, tenendo conto anche delle fluttuazioni dovute allo stile dei singoli politici, i modelli spiegano una varianza moderata all'interno del dataset (R^2 condizionale: 0.096-0.274) (Figura 2c).

Figura 2 - *a) p-values relativo al predittore anno per ciascuna delle misure di complessità considerate; b) R-quadro marginale e c) R-quadro condizionale di ciascuno dei modelli di regressione*



Seguendo i suggerimenti contenuti in Winter (2019), ci accingiamo ora a discutere il modello che predice il comportamento di MAIFC nel tempo. Segnaliamo, innanzitutto, che la distribuzione dei residui supporta l'ipotesi che le assunzioni del modello siano rispettate (distribuzione normale, con qualche outlier verso i valori predetti più alti, e omoschedasticità). Per quanto riguarda la componente del modello relativa agli effetti fissi (*fixed effects*), il modello stima un'intercetta di 4.665. Ciò significa che, nell'anno medio (circa il 2002), il valore di MAIFC è 4.665. La pendenza è negativa (-0.013), quindi con l'avanzare degli anni il modello stima una diminuzione dei valori di complessità linguistica (Figura 3). La variazione inter-individuale, catturata dalla componente degli effetti random, sembra essere moderata (dev. st. = 0.607): il modello stima il 95% dei testi nell'intervallo di valori di complessità tra 5.879 e 3.451.

Figura 3 - *Effetto marginale della variabile 'anno' (centrata) sulla variabile 'MAIFC'*



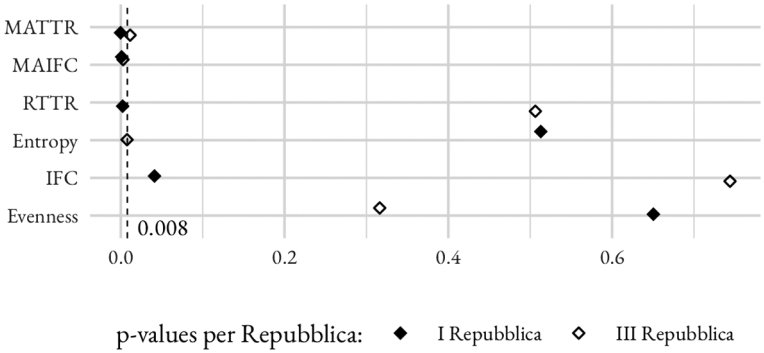
È doveroso notare che, sebbene il modello raggiunga il livello-soglia per la significatività statistica, la varianza spiegata dal modello rimane moderata (R^2 marginale: 0.105, R^2 condizionale: 0.274), suggerendo che altre variabili (es. tipo di discorso, orientamento politico, ecc.) potrebbero aver causato la variazione non catturata dal modello.

4.2 Risultati DR2: L'italiano politico repubblicano si è semplificato dalla Prima alla Seconda e dalla Seconda alla Terza Repubblica?

La seconda domanda di ricerca è stata affrontata con la medesima metodologia della prima, modificando unicamente il predittore dall'anno di enunciazione alla Repubblica in cui i discorsi sono stati pronunciati. Nella Figura 4 sono riportati i p-value sopra e sotto soglia di significatività per ciascuno dei passaggi tra Repubbliche (in nero: I-II Repubblica; in bianco: II-III Repubblica) e per ciascuna delle misure di complessità considerate. Nel passaggio dalla Prima alla Seconda Repubblica, si riscontrano differenze statisticamente significative relativamente ai valori di RTTR (p-value a = 0.003), MATTR (p-value a < 0.001) e MAIFC (p-value a < 0.001). La direzione del cambiamento è negativa. Il passaggio dalla Seconda alla Terza Repubblica è caratterizzato da un cambiamento significativo dei valori di MAIFC (p-value a = 0.003), sempre di segno negativo.

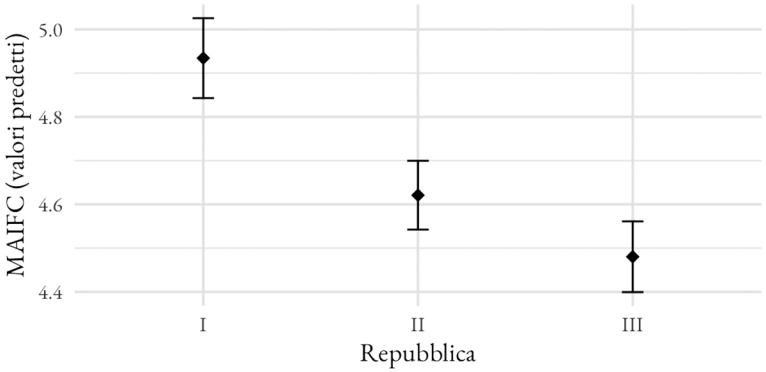
L'entropia si caratterizza per un p-value a = 0.008, sulla soglia di significatività statistica. Il cambiamento è di segno positivo, ovvero sembrerebbe esserci un incremento dell'entropia dalla Seconda alla Terza Repubblica. Tuttavia, le assunzioni (omoschedasticità e normalità dei residui) del modello non sono rispettate. Ispezionando i residui mediante istogrammi e *qq-plots*, si nota una 'coda lunga' dovuta probabilmente a valori molto bassi di entropia particolarmente concentrati nella Prima Repubblica. Questo ci ha portato a restringere l'analisi ai soli testi appartenenti alla Seconda e alla Terza Repubblica, per testare l'effettiva crescita dell'entropia tra i testi appartenenti alle due Repubbliche più recenti. Restringendo lo sguardo solo a questi due gruppi, i p-values calcolati con il metodo Satterthwaite e il *likelihood-ratio test* non raggiungono più il valore-soglia di significatività statistica di 0.008 (p-value a = 0.009; p-value b = 0.011) e i residui presentano ancora code pesanti e pattern di eteroschedasticità. Per questo è stato deciso di non considerare questo risultato come affidabile.

Figura 4 - *p-values relativi al cambiamento delle misure di complessità nel passaggio dalla Prima alla Seconda Repubblica (in nero) e nel passaggio dalla Seconda alla Terza Repubblica (in bianco)*



La distribuzione dei residui del modello che predice i valori di MAIFC, invece, supporta l'ipotesi che le assunzioni del modello siano rispettate (distribuzione normale, con qualche outlier verso i valori predetti più alti, e omoschedasticità). MAIFC è la misura di complessità lessicale la cui variazione viene meglio spiegata dalla Repubblica (R^2 marginale: 0.070, R^2 condizionale: 0.256) e il cui trend negativo si estende dalla Prima alla Seconda e dalla Seconda alla Terza Repubblica (Figura 5). Il likelihood-ratio test conferma che c'è un effetto statisticamente significativo della variabile 'repubblica' sui valori di MAIFC ($\chi 2$ (2) = 54.455, p-value b < 0.001). Per quanto riguarda la componente degli effetti fissi (*fixed effects*), il modello stima un'intercetta di 4.621, che rappresenta il valore di MAIFC medio stimato dal modello per la Seconda Repubblica. La stima per la Prima Repubblica (0.313) indica che, in media, il valore è più alto rispetto alla Seconda (MAIFC stimato medio Prima Repubblica = 4.934). Questo effetto è statisticamente significativo (p-value a < 0.001). Anche nel passaggio dalla Seconda alla Terza Repubblica il modello stima una decrescita dei valori di MAIFC di –0.141: questo effetto raggiunge il livello di significatività statistica di 0.008 (p-value a = 0.003), pur essendo meno marcato.

Figura 5 - *Andamento della variabile MAIFC in funzione della Repubblica*



5. *Risultati dell'analisi 'qualitativa'*

A partire da questi risultati, sono stati estratti 110 testi (7.3% del corpus) che rappresentano, per ciascuna Repubblica, valori di complessità (MAIFC) alta, media e bassa. Sono stati scelti dei valori soglia entro o oltre i quali un testo potesse essere considerato 'tipico' di una delle tre Repubbliche, oppure 'atipico' in positivo (per valori di complessità alti) o in negativo (per valori di complessità bassi). I testi attorno alla media sono stati estratti calcolando la media per ciascuna Repubblica e poi selezionando unicamente i testi che avevano valori di MAIFC ± 1/25 della deviazione standard della distribuzione. I testi 'atipici' invece, sono stati selezionati estraendo testi con valori di MAIFC oltre le 2 deviazioni standard (Tabella 1).

Tabella 1 - *Valori soglia per estrarre i testi con complessità alta, media e bassa in ciascuna Repubblica*

	I Repubblica	*II Repubblica*	*III Repubblica*
media	4.969	4.617	4.454
st. dev.	0.741	0.658	0.654
valori MAIFC alti	> 6.451	> 5.933	> 5.762
valori MAIFC medi	4.999 – 4.939	4.643 – 4.591	4.480 – 4.428
valori MAIFC bassi	< 3.487	< 3.301	< 3.146

I sotto-corpora così estratti sono stati usati per osservare se emergessero pattern sintattici e lessicali specifici, sia in diacronia (tra Repubbliche),

sia in termini di valori di MAIFC decrescente, indipendentemente dalla Repubblica.

In particolare, per i tratti sintattici, si è scelto di esplorare la lunghezza delle frasi (intese come segmenti testuali che vanno da punto a punto) e il numero di punti interrogativi ed esclamativi ogni 100 frasi. La lunghezza delle frasi è un indice facile da valutare, il cui legame con la complessità sintattica è stato messo in luce a partire almeno da Ferguson (1982). La frequenza di punti interrogativi ed esclamativi permette una valutazione – sommaria ma non irrealistica – della varietà di atteggiamenti illocutori adottati dai parlanti.

Per quanto riguarda i tratti lessicali, si è scelto di indagare il numero di lemmi formati coi suffissi *-ismo* e *-zione*. La scelta di considerare questi suffissi dipende dal fatto che le nominalizzazioni sono considerate un tratto che aumenta la complessità linguistica (Chafe 1985). L'analisi delle forme in *-zione*, in particolare, apre inoltre un'esplorazione molto preliminare nella direzione della valutazione della complessità lessicale in termini di ricercatezza (Kyle & Crossley 2015). Infatti, molti dei lemmi prodotti con questo suffisso appartengono a lessici specialistici o almeno al lessico comune, uscendo dai confini del Vocabolario di base.

5.1 Tratti sintattici e lessicali: complessità media vs. outlier

Il confronto fra outlier con complessità alta, testi con complessità media e outlier con complessità bassa rivela una chiara tendenza rispetto ai tratti sintattici considerati. Al crescere della complessità lessicale, la lunghezza media delle frasi aumenta, mentre diminuisce la frequenza di punti interrogativi e punti esclamativi (cfr. Tabella 2). Osserviamo dunque che i modelli di analisi della complessità lessicale adottati, benché modellati sulle interazioni tra bigrammi, correlino con tratti più macroscopici.

Tabella 2 - *Distribuzione dei tratti sintattici e lessicali in testi di complessità alta, media e bassa*

	Complessità alta	*Complessità media*	*Complessità bassa*
parole/frase	32.126	29.585	18.036
? (/100 frasi)	0.709	4.203	14.387
! (/100 frasi)	0.222	1.349	1.599
-ismo (/1000 parole)	1.145	1.007	0.689
-zione (/1000 parole)	23.832	13.720	7.879

Per quanto riguarda la frequenza delle nominalizzazioni in *-ismo* e *-zione*, l'analisi mostra che questa tende a diminuire al diminuire della complessità lessicale.

5.2 Tratti sintattici e lessicali lungo le tre Repubbliche

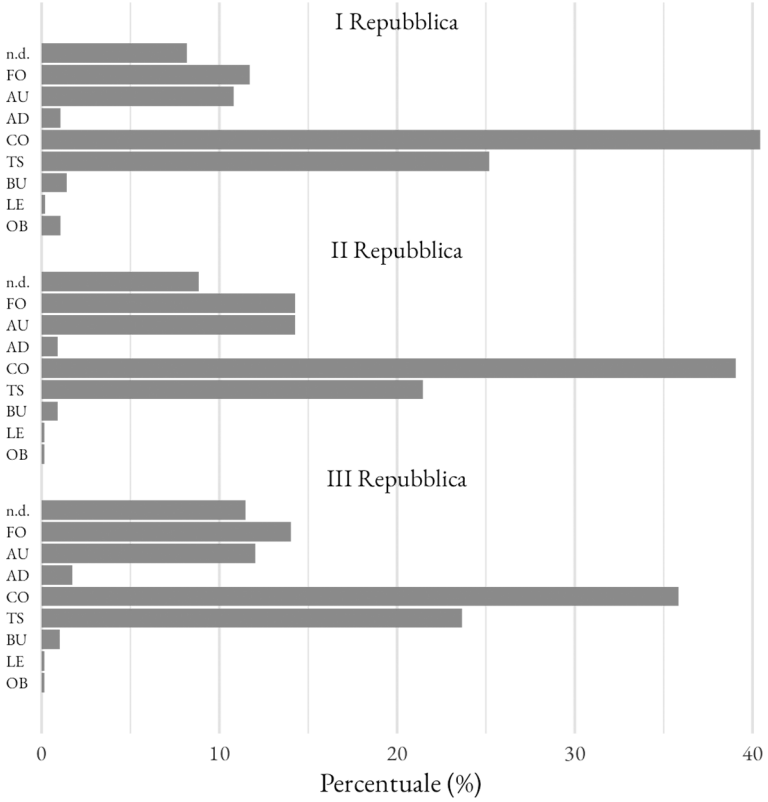
I tratti sintattici e lessicali considerati sono stati esaminati nei sotto-corpora corrispondenti ai testi di complessità bassa, media e alta, aggregati per Repubblica. Si osserva una tendenza per cui i testi della Prima Repubblica hanno frasi di lunghezza media maggiore, mentre quelli della Seconda e della Terza Repubblica hanno frasi di lunghezza media leggermente decrescente. Le frasi che terminano con punto interrogativo hanno frequenza crescente con il procedere delle Repubbliche. Al contrario, non si osserva una tendenza analoga per le frasi con punto esclamativo, che sono leggermente più frequenti nei testi della Seconda Repubblica. Non si notano variazioni di frequenza nemmeno per le nominalizzazioni in *-ismo*, mentre diminuiscono con l'avanzare delle Repubbliche quelle in *-zione*.

Tabella 3 - *Distribuzione dei tratti sintattici e lessicali lungo l'arco delle tre Repubbliche*

	I Repubblica	II Repubblica	III Repubblica
parole/frase	33.833	29.291	27.405
? (/100 frasi)	2.778	3.725	4.477
! (/100 frasi)	0.772	1.037	0.965
-ismo (/1000 parole)	0.950	1.078	1.070
-zione (/1000 parole)	21.621	17.561	14.970

Per valutare a livello esplorativo l'ambito della ricercatezza lessicale, sono state annotate con le marche d'uso del GRADIT (2ª edizione, De Mauro 2007) tutte le occorrenze delle nominalizzazioni in *-zione* presenti nei sotto-corpora comprendenti i testi di complessità media delle tre Repubbliche. I risultati, riportati in Figura 6, rivelano che nei testi della Prima Repubblica le nominalizzazioni in *-zione* sono meno frequentemente appartenenti al Vocabolario di Base, e più frequentemente appartenenti ai lessici specialistici, di basso uso o obsoleti.

Figura 6 - *Percentuale delle diverse marche d'uso dei nomi in -zione nelle tre Repubbliche*



6. *Discussione e conclusione*

In questo studio l'ipotesi che l'italiano politico si sia semplificato nel corso della storia repubblicana è stata verificata, prendendo in considerazione la componente lessicale, a partire da un corpus con ampia estensione diacronica, da cui sono state estratte misure di complessità lessicale sia classiche, come le derivate dalla *type-token ratio*, sia mutuate dalla Teoria dell'Informazione. Si è cercato di descrivere la variazione di queste misure in diacronia utilizzando modelli di regressione lineare misti, aventi come predittori principali l'anno o la Repubblica in cui il discorso è stato pronunciato.

I risultati mostrano tendenze chiare verso la semplificazione, almeno per quanto riguarda *information fluctuation complexity*, particolarmente nella sua versione corretta per la lunghezza del testo (MAIFC), così come per l'*indice di Guiraud* (RTTR) e la *moving average type-token ratio* (MATTR), in particolare nel passaggio dalla Prima alla Seconda Repubblica. Questi risultati sono ulteriormente avvalorati dalle esplorazioni qualitative su un sottocampione di testi particolarmente 'tipici' o 'atipici' per ciascuna Repubblica e livello di complessità. Le tendenze alla semplificazione si riscontrano, infatti, anche nei principali tratti sintattici e lessicali che abbiamo analizzato.

A titolo puramente esemplificativo, si considerino i passaggi qui sotto, incipit di discorsi individuati come outlier rispettivamente positivo (testo molto più complesso della media) e negativo (testo molto meno complesso della media):

(1) Onorevole Presidente, onorevole segretario. Apro questo mio breve intervento con parole di gratitudine all'onorevole Forlani per aver dato pronto consenso alla proposta da me fatta mesi fa, in seno alla direzione centrale, per organizzare un convegno destinato ad esaminare con quali decisioni, anche la Democrazia Cristiana, dovesse concorrere ad adeguarsi alle scoperte scientifiche, ai conseguenti adattamenti tecnologici, alle mutazioni sociali ormai in accentuato e rapido corso in tutto il mondo. Anche le iniziali riserve circa la natura e i tempi del convegno furono superate con il fermo concorso di Forlani. [AFAN91-P1 – A. Fanfani, 1991, congresso di partito]

(2) Signor presidente, onorevoli colleghi. Il presidente del consiglio ha scelto di riferire in parlamento, sulla situazione economica del paese, e già questa è una scelta apprezzabile. Lo ha fatto, sapendo di parlare nel luogo più alto, rappresentativo della democrazia, e sapendo di dire oggi le cose che avrebbe dovuto dire agli italiani. Ecco perché noi riteniamo le parole appena pronunziate dal presidente del consiglio delle parole oneste, serie, e affidabili per un paese che in questo momento chiede affidabilità e serietà al governo che ha voluto che governasse. [AALF11-A1 – A. Alfano, 2011, Camera dei Deputati].

Particolarmente interessante è osservare come misure lessicali, che hanno come unità di analisi la parola e che quindi sono cieche a qualsiasi nozione sintattica, sembrino comunque ben correlate con i tratti sintattici della lunghezza delle frasi, della varietà illocutiva e delle no-

minalizzazioni. Due sono le possibili interpretazioni: la prima va nella direzione di una generale correlazione tra complessità a diversi livelli di analisi. Se un testo è complesso dal punto di vista lessicale è più probabile che l'intera sua costruzione (dal punto di vista sintattico e stilistico) sia complessa. La seconda possibile interpretazione, che ha anche importanti ricadute teoriche, è che una parte importante delle relazioni sintattiche in un testo possano essere descritte alternativamente dalle relazioni di co-occorrenza tra token. Proprio la co-occorrenza o le relazioni tra token in sequenza è ciò che definisce la misura di *information fluctuation complexity*, la quale supera la modellizzazione della complessità lessicale come *bag-of-words* per includere come sua unità fondante le relazioni sintagmatiche tra elementi adiacenti.

Dunque, le risultanze delle analisi statistiche, focalizzate sul lessico, e di quelle descrittive, che ampliano lo sguardo ad altri livelli d'analisi, convergono nel supportare l'ipotesi che l'italiano repubblicano si sia effettivamente semplificato nel corso del tempo. L'unica misura di complessità che mostra una tendenza opposta è l'entropia. Già nei risultati abbiamo brevemente discusso perché sia necessario prendere questo risultato con le pinze: i modelli non sono affidabili data la violazione delle assunzioni di omoschedasticità e distribuzione normale dei residui. Tuttavia, è doveroso notare come il mancato inserimento di un predittore importante come la grandezza del vocabolario o la lunghezza del testo possa aver portato a una sotto-specificazione dei modelli per quanto riguarda le misure dipendenti dalla grandezza del vocabolario, come appunto l'entropia.

Il presente studio si colloca come primo passo nell'esplorazione della semplificazione dell'italiano politico nel corso della Repubblica. Nelle indagini future che verificheranno la medesima ipotesi sarà necessario prestare attenzione alle dipendenze tra misure di complessità e altri predittori di 'controllo' come la lunghezza del testo o la grandezza del vocabolario; includere altri predittori potenzialmente importanti per descrivere meglio la variazione in diacronia, come il tipo di discorso o l'orientamento politico; allargare lo sguardo in modo più sistematico alla morfosintassi e alla testualità. Ma anche rimanendo nel settore del lessico, molto rimane da fare nella valutazione della ricercatezza del lessico utilizzato dalle personalità politiche rappresentate nel corpus IMPAQTS.

Riferimenti bibliografici

Antonelli, Giuseppe. 2007. *L'italiano nella società della comunicazione*. Bologna: Il Mulino.

Bates, John E. 2020. Measuring complexity using information fluctuation: A tutorial. https://www.researchgate.net/publication/340284677_Measuring_complexity_using_information_fluctuation_a_tutorial

Bates, John E. & Shepard, Harvey K. 1993. Measuring complexity using information fluctuation. *Physics Letters A* 172(6). 416-425.

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67. 1-48. https://doi.org/10.18637/jss.v067.i01

Bienati, Arianna & Brasolin, Paolo & Cominetti, Federica. 2025. Il potere in parole povere: un'analisi diacronica corpus-based sulla semplificazione dell'italiano politico repubblicano (analysis notebook). Zenodo. https://doi.org/10.5281/zenodo.15019450

Bolasco, Sergio & Giuliano, Luca & Galli de' Paratesi, Nora. 2006. *Parole in libertà. Un'analisi statistica e linguistica*. Roma: Manifesto Libri.

Brasolin, Paolo & Bienati, Arianna & Cominetti, Federica. 2025a. Il potere in parole povere: un'analisi diacronica corpus-based sulla semplificazione dell'italiano politico repubblicano (data pipeline). Zenodo. https://doi.org/10.5281/zenodo.13762873

Brasolin, Paolo & Bienati, Arianna & Cominetti, Federica. 2025b. Il potere in parole povere: un'analisi diacronica corpus-based sulla semplificazione dell'italiano politico repubblicano (measures dataset). Zenodo. https://doi.org/10.5281/zenodo.15019440

Chafe, Wallace L. 1985. Linguistic differences produced by differences between speaking and writing. In Olson, David R. & Torrance, Nancy & Hildyard, Angela (ed.), *Literacy, Language and Learning. The Nature and Consequences of Reading and Writing*, 105-123. Cambridge: Cambridge University Press.

Cominetti, Federica & Gregori, Lorenzo & Lombardi Vallauri, Edoardo & Panunzi, Alessandro. 2022. IMPAQTS: un corpus di discorsi politici italiani annotato per gli impliciti linguistici. In Cresti, Emanuela & Moneglia, Massimo (a cura di), *Corpora e Studi linguistici. Atti del LIV Congresso della Società di Linguistica Italiana*, 151-164. Milano: Officinaventuno.

Cominetti, Federica & Gregori, Lorenzo & Lombardi Vallauri, Edoardo & Panunzi, Alessandro. 2024. IMPAQTS: a multimodal corpus of Italian political discourse pragmatically annotated per implicitly conveyed

questionable content. In Fišer, Darja & Eskevich, Maria & Bordon, David (ed.), *Proceedings of the ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora*, 101-109.

Corral, Álvaro & Font-Clos, Francesc. 2017. Dependence of exponents on text length versus finite-size scaling for word-frequency distributions. *Physical Review E* 96(2). 022318. https://doi.org/10.1103/PhysRevE.96.022318

Covington, Michael A. & McFall, Joe D. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17(2). 94-100.

De Mauro, Tullio. 2007. GRADIT - *Grande dizionario italiano dell'uso.* Seconda edizione. Torino: UTET.

Dell'Anna, Maria Vittoria. 2010. *Lingua italiana e politica.* Roma: Carocci.

Desideri, Paola. 1998. Metalinguaggio e retorica dell'attenuazione nel discorso politico di Aldo Moro. In Alfieri, Gabriella & Cassola, Arnold (a cura di), *La «Lingua d'Italia». Usi pubblici e istituzionali, Atti del XXIX Congresso della Società di Linguistica Italiana*, 212-252. Roma: Bulzoni.

Ferguson, Charles A. 1982. Simplified Registers and Linguistic Theory. In Loraine K. Obler & Lise Menn (a cura di), *Exceptional Language and Linguistics*, 49-66. New York: Academic Press.

Giuliano, Luca C. 2016. La parola del leader: profili di linguaggio parlamentare a confronto tra la prima e la seconda repubblica. In Librandi, Rita & Piro, Rosa (a cura di), *L'italiano della politica e la politica per l'italiano: Atti dell'XI Convegno ASLI, Associazione per la Storia della Lingua Italiana*, 131-152. Firenze: Cesati.

Gualdo, Riccardo. 2004. La comunicazione politica: novità e continuità. In Gualdo, Riccardo & Dell'Anna, Maria Vittoria (a cura di), *La faconda Repubblica, La lingua della politica italiana (1992-2004)*, 10-38. San Cesario di Lecce: Manni.

Guiraud, Pierre. 1954. *Les caractères statistiques du vocabulaire: Essai de méthodologie*. Paris: Presses universitaires de France.

Kuznetsova, Alexandra & Brockhoff, Per B. & Christensen, Rune H.B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82. 1-26. https://doi.org/10.18637/jss.v082.i13

Kyle, Kristopher & Crossley, Scott A. 2015. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly* 49. 757-786. https://doi.org/10.1002/tesq.194

Lotti, Alessandro. 2015. La lingua della 'nuova politica'. Beppe Grillo e Matteo Renzi a confronto. In Petrilli, Raffaella (a cura di), *La lingua politica. Lessico e strutture argomentative*, 103-123. Roma: Carocci.

Ondelli, Stefano. 2020. Semplicità, semplificazione e semplicismo. Su alcuni fraintendimenti nell'analisi della lingua dei politici. *Quaderns d'Italià* 25. 171-188.

Qi, Peng & Zhang, Yuhao & Zhang, Yuhui & Bolton, Jason & Manning, Christopher D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Celikyilmaz, Asli & Wen, Tsung-Hsien (ed.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics. 101-108. https://doi.org/10.18653/v1/2020.acl-demos.14

Serra-Peralta, Marc & Serrà, Joan & Corral, Álvaro. 2021. Heaps' law and vocabulary richness in the history of classical music harmony. *EPJ Data Science* 10(1). Article 1. https://doi.org/10.1140/epjds/s13688-021-00293-8

Spina, Stefania. 2012. *Openpolitica: Il discorso dei politici italiani nell'era di Twitter*. Milano: Franco Angeli.

Tavosanis, Marco. 2016. Il linguaggio della comunicazione politica su Facebook. In Librandi, Rita & Piro, Rosa (a cura di), *L'italiano della politica e la politica per l'italiano: Atti dell'XI Convegno ASLI, Associazione per la Storia della Lingua Italiana*, 677-685. Firenze: Cesati.

Wackerbauer, Renate & Witt, A., & Atmanspacher, Harald & Kurths, Jürgen & Scheingraber, Herbert. 1994. A comparative classification of complexity measures. *Chaos, Solitons & Fractals* 4(1). https://doi.org/10.1016/0960-0779(94)90023-X

Winter, Bodo. 2019. *Statistics for Linguists: An Introduction Using R*. New York: Routledge. https://doi.org/10.4324/9781315165547