Claudia Gigliotti[*], Walter Paci[*], Giovanni Acerboni[**],
Alessandro Panunzi[*], Maria Roberta Perugini[***]

VALUTAZIONE DI TECNICHE DI *PROMPT ENGINEERING* PER LA SEMPLIFICAZIONE DELL'ITALIANO BUROCRATICO E PROFESSIONALE

1. Introduzione

Le questioni relative a quali siano le parole più adeguate al contenuto da esprimere, in base ai fini che il mittente si propone e al pubblico destinatario, date delle condizioni di spazio o tempo disponibile (De Mauro 1980: 135), costituiscono il nucleo delle riflessioni che hanno stimolato il dibattito accademico sulla semplificazione linguistica e sui mezzi per ottenerla. Sulle caratteristiche della lingua burocratica e professionale si dispone di vari studi (Covino 2001; Gualdo/Telve 2011; Serianni 2005; Raso 2005; Viale 2008; Lubello 2014, 2017; Acerboni/Panunzi 2020; Cortelazzo 2021; Piemontese 2023) e manuali di scrittura (*Codice di stile* 1993; Fioritto 1997; Cortelazzo/Pellegrino 2002, 2003; Franceschini/Gigli 2003; ITTIG/Accademia della Crusca 2011). Nel panorama italiano, i reiterati richiami alla necessità di chiarezza e trasparenza nella comunicazione amministrativa si sono tradotti in interventi normativi significativi che hanno trovato espressione nelle due direttive del 2002 (Frattini 2002) e del 2005 (Baccini 2005), nella legge n. 124 del 2015 e nel disegno di legge n. 1812 del 30 aprile 2019. Nel 2013 l'Ente Italiano di normazione (UNI) ha incorporato la letteratura esistente nella norma UNI 11482: 2013 *Elementi strutturali e aspetti linguistici delle comunicazioni scritte delle organizzazioni*. Negli ultimi anni i *Large Language Models* (LLM)

[*] Università degli Studi di Firenze.
[**] L'ink. Scrittura professionale.
[***] Studio Legale Perugini.

hanno avuto un impatto significativo sulla ricerca e su varie applicazioni pratiche quotidiane. Se utilizzare l'intelligenza artificiale è ormai una pratica comune che investe diversi ambiti, sfruttare l'AI per semplificare testi tecnici (Cherubini *et al.* 2023; Acerboni 2024) è un campo emergente che mira a rendere anche i testi amministrativi più comprensibili per i cittadini.

Per mettere in dialogo i due campi di ricerca, sulla scia lasciata dal lavoro di De Mauro/Vedovelli 1999, l'esperimento si propone di valutare l'efficacia delle tecniche di *prompt* nella riscrittura di testi complessi utilizzando un test di comprensione. Le domande sono state articolate su due livelli (Tschense/Wallot 2022): a livello micro per valutare la comprensione di informazioni codificate a livello della frase; e a livello di inferenza, per indagare la capacità di comprendere informazioni non esplicitamente riferite nel testo. Le domande di comprensione sono state affiancate dall'annotazione del numero di riletture effettuate per ciascun testo. Inoltre, lo studio ha incluso quesiti finalizzati alla valutazione percettiva, con riferimento al gradimento della lettura, all'interesse per gli argomenti trattati nei testi e alla difficoltà percepita del test. Il *design* sperimentale è stato creato prendendo ispirazione dalla metodologia *eye tracking* (Godfroid 2020) e si presenta come pre-test degli stimoli sperimentali[1].

2. La semplificazione automatica dei testi

L'*Automatic Text Simplification* (ATS), o semplificazione automatica dei testi, è un *task* della linguistica computazionale che si occupa di generare testi più comprensibili per un determinato gruppo target, mantenendone il significato essenziale. Questo processo coinvolge diversi livelli (testuale, sintattico, lessicale) e richiede abilità di riscrittura (Bott/Saggion 2012). L'efficacia dell'ATS dipende sia dalla lingua del testo che dal gruppo di utenti finali. Le sue applicazioni principali includono:

1) la semplificazione di contenuti tecnico-specifici di un dominio di conoscenze – testi medici, burocratici o legali, spesso di difficile comprensione anche per gli esperti;
2) il supporto alla lettura per individui con competenze linguistiche limitate, come persone non madrelingua o con disabilità cognitive.

[1] La stesura integrale dell'articolo è stata curata da Claudia Gigliotti, che ha progettato il disegno sperimentale della ricerca. Il lavoro computazionale descritto è stato svolto da Walter Paci, che ha contribuito nella definizione dello stato dell'arte dei modelli e delle tecniche di *prompting*. Giovanni Acerboni ha messo a disposizione il materiale sperimentale e, congiuntamente a Maria Roberta Perugini e Alessandro Panunzi, ha partecipato alla valutazione delle riscritture generate dal modello. L'analisi dei dati è stata condotta da Claudia Gigliotti e Walter Paci.

La linguistica computazionale si occupa della semplificazione automatica dei testi dagli anni Novanta, arrivando nel tempo a sviluppare diversi metodi e sistemi in vari contesti (Shardlow 2014; Saggion 2017; Al-Thanyyan/Azmi 2021). La maggior parte degli studi si concentra sulla lingua inglese, ma anche in Italia sono stati compiuti progressi significativi. Inizialmente, la ricerca italiana si è focalizzata su indici di leggibilità per stimare la complessità testuale basata su caratteristiche sintattiche e semantiche. Due esempi noti sono gli indici Gulpease (Lucisano/Piemontese 1988) e READ-IT (Dell'Orletta *et al*. 2011). Il primo sistema di ATS per l'italiano è stato ERNESTA (Barlacchi/Tonelli 2013), sviluppato per semplificare storie rivolte a bambini con difficoltà di lettura. Successivamente, Scarton *et al.* 2017 hanno introdotto MUSST, uno strumento *open source* multilingue per la semplificazione sintattica che utilizza un approccio modulare. Nel panorama italiano recente, Palmero Aprosio *et al*. 2019 e Megna *et al*. 2021 hanno implementato sistemi basati su *transformer* con architettura *encoder-decoder*, raggiungendo lo stato dell'arte per il *task* di ATS. Un punto di svolta è stato segnato nel 2023 con l'introduzione dei Grandi Modelli di Linguaggio (LLMs), che hanno significativamente migliorato la qualità dei testi generati automaticamente. I primi risultati mostrano che, con un'attenta selezione del *prompt*, gli LLMs superano i sistemi precedenti sia nella semplificazione semantica che sintattica (Feng *et al.* 2023; North *et al.* 2023).

Un'analisi approfondita condotta da Nozza/Attanasio (2023) ha valutato diversi LLMs sul *task* di ATS per testi amministrativi italiani. Lo studio ha evidenziato due aspetti principali: (a) gli indici di leggibilità non sono affidabili per la valutazione della generazione automatica di testi tramite LLMs; (b) tra i modelli disponibili, ChatGPT si distingue per le migliori performance in questo ambito. Paci *et al*. 2024 ha inoltre evidenziato, attraverso la valutazione qualitativa di esperti di linguaggio legale-amministrativo di riscritture generate attraverso ChatGPT3.5 turbo, come le frasi generate siano percepite come più semplici da comprendere e che tecniche di *prompting* complesse producano semplificazioni più efficaci.

3. Premesse sperimentali

Il presente studio si fonda sul lavoro di Paci *et al*. 2024, costituendosi come un approfondimento mirato alla valutazione della qualità delle riscritture prodotte in termini di comprensione effettiva dei testi da parte di lettori non esperti. Un elemento di connessione tra i due studi è rappresentato dall'utilizzo della stessa fonte per il materiale sperimentale, ovvero il *corpus* CITPRO (Corpus dell'Italiano Professionale), una risorsa digitale privata di proprietà di Giovanni Acerboni. Il corpus raccoglie documenti burocratici e legali collezionati in oltre vent'anni di lavoro come consulente per la scrittura professionale. È composto da circa 5.000

documenti appartenenti all'amministrazione pubblica e ad aziende private, per un totale di ~100 milioni di parole, ed è suddiviso in cinque categorie principali:

1) Norme: leggi e atti nazionali e locali;
2) Processi: scritti difensivi, sentenze;
3) Comunicazioni: notizie e comunicati stampa;
4) Regolamenti: regolamenti, policy, disposizioni interne, contratti;
5) Documenti tecnico-normativi: manuali, progetti, perizie, offerte, bandi, relazioni e altri documenti tecnici.

Come ulteriore aspetto condiviso entrambi gli studi si focalizzano sull'analisi di due specifici fenomeni linguistici individuati tra i problemi evidenziati dalla norma UNI 11482:

> Periodi lunghi: definiti come frasi con una lunghezza superiore alle 40 parole;
> Cumuli nominali: definiti come frasi caratterizzate dalla presenza di più di quattro complementi indiretti per un solo verbo.

L'ultimo punto di contatto tra le due ricerche riguarda la parte computazionale di questo lavoro in quanto sono state seguite le stesse tecniche di *prompting* descritte in Paci *et al*. 2024. In questo caso per le computazioni è stato usato il modello ChatGPT3.5-turbo aggiornato a febbraio 2024 interrogato tramite API.

3.1. *Selezione preliminare degli stimoli*

Com'è noto, «i significati potenziali di un testo vengono attualizzati, e acquistano un senso univoco, all'interno di uno specifico contesto» (Palermo 2013: 27), ma nell'esperimento condotto da Paci *et al.* 2024 sono state impiegate frasi estratte da un contesto testuale più ampio. Per non compromettere la validità di un test di comprensione è stato necessario assicurarsi che il materiale sperimentale fosse idoneo alla costruzione del test. Per rimediare al problema della mancanza di contesto destro e sinistro degli *item*[2], dal dataset di Paci *et al*. 2024 sono state selezionate trentasei frasi originali non riscritte (diciotto cumuli nominali[3] e diciotto periodi lunghi[4]) e sono state sottoposte a una valutazione esterna per vagliare la capacità del ricevente di cogliere il senso complessivo del testo e determinare quali tra i possibili *item* fossero comprensibili anche in mancanza del contesto di riferimen-

[2] I termini *items* e *stimoli* sono utilizzati come sinonimi in questo contributo. La scelta di una delle due espressioni è pura questione di variazione linguistica.
[3] Da qui in poi abbreviati in Cn.
[4] Da qui in poi abbreviati in Pl.

to. La selezione delle frasi originali è stata guidata da una valutazione preliminare condotta sulla base di tre criteri ispirati alla metodologia *eye tracking* (Conklin *et al.* 2018) e alle formule di leggibilità tradizionali e basati sulle scienze cognitive (Benjamin 2012: 65-69):

1) Non familiarità del contenuto: l'informazione trasmessa dalle frasi non doveva essere inferibile da conoscenze condivise o generali. Diversamente, le risposte corrette al test non avrebbero riflettuto l'effettiva comprensibilità delle frasi, risultando invece influenzate da conoscenze pregresse condivise dai partecipanti, i quali non avrebbero necessitato di leggere il contenuto delle frasi per rispondere adeguatamente alle domande.
2) Principio di efficienza (Palermo 2013: 43): il messaggio veicolato dalle frasi era ridotto al minimo indispensabile, caratteristica che le rendeva potenzialmente comprensibili anche in assenza di un contesto esplicito.
3) Bilanciamento della lunghezza degli stimoli: nel confronto tra le frasi originali e rispettive riscritture, è stata data la priorità alle frasi che presentavano riscritture con una lunghezza quanto più possibile simile in termini di numero di parole[5]. Poiché le frasi più brevi tendono a essere più facilmente comprensibili rispetto a quelle più lunghe, questo approccio mirava a garantire un confronto che non fosse influenzato dalla diversa quantità di informazioni veicolate.

Le frasi originali selezionate sono state sottoposte a una valutazione esterna mediante un sondaggio realizzato su Google Forms e diffuso tramite passaparola sui canali social raggiungendo un campione composto da 65 annotatori, tutti madrelingua italiani, con un'età minima di 18 anni e fino a oltre 50, e un titolo di studio minimo pari al diploma. Ai partecipanti è stato richiesto di leggere tutte le trentasei frasi originali e successivamente rispondere alla domanda: «Il testo è comprensibile senza il contesto di riferimento?». Le opzioni di risposta disponibili erano «Sì» o «No». Le percentuali di risposte positive («Sì») corrispondeva alla proporzione di rispondenti che hanno espresso una preferenza favorevole rispetto al totale delle risposte raccolte. Questo dato è stato calcolato automaticamente da Google Form ed è stato organizzato e utilizzato per valutare l'idoneità degli stimoli per la costruzione del test di comprensione:

[5] Sono state considerate accettabili le riscritture con uno scarto non superiore a 30 parole rispetto all'originale.

	Cumuli nominali	Percentuale	Periodi lunghi	Percentuale
1	Cn_10	90,8%	Pl_5	92,3%
2	Cn_55	86,2%	Pl_8	87,7%
3	Cn_53	84,6%	Pl_25	81,5%
4	Cn_37	83,1%	Pl_58	81,5%
5	Cn_48	83,1%	Pl_11	80%
6	Cn_23	76,9%	Pl_41	76,9%
7	Cn_3	75,4%	Pl_22	73,8%
8	Cn_25	73,8%	Pl_19	73,8%
9	Cn_18	73,8%	Pl_60	72,3%
10	Cn_12	72,3%	Pl_10	72,3%
11	Cn_20	72,3%	Pl_7	61,5%
12	Cn_35	69,2%	Pl_45	60%
13	Cn_39	67,7%	Pl_44	56,9%
14	Cn_14	66,2%	Pl_42	55,4%
15	Cn_54	58,5%	Pl_39	53,8%
16	Cn_14	53,3%	Pl_49	53,8%
17	Cn_16	52,3%	Pl_21	50,8%
18	Cn_60	41,5%	Pl_51	43,1%

Sulla base dei risultati ottenuti, le prime sei frasi sono state selezionate come stimoli sperimentali per la realizzazione del test. Per garantire il rispetto del criterio di lunghezza in modo ancora più accurato, alcune delle riscritture sono state modificate tramite chat per perfezionare gli stimoli in termini di bilanciamento della lunghezza. Nel caso dei cumuli nominali, non sono state effettuate ricomputazioni sulle riscritture generate con il metodo *Chain-of-Thought*[6], mentre sono state prodotte nuove versioni per alcuni stimoli generati con *few-shot* a uno (1) esempio[7] e *few-shot* a tre (3) esempi[8]. Per i periodi lunghi, sono state create due nuove riscritture con il metodo *zero-shot*[9], mentre per fs1 e fs3 si è resa necessaria una ricomputazione più consistente per migliorare il bilanciamento tra gli stimoli. Gli esempi forniti in chat per guidare la ricomputazione in *few-shot* comprendeva-

[6] Da qui in poi abbreviato in CoT.
[7] Da qui in poi abbreviato in fs1.
[8] Da qui in poi abbreviato in fs3.
[9] Da qui in poi abbreviato in zs.

no il Cn_3 e il Pl_7 rispettivamente per i gruppi dei cumuli nominali fs1 e periodi lunghi fs1. Per generare i gruppi dei cumuli nominali fs3 e periodi lunghi fs3 sono stati utilizzati come esempi: i cumuli nominali Cn_3, Cn_25, Cn_12; e i periodi lunghi Pl_60, Pl_10, Pl_7. Le modifiche condotte sugli stimoli sono riassunte nella tabella che segue:

Gruppi *item*/ conteggio parole[10]		Manipolazione in chat/conteggio parole						Lunghezza media (numero parole)	
Cn_O[11]	n°	Cn_fs1[12]	n°	Cn_fs3[13]	n°	Cn_CoT[14]	n°	Riscrit.	Tot. *item*
Cn_10_O	50	No	38	No	42	No	38	39,3	42
Cn_55_O	48	Sì	44	Sì	36	No	34	38	40,5
Cn_53_O	23	No	24	No	25	No	25	24,6	24,25
Cn_37_O	33	No	31	No	28	No	26	28,3	29,5
Cn_48_O	44	Sì	34	Sì	37	No	71	47,3	46,5
Cn_23_O	37	No	37	No	34	No	33	34,6	35,25
Pl_O[15]	n°	Pl_fs1[16]	n°	Pl_fs3[17]	n°	Pl_zs[18]	n°	Riscrit.	Tot. *item*
Pl_5_O	55	Sì	59	Sì	47	Sì	56	54	54,25
Pl_8_O	59	Sì	52	Sì	50	No	58	53,3	54,75
Pl_25_O	60	No	53	No	60	No	61	58	58,5
Pl_58_O	77	Sì	89	Sì	83	No	71	81	80
Pl_11_O	79	Sì	61	Sì	59	Sì	81	67	70
Pl_41_O	48	Sì	53	Sì	54	No	48	51,6	50,75

[10] Nel grafico indicato come «n°».
[11] Cumuli nominali originali.
[12] Cumuli nominali computati con *few-shot* a uno (1) esempio.
[13] Cumuli nominali computati con *few-shot* a tre (3) esempi.
[14] Cumuli nominali computati con *Chain-of-Thought*.
[15] Periodi lunghi originali.
[16] Periodi lunghi computati con *few-shot* a uno (1) esempio.
[17] Periodi lunghi computati con *few-shot* a tre (3) esempi.
[18] Periodi lunghi computati con *zero-shot*.

4. Valutazione di tecniche di *prompt engineering*: lo studio

Per il test di comprensione il *data set* originale è stato ridotto a dodici frasi originali e rispettive riscritture per un totale di quarantotto stimoli, organizzati in quattro liste sulla base della diversa tipologia:

Lista 1 – *item* originali	Lista 2 – *item* con fs1	Lista 3 – *item* con fs3	Lista 4 – *item* con altro *prompt*	N° *item*
Cn_10_O Cn_55_O Cn_53_O Cn_37_O Cn_48_O Cn_23_O	Cn_10_fs1 Cn_55_fs1 Cn_53_fs1 Cn_37_fs1 Cn_48_fs1 Cn_23_fs1	Cn_10_fs3 Cn_55_fs3 Cn_53_fs3 Cn_37_fs3 Cn_48_fs3 Cn_23_fs3	Cn_10_CoT Cn_55_CoT Cn_53_CoT Cn_37_CoT Cn_48_CoT Cn_23_CoT	24 *item* cumuli nominali
Pl_5_O Pl_8_O Pl_25_O Pl_58_O Pl_11_O Pl_41_O	Pl_5_fs1 Pl_8_fs1 Pl_25_fs1 Pl_58_fs1 Pl_11_fs1 Pl_41-fs1	Pl_5_fs3 Pl_8_fs3 Pl_25_fs3 Pl_58_fs3 Pl_11_fs3 Pl_41_fs3	Pl_5_zs Pl_8_zs Pl_25_zs Pl_58_zs Pl_11_zs Pl_41_zs	24 *item* periodi lunghi
Totale *item*				48 *item* totali

Al fine di garantire un *set* di dati completo per tutte le condizioni ed evitare ripetizioni, l'esperimento ha adottato un *design within-subject* con controbilanciamento in cui ogni partecipante ha visto tutte le condizioni sperimentali, ma solo in una delle versioni possibili. Ciò significa che, se un partecipante vedeva un *item* nella versione originale, non lo vedeva nella versione generata e viceversa. Per prevenire che uno sforzo cognitivo eccessivo compromettesse la qualità dei risultati, sono state seguite le indicazioni di Brysbaert 2019[19] e la lunghezza del test è stato limitato a circa 3000 parole, incluse domande e risposte, per ottenere un tempo di lettura di circa 10-15 minuti[20]. Pertanto, la valutazione degli *items* è stata suddivida

[19] Il lavoro prende in considerazione la lingua inglese, ma l'italiano ha un'ortografia più trasparente e la lettura potrebbe risultare anche più veloce.

[20] La durata del test prevista di 10 minuti è stata confermata dai dati raccolti sulla piattaforma Prolific.

in otto test distinti, ciascuno contenente tre cumuli nominali e tre periodi lunghi. Si veda la tabella:

	Tipo di *items*					
TEST 1	Cn_10_O	Pl_5_fs1	Cn_55_fs3	Pl_8_zs	Cn_53_O	Pl_11_fs1
TEST 2	Cn_10_fs1	Pl_5_fs3	Cn_55_CoT	Pl_8_O	Cn_53_fs1	Pl_11_fs3
TEST 3	Cn_10_fs3	Pl_5_zs	Cn_55_O	Pl_8_fs1	Cn_53_fs3	Pl_11_zs
TEST 4	Cn_10_CoT	Pl_5_O	Cn_55_fs1	Pl_8_fs3	Cn_53_CoT	Pl_11_O
TEST 5	Cn_37_O	Pl_58_fs1	Cn_48_fs3	Pl_25_zs	Cn_23_O	Pl_41_fs1
TEST 6	Cn_37_fs1	Pl_58_fs3	Cn_48_CoT	Pl_25_O	Cn_23_fs1	Pl_41_fs3
TEST 7	Cn_37_fs3	Pl_58_zs	Cn_48_O	Pl_25_fs1	Cn_23_fs3	Pl_41_zs
TEST 8	Cn_37_CoT	Pl_58_O	Cn_48_fs1	Pl_25_fs3	Cn_23_CoT	Pl_41_O

La presentazione di ciascuno stimolo è stata randomizzata, per ogni cumulo nominale e periodo lungo (sia nella versione originale che nelle riscritture) sono state associate le stesse due domande di comprensione a scelta multipla (MCQs)[21] integrate con *wh-questions* e domande Sì/No. Ogni tipologia di domanda presentava tre opzioni di risposta[22] randomizzate per evitare effetti di ordine. Si rimanda agli esempi riportati, nei quali la risposta corretta è indicata con un asterisco (*):

> MCQs: Qual è l'ambito di ricerca? [N.B. La domanda valuta la capacità inferenziale, poiché il termine "ambito di ricerca" non è esplicitamente menzionato nel testo.]
> 1) La relazione tra operatori pubblici e privati.
> 2) La gestione energetica. (*)
> 3) La sperimentazione di nuovi modelli di prassi manageriale.
>
> *wh-question*: Chi è responsabile della rilevazione dei consumi?
> 1) Solo gli addetti.
> 2) Solo gli utenti.
> 3) Addetti e utenti. (*)

[21] *Multiple-choice questions.* Cfr. a riguardo Newton 2024.

[22] Se per le MCQs e le *wh-questions* la risposta sbagliata era formulata sul contenuto delle frasi; per le domande Sì/No si è scelto di includere come terza opzione di risposta: «Non posso dirlo con certezza». Cfr. a riguardo Sulem *et al*. 2022; Groothuis/Whitehead 2002.

Sì/No: I finanziamenti per gli investimenti vengono concessi dalla "Sezione tecnologie digitali"?
1) Sì. (*)
2) No.
3) Non posso dirlo con certezza.

Al fine di verificare il livello di attenzione del partecipante, in ogni sondaggio sono state incluse due domande *wh-* a risposta aperta – una per il set di test 1-4 e una per il set di test 5-8 – con la funzione di domande di controllo:

Test 1-4) Domanda aperta_Cn_55: «A chi è rivolto il finanziamento?»
Test 5-8) Domanda aperta_Pl_25: «Dove deve essere registrato il permesso per partecipare all'assemblea sindacale?»

A ogni stimolo sperimentale sono state associate due domande di comprensione, consentendo di ottenere un punteggio massimo complessivo di dodici punti per ciascun test, includendo nel conteggio la domanda a risposta aperta.

Durante il test, ai partecipanti era consentito rileggere il testo, pertanto ogni *item* è stato accompagnato da una domanda finale («Quante volte hai riletto il testo?»), in cui veniva richiesto di indicare il numero di riletture effettuate per ciascuno stimolo da un minimo di 0 («Non ho riletto il testo») a un massimo di 5 («Ho riletto il testo 5 o più volte»). Al termine del test è stato richiesto al partecipante di rispondere a domande personali di natura sociodemografica (genere, età, titolo di studi, provenienza geografica) e sono state poste domande di valutazione soggettiva con scala Likert 1-5 su: gradimento di lettura, interesse degli argomenti, percezione della difficoltà del test.

Scala Likert gradimento/interesse		Scala Likert difficoltà percepita	
1	Per niente	1	Molto facile
2	Poco	2	Facile
3	Discreto	3	Moderata
4	Più che discreto	4	Difficile
5	Molto	5	Molto difficile

I test sono stati realizzati su Google Form e sono stati somministrati con due metodologie: attraverso *crowdsourcing* sulla piattaforma Prolific; attraverso passaparola sui social network. In entrambi i casi, è stato effettuato uno *screening* preliminare dei partecipanti, richiedendo che fossero madrelingua italiani con un livello di istruzione minimo pari al diploma. Per i partecipanti su Prolific, sono stati utilizzati i filtri di selezione offerti dalla piattaforma, mentre per la raccolta

tramite passaparola è stata condivisa una richiesta di partecipazione volontaria, non retribuita, e ai partecipanti idonei veniva inviato via e-mail il link al test da compilare. I partecipanti sono stati suddivisi in otto gruppi, ciascuno dei quali ha completato un test diverso:

Test	Gruppo	N° partecipanti Prolific	N° partecipanti passaparola
1	1	10	15
2	2	10	14
3	3	10	14
4	4	10	14
5	5	10	11
6	6	10	16
7	7	10	11
8	8	10	14

4.1. *Analisi dei risultati*

Alla ricerca hanno aderito 189 partecipanti. Tuttavia, sono stati esclusi dall'analisi i test in cui la domanda a risposta aperta (domanda di controllo) non veniva completata, portando il totale dei partecipanti analizzabili a 160. Questo campione risulta sufficientemente bilanciato sia in termini di età sia di livello di istruzione.

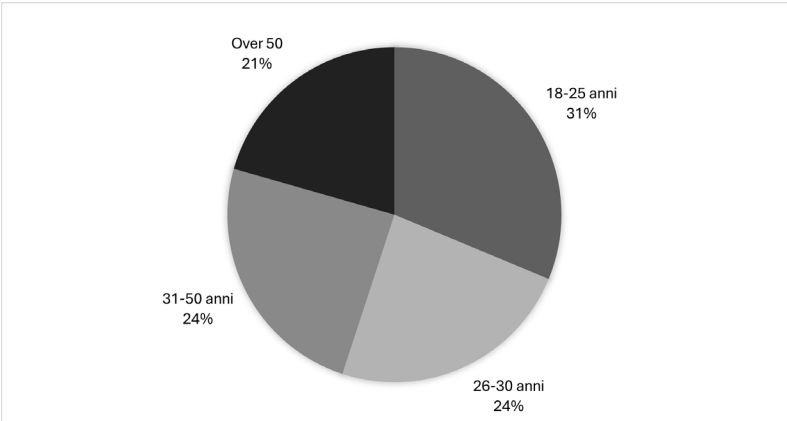

Fig. 1 - Distribuzione del campione per fasce d'età

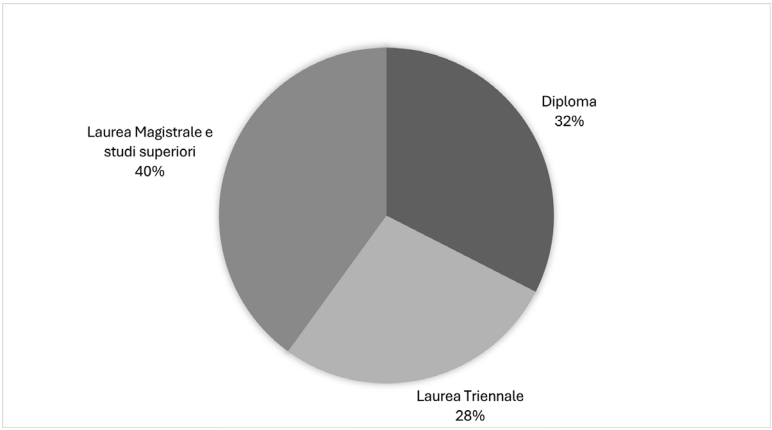

Fig. 2 - Distribuzione del campione per titolo di studi

Lo studio condotto da Paci *et al*. 2024 aveva evidenziato come, nei compiti di semplificazione di testi tecnici, l'applicazione di tecniche di *prompting* più articolate tendesse a generare risultati migliori rispetto a tecniche meno complesse. Nel presente studio, tuttavia, l'analisi dell'accuratezza delle risposte non rivela differenze statisticamente significative tra le diverse tipologie di *item*, suggerendo una comprensione relativamente uniforme. Analogamente, il numero di riletture effettuate dai partecipanti non mostra variazioni significative in relazione alla tipologia di *item*.

Gruppi *item*	Media *Accuratezza di risposta*	Dev. Stand.	Media *Riletture*	Dev. Stand.
Cn_O	0,76	0,06	1,93	1,11
Cn_fs1	0,83	0,08	1,83	1,08
Cn_fs3	0,78	0,04	1,58	1,07
Cn_CoT	0,77	0,06	1,8	1,06
Pl_O	0,76	0,16	1,83	1,23
Pl_fs1	0,8	0,09	1,65	1,17
Pl_fs3	0,78	0,13	2,01	1,25
Pl_zs	0,79	0,08	1,81	1,07

Il dato secondo cui ciascun testo richiede in media due riletture, indipendentemente dalla tipologia di *item* (Cn/Pl sia originali che riscritti), suggerisce l'esistenza di un certo grado di difficoltà nella comprensione. Questo risultato potreb-

be indicare che, sebbene la complessità dei testi non vari significativamente, il livello generale di accessibilità linguistica o cognitiva di tutti i testi rimane relativamente elevato, richiedendo uno sforzo interpretativo aggiuntivo da parte dei partecipanti. In termini di punteggio medio ottenuto nel test di comprensione, non emergono differenze significative nelle performance dei partecipanti in relazione all'età. Tuttavia, un livello di istruzione più elevato è associato a una maggiore comprensione del testo.

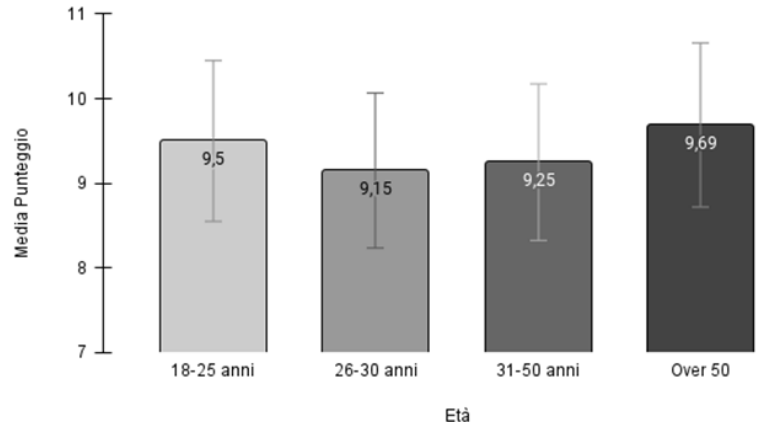

Fig. 3 - Media dei punteggi in relazione a fasce d'età

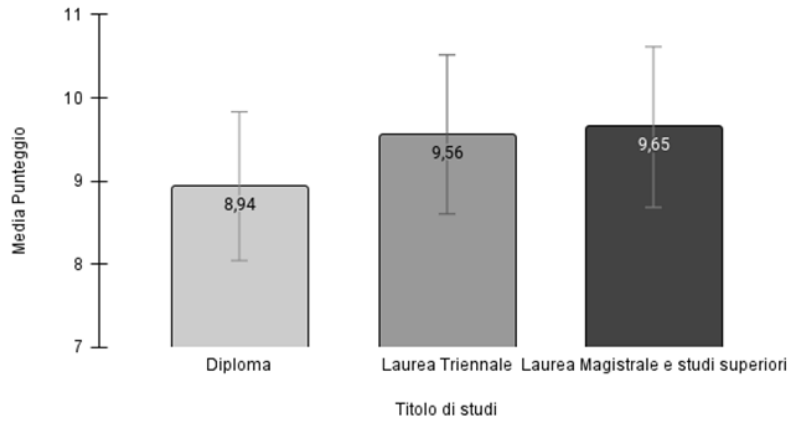

Fig. 4 - Media dei punteggi in relazione a titolo di studi

Si rileva una lieve correlazione tra le classi di punteggio ottenute e il numero di riletture effettuate, suggerendo che il processo di rilettura possa contribuire a migliorare la comprensione del contenuto testuale. Questo risultato potrebbe indicare che

l'approfondimento derivante dalla rilettura favorisce una maggiore assimilazione delle informazioni, con un impatto positivo, seppur moderato, sull'accuratezza delle risposte.

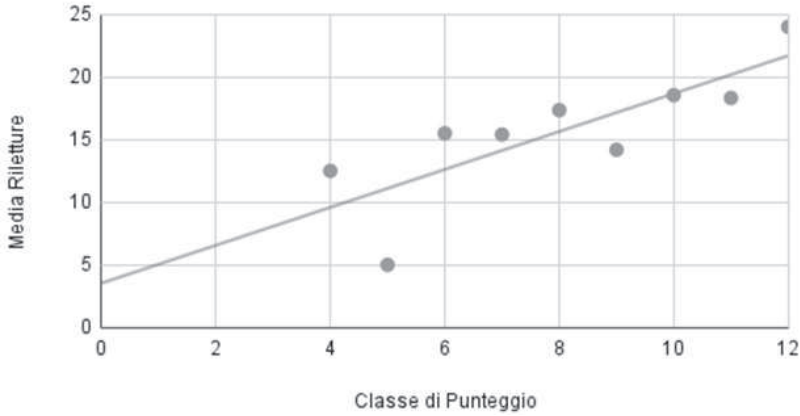

Fig. 5 - Correlazione tra classi di punteggio e media riletture

La correlazione tra un punteggio più alto nel test di comprensione – indicativo di una maggiore comprensione – e un numero maggiore di riletture è confermata dalle performance dei partecipanti, in quanto la media delle riletture rispecchia la media dei punteggi, sia in relazione alla fascia di età che al titolo di studio.

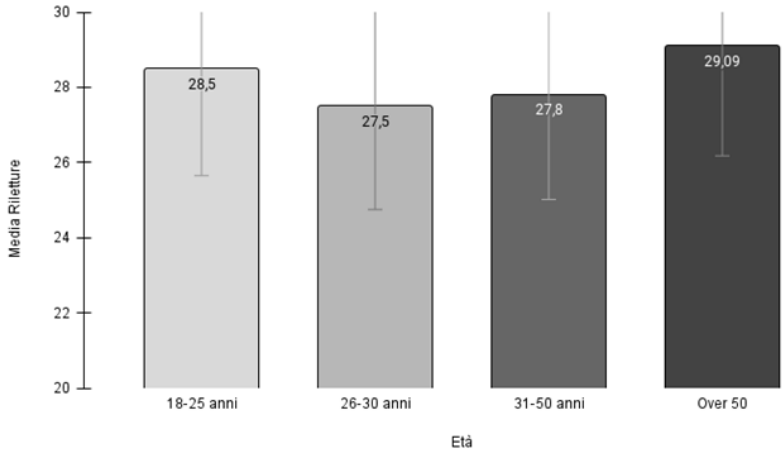

Fig. 6 - Media riletture in relazione a fasce d'età

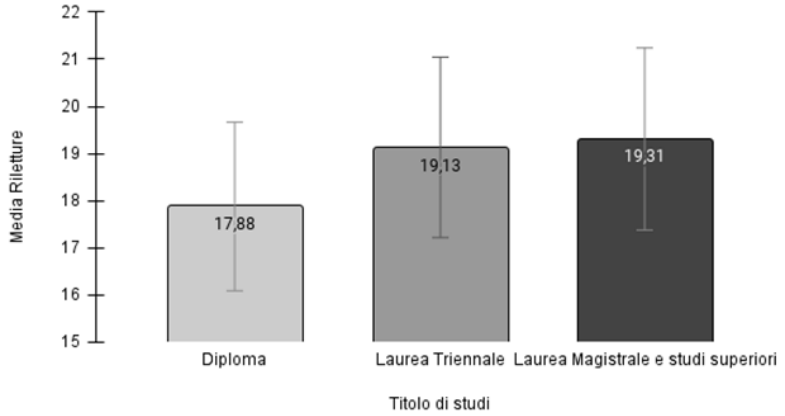

Fig. 7 - Media riletture in relazione a titolo di studi

Rispetto alle domande di valutazione soggettiva relative a gradimento, interesse e difficoltà percepita, la media di gradimento della lettura e di interesse per gli argomenti proposti risulta generalmente bassa, con un punteggio medio intorno a 2 su scala Likert 1-5, corrispondente alla valutazione "poco". Al contrario, la difficoltà del test è stata percepita a un livello medio, con una valutazione di 3 sulla stessa scala Likert, indicando una percezione di difficoltà "moderata".

	Media	Dev. Stand.
«Quanto valuti il tuo gradimento di lettura dei testi proposti?»	2,3	1,11
«Quanto valuti il tuo interesse per gli argomenti dei testi proposti?»	2,1	1,01
«Quanto valuti la difficoltà del test?»	3	0,91

Un dato particolarmente interessante emerge dalla percezione della difficoltà del test: si osserva che la media dei punteggi di comprensione tende ad aumentare quando il test è percepito come molto facile, mentre diminuisce nei casi in cui i partecipanti lo percepiscono come molto difficile. Questo risultato evidenzia una correlazione tra la percezione soggettiva della complessità del compito e le prestazioni effettive, suggerendo che la difficoltà percepita possa influire negativamente sulla capacità di elaborazione e risposta.

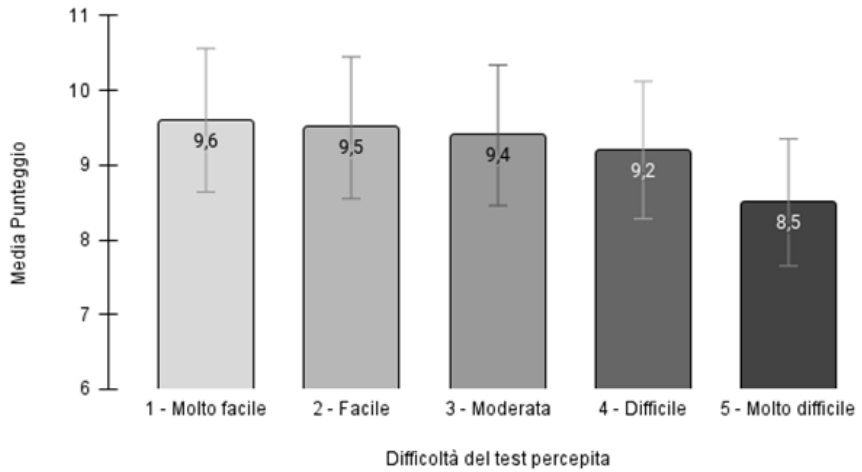

Fig. 8 - Media dei punteggi in relazione alla difficoltà del test percepita

5. Conclusioni

Il presente studio ha messo in evidenza che, sebbene l'aggiornamento tecnologico di ChatGPT permetta prestazioni soddisfacenti anche con *prompt* meno complessi, i risultati ottenuti, pur mostrando aspetti di interesse, non raggiungono ancora livelli pienamente soddisfacenti. Nonostante le differenze strutturali tra le tipologie di *item*, non emergono differenze statisticamente significative né in termini di accuratezza delle risposte né nel numero di riletture effettuate, suggerendo che le riscritture mantengono una complessità percepita simile ai testi originali. Il fatto che i testi richiedano in media due riletture per essere compresi, indipendentemente dalla tipologia di *item*, suggerisce un livello generale di difficoltà nella comprensione. Questo risultato potrebbe essere legato sia alla complessità intrinseca dei testi tecnici che a una possibile distanza linguistica o cognitiva dei partecipanti non esperti. La lieve correlazione tra il numero di riletture e i punteggi di comprensione indica che rileggere i testi contribuisce a migliorare, seppur moderatamente, l'assimilazione delle informazioni, sottolineando il ruolo del tempo e dell'attenzione dedicati alla comprensione del testo. La percezione soggettiva della difficoltà del test si rivela un fattore determinante nelle prestazioni. Una media di gradimento e di interesse generalmente bassa evidenzia una scarsa attrattività dei testi per i partecipanti, fattore che potrebbe influenzare negativamente l'*engagement* e la motivazione durante l'esecuzione del test. Punteggi più alti si osservano quando i partecipanti giudicano il test come molto facile, mentre una percezione

della difficoltà più elevata si associa a una riduzione delle performance. Ciò evidenzia l'importanza della percezione individuale nella gestione cognitiva del compito, con possibili ripercussioni sulla comprensibilità dei testi. Sebbene le misure relative al gradimento, all'interesse e alla percezione della difficoltà del test sembrino avvalorare il fenomeno sociale – comunemente identificato nell'etichetta di "burocratese" – che attribuisce una connotazione negativa al linguaggio burocratico, sarebbe opportuno approfondire l'indagine relativa al pregiudizio linguistico associato ai testi tecnici e burocratici al fine di ottenere risposte più precise. Inoltre, lo studio è stato condotto esclusivamente su estratti di testo privi di un contesto testuale di riferimento: l'uso di singole frasi limita la valutazione, restringendola ai due fenomeni specifici oggetto di indagine. Potrebbe essere opportuno, per future ricerche, considerare l'analisi di segmenti di testo più ampi al fine di ottenere risultati più significativi e completi per valutare la comprensione dei testi tecnici in modo più ampio e approfondito. Sarebbe altresì auspicabile replicare il test in una condizione di controllo sperimentale che impedisca al partecipante di rileggere il testo, consentendo così una valutazione più accurata della comprensione. Sviluppi futuri prevedo di approfondire l'indagine avvalendosi di strumenti di *eye tracking*, al fine di ottenere dati più dettagliati e affidabili sulla comprensione testuale e sui processi cognitivi coinvolti.

Riferimenti bibliografici

Acerboni 2024 = Giovanni Acerboni, *Sintesi e chiarezza degli atti processuali. Un contributo linguistico*, in Gina Gioia (a cura di), *Chiari e sintetici. Come scrivere in maniera efficace gli atti processuali secondo gli esperti*, Pisa, Pacini: 81-99.

Acerboni/Panunzi 2020 = Giovanni Acerboni / Alessandro Panunzi, *La scrittura professionale*, in Benedetta Baldi (a cura di), *Comunicare ad arte. Per costruire contenuti e promuovere eventi*, Bologna, Zanichelli: 221-236.

Al-Thanyyan/Azmi 2021 = Suha Al-Thanyyan / Aqil Azmi, *Automated text simplification: a survey*, in «ACM Computing Surveys», 54(2): 1-36.

Baccini 2005 = Mario Baccini, *Direttiva sulla semplificazione del linguaggio delle Pubbliche amministrazioni*, Dipartimento della Funzione Pubblica, www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2005-10-18&atto.codiceRedazionale=05A09847&elenco30giorni=false.

Barlacchi/Tonelli 2013 = Gianni Barlacchi / Sara Tonelli, *Ernesta: A Sentence Simplification Tool for Children's Stories in Italian*, in Alexander Gelbukh (a cura di), *Computational Linguistics and Intelligent Text Processing*. Proceedings of the14th International Conference CICLing 2013 (Samos, March 24-30, 2013), Berlin, Springer: 476-487.

Benjamin 2012 = Rebekah G. Benjamin, *Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty*, in «Educational Psychology Review», 24, 1: 63-88.

Bott/Saggion 2012 = Stefan Bott / Horacio Saggion, *Automatic simplification of Spanish text for e-accessibility*, in Klaus Miesenberger / Arthur Karshmer / Petr Penaz / Wolfgang Zagler (a cura di), *Computers Helping People with Special Needs. Lecture Notes*, Berlin, Springer: 527-534.

Brysbaert 2019 = Marc Brysbaert, *How many words do we read per minute? A review and meta-analysis of reading rate*, in «Journal of Memory and Language», 109, doi.org/10.1016/j.jml.2019.104047.

Cherubini *et. al* 2023 = Manola Cherubini / Francesco Romano / Francesco Bolioli / Nazareno De Francesco / Irene Benedetto, *La summarization di testi giuridici: una sperimentazione con GPT-3*, in «Rivista italiana di informatica e diritto» 5, 1: 191-204, doi.org/10.32091/RIID0103.

Codice di stile 1993 = *Codice di stile delle comunicazioni scritte ad uso delle amministrazioni pubbliche proposta e materiali di studio*, Roma, Istituto Poligrafico e Zecca dello Stato.

Conklin *et al.* 2018 = Kathy Conklin / Ana Pellicer-Sánchez / Gareth Carrol, *Eye-Tracking. A Guide for Applied Linguistics Research*, Cambridge, Cambridge University Press.

Cortelazzo 2021 = Micheale A. Cortelazzo, *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*, Roma, Carocci.

Cortelazzo/Pellegrino 2002 = Michele A. Cortelazzo / Federica Pellegrino, *30 regole per scrivere testi amministrativi chiari*, Università di Padova, www.maldura.unipd.it/buro/.

Cortelazzo/Pellegrino 2003 = Michele A. Cortelazzo / Federica Pellegrino, *Guida alla scrittura istituzionale*, Roma-Bari, Laterza.

Covino 2001 = Sandra Covino (a cura di), *La scrittura professionale. Ricerca, prassi, insegnamento*. Atti del I Convegno di Studi (Perugia, Università per Stranieri, 23-25 ottobre 2000), Firenze, Olschki.

De Mauro 1980 = Tullio De Mauro, *Guida all'uso delle parole*, Roma, Editori Riuniti.

De Mauro/Vedovelli 1999 = Tullio De Mauro / Massimo Vedovelli (a cura di), *Dante, il gendarme e la bolletta. La comunicazione pubblica in Italia e la nuova bolletta Enel*, Roma-Bari, Laterza.

Dell'Orletta *et al.* 2011 = Felice Dell'Orletta / Simonetta Montemagni / Giulia Venturi, *READ–IT: Assessing readability of Italian texts with a view to text simplification*, in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, Edinburgh, Association for Computational Linguistics: 73-83.

Feng *et al.* 2023 = Yutao Feng / Jipeng Qiang / Yun Li / Yunhao Yuan / Yi Zhu, *Sentence simplification via large language models*, in «arXiv», arxiv.org/abs/2302.11957.

Fioritto 1997 = Alfredo Fioritto *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*, Bologna, il Mulino (poi *Manuale di stile dei documenti amministrativi*, Bologna, il Mulino, 2009).

Franceschini/Gigli 2003 = Fabrizio Franceschini / Sara Gigli (a cura di), *Manuale di scrittura amministrativa*, Pisa-Roma, Dipartimento di Studi Italianistici (Università di Pisa)-Agenzia delle Entrate.

Frattini 2002 = Franco Frattini, *Direttiva sulla semplificazione del linguaggio dei testi amministrativi*, Dipartimento della Funzione Pubblica, www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2002-06-18&atto.codiceRedazionale=02A07864&elenco30giorni=false.

Godfroid 2020 = Aline Godfroid, *Eye tracking in second language acquisition and bilingualism. A research synthesis and methodological guide*, London-New York, Routledge.

Groothuis/Whitehead 2002 = Peter Groothuis / John Whitehead, *Does Don't Know Mean No? Analysis of 'Don't Know' Responses in Contingent Valuation Questions*, in «Applied Economics», 34(15): 1935-1940.

Gualdo/Telve 2011 = Riccardo Gualdo / Stefano Telve, *Linguaggi specialistici dell'italiano*, Roma, Carocci.

ITTIG/Accademia della Crusca 2011 = *Guida alla redazione degli atti amministrativi. Regole e suggerimenti*, Firenze, ITTIG-CNR, hdl.handle.net/2158/540886.

Lubello 2014 = Sergio Lubello, *Il linguaggio burocratico*, Roma, Carocci.

Lubello 2017 = Sergio Lubello, *La lingua del diritto e dell'amministrazione*, Bologna, il Mulino.

Lucisano/Piemontese 1988 = Pietro Lucisano / Maria Emanuela Piemontese, *GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana*, in «Scuola e città», 31, 3: 110-124.

Megna *et al.* 2021 = Angelo Luigi Megna / Daniele Schicchi / Giosuè Lo Bosco / Giovanni Pilato, *A controllable text simplification system for the Italian language*. 2021 IEEE 15th International Conference on Semantic Computing (ICSC) (Laguna Hills, January 27-29, 2021): 191-194.

Newton 2024 = Philip M. Newton, *Guidelines for Creating Online MCQ-Based Exams to Evaluate Higher Order Learning and Reduce Academic Misconduct*, in Sarah Elaine Eaton (a cura di), *Second Handbook of Academic Integrity*, Berlin, Springer: 269-285.

North *et al.* 2023 = Kai North / Tharindu Ranasinghe / Matthew Shardlow / Marco Zampieri, *Deep Learning approaches to lexical simplification: A survey*, in «arXiv», doi.org/10.48550/arXiv.2305.12000.

Nozza/Attanasio 2023 = Debora Nozza / Giuseppe Attanasio, *Is it really that simple? Prompting Language Models for Automatic Text Simplification in Italian*, in Federico Boschetti / Gianluca E. Lebani / Bernardo Nicole Novielli (a cura di), *CLiC-it 2023. Proceedings of the 9th Italian Conference on Computational Linguistics*, Venezia, CEUR Workshop Proceedings: 322-333.

Paci *et al.* 2024 = Walter Paci / Lorenzo Gregori / Giovanni Acerboni / Alessandro Panunzi / Maria Roberta Perugini, *Exploiting ChatGPT to simplify Italian bureaucratic and professional texts. AI-Linguistica*, in «Linguistic Studies on AI-Generated Texts and Discourses», 1(1), doi.org/10.62408/ai-ling.v1i1.13.

Palermo 2013 = Massimo Palermo, *Linguistica testuale dell'italiano*, Bologna, il Mulino.

Palmero Aprosio *et al.* 2019 = Alessio Palmero Aprosio / Sara Tonelli / Marco Turchi / Matteo Negri / Mattia Di Gangi, *Neural text simplification in low-resource conditions using weak supervision*, in *Proceedings of the Neural-GenWorkshop: Methods for Optimizing and Evaluating Neural Language Generation*, Minneapolis, Association for Computational Linguistics: 37-44.

Piemontese 2023 = Maria Emanuela Piemontese (a cura di), *Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile*, Roma, Carocci.

Raso 2005 = Tommaso Raso, *La scrittura burocratica. La lingua e l'organizzazione del testo*, Roma, Carocci.

Saggion 2017 = Horacio Saggion, *Automatic text simplification*, in Graeme Hirst (a cura di), *Synthesis lectures on Human-Language Technologies*, Berlin, Springer: 7-19.

Scarton *et al.* 2017 = Carolina Scarton / Alessio Palmero Aprosio / Sara Tonelli / Tamara Wanton Martín / Lucia Specia, *MUSST: A multilingual syntactic simplification tool*, in *Proceedings of the International Joint Conference on Natural Language Processing*, Tapei, Association for Computational Linguistics: 25-28.

Serianni 2005 = Luca Serianni, *Un treno di sintomi. I medici e le parole: percorsi linguistici nel passato e nel presente*, Milano, Garzanti.

Shardlow 2014 = Matthew Shardlow, *A survey of automated text simplification*, in «International Journal of Advanced Computer Science and Applications» 4(1): 58-70.

Sulem *et al.* 2022 = Elior Sulem / Jamal Hay / Dan Roth, *Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions*, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, Association for Computational Linguistics: 1075-1085.

Tschense/Wallot 2022 = Monika Tschense / Sebastian Wallot, *Modeling items for text comprehension assessment using confirmatory factor analysis*, in «Frontiers in Psychology», 13, doi.org/10.3389/fpsyg.2022.966347.

Viale 2008 = Matteo Viale, *Studi e ricerche sul linguaggio amministrativo*, Padova, CLEUP.