a cura di
Alessia Lacroce ed Enrico Marello

Linguaggio, complessità e comprensione: strumenti e strategie per la semplificazione

Atti del convegno

Ledizioni

Linguaggio, complessità e comprensione: strumenti e strategie per la semplificazione

Atti del Convegno

a cura di Alessia Lacroce ed Enrico Marello

Ledizioni

Volume pubblicato con il contributo del Dipartimento di Giurisprudenza dell'Università di Torino, fondi DG Eccellenza, MUR 2327 - progetto TuTeLa.

Attribuzione - Condividi allo stesso modo 4.0 Internazionale (CC BY-SA 4.0)

2026 Ledizioni LediPublishing
Via Boselli, 10 – 20136 Milano – Italy
www.ledizioni.it
info@ledizioni.it

Linguaggio, complessità e comprensione: strumenti e strategie per la semplificazione
a cura di Alessia Lacroce ed Enrico Marello

Prima edizione: settembre 2026

ISBN cartaceo: 9791256008698
ISBN eBook: 9791256008704
ISBN PDF Open Access: 9791256008711

Informazioni sul catalogo e sulle ristampe: www.ledizioni.it

Indice

Introduzione 5
M. Emanuela Piemontese

Leggibilità e semplificazione: Natural Language Processing per una lingua del diritto chiara 15
Giulia Venturi

Quantificare la complessità. Uno studio sulla complessità linguistica dei testi giuridico-amministrativi in italiano 35
Alessandro Buttinelli

Il Corpus TARI: costruzione di una risorsa linguistica per l'analisi della complessità del linguaggio tributario-amministrativo 47
Alessia Lacroce

La nominalizzazione nei Contratti Collettivi Nazionali di Lavoro: tra semplificazioni e blocchi sintattici 63
Annarita Miglietta

Metodi e tecniche per modernizzare i testi della pubblica amministrazione 81
Francesco Romano

L'intelligenza artificiale per semplificare i testi della pubblica amministrazione: strategie di *prompt engineering* 95
Daniele Fortis

Large Language Model per la semplificazione del linguaggio amministrativo: una valutazione comparativa tra SEMPL-IT, ChatGPT 5.1, Claude Sonnet 4.6 e Gemini 3 111
Vittorio Ganfi

Introduzione

M. Emanuela Piemontese
(Sapienza - Università di Roma)

1. Una lingua per la democrazia

Sia pure con ritardo rispetto ai paesi di lingua e cultura anglosassone, anche in Italia si è finalmente arrivati a porsi in modo sistematico il problema della qualità del linguaggio della comunicazione pubblica. Sui percorsi non lineari, frammentati e discontinui, attraverso i quali la questione della chiarezza linguistica si è andata imponendo all'attenzione pubblica italiana, possiamo farci un'idea precisa, riprendendo le riflessioni di alcuni linguisti, filosofi, giuristi, giornalisti e comunicatori, a partire per lo meno dagli anni Settanta, Ottanta. Solo per fare qualche nome ricordiamo gli interventi sul tema di Tullio De Mauro, Italo Calvino, Umberto Eco, Lucia Lumbelli, Massimo Baldini, Sabino Cassese, Sergio Lepri.

Per non ripercorrere, anche in questa sede, la storia delle tappe registrate negli ultimi tre decenni nel processo di semplificazione di testi di varia natura e destinazione, rinviamo ai lavori sul tema di R. Bombi, M.A. Cortelazzo, A. Fioritto, F. Franceschini e S. Gigli, R. Gualdo, S. Lubello, G. Melis e G. Tosatti, M.E. Piemontese, G.A. Stella, D. Vellutino, R. Zaccaria, E. Zuanelli.

La prima tappa, a mio avviso, precede addirittura la nascita e la diffusione del movimento *Plain Language* degli inizi degli anni Settanta, nei paesi anglosassoni. Mi riferisco allo sforzo compiuto dai nostri padri e dalle nostre madri Costituenti, per la preparazione, discussione e approvazione della *Costituzione* della Repubblica italiana del 1947. A ottant'anni dalla sua promulgazione, resta tuttora un modello luminoso di uso consapevole, responsabile e democratico della lingua per fini istituzionali. Infatti, una lingua semplice, chiara e comprensibile costituisce uno dei principali fondamenti della cittadinanza democratica. Nei suoi scritti e interventi pubblici, lo ricorda con insistenza Tullio De Mauro, additando la chiarezza dei testi pubblici come precondizione dell'eguaglianza sostanziale fra cittadini. La ragione di tale insistenza di De Mauro è, a nostro avviso, nella consapevolezza che la lingua delle istituzioni non descrive soltanto la realtà sociale e giuridica, ma la regola, la prescrive e la sanziona. Pertanto, se un testo istituzionale risulta ai suoi destinatari opaco, ambiguo, incomprensibile, scritto, come dice Cassese, in lingua ostrogota, produce effetti che vanno ben oltre la sfera della comunicazione.

L'oscurità dei testi istituzionali, infatti, mette a rischio la conoscenza e il rispetto stesso dei diritti e doveri da parte dei cittadini. Pertanto, soprattutto per questo tipo di testi la questione linguistica diventa un problema di natura civile e costituzionale. È ormai chiaro chiunque si interessi e occupi di questi problemi sa che alla base dell'indecifrabilità di un atto amministrativo c'è, di solito, una forte asimmetria di conoscenze tra produttori e riceventi. Contro questa asimmetria si sono mossi inizialmente e con crescente forza teorica e pratica soprattutto alcuni linguisti, ma anche giuristi molto sensibili alla qualità del linguaggio giuridico, legislativo e amministrativo. Per il superamento di modalità comunicative asimmetriche consolidate e difficili da scalfire, da oltre trent'anni sono state studiate e messe a disposizione alcune tecniche di scrittura che possano aiutare chi scrive a produrre testi più leggibili e comprensibili di quelli circolanti. Questo sforzo comune era necessario e lo è ancora anche per il fatto che, come scrive De Mauro[1], "Il paesaggio linguistico italiano è tormentato da faglie e fratture soggiacenti di cui anche chi descrive la condizione linguistica del paese e chi opera in essa, giornalista, insegnante, politico, scrittore, deve farsi consapevole". Affrontare tali faglie e fratture ancora presenti nel panorama linguistico italiano è l'obiettivo di chi oggi lavora per migliorare la qualità dei testi di rilievo pubblico italiani.

2. Dalle «Dieci tesi» (1975) al «Codice di stile» (1993): una storia continua e non lineare

Nella ricostruzione del percorso che ha seguito la crescita della consapevolezza della centralità della lingua nella vita pratica e quotidiana dei cittadini si intravede un filo rosso che dalle «Dieci tesi per l'educazione linguistica democratica» del Gruppo di intervento e studio nel campo dell'Educazione linguistica (GISCEL) del 1975 arriva al *Codice di stile* (Cassese 1993, 2023). Oggi quel filo incrocia sempre più ampi settori della società. Si tratta di una lunga storia, segnata da accelerazioni e arretramenti continui, nonostante che le prime direttive sulla semplificazione del linguaggio amministrativo dei ministri della Funzione pubblica Cassese, Frattini e Baccini, abbiano tentato di tradurre in atti di governo l'esigenza di una lingua più accettabile e accessibile. Sul versante della redazione dei testi amministrativi, infatti, dopo il *Codice di stile*, fu prodotto nel 1997 il *Manuale di stile* nato dal lavoro di un gruppo misto di linguiste e giuristi nell'ambito dei Progetti finalizzati del Dipartimento della Funzione pubblica e curato da Alfredo Fioritto. Il *Manuale di stile* forniva suggerimenti e criteri di scrittura sperimentati in numerosi corsi di formazione del personale amministrativo tenuti tra il 1993 e il 1997.

La pubblicistica linguistica ha progressivamente accompagnato questi sforzi con monografie diventate opere di riferimento ineludibili. Tra i molti lavori, ricordiamo

1 Tullio De Mauro, *Come parlano gli italiani*, in "Quaderns d'Italià", 10, 2005, pp. 133-148.

quelli di De Mauro (1994), Mortara Garavelli (2001), Cortelazzo (2021), Lubello (2017), Piemontese (1996; 2023) che rappresentano snodi significativi. Sul piano operativo, mi si perdoni se mi permetto di ricordarlo, i lavori di chi scrive hanno pionieristicamente orientato le prime occasioni di formazione dei redattori di testi di rilievo pubblico a cui hanno fatto seguito riflessioni e interventi di molti altri linguisti più giovani.

Nonostante il crescente impegno di tanti linguisti e giuristi, il bilancio complessivo non è tuttavia rassicurante. Tornando a riflettere di recente sul *Codice di stile* a distanza trent'anni dalla sua pubblicazione, linguisti e giuristi evidenziano, ciascuno dal proprio punto di vista, come le buone pratiche, pur essendo ormai ampiamente diffuse, faticano a sedimentarsi e stabilizzarsi nelle prassi redazionali quotidiane (Piemontese, 2023). Infatti, la distanza fra precetti normativi e abitudini scrittorie resta marcata. Il paradosso messo a fuoco da Buttinelli in questo volume ne è una conferma evidente: lo Stato esige il rispetto delle proprie prescrizioni, ma allo stesso tempo, non garantisce ai destinatari la possibilità di comprenderle.

3. Leggibilità, comprensibilità e comprensione: tre piani da non confondere

Perché i destinatari siano messi in condizione di capire i testi in genere, e quelli istituzionali in particolare, è necessario che chi li produce li scriva con chiarezza. Possiamo ritenere ormai pienamente acquisite e condivise le tre nozioni a cui fanno riferimento tutti i saggi di questo volume. Innanzitutto, la leggibilità: si tratta di una caratteristica di tipo quantitativo. Riguarda le proprietà superficiali e misurabili del testo: la lunghezza dei periodi, la frequenza delle parole, la struttura ipotattica, la densità lessicale. La comprensibilità è, invece, una caratteristica di tipo qualitativo che chiama in causa innanzitutto il rapporto tra il testo e il destinatario. Considera il lessico mentale, le conoscenze condivise, le inferenze attivate o da attivare per capire il testo e, soprattutto. il modo in cui il testo è organizzato dal punto di vista logico-concettuale. La comprensione è l'esito di tipo probabilistico di un processo cognitivo che si può realizzare (o non realizzare), in certe condizioni, nei singoli lettori. A questo proposito ricordiamo quello che De Mauro scriveva sul rapporto che lega il produttore e il ricevente in un articolo del 1985 ("Linguaggi", II, 3: 22-.31):

> "Se capire vuol dire ripercorrere per il possibile il percorso della progettazione d'una frase e della sua enunciazione, la complessità e la approssimativa problematicità delle operazioni di progettazione e di enunciazione rendono il comprendere linguistico una sequenza incerta e non predeterminabile di operazioni: partiamo altrettanto bene dalla traccia espressiva o dalla situazione d'utenza, dal riconoscimento di blocchi di patrimonio linguistico messo in opera o da ipotesi sul senso ecc. Ed è solo una possibilità (…) che il Pollicino comprendente rintracci l'intero percorso del Pollicino produttore".

La consapevolezza della complessità del comprendere linguistico orienta in modo sempre più fine la scelta degli strumenti di analisi e di intervento dei quali discutono gli Autori e le Autrici dei contributi raccolti in questo volume. Dalla lettura dei vari contributi emerge con chiarezza come la complessità si annidi a diversi livelli del sistema linguistico. La sua riduzione richiede strumenti differenziati e calibrati sulla base di una serie di variabili. Una formula sintetica non basta, da sola, a restituire un profilo accurato; occorrono batterie di indicatori, validati su corpora di riferimento, sensibili alle variazioni dei tipi di testo, degli obiettivi e dei destinatari.

4. La svolta computazionale: dalla descrizione alla misurazione

Negli ultimi due decenni la parte della linguistica che si occupa specificamente di leggibilità dei testi ha conosciuto una trasformazione profonda. Le formule, per così dire, storiche, tra le quali ricordiamo le due più usate in Italia e cioè l'indice di Flesch e il Gulpease (Lucisano e Piemontese, 1988) sono tarate entrambe sull'italiano. Da qualche tempo l'analisi dei testi è fatta con l'aiuto di una serie di indicatori automatici che sicuramente forniscono all'analista informazioni più articolate e dettagliate anche ai fini della riscrittura automatica. Il Trattamento Automatico del Linguaggio Naturale (NLP) consente di profilare i testi su decine di parametri. Vi rientrano dimensioni lessicali, morfosintattiche, sintattiche e testuali, organizzate gerarchicamente. La complessità è quantificata, comparata, modellata statisticamente e restituita per via algoritmica.

In questo solco si collocano i lavori dell'*Italian Natural Language Processing Lab* del CNR-ILC di Pisa, le iniziative dell'Istituto di informatica giuridica e sistemi giudiziari (IGSG) di Firenze e di numerose università italiane. Sono stati prodotti corpora annotati, ontologie di dominio, sistemi di profilazione, ambienti di simulazione della leggibilità. Più contributi di questo volume restituiscono una fotografia fedele dello stato attuale dell'arte delle ricerche di questi ultimi anni.

A questa svolta si è recentemente sovrapposta una seconda accelerazione, di natura tecnologica diversa. I modelli linguistici di grandi dimensioni *Large Language Models* (LLM) hanno aperto possibilità inedite di ri-scrittura automatica. Essi hanno però introdotto questioni nuove, di tipo metodologico, epistemico e deontologico. Quanto è affidabile, in un dominio giuridico-amministrativo, una riscrittura semplificata prodotta da un modello generalista? Come misurare la correttezza del contenuto, oltre al miglioramento in termini di leggibilità dei testi (ri)scritti? Quali strategie di *prompt engineering* possono garantire la resa migliore? Quale ruolo resta al revisore umano?

Sono domande imprescindibili che percorrono trasversalmente più contributi del volume. Esse non ammettono risposte a singole dimensioni, ma richiedono una metodologia integrata, in cui la *misurazione automatica* si combini con la *valutazione di esperti*. La cornice resta quella della linguistica quantitativa, ma con strumenti rinnovati e in continuo sviluppo. La sfida di oggi è quindi riuscire a coniugare la scalabilità delle riscritture automatiche con il rigore filologico necessario al dominio del diritto e i contesti d'uso.

5. I contributi del volume: una mappa ragionata

I sette saggi che costituiscono il volume, anche muovendo da prospettive diverse, convergono tutti su una tesi comune. Questi lavori confermano, ancora una volta, che la semplificazione è un'impresa che richiede sempre l'incontro di linguistica, diritto, informatica, comunicazione pubblica.

Procedendo in ordine logico, nel volume si passa da saggi che possono fare da cornice teorica generale ai casi di studio, fino alle applicazioni più innovative dell'intelligenza artificiale.

Il contributo di Giulia Venturi (CNR-ILC) tratteggia la cornice metodologica entro cui si possono collocare i restanti studi. L'autrice ripercorre la nascita dell'incontro fra linguistica computazionale e lingua del diritto in Italia. Ne ricostruisce le radici, dall'incontro tra Beniamino Placido e Antonio Zampolli al 1969. Mostra come la profilazione linguistica costituisca oggi una metodologia matura, capace cioè di descrivere la complessità su molti piani. La complessità del testo giuridico, scrive Venturi, è una proprietà misurabile lungo dimensioni differenziate. Solo una sua scomposizione "fine" consente interventi di semplificazione mirati e linguisticamente fondati. Il saggio offre, quindi, una sintesi chiara dello stato dell'arte e una chiave di lettura per i risultati delle altre ricerche che arricchiscono il volume.

Il lavoro di Alessandro Buttinelli sposta l'analisi sul terreno empirico, con un caso di studio significativo. L'oggetto di studio è un genere poco indagato e cioè il regolamento comunale per la riscossione della Tassa sui Rifiuti (TARI) la cui inosservanza comporta oneri e sanzioni per i contribuenti. Si tratta di una norma che riguarda l'intera cittadinanza. Ciò nonostante, il livello di complessità linguistica del testo risulta significativamente superiore a quella del corpus di controllo. L'Autore confronta, infatti, un corpus di settecentosessanta regolamenti piemontesi con un corpus Wikipedia tematicamente affine. Le metriche impiegate sono quattro: ricchezza lessicale, aderenza al Vocabolario di Base (VdB) di De Mauro (2019), la lunghezza media delle frasi, l'indice di leggibilità Gulpease. I risultati confermano, attraverso un approccio basato su tecniche di linguistica computazionale e statistica inferenziale, la perdurante inaccessibilità del linguaggio amministrativo italiano. Infatti, i testi amministrativi analizzati presentano barriere lessicali e sintattiche superiori al corpus enciclopedico di controllo, una marcata cristallizzazione formulare e un'elevata densità tecnica. Essi denunciano una scollatura tra la portata generale del tributo e l'opacità del testo che lo disciplina. Sulla base dei dati raccolti, l'Autore auspica, da una parte l'allargamento della sfera di analisi a nuove aree geografiche su scala nazionale, dall'altra nuove ricerche per definire ulteriori variabili di complessità linguistica.

Strettamente connesso al precedente è il contributo di Alessia Lacroce su un altro Corpus TARI. L'Autrice illustra la costruzione di una risorsa linguistica per l'analisi della complessità del linguaggio tributario-amministrativo. Il corpus consiste in una raccolta di atti amministrativi prodotti da comuni piemontesi nell'ambito del progetto TuTeLa dell'Università di Torino che, tra gli altri obiettivi si propone di migliorare

l'accessibilità del linguaggio amministrativo italiano. Il dato è raccolto su scala regionale, con attenzione ai sei generi documentali del ciclo di riscossione. Il progetto parte da una nozione ampia e contestuale di "vulnerabilità", mutuata dagli studi giuridici più recenti. Essa non identifica una categoria fissa di persone, ma una condizione che può riguardare chiunque, anche solo temporaneamente. Questa definizione ha il merito di spostare l'attenzione dalla persona al contesto. Il saggio illustra le modalità di costruzione del corpus e ne traccia il profilo linguistico. I dati risultanti dall'analisi mostrano un profilo morfosintattico del corpus TARI diverso da quello dell'italiano comune, un'alta specificità lessicale e una leggibilità medio-bassa. Di conseguenza gli atti TARI dei comuni piemontesi presi in considerazione risultano caratterizzati da complessità multidimensionale, lessicale e morfosintattica. Nonostante i limiti che l'Autrice stessa esplicita (copertura geografica limitata, vincoli imposti dalla natura amministrativa dei testi), il corpus TARI da lei analizzato costituisce un punto di partenza per programmi di ricerca più ampi e articolati.

Lo studio di Annarita Miglietta si concentra su un fenomeno morfosintattico cruciale, la nominalizzazione. Oggetto di studio sono i Contratti Collettivi Nazionali di Lavoro, terreno fino ad oggi poco esplorato. L'Autrice mostra come la nominalizzazione assolva funzioni ambivalenti. Da un lato condensa l'informazione e ne agevola la circolazione tecnica. Dall'altro genera blocchi sintattici impenetrabili al lettore non esperto. Il caso diventa particolarmente rilevante per i lavoratori stranieri, destinatari del recente decreto flussi 2026-2028. Il contributo di Miglietta rinnova quindi un'attenzione antica della linguistica italiana e la riconduce, con strumenti aggiornati, alle urgenze del presente. Senza sottovalutare il conflitto continuo tra necessità di maggiore chiarezza e validità giuridica dei testi, l'Autrice sottolinea che la riformulazione linguistica non è un fatto meramente linguistico ma fondamento della tecnica legislativa, protesa a garantire il diritto dei cittadini a capire.

Francesco Romano (CNR-IGSG) offre un panorama metodologico sull'ammodernamento dei testi della pubblica amministrazione. Il contributo passa in rassegna trent'anni di lavoro condotto presso l'Istituto di informatica giuridica e sistemi giudiziari di Firenze. Si va dai progetti NIR e Normattiva agli strumenti come Lexedit e Lexedit XXI. Si arriva al recente progetto PRIN «Migrants, Institutions, Translations Easy-read Law». Particolare rilievo è attribuito alle attività rivolte ai giovani, vere e proprie iniziative di alfabetizzazione civica. Romano illustra, inoltre, un metodo collaborativo di semplificazione, che integra competenze giuridiche e linguistiche. Il saggio ribadisce un dato ormai ritenuto strutturale: la necessità di un lavoro di lungo periodo e di forte intersezione disciplinare.

Daniele Fortis porta la sua attenzione, invece, sul settore sanitario, uno dei domini finora meno indagati. L'Autore presenta un'esperienza pilota maturata presso l'Azienda Socio-Sanitaria Territoriale (ASST) Santi Paolo e Carlo di Milano in cui dirige l'Ufficio Relazioni con il Pubblico e i servizi di accoglienza. Un corpus di testi amministrativi è stato sottoposto dall'Autore a processi di semplificazione utilizzando quattro tecniche di *prompt engineering* (*zero-shot*, *few-shot*, *chain-of-thought*, *tree-of-thought*) e due modelli: GPT-5

e Gemini 2.5 Flash. La valutazione dei dati si basa sull'integrazione di indici di leggibilità e di una valutazione esperta automatizzata, basata sui criteri del *Plain language*. I dati raccolti mostrano che le formule di leggibilità non danno sempre risultati concordi, mentre la valutazione automatizzata sembra individuare nella tecnica *tree-of-thought* la resa più alta. Le differenze fra i due modelli risultano, invece, poco significative. Il contributo mette in evidenza le potenzialità degli LLM e del *prompt engeneering*, aprendo interrogativi e prospettive di lavoro importanti per la validazione su lettori reali.

Vittorio Ganfi procede a un confronto fra sistemi specialistici e LLM generalisti. Nel suo contributo l'Autore presenta una versione aggiornata di SEMPL-IT, piattaforma per l'italiano amministrativo sviluppata dall'unità dell'Università del Molise coordinata da G. Fiorentino, e la confronta con ChatGPT 5.1, Claude Sonnet 4.6 e Gemini 3. Le metriche impiegate comprendono indici di leggibilità, indicatori lessicali e strutturali, *edit distance* e similarità semantica. SEMPL-IT modula la profondità della riscrittura in funzione del destinatario e mostra un rilevante rispetto del contenuto semantico. I modelli generalisti, che incrementano la leggibilità in misura variabile, in alcuni casi, lo fanno a scapito della fedeltà semantica, come emerge in particolare per Gemini 3. Il dato conferma un'esigenza ricorrente nel volume: la necessità di controlli redazionali umani in domini sensibili.

6. Linee comuni e questioni aperte

Dai sette contributi che, nel loro insieme, tracciano un disegno coerente, emergono con chiarezza cinque linee di lavoro.

La prima linea è di tipo metodologico. La complessità del testo giuridico-amministrativo richiede misurazioni multivariate. Esse devono integrare dimensioni lessicali, sintattiche e testuali, ricondotte a un quadro descrittivo unitario. Le formule storiche restano utili come prima approssimazione, ma vanno integrate da batterie di indicatori derivati dal NLP. Il riferimento al Vocabolario di Base di De Mauro continua a mantenere un ruolo centrale e fornisce tuttora un importante parametro di accessibilità lessicale.

La seconda linea è di natura multidisciplinare. Nessun intervento di semplificazione efficace è oggi pensabile senza una collaborazione strutturale tra linguisti, giuristi, informatici, comunicatori pubblici. Le esperienze documentate da Romano e dai progetti del CNR confermano la produttività di questo modello. Anche il lavoro di Lacroce muove da un consorzio interdisciplinare, dove giurisprudenza e linguistica computazionale dialogano fattivamente.

La terza linea è squisitamente etico-politica. La complessità linguistica spesso è sottovalutata perché ritenuta erroneamente un dettaglio formale. Essa è, invece, una vera e propria barriera all'esercizio della cittadinanza, particolarmente penalizzante per i soggetti in condizione di vulnerabilità. La nozione contestuale di vulnerabilità proposta da Lacroce orienta verso la generalità dei destinatari. Non si tratta di tutelare

categorie chiuse, ma di garantire l'accessibilità a chiunque possa trovarvisi, anche solo per un tempo limitato. La «lotta impari» evocata da Massimo Vedovelli, ripresa da Miglietta in questo volume, è una contesa tra la tortuosità dei testi scritti e le capacità di lettura e comprensione dei lettori. La sfida del nostro tempo è renderla finalmente un confronto leale tra le necessità di ricezione e comprensione dei destinatari e le necessità di produzione di chi scrive, in cui riteniamo di poter includere. a pieno titolo, anche la capacità di farsi capire.

La quarta linea è tecnologica. Gli LLM aprono possibilità reali di scalabilità, fino a ieri inimmaginabili. Rendono praticabile la riscrittura di interi corpora documentali in tempi compatibili con le esigenze amministrative. Tuttavia, richiedono cautele specifiche, ben evidenziate sia da Fortis che da Ganfi. Le riscritture automatiche devono essere validate, soprattutto in domini dove ogni slittamento semantico produce effetti giuridici. La revisione umana resta quindi indispensabile, soprattutto nei testi con funzione prescrittiva. Le metriche di similarità semantica e di *edit distance* entrano a far parte degli strumenti di valutazione.

La quinta e ultima linea è formativa. La diffusione della cultura della chiarezza richiede percorsi di formazione lunghi e di vario tipo. Destinatari di formazione specifica sono (o, meglio, dovrebbero essere) innanzitutto gli estensori dei testi, considerando anche la crescente quantità e complessità delle richieste sociali alle quali devono fare quotidianamente fronte i cittadini. Infatti, Romano dedica un'attenzione particolare alle attività didattiche del CNR per la scuola. Si tratta di un investimento di lungo periodo, decisivo per ribaltare prassi sedimentate e difficili da scalfire. È in questa dimensione formativa che la lezione di De Mauro sulla scuola, sull'educazione linguistica, conservano la loro centralità.

7. Una semplificazione complessa

I lavori raccolti nel volume hanno il merito di chiamare in causa una grande tensione produttiva. Già il titolo del Convegno nel quale sono stati presentati e discussi, "Linguaggio, complessità e comprensione", individua un nodo concettuale rilevante: semplificare non è banalizzare, come ingenuamente molti credono. Non significa, infatti, né appiattire il dettato normativo né eliminare precisione tecnica indispensabile al diritto. Significa, piuttosto, organizzare o riorganizzare il testo in funzione del lettore, distinguendo i piani, esplicitando i nessi, tendendo sotto controllo stretto gli automatismi e le pigrizie del burocratese, segnalando con chiarezza la gerarchia delle informazioni.

La complessità del linguaggio giuridico non si oppone alla chiarezza. La chiarezza è una forma raffinata di gestione della complessità. Essa presuppone scelte deliberate sul lessico, sulla struttura sintattica, sull'architettura testuale. I lavori di riscrittura fatta nei decenni scorsi da alcuni linguisti e giuristi italiani (Piemontese, Cortelazzo, Fioritto) hanno mostrato, per via empirica, anche la sua praticabilità, garantita dalle più avanzate

teorie della lingua e della comunicazione. La via empirica è quella che il presente volume rilancia, con strumenti non solo più aggiornati, ma anche più potenti e veloci.

Nei suoi numerosi interventi sulla necessità di maggiore trasparenza nella lingua usata da chi comunica con il pubblico (legislatori, politici, amministratori pubblici, sindacalisti, giornalisti, insegnanti ecc.) De Mauro parla di "responsabilità" tecnica ed etica. In questa prospettiva si capisce perché la chiarezza e la comprensibilità dei testi pubblici non sia riducibile a un'operazione di maquillage estetico né a un atto di pura filantropia. Si tratta di garantire a tutti i cittadini il diritto a capire per una cittadinanza attiva, come prevede la nostra Costituzione all'articolo 3 ("senza distinzione di lingua"), all'articolo 6 (tutela delle "minoranze" parlanti altre lingue) e all'articolo 21 ("libertà di espressione" inteso in senso esteso).

Il presente volume sembra raccogliere e rilanciare quell'eredità. E lo fa con gli strumenti del nostro tempo, dai corpora annotati ai modelli neurali generativi, ma preservando l'istanza civile fondante.

Ai lettori queste pagine offrono un bilancio e una prospettiva. Un bilancio dello stato dell'arte, articolato lungo le diverse linee della ricerca italiana sul tema. Una prospettiva di lavoro che vede affiancati studiosi di diversa formazione per lo stesso obiettivo: collaborare affinché la chiarezza dei testi istituzionali diventi un progetto comune e una realtà condivisa. Lo studio della leggibilità e della comprensibilità diventa terra di confine, grazie alla cresciuta consapevolezza della centralità del linguaggio nella vita civile e democratica. Non è obiettivo da sottovalutare o dare per scontato che molti accademici e ricercatori, linguisti, giuristi e informatici abbiano deciso di "varcare le frontiere", superando le separatezze disciplinari e metodologiche, ancora meno comprensibili nel cambiamento epocale che stiamo attraversando, per mettere in comune e a disposizione di tutta la società i risultati delle loro ricerche scientifiche e dei loro interventi sul campo.

Concludo, riprendendo le parole finali dell'Introduzione di Tullio De Mauro agli Atti dell'Incontro di studio "Leggibilità e comprensione" tenutosi il 26 e 27 giugno 1986, a Villa Mirafiori, a Roma (De Mauro, Piemontese, Vedovelli 1986). Ci tengo a ricordarlo perché ad esso parteciparono, arrivando da varie parti d'Italia, linguisti, psicolinguisti, psicopedagogisti, ingegneri, informatici e giuristi, filosofi del linguaggio, sociolinguisti e soci della Cooperativa "Spazio linguistico". Dopo quarant'anni esatti, credo di poter affermare che la speranza di De Mauro si sta avverando e, forse, si è già, almeno in parte, avverata:

> Credo che il nostro lavoro teorico, empirico, sperimentale di misurazione, segmentato specialisticamente, possa rientrare in un movimento teorico molto più complesso che potrebbe, se attecchisse anche in Italia – ma non è che siamo particolarmente indietro, da questo punto di vista, rispetto ad altre culture – aiutare, in Italia, a porre, dal versante della cultura scientificamente più avvertita, i problemi sociali della comunicazione e della comprensione ("Linguaggi", III/3, 1986: 5-7).

Bibliografia minima

Cassese, Sabino. "Prefazione." In *Codice di stile*, Roma: Istituto Poligrafico e Zecca dello Stato, 1993: 9.

Cassese, Sabino. "Prefazione." In *Il dovere costituzionale di farsi capire: A trent'anni dal "Codice di stile"*, a cura di M. Emanuela Piemontese. Roma: Carocci, 2023: 7–10.

Codice di stile delle comunicazioni ad uso delle amministrazioni pubbliche. Proposta e materiale di studio. Presidenza del Consiglio dei ministri. Dipartimento per la Funzione pubblica. Roma: Istituto Poligrafico e Zecca dello Stato. 1993.

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

De Mauro, Tullio. "Appunti e spunti in tema di incomprensione." In *Linguaggi* 2, n. 3. 1985: 22–31.

De Mauro, Tullio. "Introduzione." In *Leggibilità e comprensione*. Roma: Coop Spazio Linguistico, 1986: 5–7.

De Mauro, Tullio. *Capire le parole*. Roma-Bari: Laterza, 1994.

De Mauro, Tullio. *Guida all'uso delle parole*. Roma-Bari: Laterza, 2019. Prima edizione 1980.

De Mauro, Tullio, Stefano Gensini e M. Emanuela Piemontese, a cura di. *Dalla parte del ricevente: Percezione, comprensione, interpretazione. Atti del XIX Congresso internazionale di studi della SLI (Roma, 8–10 novembre 1985)*. Roma: Bulzoni, 1988.

De Mauro, Tullio, M. Emanuela Piemontese e Massimo Vedovelli. *Leggibilità e comprensione: Atti dell'incontro di studio, Roma, Istituto di Filosofia–Villa Mirafiori, 26–27 giugno 1986*. *Linguaggi* 3, n. 3. Roma: Coop Spazio Linguistico, 1986.

Fioritto, Alfredo, a cura di. *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Bologna: Il Mulino, 1997.

GISCEL. «Dieci tesi per l'educazione linguistica democratica». 1975. (https://giscel.it/dieci-tesi-per-leducazione-linguistica-democratica/).

Lucisano, Pietro, e M. Emanuela Piemontese. "Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana." In *Scuola e Città* 39, n. 3, 1988: 110–124.

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

Mortara Garavelli, Bice. *Le parole e la giustizia*. Torino: Einaudi, 2001.

Piemontese, Maria Emanuela. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli: Tecnodid, 1996.

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

Leggibilità e semplificazione: Natural Language Processing per una lingua del diritto chiara

Giulia Venturi
(Istituto di Linguistica Computazionale "A. Zampolli" - CNR-ILC, ItaliaNLP Lab)

Abstract

La complessità linguistica dei testi giuridici rappresenta una sfida rilevante per la comprensione e la semplificazione della lingua del diritto. Partendo da questa premessa, il contributo presenta alcune proposte metodologiche e applicative per la valutazione automatica della leggibilità e la semplificazione della lingua del diritto basate su strumenti di Natural Language Processing. In particolare, saranno presentati i risultati di tre linee di ricerca sviluppate dall'ItaliaNLP Lab (http://www.italianlp.it/) dell'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa. La discussione prenderà le mosse dalla presentazione di un metodo di profilazione linguistica che consente di estrarre, da corpora linguisticamente annotati, una vasta gamma di caratteristiche lessicali, morfo-sintattiche e sintattiche correlate alla complessità linguistica del testo. Verrà poi mostrato come tali caratteristiche siano impiegate come indicatori di complessità da READ-IT (http://www.italianlp.it/demo/read-it/), il primo e unico strumento di valutazione automatica della leggibilità del testo basato su metodi e strumenti di Natural Language Processing, ideato per fornire un supporto per la semplificazione guidata dei testi. Nella terza parte del contributo sarà infine illustrato come i valori di leggibilità di READ-IT siano impiegati per controllare la semplificazione automatica del testo realizzata dai più recenti Large Language Models, consentendo di calibrarne la riscrittura rispetto ai livelli desiderati di leggibilità.

1. Introduzione

La complessità linguistica è una delle principali cause di oscurità dei testi giuridici e burocratico-amministrativi, spesso criticati da più parti per la loro scarsa leggibilità e comprensibilità. La questione è ben nota a giuristi, linguisti e comunicatori pubblici che a livello nazionale e internazionale, pur da prospettive diverse, hanno nel corso degli anni proposto numerose analisi dei molteplici fattori che contribuiscono a rendere oscura la lingua del diritto.

Nel panorama italiano, da un lato, i giuristi affermano che «i problemi di significato degli enunciati giuridici possono essere affrontati solo risolvendone i problemi sintattici», problemi di significato tra cui «spiccano […] le difficoltà di accertamento della struttura sintattica degli enunciati, per esempio di comprensione del significato della

punteggiatura o del particolare ordine delle parole in enunciati spesso assai complessi e spesso poco curati da questo punto di vista 'formale' (che in questo caso vuol dire appunto sintattico-grammaticale)» (Jori e Pintore 1995, 209). Dall'altro lato, nell'ambito degli studi linguistici, è stata proposta nel tempo una ricca serie di ricerche finalizzate a delineare il profilo della lingua del diritto (Mortara Garavelli 2001; Fortis 2005; Lubello 2017) e, in particolare, a individuare i tratti lessicali, sintattici e testuali che la rendono complessa per il cittadino (Lubello 2014; Cortelazzo 2021). In un'ottica di democrazia linguistica, tali studi hanno dato origine a numerose iniziative, sia accademiche sia istituzionali, volte alla stesura di manuali e codici di regole e suggerimenti per la redazione di atti normativi e amministrativi più chiari e accessibili (Piemontese 2023).

La tematica risulta tanto più rilevante se si considera che, come osserva Mortara Garavelli (2001, 24), «ciò che si intende per linguaggio giuridico (o legale) non è dedicato esclusivamente all'esame di questioni di diritto; anzi, una parte cospicua dei testi giuridici che hanno effetti applicativi è occupata dall'esame di eventi e situazioni concrete senza il cui accertamento non è possibile individuare le norme applicabili. Il giudice che discute sulla validità delle prove offertegli, il notaio che descrive l'immobile compravenduto, il presidente della regione che stipula un contratto di appalto, i privati che redigono un contratto parlano o scrivono per esporre situazioni di fatto e dati concreti senza i quali gli effetti giuridici sarebbero inesistenti o viziati».

Fa parte delle iniziative volte a promuovere una maggiore cultura dell'accessibilità della lingua del diritto anche il filone di studi che negli anni ha definito metodi per descrivere la complessità linguistica come una proprietà misurabile del testo. Iniziato negli anni '90 in ambito anglosassone con il movimento del *Plain Language*, più recentemente, questo filone di ricerca ha trovato un terreno di sviluppo particolarmente fertile nell'ambito della Linguistica Computazionale e del *Natural Language Processing* (NLP), dove la complessità linguistica viene affrontata non solo come oggetto di analisi descrittiva, ma anche come dimensione operativa, rilevante per la valutazione automatica della leggibilità (Collins-Thompson 2014) e per la progettazione di sistemi di semplificazione del testo (Saggion 2017).

In Italia, il rapporto tra Linguistica Computazionale e studio della lingua del diritto ha radici storiche profonde, che risalgono all'incontro tra Beniamino Placido e Antonio Zampolli, primo direttore dell'Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche (CNR), oggi a lui intitolato. Così ricorda Zampolli (2003): «Il 1969 fu anche l'anno del mio primo corso di Linguistica Computazionale (il primo in assoluto in Italia). Tra i frequentatori del corso, l'allora direttore dell'ufficio di elaborazione elettronica della Camera dei Deputati, Beniamino Placido, si convinse dell'opportunità di arricchire le procedure di *information retrieval* sulle leggi al servizio dei parlamentari, con un "dizionario di macchina" che permettesse di "proiettare" in modo più efficace i termini delle domande su quelli dei testi». Questo incontro ha contribuito a finanziare le prime borse di studio dei ricercatori dell'allora Divisione Linguistica del Centro Nazionale Universitario di Calcolo Elettronico (CNUCE), poi passato al CNR, e a porre le basi per una linea di ricerca che ha visto la lingua del

diritto come un banco di prova privilegiato per lo sviluppo di metodi e strumenti linguistico-computazionali.

All'interno di questo contesto di ricerca si inseriscono le attività dell'Italian Natural Language Processing Lab (ItaliaNLP Lab)[1] dell'Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), che, all'interno delle proprie linee di ricerca, ha sviluppato nel tempo un ampio insieme di risorse, modelli e metodologie basate su NLP per l'analisi di testi giuridici e giuridico-amministrativi. Tali attività spaziano dalla modellazione della conoscenza giuridica e costruzione di ontologie di dominio, allo sviluppo di corpora e strumenti di NLP specializzati per l'analisi del dominio legale, fino alla profilazione delle diverse varietà di lingua del diritto e allo studio della complessità linguistica di testi giuridici e burocratico-amministrativi.

Il presente contributo si colloca in questo quadro e si propone di mostrare come la profilazione linguistica e la valutazione automatica della leggibilità possono costituire strumenti centrali per affrontare il problema della complessità della lingua del diritto, non solo in chiave descrittiva, ma anche applicativa. In particolare, l'articolo illustra le seguenti linee di ricerca sviluppate dai ricercatori dell'ItaliaNLP Lab che hanno portato alla definizione di:

- una metodologia innovativa di profilazione linguistica multilivello del testo basata su metodi di stilometria computazionale (*Computational Stylometry*);
- un metodo di analisi della complessità linguistica del testo che ha permesso lo sviluppo del primo e, a oggi, unico strumento di valutazione automatica della leggibilità del testo basato su parametri di profilazione linguistica automatica;
- un approccio alla semplificazione automatica del testo, basato su approcci di generazione controllata che permette di controllare a priori il livello di leggibilità del testo generato (*Readability-Controlled Text Simplification*).

Illustrando i risultati di una serie di studi condotti dai ricercatori del laboratorio, la discussione dei paragrafi seguenti intende mostrare come tali linee di ricerca possono essere impiegate con successo per:

- lo studio delle specificità di diverse varietà riconducibili alla macro-varietà della lingua del diritto;
- la semplificazione guidata di documenti amministrativi, condotta sulla base della valutazione automatica della leggibilità del testo;
- l'impiego di metodi di *Readability-Controlled Text Simplification* per la semplificazione automatica di documenti amministrativi.

2. Profilazione linguistica e complessità del testo

Il processo che permette di acquisire informazione linguistica in modo automatico a partire unicamente dalla struttura linguistica sottostante un testo è noto con il nome

1 http://www.italianlp.it/

di 'profilazione linguistica' (*linguistic profiling*). Proposto per la prima volta nel 2004 da Hans van Halteren in un articolo pubblicato in occasione della 42esima edizione della conferenza annuale dell'*Association for Computational Linguistics*, esso consiste nella rappresentazione del testo sotto forma di un vettore, detto 'profile vector', i cui elementi sono costituiti da caratteristiche linguistiche relative a molteplici livelli di descrizione. Una volta rappresentato il testo sotto forma di distribuzione statistica di un elevato numero di caratteristiche linguistiche, è possibile quantificare differenze e somiglianze tra testi in termini di vicinanza o distanza tra i rispettivi 'profile vectors' che li rappresentano[2]. Sebbene originariamente ideata per l'attribuzione di testi apocrifi o per la verifica dell'autorialità di un testo anonimo, la profilazione linguistica si è rivelata efficace in numerosi altri scenari applicativi, tra cui lo studio della variazione tra generi e registri testuali, per studi sociolinguistici o ancora per indagini stilometriche.

Dal punto di vista strettamente linguistico, questo approccio è alla base della *Multi-Dimensional Analysis* (MDA), proposta da Douglas Biber, il quale ha definito una metodologia applicativa per la categorizzazione delle varietà linguistiche alternativa a una prospettiva puramente funzionale. Secondo la MDA, le varietà linguistiche possono essere descritte in termini di distribuzioni statistiche di caratteristiche che modellano fenomeni lessicali, grammaticali e sintattico-semantici e che, nel loro insieme, contribuiscono a caratterizzare la variazione linguistica tra testi (Biber 1993)[3]. L'analisi di corpora rappresentativi di una certa varietà linguistica richiede pertanto di quantificare la frequenza con cui determinati tratti ricorrono, sulla base dell'intuizione che le differenze tra varietà emergono a livello distribuzionale: alcune caratteristiche possono essere comuni e pervasive in certe varietà e relativamente rare in altre.

In questo panorama, la disponibilità di corpora di grandi dimensioni rappresentativi dell'uso reale della lingua, unitamente allo sviluppo di strumenti di NLP sempre più affidabili, in grado di rappresentare in modo esplicito la struttura linguistica del testo (annotazione linguistica), ha favorito la diffusione della profilazione linguistica automatica come metodologia per affrontare temi tradizionali della ricerca linguistica. In particolare, un ulteriore impulso è stato dato dal progetto *Universal Dependencies* (UD)[4], iniziativa internazionale avviata nel 2013, e tuttora in corso, finalizzata alla costruzione di corpora annotati a livello morfo-sintattico e sintattico (*treebank*) sulla base

2 «In linguistic profiling, the occurrences in a text are counted of a large number of linguistic features, either individual items or combinations of items. These counts are then normalized for text length and it is determined to what extent (calculated on the basis of the number of standard deviations) they differ from the mean observed in a profile reference corpus. For each text, the deviation scores are combined into a profile vector, on which a variety of distance measures can be used to position the text relative to any group of other texts» (van Halteren 2004).

3 «[...] linguistic features from all levels function together as underlying dimensions of variation, with each dimension defining a different set of linguistic relations among registers» (Biber 1993).

4 https://universaldependencies.org/

di uno schema di annotazione condiviso (de Marneffe et al. 2021). L'idea alla base dell'iniziativa è quella di avere un unico repertorio di categorie morfosintattiche e di relazioni di dipendenza sintattica tra parole in grado di catturare sia le somiglianze sia le specificità tipologiche tra lingue anche profondamente diverse tra di loro (Nivre 2015). Attualmente il progetto include 339 *treebank* per 186 lingue diverse, con casi di lingue per le quali è disponibile più di una *treebank*, ognuna rappresentativa di una varietà testuale diversa[5]. L'annotazione del testo secondo il formalismo di rappresentazione introdotto dal progetto adotta una visione lessicalista della sintassi. In tale rappresentazione, una parola costituisce la testa sintattica dell'intera frase (denominata ROOT), mentre tutte le altre parole dipendono da un'altra parola all'interno della frase (vedi Figura 1).

In questo contesto di ricerca internazionale si colloca il metodo di profilazione linguistica sviluppato dai ricercatori dell'ItaliaNLP Lab, che prende le mosse dallo studio di Montemagni (2013), il quale ne ha posto le basi teoriche e metodologiche. Come illustrato nel contributo, esso si basa su due ingredienti principali: la disponibilità di corpora di grandi dimensioni, annotati automaticamente, rappresentativi di specifiche varietà linguistiche, e un repertorio di caratteristiche linguistiche estratte da diversi livelli di annotazione linguistica del testo. Successivamente il metodo è stato implementato dall'ItaliaNLP Lab in un'applicazione web chiamata Profiling-UD[6] (Brunato et al. 2020). L'applicazione permette, in una prima fase, di annotare linguisticamente un testo, o una collezione di testi, secondo il formalismo UD e, in una seconda fase, di estrarne un'ampia gamma di caratteristiche linguistiche, riferite ai diversi livelli di descrizione linguistica. La scelta di basare Profiling-UD sul formalismo UD rende lo strumento multilingue.

Le caratteristiche profilate da Profiling-UD. L'insieme di caratteristiche linguistiche estratte da Profiling-UD permette di modellare molteplici aspetti che contribuiscono a costruire il profilo linguistico di un testo, aspetti riconducibili a nove macro-gruppi, riassunti nella Tabella 1.

Gruppo	Caratteristica linguistica
Caratteristiche di base del testo (gruppo 1)	• Lunghezza del documento, calcolata in numero di *token* (parole ortografiche, inclusa la punteggiatura) e di frasi; • Lunghezza media della frase (in *token*); • Lunghezza media della parola (in caratteri)
Ricchezza lessicale (gruppo 2)	• Rapporto Tipo/Unità (*Type/Token Ratio*) per forme e lemmi sui primi 100 e 200 *token*
Informazione morfo-sintattica (gruppo 3)	• Distribuzione delle categorie morfo-sintattiche UD; • Densità lessicale

5 L'aggiornamento è semestrale e l'ultimo risale al 15 novembre 2025.

6 http://www.italianlp.it/demo/profiling-UD/

Morfologia verbale (gruppo 4)	• Distribuzione del tempo, modo, forma, persona e numero di verbi e ausiliari
Struttura del predicato (gruppo 5)	• Distribuzione dei verbi che svolgono il ruolo di testa sintattica all'interno di una frase; • Distribuzione di radici sintattiche della frase (ROOT) categorizzate come verbi; • Valenza verbale media (include argomenti e modificatori); • Distribuzione dei verbi per valenza
Caratteristiche relative alla struttura sintattica globale e locale di una frase (gruppo 6)	• Profondità media dell'albero sintattico; • Lunghezza media delle relazioni di dipendenza sintattica; • Lunghezza media della più lunga relazione di dipendenza per frase; • Lunghezza media di sequenze consecutive di complementi preposizionali dipendenti da un nome; • Distribuzione delle sequenze per lunghezza; • Lunghezza media delle clausole
Relazioni di dipendenza sintattica (gruppo 7)	• Distribuzione delle relazioni di dipendenza UD
Ordine degli elementi (gruppo 8)	• Distribuzione di soggetti e complementi oggetto in posizione pre e post verbale
Uso della subordinazione (gruppo 9)	• Distribuzione di frasi principali e subordinate; • Lunghezza media di sequenze consecutive di subordinate; • Distribuzione delle sequenze per lunghezza; • Distribuzione delle subordinate che precedono e seguono la frase principale

Tabella 1: Caratteristiche linguistiche estratte da Profiling-UD.

Come è possibile notare, molte delle caratteristiche contribuiscono non solo alla profilazione linguistica del testo, ma permettono anche di delinearne il livello di complessità linguistica. È il caso, ad esempio, delle caratteristiche incluse nel gruppo 1, che raccoglie tratti tradizionalmente considerati indicatori di base, ma al tempo stesso affidabili, di complessità testuale, quali la lunghezza della frase e della parola. La presenza di frasi e parole lunghe è stata infatti impiegata fin dai primi indici di leggibilità, sia per la lingua italiana (Lucisano e Piemontese 1988) sia in ambito internazionale (Kincaid et al. 1975), come approssimazione della complessità sintattica e lessicale del testo.

Per quanto riguarda gli aspetti di profilazione lessicale, numerosi studi hanno evidenziato come un'elevata densità lessicale, calcolata da Profiling-UD come rapporto tra parole piene (nomi, aggettivi, verbi e avverbi) e numero totale di parole, sia associata a un aumento del carico cognitivo per il lettore (Gibson 1993). Analogamente, valori elevati del rapporto Tipo/Unità sono stati inclusi tra i tratti rilevanti nei più recenti indici di leggibilità (Aluisio 2010), in quanto indicativi della quantità di concetti nuovi introdotti in un testo, con un potenziale effetto di appesantimento del carico cognitivo. Di fatto, valori del rapporto prossimi allo zero caratterizzano testi con un vocabolario

meno vario, mentre valori prossimi a uno sono tipici di testi particolarmente ricchi e diversificati dal punto di vista lessicale.

Allo studio del livello di complessità del testo analizzato con Profiling-UD contribuiscono aspetti più locali della struttura sintattica, quali quelli legati all'organizzazione del predicato, sia esso verbale o nominale. È in questa prospettiva che si collocano le caratteristiche incluse nel gruppo 5. Da un lato, sebbene non esista una relazione diretta tra il numero di dipendenti di un verbo e la complessità complessiva del testo, valori estremi della valenza verbale, calcolata come il numero effettivo di dipendenti di una testa verbale (siano essi argomenti o modificatori), possono incidere negativamente sulla leggibilità. Valori particolarmente bassi, come nel caso di costruzioni ellittiche, o particolarmente elevati, come nel caso di verbi con ampio numero di dipendenti, possono infatti aumentare il carico interpretativo richiesto al lettore. Simile è l'interpretazione dei valori relativi alla distribuzione delle radici sintattiche della frase (ROOT). Un basso numero di frasi il cui *token* radice è categorizzato come verbo è indicativo di una maggiore presenza di frasi nominali, caratterizzate da strutture meno canoniche dovute all'ellissi del verbo, che possono dar luogo a una maggiore ambiguità interpretativa in fase di produzione o comprensione.

Ulteriori fattori di complessità locale emergono dall'analisi delle relazioni di dipendenza sintattica (gruppo 7). Studi sperimentali hanno infatti mostrato che alcune tipologie di dipendenza risultano più onerose dal punto di vista del *processing* cognitivo, come nel caso delle frasi relative o di complementi preposizionali, esplicitamente codificati nel formalismo UD (Gibson e Pearlmutter 1994; Gibson 2000). In questa prospettiva si collocano anche le caratteristiche del gruppo 8, volte a profilare il testo rispetto all'ordine di alcuni elementi all'interno della frase. In particolare, la presenza di soggetti e complementi oggetto in posizione non canonica rispetto al verbo da cui dipendono può essere interpretata come un fattore di complessità linguistica, in linea con una delle definizioni di marcatezza proposte da Haspelmath (2006), ovvero la «*markedness as abnormality*». Assumendo che, pur trattandosi di una lingua a costituenti relativamente liberi, l'italiano presenti un ordine preferenziale Soggetto-Verbo-Oggetto, una maggiore frequenza di ordini marcati può comportare un aumento delle difficoltà di elaborazione per il lettore.

Tuttavia, come emerge dalla Tabella 1, il gruppo più articolato è rappresentato dalle caratteristiche del gruppo 6, che modellano aspetti sia globali sia locali della frase. Esso include, in primo luogo, la profondità dell'albero sintattico che rappresenta ogni frase di un testo (calcolata come il percorso più lungo, in termini di relazioni di dipendenza consecutive, dalla ROOT all'ultima foglia dell'albero) e la lunghezza delle relazioni di dipendenza (misurata come la distanza in *token* tra testa e dipendente, escludendo la punteggiatura). Entrambe le misure sono ampiamente riconosciute nella letteratura psicolinguistica come indicatori di complessità sintattica (Frazier 1985; Gibson 2000; Demberg e Keller 2008). A queste si affiancano ulteriori tratti strutturali, quali la presenza di sequenze consecutive di complementi preposizionali dipendenti da un nome, che contribuiscono alla pesantezza dei costrutti nominali complessi (Montemagni

2013). Sempre in quest'ottica si colloca la lunghezza media delle clausole, calcolata come rapporto tra il numero di *token* in una frase e il numero di teste verbali. Tale misura è impiegata in metriche consolidate per la valutazione della complessità sintattica, come la T-unit (Hunt 1966), e valori elevati sono generalmente associati a una maggiore difficoltà di comprensione per il lettore.

Un caso specifico è infine rappresentato dall'uso della subordinazione (gruppo 9), il cui impiego estensivo è ampiamente riconosciuto come un fattore di complessità strutturale e, in quanto tale, incluso tra le caratteristiche considerate dai più recenti indici di leggibilità (Aluisio 2010). Oltre alla semplice distribuzione di frasi principali e subordinate, Profiling-UD consente di analizzare la posizione delle subordinate rispetto alla frase principale da cui dipendono. Tale distinzione è rilevante poiché studi sperimentali hanno dimostrato che frasi contenenti subordinate posposte alla principale risultano, in media, più facili da comprendere rispetto a quelle con subordinate anteposte (Miller e Weinert 1998).

Per illustrare come sia possibile ricostruire il profilo linguistico di una frase a partire dalle caratteristiche estratte da Profiling-UD, in quanto segue discutiamo alcune delle caratteristiche che è possibile individuare a partire dall'annotazione della frase "Gli Stati membri comunicano alla Commissione il testo di tali disposizioni", graficamente rappresentata nella Figura 1.

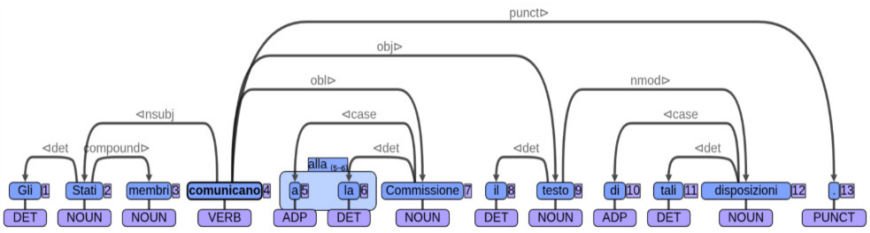

Figura 1: Esempio di annotazione linguistica secondo il formalismo UD.

Partendo dalle caratteristiche di base del testo, Profiling-UD permette di calcolare che la frase è lunga 13 *token*, includendo anche il punto fermo finale. È importante sottolineare che le unità di base dell'annotazione secondo il formalismo UD sono le parole sintattiche, il che implica che le parole morfologicamente complesse, come le preposizioni articolate, vengono scomposte in più unità. Come si può notare, la preposizione articolata 'alla' è divisa in due parole sintattiche: la preposizione (ADP) 'a' e l'articolo (DET) 'la', ciascuna annotata con la rispettiva relazione di dipendenza, ovvero *case* e *det*, entrambe collegate alla medesima testa sintattica 'Commissione'. La radice sintattica (ROOT) è il verbo 'comunicano'[7] che ha valenza 3, perché ha tre

7 Sebbene la rappresentazione grafica non permetta di visualizzarlo, a ogni *token*, quando presenti, sono associati i relativi tratti morfologici. Ad esempio, il verbo 'comunicano' è annotato come verbo di modo indicativo, tempo presente, forma finita, terza persona plurale.

dipendenti, 'Stati', 'Commissione' e 'testo', ciascuno dei quali legato al verbo da una relazione di dipendenza diversa, corrispondente a una specifica funzione sintattica: *nsubj* per il soggetto nominale 'Stati' (in posizione pre-verbale), *obl* per il complemento obliquo 'Commissione', e *obj* per l'oggetto diretto 'testo' (in posizione post-verbale).

Passando ora alle caratteristiche che modellano aspetti globali e locali della struttura sintattica, la profondità dell'albero sintattico è pari a 3, corrispondente ai tre passaggi intermedi (le relazioni di dipendenza) attraversati nel percorso che va dalla ROOT ('comunicano') alle foglie equidistanti, le parole 'di' e 'tali'. La frase ha inoltre 11 relazioni di dipendenza[8]: cinque relazioni hanno lunghezza un *token* (['Gli', '**Stati**'], ['**Stati**', 'membri'], ['la', '**Commissione**'], ['il', '**testo**'], ['tali', '**disposizioni**']); tre relazioni due *token* (['Stati', '**comunicano**'], ['a', '**Commissione**'], ['di', '**disposizioni**']); due relazioni hanno lunghezza tre *token* (['**comunicano**', '**Commissione**'], ['**testo**', 'disposizioni']); una relazione cinque *token* (['**comunicano**', 'testo']), essendo così la relazione più lunga della frase. Il valore medio della lunghezza delle dipendenze, calcolato come rapporto tra la somma delle lunghezze e il numero totale di relazioni, è pertanto pari a 2 (22/11). Infine, la frase ha una sola sequenza consecutiva di complementi preposizionali dipendenti da un nome, nello specifico da 'testo', di lunghezza 1, corrispondente alla relazione *nmod* che mette in relazione 'testo' con il complemento preposizionale 'di tali disposizioni'.

2.1 Aspetti di complessità linguistica della lingua del diritto

La profilazione linguistica di corpora rappresentativi di diverse varietà della lingua del diritto ha consentito di individuare i principali indicatori di complessità linguistica. In particolare, gli studi condotti dall'ItaliaNLP Lab hanno permesso di studiare come essi varino rispetto a quattro dimensioni principali:

- il genere testuale, studiato attraverso il confronto tra testi normativi e articoli di giornale;
- la fonte istituzionale, mettendo a confronto testi normativi di emanazione nazionale, regionale ed europea;
- la tipologia testuale interna alla lingua del diritto, confrontando testi normativi, amministrativi, la Costituzione italiana e sentenze.

Il genere testuale. La dimensione di variazione viene discussa da Montemagni (2013), Dell'Orletta e colleghi (2013) e Venturi (2013), mettendo a confronto corpora di testi normativi rilasciati da diverse istituzioni (lo stato italiano, la regione Piemonte e l'Unione Europea), e regolative di diverse materie, e una raccolta di articoli di giornale. I risultati degli studi mostrano che i testi normativi contengono frasi significativamente più lunghe rispetto ai testi giornalistici (rappresentativi della lingua italiana standard), confermando uno dei tratti tradizionalmente associati a una maggiore complessità

8 Negli esempi, la testa sintattica è segnalata in grassetto.

della lingua del diritto. A livello morfosintattico, i testi normativi mostrano una maggiore distribuzione di sostantivi rispetto ai verbi, con un conseguente elevato rapporto nomi/verbi, caratteristica associata in letteratura a testi con un'alta densità informativa e con un conseguente aumento del carico cognitivo da parte del destinatario (Voghera 2005). Tra le caratteristiche che confermano una maggiore complessità dei testi normativi, troviamo anche valori di valenza verbale mediamente inferiori in questi testi, accompagnati, da una parte, da una maggiore presenza di verbi mono-argomentali e, dall'altra, da una percentuale inferiore di verbi con un elevato numero di dipendenti. Tali evidenze sono riconducibili ad altre due specificità dei testi normativi: la bassa percentuale di soggetti espliciti (tra le tipologie di relazioni di dipendenza sintattica del formalismo UD) e l'elevata presenza di participi (rintracciati grazie alle informazioni estratte dalla morfologia verbale) che tipicamente non presentano dipendenti. Le differenze risultano ancora più evidenti sul piano della struttura sintattica globale e locale. I testi normativi contengono frasi con alberi sintattici più profondi, relazioni di dipendenza sintattiche più lunghe, e sequenze consecutive di complementi preposizionali più lunghe con una distribuzione minore di sequenze semplici (lunghezza 1) e maggiore di sequenze più lunghe. Di particolare interesse, quest'ultimo risultato sembra confermare quantitativamente quanto osservato da Mortara Garavelli (2001, 175) quando afferma che «sono specialmente i nessi, i grappoli di astrazioni concatenate in "complementi del nome" a marcare sintatticamente (e testualmente) gli enunciati.».

La fonte istituzionale. Il confronto tra testi normativi nazionali e regionali, da una parte, e testi di derivazione europea, dall'altra, (Dell'Orletta et al. 2012; Mori e Venturi 2026), ha permesso di tracciare un profilo della legislazione europea improntato a una maggiore semplicità. Gli studi mostrano che i testi europei presentano valori inferiori rispetto ai testi nazionali e regionali per molte delle caratteristiche associate, nel paragrafo precedente, a fenomeni di maggiore complessità dei testi normativi come genere testuale. In particolare, i testi normativi di emanazione europea contengono frasi mediamente più brevi, un rapporto nomi/verbi più basso, una percentuale inferiore di participi, alberi sintattici meno profondi, relazioni di dipendenza sintattica più brevi, sequenze consecutive di complementi preposizionali dipendenti da un nome meno profonde. Al contrario, si caratterizzano per valori maggiori di valenza verbale, con una percentuale minore di verbi mono-argomentali, e, come messo in luce da Mori e Venturi (2026), una percentuale minore di frasi subordinate, con una netta maggioranza di subordinate postposte alla principale da cui dipendono, caratteristica, come prima ricordato, associata a un minore carico cognitivo per il destinatario. Nel complesso, dunque, i testi normativi europei si presentano come una variante semplice del genere testuale degli atti normativi. Tale risultato può essere interpretato, da un lato, alla luce delle convenzioni redazionali tipiche della prosa normativa europea imposte dal legislatore, dall'altro, come ampiamente discusso da Mori e Venturi (2026), alla luce del fatto che i testi europei in italiano sono una traduzione di un originale inglese, che pur presentando caratteristiche tipiche della lingua del diritto, presenta caratteristiche specifiche di una lingua, l'inglese, per molti versi strutturalmente diversa dall'italiano.

L'ipotesi, empiricamente verificata nello studio, è che nel processo traduttivo rimangano caratteristiche che avvicinano i testi normativi europei più all'originale inglese che non alle caratteristiche della lingua del diritto della normativa nazionale o regionale.

La tipologia testuale. I risultati dello studio di questo livello di variazione sono riportati in modo organico da Venturi (2013), che li ha discussi anche rispetto alla precedente letteratura sul tema. L'analisi comparativa si è focalizzata su quattro tipologie testuali: testi normativi, testi amministrativi, la Costituzione italiana e un corpus di sentenze redatte da varie corti amministrative e tribunali ordinari, la Corte Costituzionale, la Corte di Cassazione Civile e la Corte Europea dei Diritti dell'Uomo. Ne è emerso un chiaro gradiente di complessità tra le diverse tipologie, con due poli opposti. Il polo della massima complessità è costituito dai testi delle sentenze, che mostrano valori comparativamente elevati per molte delle caratteristiche associate a una maggiore complessità linguistica. Dunque, le sentenze contengono frasi mediamente più lunghe, un rapporto nomi/verbi maggiore (dovuto a una percentuale inferiore di verbi), alberi sintattici più profondi, relazioni di dipendenza sintattica più lunghe, e fanno più ricorso alla subordinazione, organizzata in sequenze incassate anche molto profonde. All'estremo opposto si colloca la Costituzione italiana, che si distingue per valori nettamente inferiori delle caratteristiche menzionate. A queste aggiungiamo che gli articoli della Costituzione contengono verbi mediamente caratterizzati da una valenza verbale più alta, in conseguenza della percentuale inferiore di participi e superiore di soggetti espliciti. Il risultato sembra essere una conferma quantitativa di quanto osservato da Tullio De Mauro (2006) riguardo allo «straordinario impegno dei Costituenti», al «non comune impegno linguistico» dei padri costituenti verso la redazione di un testo leggibile e comprensibile.

Nel loro insieme, questi risultati mostrano che la complessità della lingua del diritto non costituisce una proprietà uniforme del dominio giuridico, ma emerge come configurazione variabile di tratti linguistici. In questo contributo ne sono stati descritti i principali, anche alla luce della precedente letteratura sul tema, ma la vasta gamma di caratteristiche estratte da Profiling-UD apre la strada a nuovi studi che potranno prendere in esame parametri diversi e nuove dimensioni di variazione. È interessante inoltre osservare come la natura multidimensionale della complessità della lingua del diritto renda particolarmente prezioso l'impiego di un approccio alla valutazione quantitativa della leggibilità del testo, in grado di integrare tali molteplici parametri di complessità linguistica in una misura sintetica.

3. Valutazione automatica della leggibilità per la semplificazione guidata

Se la complessità della lingua del diritto emerge come fenomeno multidimensionale dall'interazione di molteplici tratti linguistici, è ragionevole ipotizzare che la sua semplificazione passi da una fase di valutazione quantitativa e multilivello della leggibilità del testo, capace di integrare tali parametri. In quest'ottica, nel 2011 l'ItaliaNLP crea

READ-IT (Dell'Orletta et al. 2011), primo, e al momento unico, strumento in grado di calcolare in modo automatico la leggibilità del testo a partire dalla distribuzione di molteplici caratteristiche linguistiche.

L'intuizione è che uno strumento di valutazione quantitativa della leggibilità non abbia una finalità unicamente descrittiva, ma svolga un ruolo operativo nel processo di semplificazione. In questa sezione, intendiamo illustrare come, nato con questa duplice finalità, uno strumento come READ-IT può svolgere un duplice ruolo operativo. Da un lato, può supportare il redattore umano, esperto del dominio, fornendo indicatori oggettivi che orientano la revisione del testo verso livelli di complessità più adeguati al destinatario. Dall'altro, può essere impiegato per guidare i *Large Language Models* (LLMs) in compiti di semplificazione automatica (*Readability-Controlled Text Simplification*), consentendo di riportare la variabilità intrinseca della generazione automatica operata da tali modelli entro uno spazio controllato e controllabile di complessità linguistica.

3.1 READ-IT: valutazione automatica della leggibilità del testo per la lingua italiana

Nato nel 2011, grazie alla consolidata esperienza di CNR-ILC nell'ambito dello sviluppo di metodi, modelli e strumenti di NLP, READ-IT[9] (Dell'Orletta et al. 2011) permette di valutare la leggibilità di un testo in lingua italiana sulla base della distribuzione di molteplici caratteristiche linguistiche associate ad aspetti di complessità linguistica, come evidenziato in precedenti studi linguistici e psicolinguistici. Lo strumento assegna a un testo un punteggio di complessità, che esprime la probabilità di appartenenza del testo alla classe dei testi 'di difficile lettura'. La classificazione è realizzata da un classificatore statistico basato su algoritmi di *Machine Learning*, che opera dopo l'annotazione linguistica automatica del testo e lo confronta con due classi definite in fase di addestramento. Si tratta di classi formate da testi tratti dal corpus "Due Parole" (Piemontese 1996), un giornale scritto con uno stile linguistico volutamente semplice per essere compreso da persone con un basso livello di scolarizzazione o con disabilità cognitive, considerati testi di facile lettura, e dal corpus "La Repubblica", considerati testi di difficile lettura. Il punteggio assegnato riflette pertanto il grado di similarità tra la distribuzione delle caratteristiche linguistiche del testo analizzato e queste due classi. È espresso in percentuale e oscilla tra 0 (minima complessità linguistica, massima leggibilità) e 100 (massima complessità, minima leggibilità).

Basato sul principio che un testo può essere più o meno leggibile rispetto ad un sottoinsieme di caratteristiche linguistiche, READ-IT restituisce quattro diversi punteggi calcolati sulla base di quattro diverse configurazioni di caratteristiche del testo:

- Base: le caratteristiche considerate sono quelle usate nelle misure tradizionali della leggibilità di un testo (ovvero la lunghezza della frase e la lunghezza delle parole);

9 http://www.italianlp.it/demo/read-it/

- Lessicale: modello focalizzato sulle caratteristiche lessicali del testo, ovvero la presenza di parole che appartengono al Vocabolario di Base (De Mauro 2000) e la sua ricchezza lessicale.
- Sintattico: modello basato su informazione di tipo grammaticale, ovvero sulla combinazione di tratti morfosintattici (come, ad esempio, la distribuzione di nomi, verbi, aggettivi) e sintattici del testo (quali la profondità dell'albero sintattico, la lunghezza delle relazioni di dipendenza sintattica, la distribuzione delle subordinate).
- Globale: si tratta di un modello basato sulla combinazione di tutti i tratti considerati dagli altri modelli.

Ampiamente sperimentato su diverse tipologie di testi (Dell'Orletta et al. 2014), READ-IT è stato sino ad oggi utilizzato per valutare l'efficacia comunicativa di testi in diverse tipologie di comunicazione: quella tra insegnante-studente, per fornire un supporto all'insegnante nella personalizzazione di materiali didattici e test di comprensione (Bonifaci et al. 2014); operatore di call center-utente, per fornire un supporto alla redazione dei testi usati nei call center migliorando i processi di comunicazione con l'utente; medico-paziente, per assistere la redazione di consensi informati semplici e leggibili (Venturi, 2015); azienda sanitaria-cittadino, per monitorare la qualità e la trasparenza della comunicazione dei servizi offerti dalle istituzioni sanitarie ai cittadini (De Rosis et al. 2020).

3.2 READ-IT per la semplificazione guidata

L'uso di READ-IT come guida alla semplificazione è favorito dalla possibilità offerta dallo strumento di valutare la leggibilità non solo di un intero documento, ma anche delle singole frasi che lo compongono. Questo permette la riscrittura del testo in modo più consapevole e mirato alla revisione delle singole frasi, o porzioni del testo, che risultano meno leggibili perché più complesse rispetto alla distribuzione del lessico (indicazione fornita dal modello lessicale) o delle strutture sintattiche (grazie ai valori del modello sintattico).

Nel panorama delle iniziative volte a introdurre nuovi approcci per incrementare l'accessibilità della lingua del diritto, READ-IT è stato infatti recentemente proposto come strumento operativo all'interno del PRIN "Migrants, Institutions, Translations Easy-read Law"[10] (Miglietta 2025), che si propone di definire nuove strategie per migliorare l'accessibilità dei documenti utilizzati nelle procedure e nei servizi rivolti a cittadini migranti, con particolare attenzione a quelle tipologie testuali che trasmettono informazioni e richiedono una risposta scritta, al fine di individuare modelli di comunicazione istituzionale più efficaci e inclusivi. Uno degli obiettivi del progetto è pertanto quello di creare un corpus di contratti ma anche di moduli e comunicazioni istituzionali chiare, comprensibili e in formato digitale che possano essere considerati dei modelli. A questo scopo, i documenti saranno redatti in italiano e tradotti nelle principali lingue

10 www.easy-read-law.it

di immigrazione in Italia, impiegando i metodi della traduzione interculturale e sulla base di riflessioni teoriche di carattere giuridico e interculturale.

In questo quadro, READ-IT è stato impiegato come strumento di supporto alla scrittura di tali documenti come risultato di un processo di revisione guidata. La metodologia, messa congiuntamente a punto dall'Istituto di Informatica Giuridica e Sistemi Giudiziari del Consiglio Nazionale delle Ricerche (CNR-IGSG) e CNR-ILC, è illustrata in dettaglio da Romano e Venturi (2025) e si articola in tre passaggi:

- un'analisi di tipo quantitativo della complessità linguistica di un corpus di testi, realizzata con READ-IT sia a livello dell'intero documento di partenza sia di ogni singola frase che lo costituisce, con l'obiettivo di individuare testi o porzioni del testo potenzialmente poco comprensibili per il lettore;
- un'analisi di tipo qualitativo, volta a identificare gli elementi tipici del lessico amministrativo che possono ostacolare la comprensione di comunicazioni rivolte ai cittadini, ma anche le caratteristiche sintattiche, testuali e pragmatiche dei testi della pubblica amministrazione che possono rappresentare un potenziale ostacolo alla comunicazione con i cittadini;
- una fase di riscrittura del testo che mira a «modernizzarne» il linguaggio amministrativo (Cortelazzo 2021, 12), ma anche a proporre un testo più leggibile per caratteristiche sintattiche, testuali e paratestuali.

L'aspetto metodologicamente rilevante è la circolarità del processo: al termine della fase di riscrittura, il testo rivisto viene nuovamente valutato con READ-IT, per verificare l'impatto delle strategie di riscrittura. La semplificazione diventa così un'attività controllata e misurabile, nella quale l'intervento umano è guidato dalla distribuzione quantitativa di parametri linguistici.

3.3 READ-IT per il controllo della semplificazione automatica

Se nella semplificazione guidata la riscrittura dell'esperto di dominio viene verificata e orientata attraverso parametri quantitativi di leggibilità, gli stessi parametri possono giocare un ruolo diverso nel contesto della semplificazione automatica del testo. In questo caso, infatti, i punteggi forniti da uno strumento come READ-IT non si limitano a valutare ex-post il risultato prodotto dal modello, ma possono diventare vincoli espliciti ex-ante che orientano il processo generativo stesso, consentendo di controllare la complessità linguistica del testo da generare entro criteri definiti a priori.

L'intuizione si inserisce nei più recenti sviluppi della comunità scientifica internazionale in tema di generazione controllata del testo. Preso atto delle notevoli abilità linguistiche dei *Large Language Models* (LLMs), l'obiettivo comune è quello di mettere a punto dei metodi in grado di indirizzare tali abilità verso la creazione di testi sempre più personalizzati, per diversi aspetti. Tra questi, uno a nostro avviso particolarmente in linea con l'idea di democrazia linguistica è l'aspetto relativo alla personalizzazione del livello di complessità linguistica del testo generato. È questo l'ambito di ricerca che

ha come obiettivo la definizione di metodi di *Readability-Controlled Text Simplification*. Gli studi realizzati in questo contesto affondano le loro radici nella letteratura in tema di semplificazione automatica del testo, che mira a definire metodi automatici di riscrittura del testo, metodi che ne preservino il significato, rendendolo nello stesso tempo più accessibile per uno specifico pubblico di destinatari. Nel corso degli anni, i vari metodi proposti hanno dovuto tener conto dell'evoluzione degli approcci di *Natural Language Processing*. Dai primi metodi basati su regole esplicite, la frontiera è rappresentata oggi dagli approcci alla generazione controllata del testo. All'interno di questo paradigma, i modelli impiegati per la semplificazione automatica vengono controllati aggiungendo, accanto all'input di partenza (tipicamente la frase da riscrivere, più che l'intero testo), specifici token di controllo (detti appunto *control token*) che rappresentano parametri espliciti del livello di difficoltà desiderato (Scarton e Specia 2018). Il controllo può operare a un livello generale, ad esempio specificando un livello di leggibilità target (Barayan et al. 2025), oppure a un livello più dettagliato, indicando singoli parametri linguistici rispetto ai quali semplificare, come ad esempio la lunghezza della frase o il rango delle parole (Martin et al. 2020).

In questo quadro si colloca il metodo proposto dall'ItaliaNLP Lab, descritto nei dettagli da Papucci e colleghi (2026), che introduce per la prima volta l'uso di READ-IT, non solo come strumento di valutazione, ma come vera e propria variabile di controllo del processo generativo. In sintesi, lo studio riporta i risultati di una ricca serie di esperimenti condotti controllando la generazione di due LLMs[11], di dimensioni diverse in termini di parametri e dati di pre-addestramento, uno nativo italiano, l'altro multilingue, rispetto a 20 diversi livelli di READ-IT Globale. In questa fase, i due modelli sono passati attraverso una fase di ulteriore addestramento (*fine-tuning*) impiegando un corpus appositamente costruito a questo scopo dagli autori. Il corpus, chiamato IMPaCTS (*Italian Multilevel Parallel Corpus for Text Simplification*) (Papucci et al. 2025) è composto da 1.444.160 coppie di frasi in versione originale e semplificata, e include per ogni frase originale più di una versione semplificata. Le diverse versioni semplificate di ogni originale differiscono tra di loro per la serie di caratteristiche linguistiche descritte nella sezione 2 di questo contributo e per livelli di READ-IT Globale, come mostra l'esempio riportato nella tabella 2. IMPaCTS contiene due sezioni: una include frasi estratte da Wikipedia, l'altra frasi estratte da atti amministrativi[12]. Da sottolineare che IMPaCTS è stato a sua volta costruito in modo automatico.

11 I modelli considerati sono Minerva (Orlando et al., 2024), nelle versioni 1B(illions), 3B e 7B, e Qwen 3 (Bai et al., 2023), nelle versioni 1,7B, 4B e 8B.

12 Le frasi sono state estratte dal corpus PaWaC (Public Administration Web as Corpus), consultabili alla pagina seguente: https://data.europa.eu/data/datasets/elrc_1282?locale=en

		READ-IT
Frase originale	Stabilito pertanto che le concessioni rilasciate in base al presente bando dovranno intendersi come decennali e non automaticamente rinnovabili.	75
Frase semplificata	Stabilito pertanto che le concessioni rilasciate in base al presente bando dovranno intendersi come decennali e non rinnovabili.	61
	Il presente bando stabilisce che le concessioni devono essere considerate decennali e non automaticamente rinnovabili.	34
	Il presente bando stabilisce che le concessioni devono essere considerate decennali e non rinnovabili automaticamente.	31
	Quindi le concessioni basate sul presente bando devono essere considerate decennali e non automaticamente rinnovabili.	29

Tabella 2: Esempio estratto dalla sezione amministrativa di IMPaCTS, che mostra le diverse semplificazioni per una stessa frase originale, insieme ai relativi punteggi di READ-IT Globale.

Durante questa fase di *fine-tuning*, gli LLM hanno appreso che frasi con determinate caratteristiche sono associate a un certo livello di complessità, espresso dal punteggio di READ-IT. Di conseguenza, la loro abilità di generazione di frasi semplificate, data una frase in input, viene dunque condizionata dalle associazioni apprese. Ad esempio, data la frase *Dai rilievi svolti è emerso che il fabbricato è stato realizzato in una posizione, rispetto ai confini del lotto, leggermente diversa da quella autorizzata con le citate Concessioni*, la richiesta di una semplificazione con valore 35 di READ-IT conduce alla seguente riscrittura: *Il fabbricato è stato realizzato in una posizione leggermente diversa da quella autorizzata.*

È interessante in questa sede far notare che, rispetto alle metriche quantitative utilizzate per valutare la qualità della semplificazione[13], gli esperimenti condotti addestrando i modelli sulla sezione amministrativa di IMPaCTS mostrano risultati migliori che sulla sezione che contiene frasi di Wikipedia. La nostra intuizione, guidata da una dettagliata analisi linguistica di IMPaCTS, è che le trasformazioni delle versioni semplici delle frasi amministrative in IMPaCTS, rispetto alla corrispondente frase originale, siano più marcate che nelle versioni semplici delle frasi originali di Wikipedia. Questo è legato chiaramente al fatto che le frasi amministrative originali erano molto più complesse delle originali di Wikipedia e sono state dunque modificate rispetto a più fattori linguistici. In altri termini, la maggiore distanza strutturale tra originale e semplificato della sezione amministrativa di IMPaCTS ha permesso di fornire agli LLM un segnale di apprendimento più chiaro, aiutandoli pertanto a generalizzare con più efficacia.

13 Per ogni dettaglio si veda Papucci et al., 2026.

Conclusioni

L'intento di questo contributo, lungi dall'essere esaustivo, è stato quello di illustrare alcuni approcci metodologici allo studio della complessità della lingua del diritto, sviluppati da chi, nel corso degli anni, ha orientato la propria ricerca alla definizione di modelli computazionali linguisticamente motivati e alla loro sperimentazione in contesti reali. Tra questi, un ambito particolarmente significativo è quello dello studio di metodi e modelli che guidino la redazione di testi giuridici più accessibili.

Il percorso metodologico proposto è partito dunque dalla presentazione di un metodo di profilazione linguistica del testo, basato sull'annotazione automatica e sulla rappresentazione di corpora come insiemi di tratti morfo-sintattici e sintattici. Tali tratti, ricondotti, sulla base della letteratura linguistica e psicolinguistica, a fenomeni di complessità linguistica, permettono di modellare la complessità in modo multidimensionale e di integrarla in una misura sintetica attraverso strumenti di *Machine Learning* linguisticamente fondati. In questa prospettiva, la profilazione linguistica ha costituito il punto di partenza per lo sviluppo di READ-IT basato su tali principi metodologici.

In questo quadro, READ-IT non va considerato soltanto uno strumento di misurazione, ma un dispositivo di guida alla semplificazione, tanto manuale quanto automatica. Da un lato, orienta la riscrittura, rendendo la semplificazione un processo fondato su parametri espliciti e verificabili, in un ciclo virtuoso di valutazione–riscrittura–nuova valutazione; dall'altro, consente di controllare la generazione automatica, vincolando a priori i *Large Language Models* a livelli definiti di complessità linguistica e riportando la loro variabilità intrinseca entro uno spazio controllato e controllabile.

La conclusione di questo contributo vuole essere un punto di partenza per ulteriori iniziative volte a rendere la lingua del diritto più chiara e accessibile. La nostra convinzione è che semplificare non significhi semplicemente produrre un testo "più semplice", ma esercitare un controllo consapevole sui fattori linguistico-cognitivi che determinano la difficoltà di comprensione. In questa prospettiva, l'integrazione tra profilazione linguistica e valutazione automatica della leggibilità può offrire un contributo concreto, non solo alla comunità scientifica, ma anche a chi svolge una funzione pubblica. Perché, come ricorda Tullio De Mauro, «Chi è al servizio di un pubblico ha il dovere costituzionale di farsi capire.» (Piemontese 2023, 253).

Bibliografia

Aluisio, Sandra, Lucia Specia, Caroline Gasperin, and Carolina Scarton. "Readability Assessment for Text Simplification". In Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, 1–9. 2010.

Bai, Jinze, et al. Qwen Technical Report. 2023.

Barayan, Abdullah, Jose Camacho-Collados, and Fernando Alva-Manchego. "Analysing Zero-Shot Readability-Controlled Sentence Simplification". In Proceedings of the 31st International Conference on Computational Linguistics, 6762–6781. Abu Dhabi: Association for Computational Linguistics, 2025.

Venturi, Giulia, Tommaso Bellandi, Felice Dell'Orletta, Simonetta Montemagni. "NLP-Based Readability Assessment of Health-Related Texts: A Case Study on Italian Informed Consent Forms". In Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi 2015), EMNLP 2015 Workshop, 17 September, Lisbon, Portugal, pp. 131-141, 2015.

Biber, Douglas. "Using Register-Diversified Corpora for General Language Studies". Computational Linguistics 19, no. 2 (1993).

Bonifacci, Paola, Valentina Tobia, Laura Lami, Maggie Snowling. ALCE. Assessment di Lettura e Comprensione per l'Età Evolutiva. Hogrefe, 2014.

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni e Giulia Venturi. "Profiling-UD: A Tool for Linguistic Profiling of Texts". In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, 11–16 May. 2020.

Collins-Thompson, Kevyn. "Computational Assessment of Text Readability: A Survey of Current and Future Research". ITL – International Journal of Applied Linguistics 165, no. 2 (2014): 97–135.

Cortelazzo, Michele. Il linguaggio amministrativo: principi e pratiche di modernizzazione. Roma: Carocci, 2021.

De Mauro, Tullio. Il dizionario della lingua italiana. Torino: Paravia, 2000.

De Mauro, Tullio. "Introduzione. Il linguaggio della Costituzione". In *Costituzione della Repubblica Italiana (1947)*, vii–xxxii. Torino: UTET, 2006.

de Marneffe, Marie-Catherine, Christopher Manning, Joakim Nivre, and Daniel Zeman. "Universal Dependencies". Computational Linguistics 47, no. 2 (2021): 255–308.

Dell'Orletta, Felice, Simonetta Montemagni, and 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 (SLPAT 2011), 73–83. Stroudsburg, PA: Association for Computational Linguistics, 2011.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi. "Assessing Document and Sentence Readability in Less Resourced Languages and across Textual Genres". International Journal of Applied Linguistics 165, no. 2 (2014): 163–193.

Dell'Orletta, Felice, Simone Marchi, Simonetta Montemagni, Barbara Plank, and Giulia Venturi. "The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts". In Proceedings of the 4th Workshop on Semantic Processing of Legal Texts, 42–51. Istanbul, 2012.

Dell'Orletta, Felice, Simone Marchi, Simonetta Montemagni, Giulia Venturi, Tommaso Agnoloni, and Enrico Francesconi. "Domain Adaptation for Dependency Parsing at Evalita 2011". In *Evaluation of Natural Language and Speech Tools for Italian*, edited by Bernardo Magnini et al., 58–69. Berlin and Heidelberg: Springer, 2013.

Demberg, Vera, and Frank Keller. "Data from Eye-Tracking Corpora as Evidence for Theories of Syntactic Processing Complexity". *Cognition* 109, no. 2 (2008): 193–210.

De Rosis, Sabina, Elisa Guidotti, Sara Zuccarino, Giulia Venturi, Francesca Ferrè. "Waiting time information in the Italian NHS: a citizen perspective". *Health Policy* 124(8), 2020: 796–804.

Fortis, Daniele. "Il linguaggio amministrativo italiano". *Revista de Llengua i Dret* 43 (2005): 47–116.

Frazier, Lyn. "Syntactic Complexity". In Natural Language Parsing, edited by D. R. Dowty, Lauri Karttunen, and Arnold M. Zwicky. Cambridge: Cambridge University Press, 1985.

Gibson, Edward. "The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity". In Image, Language and Brain, edited by Alec Marantz and Y. Miyashita. Cambridge, MA: MIT Press, 2000.

Gibson, Edward, and Neal J. Pearlmutter. "A Corpus-Based Analysis of Psycholinguistic Constraints on Prepositional-Phrase Attachment". In *Perspectives on Sentence Processing*, 1994.

Haspelmath, Martin. "Against Markedness (and What to Replace It With)". *Journal of Linguistics* 42, no. 1 (2006): 25–70.

Hunt, Kellogg W. "Recent Measures in Syntactic Development". *Elementary English* 43, no. 7 (1966): 732–739.

Jori, Mario, and Anna Pintore, eds. *Manuale di teoria generale del diritto*. Torino: Giappichelli, 1995.

Kincaid, J. Peter, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of New Readability Formulas for Navy Enlisted Personnel. Research Branch Report 8–75. Tennessee: Chief of Naval Training, 1975.

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

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

Lucisano, Pietro, and Maria Emanuela Piemontese. "Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana". *Scuola e Città* 3 (1988): 57–68.

Martin, Louis, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. "Controllable Sentence Simplification". In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4689–4698. Marseille: ELRA, 2020.

Miglietta, Annarita. "Introduzione editoriale. La comunicazione giuridico-amministrativa interculturale: Un progetto di analisi e di riscrittura". In *Migranti Leggi Contratti: verso la chiarezza*, 1–16. 2025.

Miller, Jim, and Regina Weinert. *Spontaneous Spoken Language: Syntax and Discourse*. Oxford: Clarendon Press, 1998.

Mortara Garavelli, Bice. *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*. Torino, Einaudi, 2001.

Orlando, Riccardo, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. "Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data". In *Proceedings of the 10th*

Italian Conference on Computational Linguistics (CLiC-it 2024), 707–719. Pisa: CEUR Workshop Proceedings, 2024.

Papucci, Michele, Felice Dell'Orletta, and Giulia Venturi. "Controllable Sentence Simplification in Italian: Fine-Tuning Large Language Models on Automatically Generated Resources". In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*. Palma de Maiorca, 2026.

Papucci, Michele, Giulia Venturi, and Felice Dell'Orletta. "Generating and Evaluating Multi-Level Text Simplification: A Case Study on Italian". In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*. Cagliari, 2025.

Piemontese, Maria Emanuela. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli: Tecnodid, 1996.

Piemontese, Maria Emanuela, ed. *Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile*. Roma: Carocci, 2023.

Romano, Francesco, e Giulia Venturi. "Metodi e strumenti per la modernizzazione della lingua delle istituzioni". In *Migranti Leggi Contratti: verso la chiarezza*, 2025: 105–121.

Saggion, Horacio. *Automatic Text Simplification*. San Rafael, CA: Morgan & Claypool, 2017.

Scarton, Carolina, and Lucia Specia. "Learning Simplifications for Specific Target Audiences". In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 712–718. Melbourne: ACL, 2018.

van Halteren, Hans. "Linguistic Profiling for Author Recognition and Verification". In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

Venturi, Giulia. "Investigating Legal Language Peculiarities across Different Types of Italian Legal Texts: An NLP-Based Approach". In *Bridging the Gap(s) between Language and the Law*, 138–156. Porto, 2013.

Voghera, Miriam. "La misura delle categorie sintattiche". In *Parole e numeri. Analisi quantitative dei fatti di lingua*, edito da Isabella Chiari and Tullio De Mauro, 125–138. Roma: Aracne, 2005.

Zampolli, Antonio. "Le principali attività dell'Istituto di Linguistica Computazionale. Il punto di vista del Direttore". In *Computational Linguistics in Pisa – Linguistica Computazionale a Pisa*, VII–LI. Pisa-Roma: Istituti Editoriali e Poligrafici Internazionali, 2003.

Quantificare la complessità. Uno studio sulla complessità linguistica dei testi giuridico-amministrativi in italiano

Alessandro Buttinelli
(Università di Torino)

Abstract

La complessità del linguaggio giuridico-amministrativo rappresenta una barriera che ostacola l'accesso alla giustizia e l'esercizio dei diritti di cittadinanza. Il presente studio vuole misurare quantitativamente la complessità linguistica dei regolamenti comunali TA.RI (Tassa sui Rifiuti), la cui inosservanza comporta oneri e sanzioni per il contribuente. Attraverso tecniche di *Natural Language Processing* (NLP) e statistica inferenziale, un corpus di regolamenti è stato confrontato con un corpus di controllo (Wikipedia) su quattro metriche di complessità: ricchezza lessicale, aderenza al Vocabolario di Base, lunghezza delle frasi e Indice Gulpease. I risultati confermano che i testi amministrativi presentano barriere lessicali e sintattiche significativamente superiori e una marcata cristallizzazione formulare legata a prassi redazionali standardizzate.

1. Introduzione

La chiarezza e l'accessibilità del linguaggio istituzionale rappresentano requisiti fondamentali per il corretto funzionamento di uno Stato di diritto e per la tenuta del patto sociale tra istituzioni e cittadini. Tuttavia, il linguaggio giuridico-amministrativo italiano si distingue per una perdurante e sistematica opacità che sembra resistere a decenni di tentativi di riforma (Cassese 1994; Fioritto 2009; Alfieri et al. 2011; *inter alia*). L'uso di una sintassi involuta, caratterizzata da periodi ipotattici particolarmente lunghi, l'abuso di lessico specialistico o desueto e la tendenza pervasiva alla nominalizzazione (Cortelazzo 2021; Fiorentino e Ganfi 2024) contribuiscono a erigere una barriera tra la Pubblica Amministrazione e la cittadinanza.

Questa inaccessibilità linguistica cessa di essere una questione solo stilistica nel momento in cui impatta direttamente sulla sfera dei diritti del cittadino: un testo complesso genera una profonda asimmetria informativa che impedisce la piena comprensione delle norme, ostacolando l'accesso alla giustizia, ai servizi pubblici e alle tutele fondamentali. Il paradosso è evidente: lo Stato esige il rispetto incondizionato delle

proprie direttive, ma fallisce nel fornire gli strumenti necessari per decodificarle, ponendo il cittadino in una condizione di subalternità e vulnerabilità di fronte all'autorità pubblica.

In questo panorama, assumono particolare rilievo quelle norme prodotte a livello locale, come ad esempio i regolamenti comunali per l'applicazione della TA.RI (Tassa sui Rifiuti). A differenza di normative di settore destinate a platee ristrette di tecnici, la disciplina della TA.RI possiede una portata universale: essa vincola l'intera cittadinanza, interessando indistintamente chiunque possegga o detenga a qualsiasi titolo locali o aree scoperte. Dal punto di vista giuridico, questi regolamenti dettano rigorose prescrizioni procedurali. Essi definiscono, ad esempio, i metodi di calcolo delle tariffe, stabiliscono i requisiti formali e i termini per accedere a riduzioni o esenzioni e, soprattutto, fissano il regime sanzionatorio per i casi di inadempienza. Ne consegue che l'illeggibilità di un regolamento TA.RI può produrre un danno immediato: al cittadino viene di fatto preclusa la possibilità di conformarsi volontariamente alla norma o di far valere i propri legittimi diritti. Questa opacità lo espone al rischio concreto di subire sanzioni amministrative causate non da una deliberata volontà di evasione o elusione fiscale, ma dall'impossibilità materiale di decodificare il testo regolamentare, costringendo l'utente a rivolgersi a intermediari fiscali a pagamento per adempiere a un obbligo civico di base.

Alla luce di ciò, l'obiettivo principale del presente lavoro è misurare quantitativamente la complessità linguistica che caratterizza i testi giuridico-amministrativi in italiano, prendendo i regolamenti comunali TA.RI del Piemonte come caso di studio.

Per raggiungere questo obiettivo, il lavoro si avvale di metodologie mutuate dalla linguistica computazionale e dal *Natural Language Processing* (NLP). Attraverso l'estrazione automatizzata e il calcolo di specifiche metriche, si procederà a valutare il grado di complessità dei regolamenti. Al fine di fornire un termine di paragone valido, i valori ottenuti dall'analisi dei regolamenti TA.RI saranno confrontati con un corpus di riferimento costituito da voci estratte dall'enciclopedia Wikipedia in italiano. Quest'ultima è assunta qui come punto di riferimento ottimale di prosa informativa standard: pur trattando tematiche complesse, essa adotta convenzioni redazionali esplicitamente orientate all'accessibilità, trasparenza e comprensibilità per il grande pubblico.

Il lavoro si articola come segue: nella Sezione 2 presenteremo nel dettaglio i corpora testuali utilizzati per l'analisi, le definizioni operative delle variabili di complessità prese in considerazione e i risultati dell'indagine quantitativa condotta. Nella Sezione 3 discuteremo tali risultati alla luce delle premesse fin qui esposte, focalizzandoci sulle implicazioni che la complessità linguistica pone all'esercizio della cittadinanza e all'accesso alla giustizia. Infine, nella Sezione 4, trarremo le conclusioni dell'indagine, riflettendo in modo critico sui limiti dello studio e delineando nuove prospettive di ricerca.

2. Analisi e dati

L'analisi è stata condotta su due corpora. Il primo di questi è costituito da 760 regolamenti TA.RI dei comuni piemontesi. Il secondo è invece un corpus di controllo formato da 124 testi presi da Wikipedia in italiano, i quali raccolgono una serie di argomenti vari inerenti ai rifiuti e alla loro gestione, sia da un punto di vista materiale che regolamentare. L'affinità tematica tra i due corpora dovrebbe garantire che le eventuali differenze riscontrate nell'analisi derivino dallo stile e non dai contenuti dei testi. Inoltre, i testi di Wikipedia si prestano a essere impiegati come gruppo di controllo in virtù della loro peculiare natura ibrida. Da un lato, mostrano una marcata vocazione divulgativa, governata da linee guida editoriali che impongono una sintassi piana rivolta a un pubblico eterogeneo; dall'altro, conservano il rigore tecnico proprio dei testi specialistici, ricorrendo a tecnicismi specifici ineludibili per non compromettere l'esattezza concettuale. In questo senso, il modello linguistico di Wikipedia si configura come un registro intermedio che si approssima all'ideale della buona scrittura scientifica e professionale, epurata dalle complessità tipiche dei linguaggi specialistici. Ciò, unito alla relativa facilità di raccolta dei testi permessa dalla piattaforma, rende i testi di Wikipedia un naturale punto di riferimento per vagliare la complessità linguistica dei regolamenti TA.RI.

Quanto alle metriche di complessità qui adottate, la scelta è ricaduta su quattro variabili di complessità, tra le molte altre che si sarebbero potute adottare. In particolare, queste quattro sono: Corrected Type-Token Ratio, un indice di ricchezza lessicale (Guiraud 1954); il numero di lemmi del testo presenti nel Vocabolario di Base (De Mauro 1980, 2016), come misura del ricorso a tecnicismi specialistici o arcaismi nel testo; la lunghezza media della frase, metrica che approssima la complessità sintattica di un testo (Alfieri et al. 2011: 24; Fortis 2005: 65; Piemontese 1996: 135); infine, l'indice Gulpease (Lucisano e Piemontese 1988), una misura di leggibilità del testo che collega il punteggio al grado minimo di istruzione richiesto per la comprensione del testo.

Le metriche misurano dunque aspetti diversi della complessità linguistica. È comunque bene sottolineare come le variabili qui considerate descrivano solo parzialmente la complessità di un testo, proprietà che lo caratterizza a tutti i suoi livelli e non solo a quelli lessicali e sintattici. Per una panoramica completa sui diversi parametri di complessità linguistica, si considerino i lavori di Fortis (2005) e il più recente Fiorentino e Ganfi (2024).

Per calcolare automaticamente le variabili considerate, il calcolo è stato implementato attraverso una serie di script Python 3.11, ricorrendo alla libreria Spacy (Hannibal et al. 2020) con il modello grande per l'italiano (it_core_news_lg).

Per ogni metrica nei due corpora si sono poi calcolate le metriche descrittive (come media, mediana, deviazione standard, valori massimi e minimi) e sono stati condotti test statistici inferenziali. In particolare, a causa della deviazione dalla normalità distributiva e della marcata eterogeneità delle varianze riscontrate nelle metriche di complessità linguistica, i confronti tra i due corpora sono stati condotti utilizzando il

test non parametrico U di Mann-Whitney per campioni indipendenti a una coda. Per quantificare la magnitudo delle differenze osservate, è stata calcolata la dimensione dell'effetto tramite l'indice *r* di Rosenthal.

2.1 Corrected Type-Token Ratio

La *Corrected Type-Token Ratio* (CTTR) è una misura di ricchezza lessicale formulata per superare la tradizionale *Type-Token Ratio* (TTR), compensandone la sensibilità alle variazioni di estensione testuale (Guiraud 1954). È noto, infatti, che il valore della TTR tenda a decrescere in modo inversamente proporzionale alla lunghezza del documento, inficiando l'affidabilità dei confronti tra testi di dimensioni eterogenee. Un elevato valore di CTTR indica l'impiego di un ampio spettro di termini unici, suggerendo un vocabolario ricco e variegato. Tale varietà richiede al lettore un livello superiore di competenza linguistica per un'efficace decodifica ed è frequentemente associata a testi di natura tecnica o settoriale.

Nello specifico, per mezzo di uno script Python, il calcolo del denominatore (i *token*, N) è stato eseguito applicando un filtro di pre-processamento per escludere dal computo la punteggiatura, gli spazi vuoti e le parole funzione della lingua italiana. Parallelamente, il numeratore (i *type*, T) è stato calcolato isolando i soli lemmi unici convertiti in minuscolo; questa procedura di lemmatizzazione ha permesso di neutralizzare eventuali duplicazioni spurie derivanti da variazioni flessionali o di maiuscolizzazione del medesimo lessema.

La tabella 1 presenta le statistiche descrittive della CTTR nei due corpora.

	Wikipedia	**TA.RI**
Media	10,70	13,61
Mediana	10,53	13,66
sd	3,55	0,76
Minimo	4,33	10,93
Massimo	20,80	16,26

Tabella 1: Statistiche descrittive della CTTR nei corpora Wikipedia e TA.RI.

L'analisi rivela una marcata divergenza tra le due classi. I testi normativi del corpus TA.RI presentano valori di tendenza centrale sistematicamente superiori (M = 13,61, Mdn = 13,66) rispetto al corpus di controllo estratto da Wikipedia (M = 10,70, Mdn = 10,53). Risulta particolarmente indicativa l'analisi della dispersione dei dati: il corpus Wikipedia è caratterizzato da un'elevata eterogeneità interna (SD = 3,55, range = 4,33 - 20,80), mentre il linguaggio dei regolamenti TA.RI mostra un'estrema standardizzazione e uniformità lessicale, evidenziata da una deviazione standard nettamente inferiore (SD = 0,76, range = 10,93 - 16,26).

A causa della violazione degli assunti di normalità distributiva e della marcata eteroschedasticità tra i due campioni, l'ipotesi di una maggiore complessità del corpus amministrativo è stata verificata tramite il test non parametrico U di Mann-Whitney a una coda per campioni indipendenti. L'analisi ha confermato che i valori di CTTR nel corpus TA.RI sono significativamente maggiori rispetto a quelli di Wikipedia ($p <$ 0,05). La quantificazione di tale differenza, misurata tramite la statistica *r* di Rosenthal, indica una dimensione dell'effetto grande. Questo parametro suggerisce che lo scostamento osservato non costituisce un mero artefatto statistico, ma riflette una divergenza sostanziale e di forte rilevanza pratica nell'estensione del vocabolario impiegato.

2.2 Vocabolario di Base

L'accessibilità lessicale dei documenti è stata misurata quantificando l'incidenza del Vocabolario di Base (VdB) (De Mauro 1980, 2016) nei testi. Questo indice valuta la proporzione di lessico presente nel testo appartenente al nucleo fondamentale della lingua italiana, cioè quell'insieme di vocaboli che occorrono con la massima frequenza nei testi e nei discorsi. Il rapporto tra il numero di lemmi inclusi nel VdB e il totale del testo permette di stabilire il grado di accessibilità del documento: una minore incidenza di tale vocabolario indica che il testo richiede al lettore una competenza lessicale più estesa per essere pienamente compreso.

Dal punto di vista metodologico, il calcolo è stato eseguito tramite un algoritmo in Python che opera sul repertorio lessicale effettivo (*type*) e non sulle singole occorrenze (*token*). I testi sono stati sottoposti a una fase di pre-processamento per la rimozione della punteggiatura e degli spazi vuoti, seguita dall'estrazione dei lemmi unici. Successivamente, la metrica è stata calcolata come il rapporto tra il numero di lemmi unici identificabili nel Nuovo Vocabolario di Base e il numero complessivo di lemmi unici contenuti nel documento.

La tabella 2 riporta i risultati ottenuti dall'analisi dei corpora.

	Wikipedia	**TA.RI**
Media	0,66	0,54
Mediana	0,67	0,55
sd	0,06	0,02
Minimo	0,52	0,46
Massimo	0,82	0,62

Tabella 2: Statistiche descrittive del VdB nei corpora Wikipedia e TA.RI.

Come mostrano i dati, il corpus enciclopedico di Wikipedia presenta un ricorso maggiore al lessico di base ($M = 0{,}66$, $Mdn = 0{,}67$) rispetto al corpus normativo TA.RI ($M = 0{,}54$, $Mdn = 0{,}55$). Analizzando gli indici di dispersione, si conferma la tendenza

già emersa in precedenza: il linguaggio dei regolamenti TA.RI è caratterizzato da una marcata standardizzazione e da una variabilità interna minima (SD = 0,02, range = 0,46 - 0,62), in netto contrasto con l'eterogeneità del corpus Wikipedia (SD = 0,06, range = 0,52 - 0,82).

Il test non parametrico U di Mann-Whitney per campioni indipendenti conferma che la percentuale di lemmi appartenenti al VdB nel corpus Wikipedia è significativamente superiore rispetto a quella registrata nel corpus TA.RI ($p < 0,05$). La dimensione dell'effetto risulta grande. Ciò evidenzia come i testi amministrativi analizzati tendano sistematicamente a distaccarsi dal vocabolario di uso comune, richiedendo ai destinatari un grado di competenza lessicale specialistica nettamente superiore per l'effettiva decodifica del testo.

2.3 Lunghezza media delle frasi

La lunghezza delle frasi costituisce un indice che approssima la complessità sintattica (Alfieri et al. 2011: 24; Fortis 2005: 65; Piemontese 1996: 135). Frasi particolarmente estese comportano un grado di complessità superiore, imponendo un gravoso onere sulla memoria di lavoro e sull'attenzione del lettore e richiedendo, di conseguenza, un maggiore dispendio di risorse cognitive per la decodifica.

Dal punto di vista operativo, la metrica è stata quantificata calcolando il numero di *token* costituenti ciascuna frase mediante uno script Python. Per evitare distorsioni causate da frammenti testuali non rappresentativi o outlier strutturali (ad es. titoli o sigle), l'algoritmo ha applicato un filtro di pre-processamento escludendo i segni di punteggiatura e gli spazi, e scartando dal computo le frasi composte da tre o meno parole valide. La misura finale di complessità sintattica per ciascun documento è stata calcolata come la lunghezza media delle frasi valide così isolate.

Nella tabella 3 sono riportate le statistiche descrittive in merito alla lunghezza media delle frasi nei due corpora.

	Wikipedia	**TA.RI**
media	26,32	33,31
mediana	25,83	32,52
sd	4,33	3,26
minimo	16,16	25,99
massimo	39,20	46,21

Tabella 3: Statistiche descrittive della lunghezza media delle frasi nei corpora Wikipedia e TA.RI.

Le statistiche mostrano un marcato divario tra i due corpora. I testi amministrativi del corpus TA.RI presentano unità frasali sistematicamente più estese (*M* = 33,31, *Mdn* = 32,52) rispetto al corpus di controllo (*M* = 26,32, *Mdn* = 25,83). Anche per questa metrica, sebbene con uno scarto meno pronunciato rispetto alle variabili lessicali, i testi regolamentari denotano una minore variabilità interna (*SD* = 3,26, range = 25,99 - 46,21) rispetto all'eterogeneità del corpus enciclopedico (*SD* = 4,33, range = 16,16 - 39,20).

Il confronto statistico, condotto mediante il test non parametrico U di Mann-Whitney per campioni indipendenti, ha confermato che la lunghezza media delle frasi nei regolamenti TA.RI è significativamente superiore a quella rilevata in Wikipedia ($p < 0,05$). L'analisi della dimensione dell'effetto ha restituito un valore grande. Questo indica che la maggiore lunghezza frasale osservata nei testi normativi costituisce un tratto sistematico e predominante, in grado di imporre un carico cognitivo sostanzialmente più gravoso sulla memoria di lavoro del lettore durante i processi di decodifica e comprensione del testo.

2.4. Indice Gulpease

La valutazione della leggibilità globale dei documenti è stata condotta attraverso il calcolo dell'Indice Gulpease, una metrica quantitativa specificamente calibrata sulla struttura della lingua italiana (Lucisano e Piemontese 1988). L'indice stima il grado di accessibilità di un testo basandosi sul rapporto tra la lunghezza delle parole (espressa in lettere), la lunghezza delle frasi e il numero totale dei lessemi.

Il punteggio risultante si colloca convenzionalmente su una scala da 0 a 100, in cui valori più elevati denotano una maggiore leggibilità: un punteggio superiore a 80 indica una facile comprensibilità per lettori con istruzione primaria, un valore compreso tra 60 e 80 corrisponde a un livello di istruzione secondaria, mentre punteggi inferiori a 40 segnalano un testo la cui decodifica risulta complessa persino per lettori con un livello di istruzione superiore o universitario.

Il calcolo dell'indice sui testi è stato automatizzato per mezzo di uno script Python. Per prevenire le distorsioni statistiche causate da stringhe non rappresentative (ad es. punteggiatura isolata, spaziature, titoli brevi o sigle), l'algoritmo esclude a priori i segni di interpunzione e gli spazi vuoti, scartando contestualmente dal computo generale le frasi composte da tre o meno parole valide. Di conseguenza, i parametri della formula sono stati calcolati esclusivamente sulle frasi che soddisfano tale criterio di validità strutturale.

La tabella 4 presenta i valori delle metriche descrittive dell'Indice Gulpease nei due corpora.

	Wikipedia	**TA.RI**
Media	45,46	42,52
Mediana	45,37	42,71
sd	2,82	1,05
Minimo	36,26	38,24
Massimo	53,75	44,51

Tabella 4: Statistiche descrittive dell'Indice Gulpease nei corpora Wikipedia e TA.RI.

I risultati consolidano il quadro di maggiore complessità del linguaggio amministrativo già emerso. Sebbene entrambi i corpora si collochino su valori medi indicativi di una bassa leggibilità generale (prossimi alla soglia critica di 40, la quale segnala difficoltà di comprensione anche per lettori con istruzione superiore), il corpus enciclopedico di Wikipedia registra indici più alti (M = 45,46, Mdn = 45,37) rispetto al corpus normativo TA.RI (M = 42,52, Mdn = 42,71). L'analisi della dispersione dei dati ribadisce la forte cristallizzazione strutturale dei testi giuridico-amministrativi: i testi TA.RI mostrano una deviazione standard estremamente ridotta (SD = 1,05, range = 38,24 - 44,51), a fronte di una maggiore, benché contenuta, variabilità nel corpus enciclopedico (SD = 2,82, range = 36,26 - 53,75).

Il risultato del test U di Mann-Whitney per campioni indipendenti indica che la leggibilità del corpus TA.RI è significativamente inferiore rispetto a quella del corpus di controllo ($p < 0,05$). L'indice r di Rosenthal evidenzia una dimensione dell'effetto grande. La combinazione di un vocabolario meno accessibile e di una sintassi complessa si traduce in un indice sintetico che colloca i testi amministrativi sistematicamente verso il limite inferiore di leggibilità della lingua italiana.

3. Discussione

L'indagine ha delineato un profilo linguistico problematico per i regolamenti TA.RI. Il confronto con il corpus di controllo ha evidenziato come i testi normativi esaminati siano sistematicamente caratterizzati da un'elevata ricchezza lessicale (alta CTTR), da un vocabolario distante dall'uso quotidiano (bassa incidenza del VdB) e da una marcata lunghezza frasale (elevata lunghezza media delle frasi). La sintesi di tali criticità è confermata dall'Indice Gulpease, che colloca stabilmente questi documenti su valori prossimi alla soglia più bassa di leggibilità, indicando difficoltà di comprensione per lettori privi di un'istruzione di livello universitario. La combinazione di un lessico specialistico e di una sintassi involuta, ricca di incisi e subordinate, costringe infatti il lettore a un doppio sforzo: da un lato la decodifica semantica di termini opachi, dall'altro il mantenimento in memoria di costrutti frasali prima di riuscire a cogliere il nucleo informativo della norma.

Un dato di particolare interesse è rappresentato poi dai valori di dispersione statistica, come riportano le deviazioni standard. La varianza minima dei regolamenti registrata in tutte le metriche, in contrasto con l'eterogeneità dei testi enciclopedici, conferma empiricamente la forte cristallizzazione stilistica dell'italiano giuridico-amministrativo (Mortara Garavelli 2001). Questo dato suggerisce che la complessità non è il risultato incidentale della scrittura di un singolo legislatore locale, ma rappresenta un *pattern* strutturale ed endemico della Pubblica Amministrazione, reiterato presumibilmente attraverso pratiche di «*copy-paste*» istituzionale e l'uso di modelli testuali precompilati, senza alcun adeguato intervento di semplificazione o adattamento al destinatario. Questa prassi di riproduzione acritica affonda le radici in un'inerzia burocratica e in una diffusa avversione al rischio da parte degli enti locali: la preferenza per il mantenimento di formule tradizionali, percepite come giuridicamente inattaccabili seppur incomprensibili per i non addetti ai lavori, prevale sulla necessità di trasparenza. Ne consegue un paradosso comunicativo per cui i regolamenti, pur essendo redatti a livello municipale per rispondere alle esigenze di una specifica comunità, risultano tutti omologati a un medesimo registro oscuro e spersonalizzato.

Questa perdurante opacità linguistica assume contorni critici se letta attraverso la lente dell'accesso alla giustizia e dei diritti di cittadinanza. Il regolamento TA.RI, a differenza di un testo divulgativo o enciclopedico, possiede valore prescrittivo e vincolante per l'intera popolazione residente, indipendentemente dal livello di scolarizzazione. Il cittadino è il destinatario finale di un testo che disciplina i suoi obblighi fiscali e il cui mancato rispetto – anche se derivante da un'errata interpretazione procedurale o normativa causata dalla complessità del testo – espone a sanzioni amministrative e pecuniarie. L'esigibilità di un dovere civico dovrebbe essere direttamente proporzionale alla chiarezza con cui tale dovere viene prescritto: quando l'Amministrazione impone oneri la cui reale portata è preclusa alla maggioranza dei cittadini, la sanzione non appare più come la giusta conseguenza di una violazione volontaria, ma piuttosto come una penalizzazione derivante dall'inaccessibilità del documento. A questo proposito si può richiamare la sentenza 364 del 1988 della Corte Costituzionale, la quale stabilisce che l'ignoranza della legge scusa il cittadino quando è inevitabile. Tra i criteri di inevitabilità, la Corte sancisce esplicitamente «l'assoluta oscurità del testo legislativo».

L'impenetrabilità sintattica e l'oscurità semantica evidenziate dai dati empirici generano, in questo contesto, una profonda asimmetria informativa e di potere tra l'ente impositore e il contribuente. La complessità linguistica cessa di essere un mero problema stilistico e si configura come un ostacolo materiale all'esercizio della cittadinanza attiva (Bourdieu 1991; De Mauro 2014). Tale asimmetria colpisce in modo sproporzionato le fasce più vulnerabili della popolazione, spesso proprio coloro che avrebbero maggiore necessità di accedere alle tutele previste. Richiedendo un livello di *literacy* sproporzionato rispetto alla media nazionale, la forma del regolamento impedisce di fatto a un'ampia fetta della popolazione di muoversi autonomamente nella burocrazia. Ad esempio, l'individuazione esatta dei presupposti per richiedere un'esenzione, la comprensione del metodo di calcolo delle quote fisse e variabili, o la

corretta interpretazione dei termini perentori per presentare un ricorso diventano veri e propri percorsi a ostacoli che sempre più spesso costringono il cittadino a rivolgersi, sostenendone i costi, a intermediari e professionisti del settore.

In conclusione, i dati suggeriscono che la complessità insita in questa tipologia testuale non sia neutra, ma produca un evidente problema democratico (Mattarella 2011; Lubello 2014). Questo fenomeno mina alla base il principio di trasparenza dell'azione amministrativa (Carloni 2014), trasformando la legge da strumento di regolazione a labirinto burocratico. Il cittadino viene posto in una condizione di subalternità e vulnerabilità causata dalla barriera eretta dal codice linguistico impiegato dall'amministrazione stessa, compromettendo così l'effettività di quel diritto alla comprensione che è precondizione essenziale per ogni democrazia partecipativa.

4. Conclusioni

Il presente studio ha voluto quantificare il grado di complessità linguistica dei regolamenti comunali TA.RI del Piemonte, confermando empiricamente la perdurante inaccessibilità del linguaggio amministrativo italiano. Attraverso un approccio basato su tecniche di linguistica computazionale e statistica inferenziale, è stato dimostrato come questi testi presentino barriere lessicali e sintattiche significativamente superiori rispetto a un corpus enciclopedico di controllo. Tale opacità strutturale, unita a una marcata rigidità formulare, non costituisce un mero difetto stilistico, ma si configura come una reale limitazione all'esercizio dei diritti di cittadinanza. In questo scenario, il cittadino viene relegato a un ruolo passivo, impossibilitato a comprendere autonomamente i propri doveri fiscali o a cogliere tempestivamente le opportunità di esenzione, trovandosi spesso costretto a dipendere da intermediari professionisti per adempiere a obblighi normativi che dovrebbero, per loro natura costituzionale, essere di immediata intelligibilità.

Pur offrendo evidenze statisticamente robuste, la presente ricerca traccia il perimetro per futuri sviluppi d'indagine. In primo luogo, risulta auspicabile un allargamento della sfera d'analisi a nuove zone geografiche su scala nazionale. L'inclusione di regolamenti provenienti da diverse regioni e da comuni di differente entità demografica consentirebbe di mappare compiutamente il fenomeno, verificando se la complessità osservata sia omogenea su tutto il territorio o se sussistano variazioni legate a specifiche prassi amministrative locali. Ad esempio, sarà cruciale indagare se le grandi aree metropolitane, tendenzialmente dotate di uffici legali e di comunicazione più strutturati, producano testi con indici di leggibilità differenti rispetto ai piccoli o medi comuni, i quali potrebbero risultare maggiormente esposti alla reiterazione acritica di modelli precompilati, accentuando così il fenomeno del «copy-paste» istituzionale.

In secondo luogo, i futuri filoni di ricerca dovranno prendere in esame ulteriori variabili di complessità linguistica. Oltre alle metriche qui impiegate, l'uso avanzato dell'analisi automatica del testo permetterà di indagare altre dimensioni linguistiche

connesse alla complessità. La misurazione della densità nominale, dell'indice di subordinazione, dell'uso di forme passive o impersonali e della coesione referenziale potrà fornire un quadro ancora più accurato delle difficoltà presenti in questa tipologia testuale. Si pensi, a titolo esemplificativo, all'abuso della nominalizzazione, che priva la frase del suo naturale dinamismo, o all'impiego pervasivo del passivo, che occulta sistematicamente l'agente responsabile dell'azione amministrativa. Quantificare l'incidenza di tali costrutti permetterebbe di comprendere non solo quanto il testo sia gravoso da leggere, ma anche in che misura esso risulti linguisticamente depersonalizzato e deresponsabilizzante.

Infine, la quantificazione del problema costituisce il presupposto necessario per una successiva fase applicativa. Una volta isolati i tratti problematici che caratterizzano la complessità dei testi normativi, l'obiettivo ultimo della ricerca linguistica in questo settore deve tradursi nella proposta di soluzioni da attuare.

Bibliografia

Alfieri, Gabriella, Marco Biffi, Agostino Bultrini, Barbara Cacelli, Paolo Cappelletto, Massimo Carli, Andrea Chiari, et al. *Guida alla redazione degli atti amministrativi: regole e suggerimenti*. Firenze: Ittig, 2011.

Bourdieu, Pierre. *Language and symbolic power*. Oxford: Polity Press, 1991.

Carloni, Enrico. *L'amministrazione aperta. Regole strumenti limiti dell'open government*. Santarcangelo di Romagna: Maggioli editore, 2014.

Cassese, Sabino, a cura di. *Codice di stile delle comunicazioni scritte ad uso delle amministrazioni pubbliche*. Roma: Istituto poligrafico e zecca dello stato, 1994.

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

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

De Mauro, Tullio. *Storia linguistica dell'Italia repubblicana. Dal 1946 ai nostri giorni*. Roma-Bari: Laterza, 2014.

De Mauro, Tullio. "Il nuovo vocabolario di base della lingua italiana". *Internazionale*, (dicembre 2016). https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

Fiorentino, Giuliana, e Vittorio Ganfi. "Parametri per semplificare l'italiano istituzionale: revisione della letteratura". *Italiano LinguaDue* 16, n. 1 (2024): 220–237.

Fioritto, Alfredo. *Manuale di stile dei documenti amministrativi*. Bologna: il Mulino, 2009.

Fortis, Daniele. "Il linguaggio amministrativo italiano". *Revista de Llengua i Dret* 43, (2005): 47–116.

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

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, e Adriane Boyd. *spaCy: Industrial-*

strength Natural Language Processing in Python. Zenodo, 2020. https://doi.org/10.5281/zenodo.1212303.

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

Lucisano, Pietro, e Maria Emanuela Piemontese. "Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana". *Scuola e Città* 3, (1988): 110–124.

Mattarella, Bernardo Giorgio. *La trappola delle leggi. Molte, oscure, complicate*. Bologna: il Mulino, 2011.

Mortara Garavelli, Bice. *Le parole e la giustizia: divagazioni grammaticali e retoriche su testi giuridici italiani*. Torino: Einaudi, 2001.

Piemontese, Maria Emanuela. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli: Tecnodid, 1996.

Il Corpus TARI: costruzione di una risorsa linguistica per l'analisi della complessità del linguaggio tributario-amministrativo

Alessia Lacroce
(Università di Torino)

Abstract

Questo contributo presenta il Corpus TARI, una raccolta di atti amministrativi tributari prodotti da comuni piemontesi durante il ciclo di riscossione della Tassa sui Rifiuti. La risorsa è stata sviluppata nell'ambito del progetto TuTeLa dell'Università di Torino per studiare la complessità del linguaggio amministrativo e sostenere la progettazione di strumenti di semplificazione rivolti a una platea eterogenea di destinatari. Dopo una descrizione della costruzione e annotazione del corpus, il paper presenta un'analisi delle proprietà del linguaggio amministrativo usato negli atti TARI. Il Corpus TARI è analizzato in contrasto con il corpus di italiano generale itTenTen20. Per l'analisi sono state considerate la leggibilità, la distribuzione delle categorie grammaticali, keywords e collocazioni. I risultati mostrano una leggibilità medio-bassa, uno stile fortemente nominale, una ridotta presenza di verbi, pronomi e avverbi e un'elevata densità di numeri. La specificità del dominio emerge sia da tecnicismi e acronimi sia dall'uso specialistico di parole comuni. L'analisi collocazionale evidenzia, inoltre, un discorso proceduralizzato, depersonalizzato e caratterizzato da formule ricorrenti. Pur con limiti di rappresentatività geografica e di accessibilità dei testi integrali, il Corpus TARI costituisce una base empirica per lo studio computazionale e la semplificazione della comunicazione tributario-amministrativa italiana.

1. Introduzione

Il presente contributo descrive la costruzione del Corpus TARI, una raccolta di atti amministrativi tributari prodotti da comuni piemontesi nel ciclo di riscossione della Tassa sui Rifiuti (TARI). Il corpus è stato sviluppato nell'ambito del progetto TuTeLa (*Turin Tech and Law: a Lab for Social Good*) dell'Università di Torino, che si propone, tra gli altri obiettivi, quello di migliorare l'accessibilità del linguaggio amministrativo italiano attraverso strumenti di linguistica computazionale, con attenzione ai bisogni delle persone vulnerabili. Il progetto muove da una nozione di vulnerabilità ampia, presa in prestito dall'ambito giuridico: vulnerabile non è necessariamente chi appartiene a un gruppo svantaggiato, ma chiunque si trovi, anche solo temporaneamente, in una

condizione in cui non dispone delle risorse o della resilienza per prevenire o affrontare un danno e che, per tale motivo, richiede una particolare protezione (Conti 2014, cit. in Lorubbio 2020: 662). In altre parole, la vulnerabilità non è una caratteristica di un gruppo specifico, ma una condizione che può riguardare chiunque, anche solo temporaneamente.

In virtù di questa definizione, la scelta di lavorare sugli atti TARI si presenta come particolarmente motivata. La TARI è il secondo tributo comunale per importanza dopo l'IMU. Per legge coinvolge chiunque occupi o detenga un immobile suscettibile di produrre rifiuti, indipendentemente dal loro titolo (proprietari, affittuari, residenti e non residenti). Questa "universalità" significa che gli atti prodotti nel corso del ciclo di riscossione raggiungono una platea che attraversa livelli di istruzione, fasce d'età e condizioni sociali molto diverse. Se colleghiamo questa osservazione alle ancora recenti evidenze sulla complessità del linguaggio amministrativo italiano, come quelle discusse da Buttinelli (in questo volume) sui regolamenti TARI, viene da chiedersi se l'iter di richiesta del pagamento poggi su documenti effettivamente comprensibili per i propri destinatari e in che misura. Il Corpus TARI nasce proprio per colmare questa lacuna e fornire la base empirica necessaria per sviluppare strumenti di analisi e semplificazione calibrati su questo dominio.

La relazione è organizzata come segue. La sezione 2 illustra i criteri di selezione e le ragioni che rendono gli atti TARI un caso di studio privilegiato e descrive la costruzione del corpus, dal protocollo di raccolta alla pipeline di annotazione automatica, con la relativa struttura dei metadati. La sezione 3 presenta alcune prime evidenze quantitative e qualitative sul profilo linguistico del corpus. La sezione 4 discute le conclusioni e i limiti del lavoro.

2. Costruzione del corpus

2.1 Selezione del dominio

Sul piano della riscossione, i dati disponibili mostrano che il ciclo di comunicazione tra amministrazione e contribuente presenta criticità diffuse. Secondo il rapporto dell'Ufficio Parlamentare di Bilancio (UPB, Focus n. 5/2024), nel triennio 2021-2023 gli incassi complessivi della TARI sono stati mediamente pari all'85% degli importi accertati, con valori che decrescono passando dal Nord (94%) al Centro (86%) al Sud (77%). Bordignon et al. (2024), sulla base dell'analisi dei bilanci di tutti i comuni italiani tra il 2016 e il 2022, stimano che solo il 60% di quanto accertato della TARI venga effettivamente riscosso, e che i comuni italiani riescano a incassare in media il 78% dei propri tributi, con una forte eterogeneità territoriale. In questo quadro, il Piemonte si configura come un caso di studio di particolare interesse. Secondo i dati ISTAT (2025), l'89% dei suoi 1.180 comuni ha meno di 5.000 abitanti, con oltre la metà che scende al di sotto dei 1.000 residenti. Questa frammentazione demografica si riflette

sull'organizzazione amministrativa: nei comuni più piccoli non esiste un ufficio tributi dedicato, e la gestione della fiscalità è affidata a funzionari che si occupano anche di molti altri servizi, il che può incidere sulla capacità di gestire efficacemente l'intero ciclo di riscossione. Il corpus TARI raccoglie documenti che provengono prevalentemente da questo tessuto di piccoli comuni, in cui la complessità linguistica degli atti prodotti si intreccia con risorse umane limitate. Studiare questi testi significa quindi affrontare un problema che può avere ricadute concrete sia sulla comprensibilità per i destinatari, sia sulla qualità redazionale da parte dei produttori.

2.2 Raccolta dei dati

La raccolta dei dati si inserisce nell'ambito di un'iniziativa di sensibilizzazione più ampia, volta a coinvolgere tutti i 1.180 comuni piemontesi. A ciascun ufficio tributario comunale è stata inviata una lettera di interesse in cui si illustrava il progetto e si richiedevano i modelli anonimizzati utilizzati per notificare o ricordare ai cittadini gli obblighi di pagamento della TARI nel corso del ciclo standard di riscossione. La partecipazione era interamente volontaria e gratuita. La raccolta si è svolta tra luglio e ottobre 2025.

Sono state individuate sei tipologie di atti nel processo di notifica di pagamento, che corrispondono a funzioni comunicative distinte e a un registro progressivamente più formale e coercitivo:

1. La comunicazione di pagamento è la prima comunicazione inviata ai contribuenti e ha funzione principalmente informativa. Essa informa dell'importo dovuto per l'anno in corso e facilita il pagamento volontario tramite il modulo F24. Nella comunicazione sono incluse informazioni sui presupposti della TARI e la composizione delle tariffe, oltre a informazioni accessorie rispetto al gestore dei rifiuti.
2. Il sollecito di pagamento è un richiamo formale relativo a un debito non saldato che richiede il pagamento entro 30 giorni.
3. L'atto istruttorio è un provvedimento pre-accertativo con cui l'amministrazione segnala anomalie e invita il contribuente a presentarsi presso l'ufficio delle imposte prima dell'emissione di eventuali sanzioni.
4. Lo schema di atto ex art. 6-bis L. 212/2000 è un documento preliminare che offre al contribuente la possibilità di contestare o raggiungere una risoluzione anticipata prima dell'emissione formale di un avviso di accertamento.
5. L'avviso di accertamento è l'atto con cui si notifica al contribuente un credito a seguito di una verifica sostanziale.
6. L'ingiunzione di pagamento, infine, è l'ultimo atto formale prima dell'esecuzione coattiva.

La distribuzione dei sei sottotipi nel corpus è riportata nella Tabella 1.

Tipo di atto	N. atti
Comunicazioni di pagamento	24
Solleciti di pagamento	22
Atti istruttori	6
Schemi di atto	4
Avvisi di accertamento	24
Ingiunzioni di pagamento	3
Totale	83

Tabella 1: Distribuzione dei tipi di atto nel corpus.

È opportuno precisare che nel sistema tributario italiano non esiste un ciclo di riscossione TARI standardizzato a livello nazionale. L'unico atto formalmente obbligatorio è l'avviso di accertamento, che il comune è tenuto a emettere nei casi in cui intenda rettificare quanto dichiarato dal contribuente o recuperare somme non versate (Legge 160/2019, art. 1, comma 792). Tutti gli altri atti che compongono il ciclo – dalla comunicazione di pagamento iniziale al sollecito, dall'atto istruttorio allo schema di atto ex art. 6-bis L. 212/2000 – non sono imposti da alcuna norma nazionale e riflettono scelte autonome dei singoli comuni. Questo significa che i documenti raccolti non rappresentano un insieme uniforme, ma una varietà di pratiche comunicative locali che differiscono per struttura, registro e grado di formalità. Dei 1.180 comuni contattati, 19 hanno accettato di partecipare, pari a circa l'1,6% del totale. Questo tasso di risposta, contenuto ma non inatteso per un primo contatto esplorativo, riflette i vincoli temporali del progetto e la natura volontaria della partecipazione. Il campione risultante non è rappresentativo in senso statistico, ma è comunque utile per una prima esplorazione sistematica della complessità linguistica degli atti TARI prodotti da comuni piemontesi di piccole e medie dimensioni.

I comuni aderenti hanno fornito un totale di 121 testi. Da questo insieme sono stati esclusi tutti i documenti la cui direzione comunicativa è invertita, ovvero che vanno dal cittadino all'amministrazione anziché dall'amministrazione al cittadino. Rientrano in questa categoria la modulistica e le istanze di rateazione, che non costituiscono atti di comunicazione amministrativa verso il contribuente. Dopo questa esclusione, il corpus finale è composto da 83 atti, la cui distribuzione per comune è riportata nella Tabella 2.

Comune	Provincia	N. atti	Token totali
Bricherasio	TO	7	12.039
Brossasco	CN	3	2.598
Chiusa di Pesio	CN	5	3.963
Cuceglio	TO	3	2.693
Ivrea	TO	6	10.720
Lozzolo	VC	4	7.800
Marano Ticino	NO	5	4.776
Pavone Canavese	TO	4	3.658

Comune	Provincia	N. atti	Token totali
Perosa Argentina	TO	2	3.020
Pinerolo	TO	4	7.448
Roccaforte Mondovì	CN	3	2.164
Rocchetta Belbo	CN	2	556
Rubiana	TO	2	1.165
Sandigliano	BI	3	3.234
Savigliano	CN	5	8.847
Settimo Torinese	TO	2	4.860
Sinio	CN	3	4.403
Verbania	VB	12	18.841
Vigone	TO	8	11.063
Totale	-	83	113.848

Tabella 2: Comuni e atti nel Corpus TARI.

Come emerge dalla Tabella 2, la distribuzione degli atti tra i sottotipi è disomogenea. Comunicazioni, solleciti e avvisi di accertamento contano almeno 20 testi ciascuno, mentre ingiunzioni (n = 3) e schemi di atto (n = 4) sono rappresentati da campioni molto ridotti. Questa asimmetria riflette in parte la struttura stessa del ciclo di riscossione, in cui alcuni atti sono più frequentemente prodotti dai comuni, e in parte i limiti della raccolta volontaria, che non ha consentito di insistere sul completamento del ciclo per tutti i partecipanti. Per mitigare questo squilibrio nelle analisi comparative, i sei sottotipi sono stati raggruppati in tre macro-categorie in base alla funzione comunicativa e al grado di coercitività (Tabella 3).

Tipo di atto	N. atti	Macro-tipo	N. macro-tipo
Comunicazioni di pagamento	24	Atti comunicativi	24
Solleciti di pagamento	22	Atti procedurali	32
Atti istruttori	6		
Schemi di atto	4		
Avvisi di accertamento	24	Atti impositivi	27
Ingiunzioni di pagamento	3		
Totale	83		83

Tabella 3: Distribuzione dei tipi di atto nel corpus.

Gli atti comunicativi (comunicazioni di pagamento, n = 24) hanno funzione prevalentemente informativa e descrittiva. Essi illustrano l'importo dovuto e le modalità di pagamento, senza ancora esercitare pressione sul destinatario. Gli atti procedurali (solleciti, atti istruttori e schemi di atto, n = 32) sono focalizzati sull'azione pratica che il contribuente è chiamato a compiere e svolgono una funzione di mediazione, offrendo la possibilità di regolarizzare la propria posizione prima che si producano

conseguenze più gravi. Gli atti impositivi (avvisi di accertamento e ingiunzioni, n = 27) esplicitano la pretesa tributaria in forma vincolante e sono direttamente seguiti, in caso di mancato adempimento, dalla riscossione forzata. Questa distribuzione in macro-categorie consente di compensare parzialmente lo sbilanciamento numerico tra i gruppi e di rendere le stime comparative più affidabili.

2.3 Pre-processing e pipeline NLP

I documenti ricevuti in formato PDF, DOCX e RTF sono stati convertiti automaticamente in testo semplice, seguiti da una fase di pulizia per correggere errori di estrazione automatica: rimozione di intestazioni e piè di pagina ricorrenti, numeri di pagina e spazi ridondanti. Le date e gli orari sono stati normalizzati in un formato uniforme. I dati personali sono stati eliminati o sostituiti con segnaposto anonimi. Un caso particolare è rappresentato dalle tabelle, che nei testi amministrativi tributari costituiscono un elemento rilevante e fondamentale per raggiungere le finalità del testo. Esse riportano importi, scadenze, codici di pagamento e dettagli tariffari che il contribuente è direttamente chiamato a leggere e utilizzare. Per questa ragione si è scelto di non eliminarle, ma di sistematizzarle. Nei casi in cui il contenuto tabellare fosse eccessivamente rumoroso, come ad esempio per la presenza di segnaposto, valori numerici specifici del singolo contribuente o formattazione residua dalla conversione, si è mantenuta la struttura con le intestazioni di colonna e un esempio di contenuto ripulito. Nei casi di maggiore complessità si è conservata la sola intestazione. In entrambi i casi, ogni tabella è segnalata nel testo pulito da un segnaposto esplicito nella forma [TABELLA_N], che ne preserva la posizione nella sequenza testuale e consente di tenerne conto nelle analisi sulla struttura del documento. Le operazioni più sistematiche sono state supportate da script Python, mentre la revisione finale, in particolare quella delle tabelle, è stata condotta manualmente su tutti i testi. I file sono stati infine esportati in formato CSV, con il testo grezzo e la versione pulita come colonne separate, e preparati per l'annotazione automatica.

L'annotazione linguistica è stata effettuata con Stanza (Qi et al. 2020), un framework neurale addestrato su *Universal Dependencies*, la cui adeguatezza per l'annotazione di testi istituzionali italiani è stata dimostrata in studi precedenti (Agnoloni et al. 2022). Stanza è stato preferito a spaCy (Honnibal et al. 2020) per la sua maggiore affidabilità nella segmentazione delle frasi in presenza di abbreviazioni, un fenomeno frequente nei testi amministrativi. Il modello pre-addestrato "it_combined", specifico per l'italiano, copre tokenizzazione, segmentazione, lemmatizzazione, *POS-tagging* con tagset *Universal Dependencies* e analisi delle dipendenze. Prima dell'annotazione è stata applicata una funzione di *pre-processing* per evitare una suddivisione errata delle frasi in corrispondenza delle abbreviazioni, sostituendo temporaneamente i punti abbreviativi con trattini bassi e ripristinandoli dopo la segmentazione. Per le analisi sintattiche sono state mantenute solo le frasi con almeno una forma verbale e una lunghezza minima di

tre token, per escludere intestazioni, etichette e frammenti formulari che non costituiscono frasi grammaticali complete.

Per verificare l'affidabilità dell'annotazione automatica è stato estratto un campione casuale stratificato di 300 token dall'intero corpus annotato, escludendo punteggiatura, simboli e segnaposto. Il campionamento è stato condotto con *seed* fisso (*random_state* = 42) per garantire la riproducibilità. Il confronto con l'annotazione manuale ha rilevato un'accuratezza del 96,7% per il *POS-tagging* (IC 95%: 94,6–98,7%) e del 95,0% per il *dependency parsing* (IC 95%: 92,5–97,5%). Le criticità emerse riguardano principalmente la classificazione errata di nomi o preposizioni come nomi propri, poiché il modello tende a etichettare come PROPN le parole scritte in maiuscolo, una convenzione grafica tipica dei testi amministrativi, e il mancato riconoscimento di locuzioni preposizionali fisse come «ai sensi di» o «in ossequio a». Per ridurre gli errori di classificazione PROPN è stato applicato un filtro di correzione automatica sui token più frequentemente classificati in modo errato.

A ciascun documento del corpus sono associati metadati che ne descrivono le caratteristiche contestuali e linguistiche. Sul piano contestuale, ogni atto è identificato dal comune di provenienza, dalla provincia, dal tipo di atto e dalla macro-categoria funzionale di appartenenza. Sul piano linguistico, per ciascun documento sono disponibili la lunghezza in token, l'indice di leggibilità Gulpease, la lunghezza media delle parole in caratteri e un insieme di misure morfosintattiche e sintattiche estratte dalla pipeline descritta nella sezione precedente. I testi integrali del corpus non sono pubblicamente disponibili per vincoli amministrativi legati alla natura dei documenti.

3. Profilo linguistico del corpus TARI: un'analisi esplorativa

Per una prima esplorazione del corpus abbiamo caricato i dati sulla piattaforma Sketch Engine per ricavare le analisi generali sul lessico. In particolare, l'analisi quantitativa è guidata da tre domande:

a. Come si distribuisce il lessico del corpus rispetto alle categorie dell'italiano generale?
b. Quali termini caratterizzano specificamente il dominio tributario-amministrativo rispetto al lessico di uso comune?
c. Come si comportano sintatticamente i termini chiave del dominio e con quali parole tendono a co-occorrere e a combinarsi?

Per rispondere alla prima domanda abbiamo estratto la distribuzione delle parti del discorso (POS) del corpus TARI rispetto a quelle di itTenTen20, un corpus di riferimento dell'italiano web di oltre 11 miliardi di token disponibile sulla piattaforma. Per la seconda domanda abbiamo effettuato un'analisi delle *keywords* (*keyness analysis*) per identificare i lemmi statisticamente sovrarappresentati nel corpus TARI rispetto al corpus di controllo di italiano generale. La misura utilizzata, calcolata in automatico

da Sketch Engine, è il Simple *Math score* (Kilgarriff et al. 2009)[1]. Infine, per la terza domanda abbiamo usato la funzione *Word Sketch*, che restituisce i collocati sintatticamente significativi di un termine attraverso il *logDice* score (Rychlý 2008).

3.1 Caratteristiche generali e di leggibilità

I testi nel corpus presentano una lunghezza media di 1.728 token, con una mediana di 1.608 e un intervallo interquartile molto ampio (488–2.578), che riflette una variabilità interna marcata tra sottotipi. I testi con funzione procedurale, come solleciti e atti istruttori, tendono a non superare i 1.000 token, mentre gli avvisi di accertamento raggiungono una lunghezza mediana di 2.960 token. La Tabella 4 riepiloga le caratteristiche superficiali per sottotipo.

Tipo di atto	N.	Lunghezza mediana (IQR)	Gulpease (media ± DS)	AWL (media ± DS)
Atti istruttori	6	374 [284–396]	54,60 ± 6,77	5,11 ± 0,17
Solleciti di pagamento	22	444 [319–812]	56,06 ± 6,02	5,36 ± 0,41
Avvisi di accertamento	24	2.960 [2.086–3.569]	49,34 ± 1,46	4,99 ± 0,11
Schemi di atto	4	1.832 [1.630–2.324]	50,92 ± 2,28	5,00 ± 0,11
Comunicazioni	24	1.650 [1.019–2.411]	49,42 ± 1,83	5,03 ± 0,11
Ingiunzioni di pagamento	3	1.475 [890–1.754]	57,60 ± 8,82	4,88 ± 0,04
CORPUS	83	1.608 [488–2.578]	51,90 ± 5,08	5,10 ± 0,28

Tabella 4: Caratteristiche superficiali per sottotipo di atto (LMT = lunghezza media in token; AWL = lunghezza media delle parole).

L'indice di leggibilità Gulpease medio del corpus è pari a 51,90 (DS = 5,08), un valore che colloca gli atti TARI nella fascia medio-bassa, al di sotto della soglia di 60 che la letteratura indica come discriminante per i lettori con scolarizzazione media (Lucisano e Piemontese 1988). Il valore più alto appartiene alle ingiunzioni di pagamento (57,60), ma è calcolato su soli tre documenti. Gli avvisi di accertamento, che

1 Il Simple Math score è definito come in a). Le frequenze relative (freq_rel) sono espresse in ipm (istanze per milione di token). Score = (freq_rel_corpus_target – freq_rel_corpus_riferimento) / (freq_rel_riferimento + 1). Lo *score* misura di quanto la frequenza relativa di un lemma nel corpus di interesse supera quella del corpus di riferimento, normalizzando per la frequenza di base in itTenTen. Termini con score prossimo a zero sono equamente distribuiti tra i due corpora; termini con lo score negativo sono sottorappresentati nel corpus TARI. Complementare allo score è il rapporto diretto tra le due frequenze relative (freq_rel_focus / freq_rel_ref), che misura di quante volte un termine è più frequente nel corpus TARI rispetto all'italiano generale, ed è particolarmente utile per identificare parole comuni che nel dominio tributario assumono un'intensità d'uso anomala.

costituiscono quasi un terzo del corpus, mostrano il valore medio più basso (49,34 ± 1,46), con una variabilità interna molto ridotta che riflette la natura formulare e standardizzata di questo sottotipo.

La lunghezza media delle parole (AWL) si attesta a 5,10 caratteri per token lessicale (DS = 0,28), con valori relativamente omogenei tra i sottotipi. Il sottotipo con AWL più elevata è quello dei solleciti (5,36), che mostrano anche la maggiore dispersione interna (DS = 0,41), mentre le ingiunzioni presentano il valore più basso (4,88). Questa uniformità è coerente con quanto osservato da Dell'Orletta et al. (2013) per altri generi testuali italiani, nei quali l'AWL tende a variare poco tra corpora pur in presenza di differenze sostanziali su altre dimensioni.

3.2 Profilo morfo-sintattico

La distribuzione delle categorie grammaticali del corpus TARI, calcolata su 103.480 token (escludendo punteggiatura e simboli), è stata confrontata con quella di itTenTen20 (14.514.566.714 token totali). Per ogni categoria è stato calcolato il *log-likelihood ratio* (LL) e il *Log Ratio* come misura della dimensione dell'effetto (Hardie 2014), utilizzando il *Lancaster Log-Likelihood and Effect Size Calculator* (Rayson 2008). I risultati sono riepilogati nella Tabella 5.

Categoria	TARI %	itTenTen %	LL	%DIFF	Log Ratio	p
Nomi	37,88	32,8	842,68	15,33	0,21	< .001
Verbi	11,25	16,2	−1.672,87	−30,73	−0,53	< .001
Aggettivi	10,57	11	−99,69	−3,59	−0,05	< .001
Avverbi	2,72	6	−2.278,45	−54,56	−1,14	< .001
Preposizioni	23,47	19,9	799,71	17,65	0,23	< .001
Congiunzioni	3,78	5,1	−293,47	−26,09	−0,44	< .001
Pronomi	2,12	6,4	−3.912,31	−66,97	−1,60	< .001
Numeri	8,21	2,5	8.393,18	230,22	1,72	< .001

Tabella 5: Distribuzione delle POS nel corpus TARI.

Il corpus TARI mostra un profilo che si distanzia in modo significativo da quello dell'italiano generale. Il test del chi-quadrato rivela delle differenze statisticamente significative per tutte le categorie principali (tutti i $p < 0,0001$). I nomi rappresentano il 37,9% dei token lessicali nel corpus TARI, contro il 32,8% di itTenTen (LL = +842,68; Log Ratio = +0,21), mentre i verbi si attestano all'11,25%, un valore nettamente inferiore al 16,24% dell'italiano generale (LL = −1.672,87; Log Ratio = −0,53). Il corpus TARI presenta uno stile altamente nominale.

Particolarmente rilevanti sono i dati sui pronomi e sugli avverbi. I pronomi rappresentano il 2,12% dei token nel corpus TARI, contro il 6,4% dell'italiano generale (LL = −3.912,31; Log Ratio = −1,60). Questo risultato segnala una tendenza all'uso

di un linguaggio depersonalizzato, in cui gli attori partecipanti (il «tu» esplicito rivolto al destinatario, il «noi» istituzionale e i soggetti agenti identificabili) non emergono in modo netto. Anche gli avverbi si attestano al 2,72% contro il 5,99% di itTenTen (LL = −2.278,45; Log Ratio = −1,14), il che indica una scrittura maggiormente centrata sulla presentazione diretta di obblighi, istruzioni e procedure amministrative, in cui viene dato poco spazio a elementi che servono a precisare, modulare o contestualizzare ciò che viene detto. Al contrario, i numeri rappresentano l'8,2% dei token del corpus TARI contro il 2,5% di itTenTen (LL = +8.393,18; Log Ratio = +1,72), un tratto atteso considerato il dominio tributario, che riflette la densità di date, importi, codici normativi e riferimenti numerici che riempiono ogni atto tributario e che presuppongono capacità di elaborazione numerica non garantite per tutti i destinatari.

I dati in questa sezione mostrano che il corpus TARI è costruito attorno a oggetti amministrativi (tributi, importi, scadenze, atti, codici), procedure e riferimenti normativi, mentre risultano marginali le strategie linguistiche normalmente utilizzate per modulare il discorso e sviluppare un discorso orientato verso il destinatario. Questa configurazione è coerente con la funzione istituzionale degli atti tributari, ma è anche indice di una tensione tra precisione amministrativa e accessibilità comunicativa.

3.3 Lessico di dominio: analisi delle keywords

L'analisi delle *keywords* restituisce 3.056 lemmi sovrarappresentati nel corpus TARI rispetto a itTenTen20, dopo la rimozione di rumore dovuto a refusi già presenti nei testi e segnaposto. La distribuzione degli *score* è asimmetrica. Il 63% delle *keywords* ha *score* inferiore a 5, il 33,9% si colloca nella fascia compresa tra 5 e 49, mentre solo una quota molto ridotta raggiunge valori elevati di salienza. In particolare, 52 lemmi (1,7%) presentano uno *score* compreso tra 50 e 99, e solo 44 termini (1,4%) raggiungono uno *score* ≥ 100, e quindi una sovrarappresentazione più marcata nel corpus TARI. Questo indica che il lessico del corpus TARI è in gran parte condiviso con l'italiano generale, mentre la specificità del dominio emerge da un nucleo ristretto ma altamente distintivo di lemmi specialistici. Tra questo nucleo ristretto di termini troviamo lemmi di carattere tributario, procedurale, sigle, enti e l'infrastruttura per la gestione del servizio rifiuti: *tributi* (score: 984), *arera* (594), *perequativo* (474), *tefa* (333), *riscossione* (233), *coattivo* (223,9), *irrogazione* (212,4), *contribuente* (133).

Passando al confronto tra le frequenze relative dei due corpora, emerge che 97 termini hanno frequenza relativa inferiore a 0,1 istanze per milione (ipm) in itTenTen, e quindi appartengono quasi esclusivamente al dominio tributario-amministrativo. Tra questi si trovano tecnicismi del sistema tariffario (tefa, tqrif, sorit, perequativo, spazzamento), acronimi istituzionali (cbill, csea, iuv, rid, pagopa) e lessico formulare burocratico (relata, notificatore, sottoindicata, insoluto). Si tratta di parole che un parlante dell'italiano generale non incontra praticamente mai al di fuori di questo contesto, e la cui comprensione presuppone un repertorio concettuale che non può essere dato per scontato nella platea dei destinatari.

Infine, osservando il rapporto tra la frequenza relativa del corpus TARI e di itTenTen vediamo che i termini con un rapporto superiore a 100 non sono solo parole settoriali ma anche parole comuni che tuttavia assumono nel corpus una funzione specialistica, come mostrato in dettaglio nel § 3.4. Questo aspetto è particolarmente rilevante in prospettiva di semplificazione automatica poiché mostra che intervenire solo sui termini tecnici più rari non sarebbe sufficiente, dal momento che una parte consistente della complessità del dominio è prodotta da un uso specializzato di parole comuni.

3.4 Profilo collocazionale

Per comprendere come i termini si combinino sintatticamente nel corpus, abbiamo utilizzato la funzione *Word Sketch* di *Sketch Engine*, che consente di identificare i collocati grammaticalmente salienti di un lemma attraverso il *logDice score* (Rychlý 2008). L'analisi collocazionale permette di andare oltre la frequenza e osservare come questi termini si comportano sintatticamente, con quali parole tendono a combinarsi e quali profili semantici costruiscono rispetto all'italiano generale.

Per l'analisi collocazionale sono stati selezionati alcuni lemmi che coprono tre funzioni centrali nelle procedure amministrativo-tributarie: i soggetti istituzionali (*comune*, *amministrazione*, *ufficio*, *contribuente*), gli atti e le procedure (*notifica*, *accertamento*, *ricorso*) e i termini legati all'adempimento tributario (*pagamento*, *importo*). La loro frequenza e il grado di *keyness* sono riportati in Tabella 6.

Categoria funzionale	Lemma	Frequenza nel corpus TARI	Keyness score
Soggetti	contribuente	295	132,6
	ufficio	486	19,92
	comune	430	6,52
	amministrazione	51	2,23
Procedure amministrative	accertamento	336	82,8
	ricorso	354	42,04
	notifica	358	156,87
Lessico tributario e regolatorio	importo	421	59,15
	pagamento	684	46,28

Tabella 6: Lemmi selezionati per l'analisi collocazionale.

Un primo risultato riguarda il profilo collocazionale dei soggetti istituzionali nel loro insieme. I termini *comune*, *amministrazione* e *ufficio* presentano un numero di collocati statisticamente significativi molto ridotto nel corpus TARI rispetto a itTenTen. *Comune* ha solo due collocati significativi, *facente* e *competente*, entrambe formule fisse che lo qualificano come ente di riferimento senza attribuirgli un ruolo agentivo. In itTenTen,

al contrario, *comune* appare come soggetto di verbi come *accorpare*, *aderire*, *coinvolgere*, *obbligare* e porta con sé modificatori come *limitrofo*, *montano*, *virtuoso*, *problematico*, che ne restituiscono una rappresentazione geografica e valutativa del tutto assente nel corpus TARI. Anche *amministrazione* mostra una ridotta agentività. Nel corpus TARI compare come soggetto principalmente con il modale *potere*, mentre in itTenTen si combina con verbi come *provvedere*, *decidere*, *procedere*, *adottare*. Questa differenza suggerisce che negli atti TARI l'amministrazione viene evocata come entità istituzionale ma non come soggetto esplicito dell'azione, mentre nel discorso pubblico è costruita come un soggetto attivo che prende decisioni. Curiosamente, nella relazione di coordinazione si combina con *errore*, un collocato che riflette la possibilità di contestare l'atto per vizio amministrativo.

Ufficio presenta un profilo leggermente più ricco. Nel corpus TARI, con funzione di soggetto, si colloca con verbi come *procedere*, *ricevere*, *provvedere*, rispetto a itTenTen in cui i collocati includono verbi come *arredare*, *affittare*, *ospitare*, che rimandano allo spazio fisico piuttosto che alla funzione istituzionale.

Il *contribuente* si distingue dagli altri soggetti istituzionali per ricchezza collocazionale e complessità semantica, ed è il termine che mostra il contrasto più netto con l'italiano generale. Come oggetto del verbo, nel corpus TARI è retto quasi esclusivamente da *invitare*, *informare*, *accertare* e *agevolare*, verbi procedurali e con connotazione neutra che costruiscono il contribuente come destinatario di azioni amministrative definite. In itTenTen lo stesso ruolo sintattico è retto da verbi con connotazione negativa come *tartassare*, *vessare*, *penalizzare*, *gravare*, *spremere,* *spennare*, *derubare*, *tassare* che riflettono il dibattito pubblico (itTenTen raccoglie articoli web) sul rapporto tra fisco e cittadino. Come soggetto, il contribuente nel corpus TARI è descritto da un ruolo attivo, pur all'interno di procedure circoscritte; infatti, esso *può*, *deve*, *aderisce*, *richiede*, *riceve*. In itTenTen il suo ruolo come soggetto è semanticamente molto più ricco: *impugna*, *versa*, *paga*, *omette*, *contesta*, *ricorre*, *usufruisce*, *adempie*, un repertorio che copre sia l'adempimento sia la contestazione.

Un dato interessante da osservare riguarda i modificatori aggettivali. Nel corpus TARI il contribuente è qualificato da soli cinque aggettivi statisticamente significativi, nessuno dei quali è di tipo valutativo: *gentile*, quasi certamente dalla formula di apertura "Gentile Contribuente", *riconoscibile*, *intestato*, *stesso* e *presente*. Al contrario, in itTenTen il contribuente porta con sé un intero repertorio di giudizi morali e sociali come *moroso*, *infedele*, *inadempiente*, *disonesto*, *virtuoso*, *diligente*, *leale*, *benestante*, *povero*, *ignaro*, *inconsapevole*, *indifeso*. Il corpus TARI non costruisce una rappresentazione del contribuente come persona, ma lo riduce a un'entità procedurale identificata da dati anagrafici e fiscali, priva di profilo morale o sociale.

Per le procedure, nel corpus TARI il *ricorso* si *ammette*, si *propone*, si *presenta* ed è *procedibile*. In itTenTen si *respinge*, si *rigetta*, si *accoglie* ed è *inammissibile*, *infondato*, *tardivo*. Queste collocazioni mostrano due punti di vista diversi rispetto a questo istituto: mentre in TARI il punto di vista è quello del contribuente che esercita un diritto, in itTenTen il punto di vista è quello dell'organo giudicante che valuta. Nella costruzione

"ricorso contro + nome", in TARI il ricorso è quasi esclusivamente contro *l'avviso* o *l'atto* tributario specifico; in itTenTen è contro *decisione*, *sentenza*, *provvedimento*, un universo giuridico molto più ampio.

Notifica presenta la differenza semantica più marcata tra i due corpora. Nel corpus TARI è l'atto giuridico di comunicazione ufficiale che si combina con *avviso*, *atto*, *ingiunzione*, *ricorso*, *cartella*. I suoi modificatori sono *preventivo*, *telematico*, *esecutivo*. In itTenTen la stessa parola designa prevalentemente la notifica digitale dello smartphone e co-occorre prevalentemente con verbi come *disattivare*, *silenziare*, *visualizzare* e modificatori quali *push* e *istantaneo*.

Accertamento in TARI *diventa* definitivo o esecutivo ed è strettamente legato ad *adesione*, il che riflette la possibilità di definizione bonaria prevista dal procedimento tributario. In itTenTen si *dispone*, si *esperisce*, si *esegue*, un registro più astratto che riflette un uso del termine distribuito su contesti molto diversi da quello tributario.

Infine, *pagamento* e *importo* mostrano come il lessico dell'adempimento nel corpus TARI costruisca un universo procedurale chiuso. *Importo* è modificato da *versato*, *dovuto*, *insoluto*, *complessivo*, stati dell'importo nel ciclo di riscossione, e si combina con *rettificare*, *visualizzare*, *corrispondere*. *Pagamento* si lega fortemente a *omettere*, una delle collocazioni più alte dell'intero corpus, cristallizzata nella formula "omesso pagamento", e a *ripartire* nel senso di suddividere in rate. I modificatori nominali di *pagamento* in TARI sono *anno*, *ingiunzione*, *accertamento*, *PagoPA*; in itTenTen sono *contactless*, *PayPal*, *cash*. In TARI il pagamento è un adempimento procedurale con scadenze e conseguenze definite; in itTenTen è una transazione commerciale tra parti.

Nel complesso, l'analisi collocazionale mostra che il corpus TARI costruisce un discorso fortemente proceduralizzato e istituzionale, nel quale i ruoli degli attori sono rigidamente definiti, le azioni risultano circoscritte a procedure specifiche e il linguaggio tende a cristallizzarsi in formule relativamente stabili. Il confronto con itTenTen rivela che gli stessi lemmi nell'italiano generale occupano posizioni sintattiche molto più varie e costruiscono profili semantici radicalmente diversi. Questa distanza è particolarmente rilevante per i destinatari degli atti, che portano con sé le associazioni dell'italiano generale e possono trovarsi disorientati di fronte a un uso così specializzato e circoscritto delle stesse parole.

4. Conclusioni

Il presente contributo ha descritto la costruzione del corpus TARI e ha presentato una prima analisi esplorativa delle sue proprietà linguistiche superficiali (leggibilità, lunghezza dei testi e delle parole, lessico), guidata da tre domande di ricerca.

Rispetto alla prima domanda, su come si distribuisce il lessico del corpus rispetto alle categorie dell'italiano generale, i dati mostrano che il corpus TARI presenta un profilo morfosintattico significativamente diverso da quello dell'italiano generale. Lo stile è fortemente nominale, con una sovrarappresentazione dei nomi e una

sottorappresentazione marcata di verbi, pronomi e avverbi. La quasi assenza di pronomi segnala un testo depersonalizzato, in cui gli attori comunicativi non emergono esplicitamente. La fortissima presenza di numeri riflette la densità di importi, date e codici normativi che caratterizzano ogni atto tributario.

In riferimento alla seconda domanda, su quali termini caratterizzano specificamente il dominio, l'analisi delle parole chiave mostra che la specificità lessicale del corpus emerge sia attraverso tecnicismi altamente specialistici sia attraverso la ri-specializzazione di parole appartenenti all'italiano comune. Nel primo caso rientrano 97 termini quasi esclusivi del dominio, tra cui tecnicismi tariffari come *tefa*, *tqrif* e *sorit*, acronimi istituzionali come *cbill*, *iuv* e *pagopa*, e lessico formulare burocratico come *relata*, *notificatore* e *insoluto*. Rientrano anche 44 termini con *keyness score* ≥ 100, che delineano le dimensioni tematiche centrali del corpus tra fiscalità locale, procedure di riscossione e istituti giuridici tributari. Nel secondo caso, forse il più rilevante in prospettiva di semplificazione, rientrano i 161 termini con rapporto "corpus TARI/corpus di riferimento" superiore a 100, parole comuni dell'italiano generale che nel corpus TARI assumono un'intensità d'uso elevata, tra cui *omettere*, *riscossione*, *notifica*, *sollecito* e *contribuente*. Intervenire solo sul lessico tecnico più raro sarebbe insufficiente, dal momento che una parte consistente della complessità del dominio è prodotta da un uso specializzato di parole apparentemente familiari.

Quanto alla terza domanda, su come si comportano sintatticamente i termini chiave e con quali parole si combinano, il confronto con itTenTen evidenzia che gli stessi lemmi, nell'italiano generale, occupano posizioni sintattiche più varie e attivano profili semantici differenti. *Amministrazione*, *comune* e *ufficio* nel corpus TARI compaiono soprattutto in formule istituzionali fisse e raramente come soggetti di azioni esplicite, mentre nell'italiano generale si combinano più frequentemente con verbi che descrivono decisioni, attività e relazioni sociali. *Contribuente*, nel corpus TARI, è costruito come entità procedurale, mentre nell'italiano generale si associa più spesso a valutazioni morali e sociali; *notifica* indica un atto giuridico-formale nel corpus TARI, mentre in itTenTen rimanda prevalentemente all'ambito digitale; *ricorso* è osservato dalla prospettiva del contribuente che agisce, mentre nell'italiano generale emerge più spesso dal punto di vista dell'organo che valuta.

Infine, sul piano della leggibilità, il corpus si colloca in una fascia medio-bassa, con un indice Gulpease medio di 51,90, al di sotto della soglia di 60 comunemente associata a una buona comprensibilità per lettori con scolarizzazione media.

Nel complesso, le evidenze quantitative e collocazionali convergono nel descrivere gli atti TARI come testi caratterizzati da complessità multidimensionale, lessicale per la presenza di tecnicismi esclusivi e di parole comuni usate in accezioni specialistiche, morfosintattica per lo stile nominale e la depersonalizzazione, collocazionale per la cristallizzazione formulare che rende i testi prevedibili per chi conosce il dominio e opachi per chi non lo conosce.

Il corpus presenta alcuni limiti che è opportuno discutere esplicitamente. Il primo riguarda la copertura geografica: i 19 comuni partecipanti rappresentano circa l'1,6%

dei comuni piemontesi e appartengono esclusivamente al Nord Italia. Il campione non è rappresentativo in senso statistico e le distribuzioni osservate potrebbero differire da quelle di altri contesti regionali o di altri tributi. Sviluppi futuri dovranno cercare di coinvolgere un numero maggiore di amministrazioni e di estendere la raccolta ad altri ambiti geografici e ad altri settori della comunicazione tributaria. Un secondo limite riguarda la natura amministrativa dei documenti, che ha imposto vincoli di anonimizzazione e ha reso impossibile la pubblicazione dei testi integrali. Il corpus è quindi accessibile nella sua componente quantitativa e nei metadati, ma non come risorsa testuale aperta. Nonostante questi limiti, il corpus TARI rappresenta una raccolta sistematica di atti amministrativi tributari italiani costruita con finalità di analisi linguistica computazionale. La costruzione di risorse linguistiche di questo tipo è un prerequisito per qualsiasi intervento di semplificazione basato su evidenza empirica, e il corpus TARI costituisce un punto di partenza per un programma di ricerca più ampio sulla leggibilità e la comprensibilità del linguaggio amministrativo di dominio tributario italiano.

Bibliografia

Agnoloni, T., R. Bartolini, F. Frontini, S. Montemagni, C. Marchetti, V. Quochi, M. Ruisi e G. Venturi. "Making Italian Parliamentary Records Machine-Actionable: The Construction of the ParlaMint-IT Corpus." In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 117–24. ELRA, 2022.

Bordignon, M., D. Cipullo, I. Valpreda e L. Virgadamo. *La riscossione dei tributi municipali. Osservatorio sui Conti Pubblici Italiani*, Università Cattolica del Sacro Cuore, 2024. https://osservatoriocpi.unicatt.it.

Buttinelli, A. "Quantificare la complessità. Uno studio sulla complessità linguistica dei testi giuridico-amministrativi in italiano." In *questo volume*, 2026.

Conti, R. "Diritti fondamentali, soggetti vulnerabili: tappe e obiettivi di un articolato 'cammino' interno." Relazione presentata al *Convegno dell'Associazione Cammino*, Senato della Repubblica, Roma, 24 gennaio 2014.

Dell'Orletta, F., S. Montemagni e G. Venturi. "READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification." In *Proceedings of the Workshop on Automatic Text Simplification*, 73–83. ACL, 2011.

Dell'Orletta, F., S. Montemagni e G. Venturi. "Linguistic Profiling of Texts across Textual Genres and Readability Levels." In *Proceedings of Recent Advances in Natural Language Processing*, 189–97, 2013.

Fioritto, A., a cura di. *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Bologna: Il Mulino, 1997.

Hardie, A. "Log Ratio: An Informal Introduction." *ESRC Centre for Corpus Approaches to Social Science*, 2014. http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/.

Honnibal, M., I. Montani, S. Van Landeghem e A. Boyd. *spaCy 2: Natural Language Understanding*

with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. Explosion AI, 2020.

Kilgarriff, A. "Simple Maths for Keywords." In *Proceedings of the Corpus Linguistics Conference CL2009*, a cura di M. Mahlberg, V. González-Díaz e C. Smith. Liverpool: University of Liverpool, 2009.

Lorubbio, V. "La tutela dei soggetti vulnerabili." *DPCE Online 2020*, n. 1 (2020): 661–78.

Lucisano, P. e M. E. Piemontese. "GULPEASE: Una formula per la predizione della difficoltà dei testi in lingua italiana." *Scuola e Città*, n. 3 (1988): 110–24.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton e C. D. Manning. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–8. ACL, 2020.

Rayson, P. *Wmatrix: A Web-Based Corpus Processing Environment*. Computing Department, Lancaster University, 2008. http://ucrel.lancs.ac.uk/wmatrix/.

Rychlý, P. "A Lexicographer-Friendly Association Score." In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, 6–9, 2008.

Ufficio parlamentare di bilancio. La tassa sui rifiuti: carico fiscale, riscossione e implicazioni sui bilanci dei comuni. *Focus n.* 5/2024. 2024. https://www.upbilancio.it/it/focus-n-5-2024-la-tassa-sui-rifiuti-carico-fiscale-riscossione-e-implicazioni-sui-bilanci-dei-comuni/. Consultato il 22 maggio 2026.

La nominalizzazione nei Contratti Collettivi Nazionali di Lavoro: tra semplificazioni e blocchi sintattici

Annarita Miglietta
(Università del Salento)

Abstract

La presente ricerca focalizzerà l'attenzione sul ruolo della nominalizzazione in un corpus costituito da quattro Contratti Collettivi Nazionali di Lavoro, quelli relativi ai settori dell'agricoltura, edilizia, turismo e lavoro domestico, settori in cui vengono impiegati maggiormente gli immigrati. In particolare, con strumenti statistici, Read.it e T-Lab, si calcoleranno la densità semantico-informativa e le frequenze delle strutture nominali, nonché le concordanze di alcune parole, deverbali, per cercare di individuare come queste stabiliscono reti di relazioni con altri costituenti frasali, dando vita ad uno snodo catalizzante di strutture nominali, abusate negli scritti giuridico-istituzionali e note per astrattezza e complessità comunicativa. L'obiettivo è quello di provare se è sempre possibile una riscrittura – con cancellazione/sostituzione di alcune catene nominali – al fine di rendere i contratti più accessibili, pur mantenendone validità e *vis* giuridiche.

Introduzione

L'Italia ha storicamente sofferto di una comunicazione istituzionale poco chiara, complessa e poco trasparente, unidirezionale e sbilanciata per i cittadini che, si sa, spesso non posseggono un'adeguata cultura giuridica. Questa situazione ha portato il nostro Paese a un ritardo rispetto ad altri (si pensi a quelli anglofoni[1]), sebbene la questione

1 In America, per esempio, c'è ormai una lunga e consolidata normativa, che impone l'uso di un linguaggio chiaro nei documenti pubblici e commerciali. In 10 degli Stati Uniti è richiesto l'uso di un linguaggio semplice nei contratti con i consumatori. Nello stato di New York sin dal 1978 una legge impone che i contratti di locazione e quelli rivolti ai consumatori siano scritti in "a clear and coherent manner using words with common and every day meanings", nonché divisi per paragrafi, "Appropriately divided and captioned by its various sections", e che "Any creditor, seller or lessor who fails to comply with this subdivision shall be liable to a consumer who is a party to a written agreement governed by this subdivision in an amount equal to any actual damages sustained plus a penalty of fifty dollars" (cfr. New York General Obligations (GOB) CHAPTER 24-A, ARTICLE 5, TITLE 7). "A Florida law goes further,

oggi sia molto sentita per la realtà multietnica e plurilingue che connota l'Italia, grazie a flussi migratori che di anno in anno assumono sempre più vaste proporzioni.

Nonostante numerosi tentativi – che risalgono agli anni Ottanta del Novecento – le soluzioni a un tale problema sembrano ancora lontane. Il 16 novembre 1979, nel *Rapporto sui principali problemi dell'amministrazione dello Stato* presentato alle Camere, il ministro Massimo Severo Giannini metteva in risalto i pericoli legati alla produzione normativa che rendeva complicata l'individuazione delle norme applicabili e delle modalità di attuazione. Si legge ad esempio: "l'oscurità delle norme dà origine a continui conflitti interpretativi", "la norma è tra quelle difficilmente applicabili", "vi è dunque un rincorrersi di norme che parlano di ruoli unici […]" (§ 4.9), "norme che si accavallano", "pur nei confronti dei pochi enti, le norme hanno disposto regole troppo rigide, spesso inapplicabili" (§ 5.9). In seguito al *Rapporto* si registrò in quegli anni il diffondersi di una cultura sulla normazione grazie ai lavori della Commissione Barettoni Arleri[2] (presentata alla Camera dei Deputati il 17 giugno 1981) e grazie al seminario organizzato dalla Regione Toscana nel 1983-1984, che portò alla realizzazione di un "manuale per la redazione dei testi legislativi" e all'adozione di circolari regionali sulla formulazione degli atti normativi, cui seguì il manuale *Regole e suggerimenti per la redazione dei testi normativi,* approvato, il 24 gennaio 1992, dalla Conferenza dei Presidenti delle Assemblee legislative delle Regioni e delle Province autonome con il supporto scientifico dell'Osservatorio Legislativo Interregionale (OLI), la cui quarta edizione è del 2024. Nella penultima edizione del 2007, il documento presenta un Preambolo "che codifica una serie di principi generali per la redazione dei testi normativi (anteposti alle regole puntuali contenute nelle parti successive del manuale). Si tratta di principi relativi all'uso del linguaggio e precisamente: chiarezza, precisione, uniformità, semplicità, economia" (Fioritto 2023: 130).

Per scongiurare *i pericoli* ai quali si va incontro quando *si scrive male una legge,* a partire dagli anni Novanta del secolo scorso, cominciarono a diffondersi pubblicazioni di guide di stile recanti precise regole e raccomandazioni di "buona scrittura": dalla pubblicazione del *Codice di stile* (1993), promosso dal Dipartimento della Funzione Pubblica sotto la guida del ministro Sabino Cassese, al *Manuale di stile* (1997), redatto

requiring insurances to meet a minimum score (of 45) using the Flesch Reading Ease Formula" (Sutherland 2013: 357).

2 Nel *Rapporto* la Commissione poneva all'attenzione "la questione della forma letteraria delle norme non solo per il semplicistico motivo che la redazione di esse in forma linguisticamente buona e corretta ne agevola l'applicazione, bensì anche per altra più consistente ragione. Esiste ormai una forma, pur corretta letterariamente, che è specifica e propria del linguaggio tecnico-giuridico. La secolare tradizione giuridica comporta che nello scibile giuridico moltissime parole abbiano ormai assunto un significato tecnico convenzionale preciso, cui il legislatore dovrebbe conformarsi, ovvero consapevolmente ed espressamente derogare" e lamentava tra l'altro lo scadimento della forma testuale: "Il problema si è indubbiamente acutizzato negli ultimissimi anni della produzione legislativa, nei quali il linguaggio, sia tecnico-giuridico sia genericamente di buon italiano, è apparso decisamente scadere" (Barettoni Arleri 1981: 18).

dal giurista Alfredo Fioritto, fino alle Direttive Frattini (2002) e Baccini (2005), e alle tante varie opere di linguisti e giuristi. Ma, nonostante questo proliferare di attenzioni e tentativi di soluzione, gli ostacoli persistono: sembra proprio che semplificare sia complesso. I documenti istituzionali in Italia, spesso scritti in un linguaggio confuso, difficile, talvolta disordinato, poco trasparente, rappresentano un ostacolo significativo soprattutto per gli immigrati che cercano di accedere – spesso vanamente – a servizi e diritti. I mancati interventi in materia di chiarezza per favorire la piana fruibilità dei documenti creano quella disuguaglianza che è stata definita da Massimo Vedovelli una "lotta impari", tra la capacità di comprensione[3] degli immigrati, o dei fruitori dei testi, in generale, e la complessità[4] della scrittura.

I testi poco chiari, connotati per una formulazione linguistica spesso ambigua, limitano, lo sappiamo bene, i diritti dei cittadini e ancor di più di coloro che arrivano nel nostro Paese, non ne garantiscono l'integrazione, o quanto meno l'accoglienza. È noto che i diritti concreti della cittadinanza, della salute e del lavoro possono essere realmente realizzati solo se il contributo offerto all'immigrato è proporzionato alla capacità dell'immigrato – spesso contraente debole nei rapporti contrattuali e privo di alfabetizzazione giuridica - di dare un senso ai testi istituzionali scritti. Questo è un aspetto di importanza rilevante se si considera come la carenza di informazioni, spesso oltretutto impenetrabili e inadeguate, non solo penalizzi l'immigrato, ma danneggi anche il paese ospitante e i suoi operatori in ogni settore. L'Italia ormai si trova quindi di fronte a una sfida fondamentale in materia di immigrazione e deve prendere un impegno per garantire agli immigrati la partecipazione attiva in tutti i settori pubblici (occupazione, istruzione, assistenza sanitaria, alloggi, ecc.), rimuovendo gli ostacoli primari imposti da una comunicazione inefficace.

E questo diviene ancora più urgente se pensiamo che con il DPCM del 2 ottobre 2025 per il triennio 2026 – 2028 sono stati programmati in Italia nuovi ingressi di lavoratori stranieri.

3 La leggibilità dipende da fattori quantitativi (lunghezza della frase, lunghezza delle parole, ecc.) ed è quindi misurabile attraverso indici di leggibilità rappresentabili lungo una scala di valori che permette di confrontare la leggibilità di testi differenti; mentre la comprensibilità di un testo dipende dalle competenze linguistiche e cognitive del lettore, dal suo retroterra culturale e non è misurabile. Per approfondimenti cfr. Vena (2022).

4 Mentre Kusters (2003: 403) osserva che "Linguistic phenomena are considered complex when they are difficult to grasp by an outsider who is not familiar with the speech community", per il nostro Paese, visti i dati delle indagini OCSE-PISA e INVALSI che registrano bassi livelli di competenza in lettura, il concetto si può estendere anche ai madrelingua.

2. Il Decreto Flussi

Il *Decreto del Presidente del Consiglio dei Ministri 2 ottobre 2025 "Programmazione dei flussi di ingresso legale in Italia dei lavoratori stranieri per il triennio 2026-2028"* nell'art. 5, capo 1 recita:

> "Sono ammessi in Italia, per motivi di lavoro subordinato stagionale e non stagionale e di lavoro autonomo, i cittadini stranieri residenti all'estero entro le seguenti quote complessive:
> a) 164.850 unità per l'anno 2026;
> b) 165.850 unità per l'anno 2027;
> c) 166.850 unità per l'anno 2028.
> I lavoratori proverranno da Paesi che hanno firmato accordi di cooperazione con l'Italia: Albania, Algeria, Bangladesh, Bosnia-Erzegovina, Corea (Rep. di Corea), Costa d'Avorio, Ecuador, Egitto, El Salvador, Etiopia, Filippine, Gambia, Georgia, Ghana, Giappone, Giordania, Guatemala, India, Kirghizistan, Kosovo, Mali, Marocco, Mauritius, Moldova, Montenegro, Niger, Nigeria, Pakistan, Perù, Repubblica di Macedonia del Nord, Senegal, Serbia, Sri Lanka, Sudan, Thailandia, Tunisia, Ucraina, Uzbekistan.
> E saranno impiegati nei seguenti settori:
> - Agricoltura, silvicoltura e pesca
> - Industrie alimentari, delle bevande e del tabacco
> - Industrie tessili, dell'abbigliamento e calzature
> - Industrie metallurgiche e dei prodotti in metallo
> - Altre industrie
> - Costruzioni
> - Commercio all'ingrosso e al dettaglio
> - Servizi di alloggio e ristorazione
> - Servizi turistici
> - Servizi di trasporto, logistica e magazzinaggio
> - Servizi operativi di supporto alle imprese e alle persone
> - Sanità, assistenza sociale e servizi sanitari privati
> - Altri servizi"

Inoltre, visto che sempre nel DECRETO-LEGGE 3 ottobre 2025, n. 146 nell'art. 1, nel Preambolo che giustifica i presupposti, i motivi dell'entrata in vigore del provvedimento, tra l'altro, si legge:

> "Ritenuta la straordinaria necessità e urgenza di intervenire sulla disciplina dell'immigrazione e della condizione dello straniero, con particolare riferimento alla disciplina del rapporto di lavoro dello straniero, anche in relazione al settore dell'assistenza familiare o sociosanitaria per l'assistenza di persone con disabilità o grandi anziane; ritenuta, inoltre, la straordinaria necessità e urgenza di incrementare l'azione di contrasto al caporalato e allo sfruttamento lavorativo in agricoltura, oltre che al reclutamento illegale di manodopera straniera".

Possiamo pensare ormai improrogabile una soluzione urgente in termini di chiarezza dei contratti di lavoro.

3. Il corpus e la nominalizzazione

In questa sede si riportano alcuni dei risultati del Progetto Prin 2022 *Migrants, Institutions, Translations Easy-read law,* che si propone di focalizzare l'attenzione proprio sul mondo del lavoro, sui contratti di lavoro inaccessibili per gli italiani e ancor più "impenetrabili" per gli immigrati. Nello specifico siamo partiti dal Contratto Collettivo Nazionale di Lavoro (d'ora in avanti CCNL), cioè dall'"Accordo stipulato fra un'associazione di lavoratori e un datore di lavoro (o associazione di categoria) mediante il quale vengono stabilite le regole generali riguardanti il trattamento economico e normativo cui devono uniformarsi i contratti individuali di lavoro (art. 39 Cost.)[5].

I contratti collettivi sono contratti di diritto comune, regolati dal Codice Civile, e vincolano gli iscritti alle associazioni sindacali che li stipulano.

I CCNL, che durano in genere quattro anni, contengono 2 parti:

- la parte normativa, con le tabelle retributive e le regole fondamentali del rapporto di lavoro (orario, permessi, straordinario, ferie, ecc.);
- la parte obbligatoria, con le regole che andranno a disciplinare i futuri rapporti tra le controparti (collettive) del contratto, cioè i sindacati e le associazioni di imprenditori firmatarie dello stesso.

Tra le peculiarità (lessicali, morfosintattiche) del testo giuridico si è scelto di analizzare la nominalizzazione lessicale[6], ossia quella che Castelli ha definito come "un nome morfologicamente derivato, inserito in un SN che esprime [sinteticamente] il contenuto di una intera frase" (Castelli 1988: 333), e che, come osserva Cortelazzo, è "il tratto sintattico più rilevante" per le lingue specialistiche, in generale (Cortelazzo 1990: 17), e per la lingua giuridica, potremmo aggiungere, in particolare. Il fenomeno si basa soprattutto sul passaggio da un aggettivo o un verbo, che indica un processo (espresso dal modo, dal tempo, a volte, dalla persona verbale[7]) ad un nome astratto - che non sempre, tuttavia, riguarda il lessico giuridico, quello espressamente tecnico-specialistico – preferito per il principio di universalità cui ambisce il testo formale, per garantire un buon livello di astrazione e sintesi. Bice Mortara Garavelli (2001: 172) osserva: "L'uso degli astratti è tipico di ogni discorso che verta su principî, su categorie piuttosto che su accadimenti particolari, su definizioni prima che sulle entità definite, su caratteristiche, generali o particolari, prima che su individui e oggetti caratterizzati: come accade in tutti i rami del sapere dove si trovino enunciazioni teoriche, dove si formulino criteri generali", e "la scrittura giuridica è il trionfo dell'astrazione" (Mortara-Garavelli 2001:

5 https://www.brocardi.it/dizionario/2432.html. Cfr. per la definizione di CCNL anche Carnelutti "un ibrido che ha il corpo del contratto e l'anima della legge", poiché "attraverso il meccanismo contrattuale gioca una forza che trascende il diritto soggettivo, e si sprigiona un movimento che va oltre il rapporto giuridico tra le parti" (Carnelutti 1927: 108).

6 Non vengono esaminate le nominalizzazioni sintattiche che di solito utilizzano forme indefinite dei verbi (ad esempio, gli infiniti).

7 Il verbo può anche essere espresso nei modi indefiniti, molto frequenti in testi specialistici, come quelli del diritto.

171). Riccardo Orestano parla di una suggestione quasi aristotelica, di una volontà di definire, di associare a ogni sostantivo una "«definizione» che riveli la natura o essenza di ciò che è designato con quella parola, una «definizione» che colga conoscitivamente ciò che contraddistingue la sua sostanza ontologica e con ciò stesso il suo carattere identico" (Orestano 2008: 99).

4. Alcune analisi

In questa sede si focalizzerà l'attenzione sul ruolo della nominalizzazione nel corpus Contratti Collettivi Nazionali di Lavoro (CCNL) relativi ai settori dell'agricoltura, edilizia, turismo e lavoro domestico, settori in cui vengono impiegati maggiormente gli immigrati[8], come si può osservare in figura (Fig. 1):

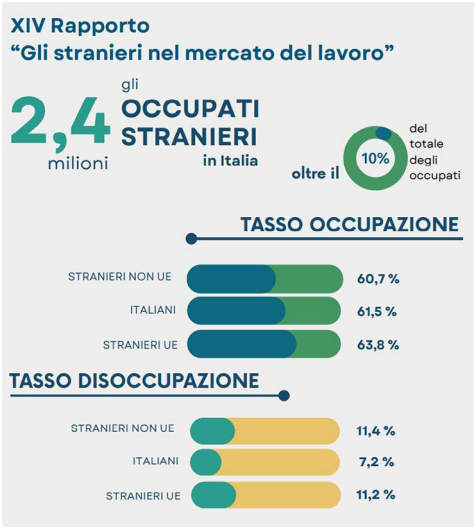

Figura 1: (da: https://www.integrazionemigranti.gov.it/AnteprimaPDF.aspx?id=6304)

In particolare, si calcoleranno la densità semantico-informativa e le frequenze delle strutture nominali, nonché le loro co-occorrenze per individuare possibili reti di relazioni che alcuni nomi stabiliscono con altri costituenti frasali, contribuendo così all'intensificazione dell'astrattezza degli enunciati e alla complessità comunicativa. L'obiettivo è quello di provare se è sempre possibile una riscrittura – con cancellazione/sostituzione di alcune catene nominali – al fine di rendere i contratti più accessibili, pur mantenendone la validità giuridica.

8 Cfr. XIV RAPPORTO ANNUALE a cura del Dipartimento per le politiche sociali, del terzo settore e migratorie. Direzione Generale dell'Immigrazione e delle Politiche di Integrazione, 2024.

I nostri testi[9] presentano, tutti e quattro, un'alta densità semantica e una complessità sintattica che è pari al 100% per i CCNL di Turismo e Agricoltura, mentre per Edilizia si registra un 99,7% e per Badanti e collaboratori domestici un 98,2%.

Tra le parole chiave più frequenti, secondo le attese, si registrano:

LEMMA	OCCORRENZE
LAVORATORE	1.096
LAVORO	877
CONTRATTO	433
PERIODO	426
AZIENDA	375

Tabella 2

Per incontrare i verbi – la cui funzione in certi tipi di testo è indebolita – nella lista di frequenza bisogna arrivare a *corrispondere* che registra 145 occorrenze, seguito da *svolgere* con 124 occorrenze, *stipulare* con frequenza 108. I verbi che registrano maggiori frequenze d'uso sono *essere* con 559 esiti per la forma della terza persona dell'indicativo presente sia come ausiliare che come copula, con 114 esiti nella forma di futuro e *dovere* con 106 occorrenze alla terza persona dell'indicativo futuro. Basti osservare il diagramma della *co-word analysis*, ottenuto con T-Lab[10], in cui vengono rappresentate 100 parole chiave del corpus e le loro reti di relazione all'interno del testo, considerando che la distanza tra le bolle e la distribuzione nei differenti quadranti indicano il loro grado di prossimità:

9 I singoli CCNL si differenziano per lunghezza e pesi diversi: tuttavia questo dato non inficia i risultati dei ragionamenti qui riportati. Nello specifico i pesi in termini statistici sono i seguenti: Lavoro edile 36,88%; Turismo 28,14%; Badanti 10,43%; Agricoltori 24,55%.

10 Ricordiamo che T-Lab è un software utilizzato per moltissimi campi, da quelli scientifici (medicina, fisica biomedica) a quelli umanistici (sociologia, letteratura, linguistica). Il programma mette a disposizione un insieme di strumenti linguistici, statistici e grafici per l'analisi dei testi. Con i suoi pacchetti, nello specifico, è possibile fare un'analisi di *text mining*: dalla tokenizzazione alla lemmatizzazione (come primo passo fondamentale per rendere un testo pulito e accessibile alla lettura automatica), alla fase di *processing*, attraverso l'individuazione delle *stop words* (e loro rimozione), l'eliminazione degli spazi bianchi, fino a giungere alla riduzione del testo in una *bag-words* (in cui saranno state individuate anche le polirematiche, trattate come singole parole). Si possono misurare le frequenze delle parole del vocabolario così ottenuto e distinguere i lemmi in relazione alla categoria grammaticale di appartenenza. Inoltre, si possono costruire matrici per calcoli statistici, che in un'analisi multidimensionale permettono di fare, tra l'altro, l'analisi delle corrispondenze (oltre all'analisi dei cluster e alle reti semantiche).

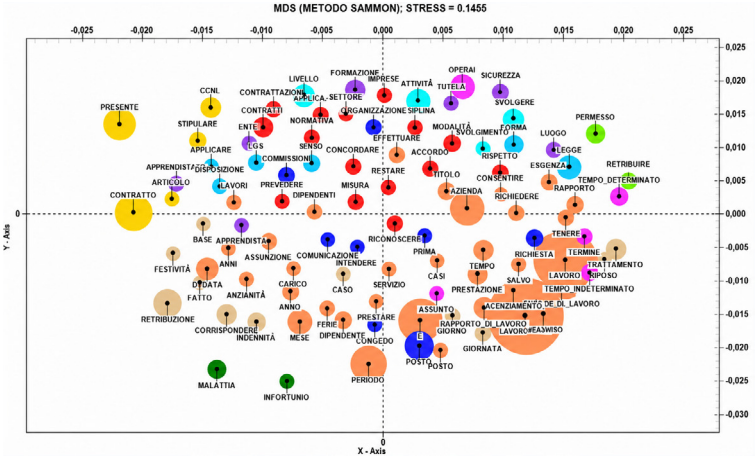

Figura 2: Co-word analysis.

In particolare, le bolle, più grandi, di color salmone nel quadrante inferiore destro rappresentano le parole più frequenti e un cluster molto denso. Il verbo *corrispondere* (bolla beige) si trova nel quadrante inferiore di sinistra e dalla sua grandezza è facile evincere quanto non sia rilevante il suo peso (in termini statistici) rispetto non solo ai nomi presenti nel quadrante destro, ma anche rispetto a quelli riportati negli altri due quadranti superiori[11]. Così com'è intuitiva la scarsa presenza di verbi, in generale, tra i 100 lemmi totali che vengono qui rappresentati.

Questi dati confermano i valori registrati sotto la voce *Profondità media di strutture nominali complesse* di Read.it[12] per i singoli testi CCNL e relativi alle strutture nominali complesse ossia alle strutture "costituite da una testa nominale modificata da aggettivi e/o complementi preposizionali: [per cui] un'alta frequenza di "catene" particolarmente "pesanti" di modificatori nominali rappresenta senza dubbio un fattore di complessità sintattica"[13]:

11 La sovrapposizione delle bolle indica, insieme a uno stesso colore, una vicinanza molto stretta tra le parole. Colori differenti e distribuzione in differenti quadranti ne indicano la distanza.

12 Read.it è il programma sviluppato dall'Italian Natural Language Processing Laboratory (ItaliaNLP Lab) dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa, che fino a oggi risulta, come osservano Dominique Brunato e Giulia Venturi (2016), l'unico strumento in grado di valutare la leggibilità dei testi italiani e individuare i luoghi di maggiore complessità lessicale e/o sintattica.

13 http://www.italianlp.it/wp-content/uploads/2016/01/Documentazione-READ-IT.pdf

CCNL	Profondità media di strutture nominali complesse
Edilizia	1,527
Turismo	1,482
Badanti	1,486
Agricoltori	1,457

Tabella 3

4.1 Assunzione

Da una prima analisi delle occorrenze del deverbale *assunzione*[14] è emerso che, accanto ai casi di possibile riscrittura con una frase verbale, sebbene con un cambio di prospettiva semantica, ci sono anche casi di nominalizzazione refrattaria a una possibile rielaborazione.

Partiamo dai casi in cui è possibile sostituire il nome con un verbo:

1) ove non espressamente previsto in sede di ASSUNZIONE[15]

Riscr.: dove non è espressamente previsto *quando il lavoratore viene assunto*

In questo caso il lavoratore diventa oggetto-logico, il paziente di un verbo che indica un atto e non un evento astratto.

2) Le Parti convengono che in caso di ASSUNZIONE di un lavoratore a tempo indeterminato

Riscr.: le Parti convengono che se si assume un lavoratore a tempo indeterminato

Anche qui la locuzione può essere sostituita con una condizionale *se si assume.*

Ci sono, invece, casi, come quello seguente, in cui la rielaborazione del nominale toglierebbe incisività al testo:

3) inquadrato al medesimo livello iniziale di ASSUNZIONE dell'apprendista

Se si provasse a rendere *assunzione* con una proposizione relativa si perderebbero efficacia e senso: *inquadrato allo stesso livello iniziale di quello al quale si assume un apprendista.* Le due proposizioni inserite nella riscrittura non soltanto «diluiscono» la concisione del concetto originale, ma ne fanno perdere l'astrazione, svuotando, probabilmente, di quella pretesa «ufficialità» il testo.

Non mancano strutture che presentano un blocco sintattico che ne impedisce la riformulazione immediata:

4) in caso di nuova ASSUNZIONE con contratto a tempo parziale

14 I dati e le analisi complete sono stati presentati in occasione del Convegno *Il nome in italiano. Morfologia, sintassi e testualità*, tenutosi a Losanna, 3-4 novembre 2025.

15 I contesti di occorrenza dei lemmi oggetto di analisi sono riportati così come vengono resi da T-Lab, utilizzando la funzione *Concordanze*.

L'aggettivo qualificativo, sia che preceda sia che segua il nome, ostacola la resa di una frase verbale.

Nel frammento successivo, invece, si potrebbe tentare una rielaborazione sacrificando nuclei semantici e scendendo a compromessi (non sempre, tuttavia, risolutivi):

5) servizi speciali in occasione dei quali è consentita l'ASSUNZIONE diretta di manodopera per una durata non superiore a tre giorni

Riscr.: *servizi speciali in occasione dei quali è consentito assumere direttamente manodopera per una durata non superiore a tre giorni.*

A parte il determinante *diretta* che diventa avverbio, *manodopera* che passa a oggetto, si profila il mutamento della forma verbale da passiva a impersonale è consentito. Il concetto, in linea generale, rimane, con prospettive differenti: il focus si sposta dall'atto, *l'assunzione*, all'agire *consentire.* Quello che sembra rimanere immutato è la spersonalizzazione della concessione: non è esplicitato mai l'agente.

4.2. Richiesta

Proviamo ad analizzare *richiesta* (nel corpus si contano 117 occorrenze). Dal grafico relativo all'*Analisi delle sequenze* di T-Lab, che fornisce una rappresentazione delle associazioni di parole – lo spessore dei segmenti è direttamente proporzionale all'indice di prossimità e frequenza, mentre per quanto riguarda i colori: il blu indica costituenti che precedono e l'arancione quelli che seguono il lemma – tra tutte le parole che cooccorrono con *richiesta* solo sei sono verbi: *presentare, provvedere, anticipare, contenere, è* (nella forma ausiliare), *intendere.*

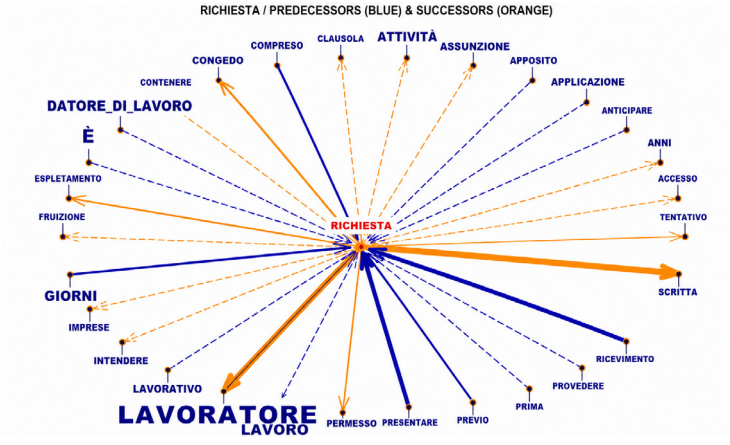

Figura 3: Analisi delle sequenze.

Proviamo, ora, ad analizzare alcune frasi, grazie allo strumento *Concordanze* di T-Lab. Partiamo dalla semplice frase:

1) su richiesta del lavoratore

Riscr.: se il lavoratore lo richiede

Come si può osservare con la riscrittura si passa dalla forma astratta, densa di ufficialità della struttura nominale, ad una forma più esplicita e meno formale, che sottolinea la condizione che l'evento può verificarsi solo se e quando il lavoratore lo richiede. Insomma, come osserva Daniele Fortis, la nominalizzazione, come il *si* impersonale e il passivo, è una risorsa che viene utilizzata con «l'intento di conferire al documento un tono ufficiale, autorevole, solenne. Si cerca di mettere il più possibile in ombra gli esseri umani e di suscitare l'impressione che a parlare sia non la persona, bensì l'istituzione» (Fortis 2005: 69).

Richiesta si può sostituire senza che si verifichino alterazioni sintattico-semantiche, semplificando gli pseudo-tecnicismi formali *fare/formulare richiesta* in: il lavoratore *che lo richiede*.

Così come per i seguenti frammenti:

2) L'azienda può concedere, oltre a quelli già previsti dalla normativa vigente, periodi di aspettativa al lavoratore che ne faccia RICHIESTA e che abbia almeno quattro anni di anzianità per comprovate e riconosciute necessità personali o familiari per un periodo massimo di sei mesi.

3) può riguardare un numero di giornate maturato dal cedente non superiore a 10 nell'anno. I lavoratori interessati e in possesso dei requisiti di cui all'art. 24 del d.lgs. n. 151/2015 e nel rispetto delle condizioni sopraindicate possono formulare al datore di lavoro la RICHIESTA di ferie solidali per un massimo di 30 giorni l'anno.

O ancora meglio si può riscrivere agevolmente, con tecnica mininvasiva, il periodo seguente:

4) L'azienda è tenuta, non oltre i 10 giorni dalla RICHIESTA del congedo, a esprimersi sulla stessa e a comunicarne l'esito al dipendente. L'eventuale non accoglimento, la proposta di rinvio ad un periodo successivo e determinato

Riscr.: L'azienda è tenuta a esprimersi e a comunicarne l'esito al dipendente, non oltre i 10 giorni da quando è richiesto il congedo […]

La riscrittura, inoltre, risulta «senza controindicazioni» quando *richiesta* è seguito da un argomento preposizionale che diventa argomento diretto: *richiesta di ferie, richiesta del congedo* > *richiedere il congedo, richiedere le ferie*.

Così come si possono semplificare: *su richiesta, a richiesta* con delle condizionali *se lo richiede*, *ne fa richiesta* > *lo richiede*. Si potrebbero, inoltre, «abbattere» molti sintagmi:

5) Il datore di lavoro in caso di mancata RICHIESTA di fruizione dei riposi compensativi per le ore depositate in banca ore potrà, entro il 31 dicembre di ogni anno (Indice Gulpease[16] 46)

Riscr.: Se il lavoratore non chiede di utilizzare le ore della banca ore, il datore di lavoro potrà (Indice Gulpease 62).

In questo frammento il sovraccarico nominale è catalizzato da *richiesta* modificata dal participio *mancata*, preceduta da *in caso di,* e seguita da una nominalizzazione a cascata. Probabilmente, in questo caso il focus della condizionale sarà ancora una volta sull'azione e non sull'atto, ma i vantaggi in termini di leggibilità calcolati con l'Indice Gulpease sono alquanto apprezzabili: si passa dal testo originale che risulta di media leggibilità (valore pari a 46), rivolto a persone che hanno un'istruzione superiore, ad un testo facilmente leggibile da parte di chi ha una licenza media (valore pari a 62).

Ma non sempre si può ricorrere a forme più accessibili, snelle e più aderenti al contesto, senza perdere parte delle informazioni. Infatti, se provassimo a proporre il concetto riformulato del frammento seguente:

6) dal deposito del ricorso nella cancelleria del tribunale in funzione di giudice del lavoro o dalla comunicazione alla controparte della RICHIESTA di tentativo di conciliazione o arbitrato, ferma restando la possibilità di produrre nuovi documenti formatisi dopo il deposito del ricorso.

Potremmo avere, ad esempio:

Riscr.: Da quando il ricorso viene presentato in tribunale o si chiede all'altra parte di tentare la conciliazione o l'arbitrato, non si possono più aggiungere nuovi documenti, a meno che siano documenti creati dopo quel momento.

Il sintagma ad alta densità lessicale, sì, verrebbe semplificato, ma si rinuncerebbe a *richiesta di conciliazione* che sembra uno specialismo reso per analogia a "*tentativo* di conciliazione che il giudice compie allo scopo di pervenire al componimento della controversia, sì da determinare la cessazione della materia del contendere"[17], o su calco di *verbale di conciliazione*. Inoltre, riformulando si cambia il punto di vista, il focus viene trasferito dalla *richiesta* al verbo *tentare*, ossia all'azione: si passa da un fatto ad un processo.

Non c'è, invece, possibilità di riscrittura quando *richiesta* è seguita da participio con funzione aggettivale a meno che, come nel caso seguente, non si utilizzi il sintagma *richiedere per iscritto*:

16 L'indice, il primo tarato sulla lingua italiana, verifica la comprensibilità dei testi attraverso la formula Gulpease: dove LP è il numero di lettere in 100 parole e FR è il numero di frasi in 100 parole. I valori restituiti dalla formula sono messi in relazione con la leggibilità correlata al grado di scolarizzazione del lettore. Cfr. pro.corrige.it/ortografia/lindice-gulpease.

17 https://dizionari.simone.it/1/conciliazione

7) giorni prima, RICHIESTA scritta al datore di lavoro

Né la rielaborazione si rende praticabile se *richiesta* è seguita da participi passati con funzione di relativa implicita:

8) Durante il periodo di concertazione sarà compito delle OO.SS. firmatarie del contratto valutare l'esistenza o meno delle motivazioni che hanno indotto l'azienda a richiedere un "Accordo per Stato di Crisi".

Le OO.SS. potranno chiedere una sola proroga di 15 giorni di calendario nel caso che intendano ulteriormente approfondire la sussistenza dei motivi oggettivi della RICHIESTA avanzata dall'azienda.

Riscr.: I sindacati possono chiedere una sola proroga, di 15 giorni, se vogliono approfondire i motivi oggettivi per i quali l'azienda ha fatto la richiesta.

Se si vogliono mantenere le due fasi distinte, quella di valutare "l'esistenza o meno dello Stato di Crisi" e il successivo, ulteriore accertamento, come si può osservare, la frase non accoglie la riformulazione di *richiesta* del secondo periodo in una forma del verbo *richiedere* così com'è incassata in un susseguirsi a cumulo di sintagmi nominali innescati in una sorta di reazione a catena, dal primo sostantivo *sussistenza,* che tuttavia non costituisce l'unico ostacolo per una possibile rielaborazione. Si pensi ad *approfondire* che nell'esplicitazione reggerebbe una relativa. La frase dovrebbe essere anticipata da una condizionale introdotta da *se* e realizzata con un più comune *volere* al posto di *intendere*.

Visto che l'AI in studi anglosassoni ha dato risultati interessanti[18], è stato fatto un tentativo di riscrittura anche interpellando ChatGpt5, utilizzando un prompt molto semplice: "Semplifica il frammento, rendendolo più semplice e lineare". Si è ottenuto:

Riscr.: Durante il periodo di concertazione, le OO.SS. firmatarie del contratto valuteranno se sussistono le motivazioni che hanno spinto l'azienda a richiedere un Accordo per Stato di Crisi. Le OO.SS. potranno chiedere una sola proroga di 15 giorni di calendario se riterranno necessario approfondire ulteriormente tali motivazioni.

Come si può osservare, la rielaborazione è totale e comprende l'intero frammento di testo, l'intervento – per evitare *richiesta* – non è stato puntuale, non ha potuto interessare il singolo lemma. Inoltre, ancora una volta, la riscrittura porta a focalizzare l'attenzione sull'azione, perdendo in termini di tono denso, conciso e sentenzioso, di pretesa ufficialità: nel secondo periodo viene oscurata l'*azienda* e la sua *richiesta*: sono solo le OO.SS. che possono richiedere una proroga per "approfondire ulteriormente tali motivazioni". Inoltre, cade l'oggettività della motivazione (motivi oggettivi)

18 Basti pensare ai risultati riportati in McMinn, D. *et al.*, "Seeing it in the Flesch: Comparing Readability Between AI-Generated and Human-Written Plain-Language Abstracts", *Value in Health*, Volume 26, Issue 12, S11, per cui l'AI produce testi più leggibili rispetto a quelli prodotti dall'intelligenza umana.

dell'originale. Un compromesso che meriterebbe di essere testato presso i reali utenti, i lavoratori, gli immigrati.

6. Applicazione

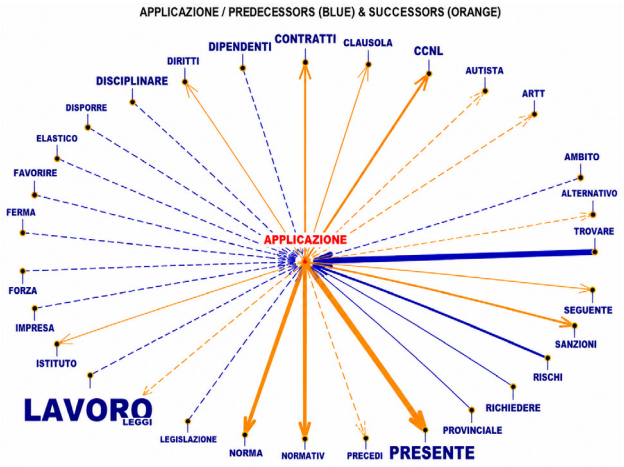

Figura 4: Analisi delle sequenze.

Proviamo a leggere i risultati per *applicazione* (83 occorrenze). Anche in questo caso la percentuale di verbi che hanno probabilità di occorrere con *applicazione* è bassa. Sono: *richiedere, trovare, disciplinare, disporre* e *favorire*.

Dalle analisi è risultato che il lemma può essere sostituito nelle frasi in cui occorre come sintagma preposizionale: *in applicazione* degli articoli > applicando, negli pseudo-tecnicismi *trova applicazione* > *si applica*, *nella sfera/nel campo/nell'ambito di applicazione del presente CCNL* > *là dove si applica il presente CCNL*.

O in frasi come la seguente, dove può occorrere un'interrogativa indiretta:

1) i lavoratori, mediante loro rappresentanze aziendali, ovvero, in mancanza di queste, mediante i rappresentanti delle organizzazioni sindacali firmatarie del CCNL hanno diritto di controllare l'APPLICAZIONE delle norme per la prevenzione degli infortuni e delle malattie professionali e di promuovere la ricerca

Riscr.: Hanno diritto di verificare se le norme per la prevenzione degli infortuni e delle malattie professionali *vengono applicate*, e di promuovere la ricerca

In altri contesti, invece, è necessario fare qualche ulteriore sostituzione:

2) sostegno al reddito o contratto di solidarietà, che interessano lavoratori adibiti alle mansioni cui si riferisce la stipula del contratto di cui al presente articolo; c. ai datori di lavoro che non hanno effettuato la valutazione dei rischi in APPLICAZIONE della normativa di tutela della salute e della sicurezza dei lavoratori.

Qui *in applicazione della normativa* non può essere riscritto con una forma del verbo *applicare*, ma con una proposizione modale del tipo *come prevede/richiede la normativa*. Ancora una volta, si evita l'astrattezza, il "tono sentenzioso" di disposizioni imparziali, per una minore densità formale, una comunicazione più dinamica che sposta il focus dell'attenzione: dall'applicazione alla norma e da un'azione (applicare) ad un dato (la normativa). Insomma, per citare Eco, si dice *quasi* la stessa cosa.

Anche in casi in cui il nome sia modificato da due aggettivi la riscrittura, dal punto di vista sintattico, perde di concisione ed efficacia, mantenendo tuttavia lo stesso grado di spersonalizzazione affidato all'uso degli infiniti:

3) Se la controversia discende dal riconoscimento del profilo professionale in rapporto alle mansioni effettivamente svolte dal lavoratore e dalla mancata od erronea APPLICAZIONE dell'art. 31 del presente contratto, il tentativo di amichevole componimento sarà espletato con l'assistenza di due esperti

Riscr.: Se la controversia deriva dal non aver riconosciuto il profilo professionale in base al lavoro che il dipendente ha svolto realmente o *dal non aver applicato o applicato in maniera erronea* l'articolo 31 di questo contratto, si tenterà una soluzione amichevole con l'aiuto di due esperti.

In altri casi, invece, la riscrittura sembra non praticabile:

4) o PEC, di un periodo di aspettativa debitamente certificato di durata non superiore a 6 mesi durante il quale non decorrerà retribuzione. 2. A tal fine la direzione aziendale, su richiesta del dipendente o dei suoi familiari, fornirà lo stato di APPLICAZIONE della norma contrattuale.

A meno che non si accetti una riformulazione in cui il concetto non coincide con quello del testo originale: la direzione dell'azienda informerà *su come quella norma del contratto viene applicata*. Qui il *modo* di applicazione non può sostituire sinonimicamente lo *stato*.

Altre volte se si volesse intervenire con una rielaborazione del testo, incorreremmo in alterazioni molto profonde del significato, con pericolose ambiguità. Per esempio, se tentassimo una riscrittura di:

5) Sono lavoratori discontinui o di semplice attesa o custodia quelli sotto indicati ai quali non sia richiesta un'APPLICAZIONE assidua e continua:

Riscr.: i lavoratori discontinui o di semplice attesa o custodia sono quelli *ai quali non si richiede di applicarsi in maniera assidua e continua.*

Il significato sarebbe stravolto, in quanto qui l'estensore vorrebbe intendere che *al lavoratore non si chiede di lavorare in maniera costante e continua* e non che non deve impegnarsi, non deve applicarsi all'attività che svolge.

7. Conclusioni

Dall'analisi delle occorrenze dei pochi lemmi analizzati e dalle osservazioni di carattere sintattico-semantico che sono state fatte si può notare come – probabilmente – non si possano fare generalizzazioni sulle nominalizzazioni né in termini di preferenze, né in termini di vantaggi. Se è vero che, come si legge nel *Manuale di Stile* del 1994, il linguaggio amministrativo, ma potremmo aggiungere, anche quello giuridico, è utilizzato più nell'ottica della legittimità formale degli atti che in quella, sostanziale, della comprensibilità del messaggio" (1994: 20) e si riporta tra i caratteri che "segnano la distanza dalla lingua comune" la preferenza di nomi derivati da verbi (bonifico, subentro, storno, condono), si può, tuttavia, osservare proprio a proposito della nominalizzazione che è in parte accettabile quanto sostiene Dolorosa Maria Ross, ossia che la struttura sintattica astratta "comporta una semplificazione della sintassi e una schematizzazione del periodo. Offre vantaggi in termini di coesione testuale, perché è più semplice far riferimento a sintagmi nominali nel prosieguo del discorso, e facilita "la progressione a livello testuale" (Gotti 1991: 78)" (Ross 2004: 116). A proposito anche Elisa Mattiello osserva che

> in spite of the abstractness that nominalization obtains by the use of non-concrete agents (e.g. the protection of human rights), and in spite of the difficulty of disambiguation it provokes because of the richness and semantic density of complex NPs, [….], it gives actions a stable semantic reference (e.g. developing the overall policy "policy development"), thus facilitating textual progression (Gotti 1991: 78) and assuring continuity and stability of reference. (Mattiello 2010: 130).

Si è quindi rilevato come non sempre si possano realizzare commutazioni di frasi nominali con frasi verbali, pena non soltanto l'agrammaticalità di alcune soluzioni, ma anche la perdita del significato originale del contenuto. Non sempre l'astrattezza veicolata da forme nominali complesse dà adito a concetti vaghi ed ambigui, tutt'altro.

La rielaborazione di alcune frasi o periodi dei CCNL esaminati ha dimostrato come, talvolta, faccia perdere quella *vis* giuridica propria della testualità contrattuale con la quale bisogna fare i conti e scendere – quando possibile – a compromessi: meglio un'asserzione forte che presenta un'alta complessità di decodifica, visto che come osserva Sabatini (2011) la preferenza per le costruzioni nominali costituisce un fattore di maggiore complessità linguistica[19], o un messaggio meno incisivo, più ridondante, ma piano in termini di leggibilità? Purtroppo, la scelta non è così pacifica e scontata come sembrerebbe a un'analisi delle strutture superficiali e così come spesso si consiglia nelle ormai innumerevoli guide stilate per la semplificazione dei testi giuridico-amministrativi che vorrebbero testi assemblati sulla base di liste di istruzioni[20]. Non sempre le soluzioni alternative forniscono risultati migliori in termini di leggibilità, linearità e fluidità sintattica: spesso le alternative sono l'uso di impersonali, infiniti, passivi deagentivizzati, tutte risorse che non giovano alla chiarezza, ma che dovrebbero essere

19 Cfr. Sabatini 2011.

20 Cfr. Frosali 2025.

evitate. Pertanto le decisioni vanno ponderate e prese considerando i singoli casi: riscrivere quando la leggibilità non compromette, anzi concorre a migliorare i concetti, evitare là dove la riscrittura comporterebbe perdita semantica, metterebbe a rischio la grammaticalità e, ancor di più, la validità giuridica. Sebbene occorra tener conto dell'osservazione di Frosali per cui "l'obiettivo di produrre testi ad alta leggibilità è un risultato auspicabile ma assolutamente ancillare, essendo la legittimità e la precisione il primo e fondante motivo dell'esistenza stessa dell'enunciazione burocratica" (Frosali 2025: 141), ma soprattutto giuridica, vale la pena, tuttavia, ricordare che la riformulazione linguistica non è un fatto meramente estetico, bensì il "fondamento della stessa tecnica legislativa"[21] che deve garantire, nel rispetto del rigore e della precisione, il diritto all'informazione di tutti i cittadini.

Bibliografia

Barettoni Arleri, Alberto. *Relazione della Commissione di studio per la semplificazione delle procedure e la fattibilità e l'applicabilità delle leggi nonché l'approntamento dei conseguenti schemi normativi*, Camera dei Deputati, 7 giugno 1981.

Brunato, Dominique, e Venturi, Giulia. *Le tecnologie linguistico-computazionali per la leggibilità della comunicazione istituzionale*. In S. Panizza (a cura di), *Profili attuali di qualità degli atti normativi e amministrativi*. Pisa: University Press, 2016, pp. 119-157.

Callon, Michel, Law, John e Rip, Arie. *Mapping the Dynamics of Science and Technology*. Londra: Palgrave Macmillan, 1986.

Carnelutti, Francesco, *Teoria del regolamento collettivo dei rapporti di lavoro*. Padova: Cedam, 1927.

Cassese, Sabino. *Codice di Stile*. Presidenza del Consiglio dei Ministri. Dipartimento per la Funzione Pubblica. Istituto Poligrafico e Zecca dello Stato. Libreria dello Stato, 1993.

Castelli, Margherita. *La nominalizzazione*. In G. Renzi, Salvi & A. Cardinaletti (a cura di), *Grande grammatica italiana di consultazione*. Bologna: il Mulino, vol. 1° (1988), pp. 333-356.

Cortelazzo, Michele. *Lingue speciali: la dimensione verticale*. Padova: Unipress, 1990.

Dipartimento per le politiche sociali, del terzo settore e migratorie - Direzione Generale dell'Immigrazione e delle Politiche di Integrazione. *XIV Rapporto Annuale, Gli stranieri nel mercato del lavoro in Italia,* 2024.

Eggins, Suzanne. *An introduction to systemic functional linguistics*. London: Pinter, 2004[2].

Fabricius-Hansen, Cathrine. "Informational density: a problem for translation and translation theory". *Linguistics*, vol. 34, n. 3 (1996): 521-566. https://doi.org/10.1515/ling.1996.34.3.521.

Fabricius-Hansen, Cathrine. *Informational density and translation, with special reference to German – Norwegian – English*. In: S. Johansson, S. Oksefjell (a cura di) *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi, 1998: 197-234.

21 https://www.parlamentiregionali.it/wp-content/uploads/2024/03/Manuale-di-tecnica-legislativa-per-le-Regioni_11.03.2024.pdf, p. 16.

Fioritto, Alfredo. *Semplicità e semplicismo nell'attività amministrativa,* in M. E. Piemontese (a cura di), *Il dovere costituzionale di farsi capire*. Roma: Carocci, 2023: 123–149.

Fioritto, Alfredo. *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Presidenza del Consiglio dei Ministri. Dipartimento della Funzione Pubblica. Bologna: Il Mulino, 1997.

Fortis, Daniele. "Il linguaggio amministrativo italiano". *Revista de Llengua i Dret*, n. 43 (2005), pp. 47-117.

Frosali, Alessio. "Semiotica e semantica del Burocratese: non solo lessico". *Qulso* 11 (2025): 129-146. http://dx.doi.org/10.13128/QULSO-2421-7220-18665.

Gotti, Maurizio. *I linguaggi specialistici: Caratteristiche linguistiche e criteri pragmatici*. Firenze: La Nuova Italia, 1991.

Kusters, Wouter. *Linguistic complexity: The influence of social change on verbal inflection*. Utrecht: LOT, 2003.

Mattiello, Elisa. "Nominalization in English and Italian normative legal texts". *ESP Across Cultures*, Special issue: *Legal English across cultures* a cura di M. Gotti and C. Williams (2010): 129-146. https://edipuglia.it/wp-content/uploads/2020/12/Indice7.2010.pdf#page=125

McMinn, David, Valena,Tomas, e Bender, Walter. *Seeing it in the Flesch: Comparing Readability Between AI-Generated and Human-Written Plain-Language Abstracts. Value in Health*, vol. 26, n. 12 (2023): S11.

Mortara-Garavelli, Bice. *Le parole e la giustizia*. Torino: Einaudi, 2001.

Orestano, Riccardo. "Il problema delle fondazioni in diritto romano". *Rivista di Diritto Romano*, Led on line, vol. VIII (2008): 1-140.

Ross, Dolorosa Maria. *Profili morfologici della lingua medica: contrasti in ambito germanico-romanzo*, 2004. In https://www.openstarts.units.it/server/api/core/bitstreams/14184727-7b19-425a-bc3a-ed9294dbc2c4/content

Sabatini, Francesco. *L'italiano nel mondo moderno*. Tomo II, Napoli: Liguori, 2011.

Sutherland, Carolyn. *The Elusive Quest for Simplicity: Measuring and Assessing the Readability of Enterprise Agreements*, 1993 to 2011 (March 1, 2013). In https://ssrn.com/abstract=2296749

Ure, Jean. *Lexical density and register differentiation*. In G. Perren e J. L.M. Trim (a cura di), *Applications of Linguistics*. London: Cambridge University Press, 1971, pp. 443-452.

Vena, Debora. *Leggibilità e comprensibilità: un binomio necessario. Italiano LinguaDue*, vol. 14, n. 1 (2022): 471–491. https://doi.org/10.54103/2037-3597/18298.

Sitografia

http://www.italianlp.it/wp-content/uploads/2016/01/Documentazione-READ-IT.pdf

https://dizionari.simone.it/1/conciliazione.

https://www.brocardi.it/dizionario/2432.html

https://www.integrazionemigranti.gov.it/AnteprimaPDF.aspx?id=6304

https://www.parlamentiregionali.it/wp-content/uploads/2024/03/Manuale-di-tecnica-legislativa-per-le-Regioni_11.03.2024.pdf.

https://www.parlamentiregionali.it/wp-content/uploads/2024/03/Manuale-di-tecnica-legislativa-per-le-Regioni_11.03.2024.pdf.

Metodi e tecniche per modernizzare i testi della pubblica amministrazione

Francesco Romano
(CNR/IGSG)

Abstract

L'intervento, dopo aver passato in rassegna brevemente le metodologie, le tecnologie, i progetti sviluppati negli anni dal CNR per la modernizzazione dei testi istituzionali, illustrerà un metodo collaborativo per la semplificazione di tali documenti sviluppato nell'ambito del progetto PRIN "Migrants, Institutions, Translations Easy-read Law". Il linguaggio del diritto, infatti, seppure annoverabile tra i linguaggi specialistici, è uno di quelli più rilevanti dal punto di vista sociale. Come noto, la sua oscurità può creare esclusione sociale eliminando la possibilità pratica per cittadini e cittadine di conoscere i propri diritti e doveri e accedere a servizi. Per questo motivo nell'ultima parte del saggio sono anche riportate le molte attività del Consiglio nazionale delle ricerche rivolte ai giovani e dedicate a fornire loro strumenti per decifrare e "tradurre" il linguaggio burocratico.

Contesto

La comunicazione assume un ruolo sempre più rilevante in società nelle quali la cittadinanza «appare come una trama diversificata di formule di inclusione, di partecipazione attiva, di identità multiple». In questo scenario la «tecnologia *dell'infocommunication*» si pone come ulteriore elemento di complessità (Ceccarini 2015: 15). Anche di questi scenari deve tenere conto una pubblica amministrazione il cui processo di ammodernamento è andato di pari passo con «lo sviluppo della comunicazione pubblica» e per la quale le esigenze di trasparenza e visibilità, richieste dalla normativa, si sono intrecciate con la necessità di «comunicare di più e meglio» anche al fine di trovare una legittimazione che «nella società democratica va conquistata e riconquistata ogni giorno attraverso il dialogo» (Solito, 2018: 80).

In questo dialogo aperto con i cittadini «l'uso della lingua italiana diventa un fattore abilitante per accedere alle informazioni istituzionali, ai documenti amministrativi e ai dati pubblici». Tale lingua dovrà essere chiara e semplice oltre che per i cittadini, anche «per i dispositivi digitali in cerca di parole chiave attraverso motori di ricerca e di metadati per il web semantico» (Vellutino, 2018: 13). Tali contenuti «utili, per risolvere un problema specifico degli utenti» dovranno essere «facilmente esperibili» e condivisi tramite «l'uso di canali di comunicazione appropriati» (Talamo, Zarriello, 2022: 59).

Proprio il tema dell'accessibilità del dato giuridico, anche dal punto di vista della chiarezza della lingua del diritto e comunque della comprensibilità della comunicazione giuridica, oltre che dell'interoperabilità e dell'apertura dei dati digitalizzati, è uno dei temi che da molti anni vede impegnato il CNR e segnatamente l'Istituto di informatica giuridica e sistemi giudiziari (IGSG).

L'esperienza che è stata maturata nel corso di questi anni è classificabile nelle seguenti aree tematiche: a) la definizione di standard linguistici, che possano anche divenire standard informatici e quindi agevolare non solo la comprensibilità dei testi giuridici ma anche il loro trattamento informatico; b) la sperimentazione di strumenti informatici in grado di favorire l'adozione di tali standard; c) la formazione dedicata alla comunità dei redattori di documenti giuridici; d) la divulgazione e la sperimentazione di attività didattiche per studenti e studentesse delle scuole, in grado di fornire strumenti ai cittadini di domani per decifrare e depotenziare le insidie del linguaggio giuridico e della comunicazione delle istituzioni, anche di quella online.

Standard linguistici e informatici e strumenti per la loro applicazione

I ricercatori e le ricercatrici del CNR da molti anni sono impegnati nell'ambito di gruppi di lavoro che hanno predisposto le linee guida per la redazione degli atti normativi[1]. Le regole di *drafting* normativo, che nel corso degli anni sono state sempre più osservate dai redattori di testi legislativi, sia a livello nazionale che regionale, hanno stabilito una precisa strutturazione del testo di leggi, decreti ecc. e hanno quindi permesso di definire e applicare standard informatici per la mappatura di tali documenti e per la loro gestione digitale[2].

Il riferimento è ovviamente agli standard individuati nell'ambito del progetto NIR (Accesso alle norme in rete) che grazie al linguaggio di marcatura XML (eXtensible Markup Language) e allo standard URN (Uniform Resource Name) hanno dato vita a una serie di studi e applicazioni per la marcatura strutturale e semantica di testi legislativi e delle singole disposizioni (come le modifiche testuali esplicite) e dei nessi (come i riferimenti testuali) in essi contenuti (Marchetti et al. 2001, Spinosa 2001 e Bolioli et al. 2002). Tali metodologie sono state adoperate per la conversione in formato XML di vasti corpora legislativi anche delle regioni italiane.

Un'esemplificazione chiara degli esiti di questo processo di digitalizzazione dell'informazione normativa si ha nel Portale Normattiva, che delle metodologie e tecnologie del progetto NIR ha fatto tesoro e che oggi si pone come punto unico di accesso alla normativa italiana.

1 Per una di queste esperienze si veda (Pietrangelo 2024).

2 Per una verifica storica di come questo processo si è sviluppato nel tempo si veda il fondamentale Legimatica e redazione delle leggi (Mercatali 1995).

In precedenza, sono stati progettati e realizzati software in grado di verificare se le regole di tecnica legislativa fossero correttamente applicate nella stesura di una proposta di legge.

Gli addetti alla stesura di un testo di legge avevano, infatti, a disposizione un software come Lexedit abilitato alla correzione sintattica e stilistica dei testi normativi (Biagioli, Mercatali 1993). In seguito, la re-ingegnerizzazione del software Lexedit ha portato allo sviluppo del tool Lexedit XXI per il controllo di qualità dei testi legislativi.

Una volta terminata la fase di digitalizzazione dei documenti normativi dello Stato e delle Regioni è parso utile avviare un percorso analogo che coinvolgesse gli atti della pubblica amministrazione. Prendendo spunto proprio dal Portale Normattiva, alcuni ricercatori e ricercatrici dell'Istituto di teoria e tecniche dell'informazione giuridica (ITTIG) tentarono di sviluppare un portale prototipale che riproducesse le funzionalità di Normattiva per un determinato set di documenti amministrativi (delibere di giunta, delibere di consiglio, regolamenti comunali). La grande eterogeneità dei documenti analizzati e caricati sul portale evidenziò subito la necessità di definire chiari standard linguistici prima di poter utilizzare adeguati standard informatici in grado di mappare la struttura formale dei provvedimenti amministrativi o di riconoscere automaticamente la scrittura molto eterogenea delle citazioni di atti normativi e amministrativi presenti in tali documenti della pubblica amministrazione.

A tale fine è stato quindi proposto, nel febbraio 2011, un manuale che, oltre a sistematizzare regole già presenti in altre linee guida, ha tentato di definire una struttura standard dei provvedimenti amministrativi e delle citazioni di altri atti amministrativi. Il riferimento è alla "Guida per la redazione degli atti amministrativi. Regole e suggerimenti" (ITTIG, Accademia della Crusca 2011), manuale predisposto da un gruppo di lavoro promosso da Accademia della Crusca e ITTIG CNR che, nel corso degli anni, ha conosciuto numerose esperienze di condivisione e anche la sperimentazione di specifici software in grado di favorirne l'adozione. In particolare, nelle parti seconda e terza della Guida viene proposto uno standard per la strutturazione dei provvedimenti amministrativi ma anche una formalizzazione delle citazioni dei documenti amministrativi, in grado sia di rendere più comunicativi i documenti della PA, sia di agevolarne il trattamento informatico. Per quanto riguarda la struttura del provvedimento amministrativo lo schema proposto dalle regole della seconda parte della Guida è il seguente (Mercatali, Romano, 2013, 155-177):

PARTE INIZIALE o INTESTAZIONE
denominazione formale del tipo di provvedimento (es. decreto, ordinanza, determinazione, deliberazione ecc.)
autorità emanante (dal generale al particolare)
oggetto [funzione del provvedimento nel caso concreto)
estremi (codice di protocollo o numero assegnato dall'autorità emanante)
URN (nel caso di provvedimento in formato digitale)
PARTE CENTRALE (preambolo - motivazione - dispositivo)

a. Visto abc abc;
b. Visto def def:
  b.1. ghi ghi:
    b.1.1. lmn lmn;
    b.1.2. opq opq;
c. Visto rst rst;
d. Considerato uvz uvz;
e. Considerato xyk xyk.
verbo introduttivo (delibera, decreta, ordina...)
1. Abc abc abc.
2. Cde cde cde:
  2.1. fgh fgh fgh;
  2.2. jkl jkl jkl:
    2.2.1. abc abc abc;
    2.2.2. fgh fgh fgh.
2.3. lmn lmn.
3. Tuv tuv.
FORMULE INSERITE DOPO IL DISPOSITIVO
- clausola enunciativa del regime contenzioso
- esecutività del provvedimento
- soggetti cui il provvedimento va comunicato, notificato, trasmesso
PARTE FINALE o CONCLUSIONE
luogo
data
sottoscrizione
ALLEGATO A - rubrica (paragrafo n.).

Le regole della terza parte della Guida propongono invece uno standard innovativo per la citazione dei documenti amministrativi, mentre confermano le regole già in uso nelle regole di tecnica legislativa, per la citazione dei testi normativi presenti in un atto amministrativo.

Lo standard proposto per citare atti e provvedimenti amministrativi è quindi il seguente:

1. denominazione formale dell'atto;
2. autorità emanante;
3. data e numero di protocollo (o numero assegnato all'atto dall'ente emanante) separati da una virgola;
4. oggetto, racchiuso tra parentesi.

La formazione dei redattori di testi istituzionali

Gli standard proposti, oltre a essere stati diffusi attraverso software e strumenti informatici in grado di applicare tali standard in maniera automatica o semi-automatica

(Romano 2016), sono stati oggetto anche di un'ampia fase di formazione che vede il CNR impegnato ininterrottamente dal 2011 a oggi.

Le attività formative costituiscono infatti un momento fondamentale per diffondere gli standard linguistici e informatici nella comunità dei redattori dei testi istituzionali. Così negli ultimi anni sono stati progettati e attuati corsi di formazione basati sulle reali esigenze dei corsisti. Tali esigenze si sono concretizzate nel prevedere laboratori con i redattori degli atti progettati proprio per fare esercitare i corsisti in maniera pratica e diretta.

Al termine di queste fasi laboratoriali sono state testate le nozioni acquisite mettendo a confronto i testi prodotti dai partecipanti alle attività formative prima del corso, con quelli prodotti dagli stessi all'esito delle attività laboratoriali. Questa esperienza ha mostrato come l'attività di formazione rimanga fondamentale e abbia bisogno di ascolto preventivo dei bisogni degli utenti, di una certa flessibilità nell'uso delle raccomandazioni fatte nel corso, in base al contesto specifico e all'utenza di riferimento, di dare molto spazio ai laboratori con i redattori degli atti e, infine, di creare un processo formativo continuo che coinvolga e metta in rete i vari attori di questo processo (Fioravanti et al 2025: 13-35).

Lo sforzo profuso, con gli strumenti e le attività brevemente descritte, potrebbe far pensare che la comprensibilità e l'accessibilità dei testi delle istituzioni non costituiscano più un problema. In realtà, più di un autore nota come, in relazione agli sforzi profusi negli ultimi anni, i risultati raggiunti siano stati piuttosto modesti (Lubello 2025: 39 e Piemontese 2023), se non fallimentari (Cortelazzo 2021: 11).

Il progetto Easy-read-law

Nell'ambito del Progetto PRIN "Migrants, Institutions, Translations Easy-read Law", l'unità CNR ha svolto attività tese alla definizione di criteri operativi per la semplificazione linguistica dei testi amministrativi[3].

Tale approccio aveva come finalità ultima quella di rendere i testi trattati più inclusivi e capaci di rispondere alle esigenze di cittadini con diverse competenze linguistiche e background culturali eterogenei. Infatti il progetto, oltre a lavorare sulla dimensione linguistica e testuale dei diversi documenti presi in considerazione, ha voluto considerare, oltre ad un certo grado di entropia semantica insita nel linguaggio giuridico, anche quella conoscenza multidimensionale necessaria a comprendere il significato di una parola, specie in un dominio come quello giuridico in cui i significati di parole polisemiche sono spesso "legati a ciò che definiamo cultura in senso ampio" (Cacia et al 2013: 225-226). In particolare è stata individuata e resa disponibile una metodologia

3 Il progetto ha visto la collaborazione di Università del Salento, Università di Bari Aldo Moro, Università di Roma Tre e Consiglio Nazionale delle Ricerche.

che potrà essere utilmente riusata in altre esperienze di modernizzazione dei testi della PA.

Tale metodologia ha previsto, in una prima fase, la selezione di un corpus di risorse digitali, ad esempio testi di pagine web di pubbliche amministrazioni funzionali alla comunicazione di procedimenti amministrativi, moduli, relativi a diverse pratiche amministrative, testi in materia di lavoro (schede informative, contratti, lettere di assunzione). Tutti questi documenti sono stati scelti in relazione alla tematica del progetto e cioè l'inclusione delle persone con background migratorio sul nostro territorio. Tali documenti sono stati sottoposti a una prima analisi dello strumento READ-IT (Dell'Orletta et al., 2011). Dopo questa prima analisi, che ha fornito chiare indicazioni sulle problematiche connesse al lessico e alla sintassi di tali testi, l'unità CNR ha concentrato il proprio lavoro su alcune tipologie testuali, che sono state oggetto di specifiche analisi e di riscritture. Le tipologie testuali identificate sono state: comunicazioni online della PA e moduli. Come detto, questi testi sono stati sottoposti a una accurata analisi qualitativa tesa a verificare la presenza di quei parametri descritti in letteratura come tipici del linguaggio burocratico e di difficile comprensione (Fortis 2005). Da queste analisi sono emerse le caratteristiche che si riportano[4]. Per quanto riguarda il lessico è stata verificata la presenza di sostantivi deverbali (rilascio), di verbi denominali (contattare), di aggettivi in -ale (parametro dimensionale), di locuzioni congiuntive e preposizionali (ai sensi del dpr; si dovrà allegare una dichiarazione sostitutiva e/o dichiarazione asseverata), di locuzioni verbali (procedendo con la compilazione dei campi; avere valenza); di arcaismi (previo), di burocratismi (attestazione), di tecnicismi collaterali (richiedente; rilasciare), di termini dal registro elevato (utilizzare; esplicitare), di formule rituali (smi - acronimo per «e successive modificazioni»), di precisazioni ridondanti e pleonastiche (apposito modello; si dovrà allegare obbligatoriamente; tale dichiarazione ha valenza esclusivamente finalizzata a questo tipo di pratica), di acronimi e abbreviazioni (uso di citazioni normative abbreviate DPR, DM, Dlgs che, in base alle norme di tecnica legislativa, dovrebbero essere in forma esplicita nella prima citazione e comunque formalizzate, secondo una sintassi standard). Per quanto riguarda gli elementi sintattici è stato possibile riscontrare: una notevole complessità di alcuni periodi, sia per quanto riguarda la lunghezza (periodi di 80 parole), sia per quanto concerne la presenza di incisi, l'uso del gerundio (allegando), l'uso dell'infinito (utilizzare in questo caso l'allegato 2 del modulo); l'uso del participio, anche con funzione aggettivale (richiedenti; avvenuto pagamento), la presenza del passivo (la domanda può essere presentata) e della nominalizzazione (rilascio; compilazione), l'uso del futuro deontico (si dovrà allegare). Dal punto di vista della testualità è stato possibile rilevare, infine, una certa incoerenza nell'uso di elementi paratestuali, per esempio, nelle elencazioni, dove a volte i punti elenco sono introdotti da varie forme grafiche (puntini, lineette, talora nessun elemento grafico) e dove talora i paragrafi non sempre si concludono con la punteggiatura. Un elemento ulteriore è quello

4 L'analisi si riferisce a una comunicazione online relativa alla richiesta di idoneità alloggiativa.

dell'eccessiva concisione (mancanza degli articoli) degli elementi presenti negli elenchi (L'attestazione di idoneità alloggiativa serve per ottenere: - permesso di soggiorno per motivi familiari; - nulla osta al ricongiungimento familiare; ecc.). Una menzione particolare merita il parametro dell'intertestualità. Ovviamente un testo giuridico fa spesso riferimento ad altri testi dell'ordinamento ma, in base al tipo di testo[5], alcuni di questi riferimenti possono essere tranquillamente collocati in una zona in cui la loro lettura sia solo eventuale (Romano, Venturi 2025). Una volta verificate le caratteristiche qualitative e quantitative di tali documenti è stata proposta una prima riscrittura delle comunicazioni online delle procedure per la idoneità alloggiativa e per la dichiarazione di residenza e di alcuni moduli.

Prima di iniziare l'attività di semplificazione è stata verificata la normativa che riguarda le procedure relative ai moduli e alle comunicazioni oggetto della riscrittura. Così, ad esempio, per il modulo per la dichiarazione di residenza è stata individuata la normativa applicabile, rilevando in particolare che il d.P.R. n. 223/1989 (artt. 6, 12 e 14) lascia ampi margini di manovra nella creazione del modulo e non prescrive alcun elemento obbligatorio, mentre richiede solo che «chi trasferisce la residenza dall'estero deve comprovare all'atto della dichiarazione […], la propria identità mediante l'esibizione del passaporto o di altro documento equipollente».

Per le attività di semplificazione è stato fatto uso, ove possibile, dei seguenti criteri:

- rispetto al lessico: scegliere e usare parole del Lessico Fondamentale del NVdB (De Mauro e Chiari 2016), con il caveat di produrre glosse esplicative nel caso in cui ricorrano tecnicismi specifici che non possono essere sostituiti; eliminare termini e locuzioni ridondanti; sciogliere abbreviazioni e sigle la prima volta che sono usate; scrivere correttamente le citazioni di atti normativi e amministrativi; evitare le locuzioni congiuntive e preposizionali; evitare le locuzioni verbali; evitare arcaismi; evitare burocratismi e tecnicismi collaterali; evitare parole dal registro elevato; evitare precisazioni ridondanti e pleonastiche; evitare espressioni rituali e formule.
- Rispetto alla sintassi: per i verbi limitare l'uso del passivo e dell'impersonale; limitare le forme implicite del verbo, come gerundi o participi, quando è possibile usare le corrispondenti forme esplicite; evitare il futuro deontico; evitare le nominalizzazioni; evitare i periodi lunghi.
- Rispetto all'organizzazione del testo: rendere omogenei gli elenchi; spostare i riferimenti ad atti normativi e amministrativi in fondo al testo; creare i titoli delle sezioni del testo funzionali alla lettura dell'utente; rivolgersi direttamente all'utente per indicare le azioni da svolgere.

La fase successiva ha previsto un test con l'utenza di riferimento. Così i modelli dei moduli sopra descritti sono stati inviati ai colleghi e alle colleghe dell'Università

5 Data la distinzione dei testi in rigidi, semirigidi ed elastici (Sabatini 2016: 140-141), i testi che danno informazioni sintetiche, come i siti pubblici, possono essere considerati come semirigidi e quindi non richiedono la precisione che si è tenuti a usare in testi rigidi quali leggi, codici, regolamenti, contratti, sentenze, testi di definizione.

di Bari Aldo Moro. Il risultato sono state 20 interviste con persone immigrate nel nostro territorio che hanno reso informazioni sul lessico dei moduli e sul loro grado di comprensibilità. Analogamente altri moduli sono stati testati in alcune scuole genovesi – con alta presenza di studenti di seconda generazione – a cura delle colleghe di CNR IRCrES.

I risultati sia nell'uno che nell'altro caso hanno evidenziato alcune difficoltà connesse alla leggibilità dei moduli per la presenza di parole tecniche (comodato, residenza ecc.) che non avevamo ritenuto di poter sostituire con parole più semplici. Le indicazioni delle attività di test con gli utenti ci hanno quindi suggerito di creare glossari per le parole tecniche da inserire in calce ai diversi moduli[6]. La creazione di tali glossari è stata curata con strumenti di intelligenza artificiale, così come testato in recenti studi (Bolioli et al 2025).

Il progetto, infine, ha individuato come strumento per la diffusione dei prodotti della ricerca, ma anche come laboratorio per l'eventuale continuazione delle attività che il consorzio di ricerca intendesse sviluppare, un sito web. Il sito di progetto è concepito come uno spazio digitale centralizzato dedicato alla raccolta, organizzazione e pubblicazione dei materiali prodotti nell'ambito dell'iniziativa. La piattaforma ha l'obiettivo di garantire accessibilità, trasparenza e fruibilità dei contenuti, agevolando così la diffusione delle informazioni e la condivisione dei risultati. Il sito è organizzato in sezioni tematiche, progettate per rendere intuitiva la navigazione e agevole il reperimento dei materiali[7]. La progettazione, che è stata portata avanti e validata da tutte le unità di ricerca del PRIN, ha previsto la realizzazione di due pagine principali: una dedicata ai prodotti della ricerca e una dedicata alla divulgazione. Nella prima sono state previste varie categorie di prodotti, utili per reperire facilmente i contributi di ciascuna unità. In particolare, sono state previste le seguenti categorie:

1. Contratti (per modelli di contratto e di lettere di assunzione).
2. Schede informative (per modelli di schede informative per i lavoratori).
3. Materiale divulgativo (per il materiale con illustrazioni facilitate per i lavoratori).
4. Moduli (per i modelli di moduli della Pubblica amministrazione).
5. Modelli di pagine informative della PA (con schede informative per i lavoratori sul web).
6. Focus group (per evidenziare i risultati dei focus group condotti con gli utenti).
7. Report (per i report su singoli temi prodotti dalle diverse unità).

L'attività di ricerca evidenziata ha contribuito a rafforzare la riflessione scientifica sul rapporto tra lingua, cittadinanza e inclusione e ha offerto spunti per l'elaborazione di strumenti pratici di formazione e di supporto alla redazione di testi chiari e trasparenti. In definitiva il progetto ha reso disponibile un modello di analisi utile alla valutazione

6 Il modulo per la dichiarazione di residenza è stato oggetto anche di un focus group svolto durante il workshop "Semplificazione del linguaggio giuridico", tenuto online con le studentesse e gli studenti dell'Università di Torino nel corso del professor Enrico Marello.

7 Il sito è online su https://easy-read-law.it/

della leggibilità e dell'accessibilità comunicativa di testi istituzionali e ha anche realizzato e testato modelli di testi semplificati in base ad un approccio interculturale.

Le attività didattiche e di terza missione per i giovani

I progetti per la modernizzazione dei testi istituzionali e la formazione mirata, anche con momenti laboratoriali, per i redattori di tali documenti, rimangono momenti imprescindibili per avere risultati anche nel lungo periodo per la comprensibilità di questo tipo di documento e per tentare di rendere le parole delle istituzioni «parole di tutti e per tutti», come ebbe a dire Tullio De Mauro riferendosi a quelle della Costituzione in un suo commento del 2015 di recente ripubblicato (De Mauro 2024: 77).

Per raggiungere questo fine, da qualche anno, l'attenzione del CNR è rivolta anche alla didattica dedicata ai più giovani. Per questo motivo sono state ideate e realizzate molte iniziative, per lo più a carattere laboratoriale, destinate ai più giovani.

Una di tali iniziative è quella costituita dal laboratorio per le studentesse e per gli studenti delle scuole superiori chiamato "Comunicare contenuti delle istituzioni pubbliche semplici e inclusivi". Tale proposta nasce dalla collaborazione con il Gender Equality Team del CNR che negli scorsi anni ha elaborato le preziose ed utili *Linee guida per un linguaggio inclusivo* (Di Tullio et al 2024). Il lungo cammino per l'uguaglianza delle donne nella Repubblica Italiana, iniziato con il contributo delle 21 deputate elette nell'Assemblea Costituente, si sta realizzando a piccoli passi, ad esempio garantendo pari diritti sul lavoro, asili nido, riforme del diritto di famiglia (Mattiangeli 2024: 151), ma per «raggiungere una parità di fatto» tra i sessi è indispensabile che la società si liberi dai residui «pregiudizi negativi nei confronti delle donne», molti dei quali spesso camuffati e nascosti «sotto forme di apparente valore oggettivo» e veicolati anche «attraverso la lingua, in modo spesso subdolo e ripetitivo» (Sabatini 1987: 19). Il percorso didattico prevede anche una parte dedicata all'illustrazione delle regole destinate alla produzione di testi amministrativi chiari ed efficaci e vuole coinvolgere i partecipanti attraverso lezioni frontali che illustrano gli strumenti a disposizione per la stesura e la revisione di contenuti che siano chiari e con laboratori che facciano sperimentare le categorie di semplicità, chiarezza, trasparenza e inclusività alla comunicazione verbale e visiva.

Queste ed altre attività sono proposte nell'ambito dei percorsi didattici presenti al CNR e messi a disposizione delle scuole nei programmi annuali di *Formazione Scuola Lavoro*. Come noto, anche la comunicazione in rete dei soggetti pubblici vede riproporsi molte delle caratteristiche tipiche del burocratese quali «oscurantismo, ambiguità, vaghezza» (Lubello 2018: 99). Queste peculiarità sono state confermate anche da recenti indagini, svolte su campioni significativi di pagine web di amministrazioni locali che hanno messo in evidenza come queste comunicazioni online continuino «a privilegiare un approccio orientato all'amministrazione piuttosto che all'utente, producendo testi non adeguati al mezzo digitale». Tali caratteristiche sono state verificate su un

corpus di testi provenienti da pagine web di comuni italiani che sembrano «ancorati a lunghe e pesanti tradizioni linguistiche» e i cui tratti «assumono sul web un carattere ancor più inappropriato» (Romano, Tombesi 2024: 43-44). Per questo motivo è stato proposto alle scuole un percorso didattico dedicato a "La comunicazione di informazioni giuridiche in Rete" che vede impegnati studenti e studentesse nel reperimento di informazioni delle istituzioni pubbliche sui loro siti web per poi verificarne i contenuti informativi e testarne la comprensibilità, al fine di proporre soluzioni per una loro maggiore comprensibilità. L'obiettivo formativo di questi percorsi è quello di insegnare come verificare la provenienza di notizie e informazioni rilevanti e di imparare a conoscere le insidie del linguaggio burocratico veicolato dal web.

Questa esperienza si inserisce in una tradizione inaugurata alcuni anni fa nell'ambito di progetti FAMI e destinati anche ai giovani di seconda generazione presenti sui nostri territori. In essi i giovani delle scuole medie superiori erano coinvolti nella progettazione e realizzazione di testi (spesso pagine web rilevanti per i loro interessi come quelle sugli abbonamenti ai mezzi di trasporto pubblico, quelle sulla iscrizione ai corsi universitari, ecc.) chiari, efficaci e che tenessero conto delle esigenze dell'utenza di riferimento (Romano 2022 e Salvi et al 2019).

Le risorse linguistiche messe a disposizione del CNR in banche dati online messe liberamente a disposizione di studiosi e cittadini, sono il fulcro anche di un percorso didattico dedicato a "La storia del linguaggio giuridico attraverso i documenti di alcune banche dati". Attraverso i documenti presenti in alcune banche dati CNR (LLI - Lingua legislativa italiana e LGI – Lessico giuridico italiano), gli studenti e le studentesse sono chiamati a verificare il significato di una o più parole presenti in testi giuridici antichi e contemporanei, nonché a proporre delle vere e proprie voci di un vocabolario giuridico attraverso la piattaforma IS-LeGI, che da risorsa utile per gli studi di linguisti, giuristi e storici è divenuta anche strumento per la didattica digitale (Turchi, Romano 2023)[8].

Queste risorse linguistiche, assieme ad altre banche dati di documenti giuridici (ad esempio il Portale Normattiva), sono utili per creare percorsi formativi, anche interdisciplinari, nei quali la polisemia del linguaggio giuridico, l'attitudine dei testi del diritto di trattare tutti gli ambiti che riguardano la vita delle nostre società, la presenza di risemantizzazioni che rendono di difficile interpretazione alcuni testi giuridici, diventano temi da sperimentare in modo pratico per poter disporre di strumenti per una traduzione semplificata di testi altrimenti oscuri. È il caso dei laboratori didattici "*Semplificare il lessico delle istituzioni"* e "*Cerca le parole della Scienza nella Legge. Un gioco fra discipline tecniche e umanistiche"* che le ricercatrici e i ricercatori del CNR hanno organizzato nell'ambito di convegni e fiere dedicate alla divulgazione scientifica e alle attività ludico-didattiche[9].

8 Le banche dati sono online su https://www.igsg.cnr.it/banche-dati/

9 I due laboratori didattici sono stati sperimentati durante il XXII Convegno "Il linguaggio della Ricerca" che si è svolto a Bologna presso l'area della ricerca CNR il 20 novembre 2025, ma anche al Festival della cultura ludica *Play in the city* che si è svolto a Modena il 22 novembre 2025. Vedi https://www.igsg.cnr.it/2025/11/igsg-a-play-in-the-city/

Lavorare con i giovani perché apprendano un numero sempre maggiore di parole e con il loro significato imparino anche a preservarne la qualità, rispettando i concetti che vogliono veicolare, sembra uno dei modi per continuare ad avere cura della nostra democrazia (Zagrebelsky 2007: 36). Si devono, quindi, sperimentare nuovi percorsi per fare in modo che le persone possano decifrare le parole delle istituzioni, in modo da potersi considerare pienamente cittadini e cittadine e non persone per le quali il «non capire – anzi, l'impossibilità di capire» toglie «l'alternativa del decidere» (Irti 2021: 38).

Bibliografia

Biagioli, Carlo, Mercatali, Pietro, "Strumenti automatici per redattori di testi legislativi: Lexedit2 in ambiente di normazione". In *Informatica e diritto*, I, 1993, pp. 107-122.

Bolioli, Andrea, Cherubini, Manola, Romano, Francesco, "L'intelligenza artificiale generativa per la modernizzazione del linguaggio giuridico". In *Natale Rampazzo (a cura di), Dialoghi sull'intelligenza artificiale. Atti dei seminari del Cnr-Iriss (marzo-maggio 2024)*, 6, 2025, Roma: Cnr Edizioni, pp. 27-40.

Bolioli, Andrea, Dini, Luca, Mercatali, Pietro, Romano, Francesco, "For the Automated Mark-up of Italian Legislative Texts in XML". In *Legal Knowledge and Information Systems. Jurix 2002: The Fifteenth Annual Conference*, (a cura di) T.J.M. Bench-Capon, A. Daskalopulu, R.G.F. Winkels, IOS Press: Amsterdam, 2002.

Cacia, Daniela, Papa, Elena, Verdiani, Silvia, *Dal Mondo alle parole. Definizioni spontanee e dizionari d'apprendimento*, Roma: Società editrice romana, 2013.

Ceccarini, Luigi, *La cittadinanza online*, Bologna: Il Mulino, 2015.

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

Dell'Orletta, Felice, Montemagni, Simonetta, Venturi, Giulia, "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 (SLPAT)*, Edimburgo, 2011, pp. 73-83.

De Mauro, Tullio, *Passione civile*, Roma-Bari: Laterza, 2024.

De Mauro, Tullio, Chiari, Isabella, *Il Nuovo vocabolario di base della lingua italiana*, https://intern.az/1w1U, 2016.

Di Tullio, Ilaria, Mattiazzi, Martina, Presto, Sabrina (a cura di), *Linee guida per il linguaggio inclusivo rispetto al genere*, Roma: CNR edizioni, 2024.

Fioravanti, Chiara, Lombardi, Giulia, Pietrangelo, Marina, Romano, Francesco, Formare alla scrittura di atti amministrativi più chiari e comprensibili: un caso di studio, in LTO. Lingua e testi di oggi, 6, 1, 2025, pp. 13-35.

Fortis, Daniele, "Il linguaggio amministrativo italiano". In *Revista de Llengua i dret*, 43, 2005: 47-116.

Gruppo di lavoro promosso da ITTIG e Accademia della Crusca, *Guida alla redazione degli atti amministrativi. Regole e suggerimenti*, Firenze: ITTIG-CNR, 2011.

Irti, Natalino, *Viaggio tra gli obbedienti*, Milano: La nave di Teseo, 2021.

Lubello, Sergio, "Una storia che viene da lontano e non ancora finita: i mille volti del burocratese". In Giuliana Fiorentino, Alessandro Cioffi, Maria Ausilia Simonelli (a cura di), Amministrazione attiva. Semplicità e chiarezza per la comunicazione amministrativa, Firenze: Franco Cesati Editore, 2025.

Lubello, Sergio, *L'e-italiano. Scriventi e scritture nell'era digitale*, Firenze: Franco Cesati Editore, 2018.

Marchetti, Andrea, Megale, Fabrizio, Seta, Enrico, Vitali, Fabio, "Marcatura XML degli atti normativi italiani. I DTD di Norme in Rete". In *Informatica e diritto*, XXVII Annata, Vol. X, 2001, n. 1: 123-148.

Mattiangeli, Susanna, *La Costituzione nelle parole*, Roma: Lapis edizioni, 2024.

Piemontese, Maria Emanuela (a cura di), *Il dovere costituzionale di farsi capire*, Roma: Carocci, 2023.

Pietrangelo, Marina, "Un nuovo approccio metodologico nel Manuale per la redazione delle leggi regionali: riannodare il drafting formale e il drafting sostanziale". In Consulta online, 2, 2024: 936-947.

Romano, Francesco, Tombesi, Elena, "Come comunicano i Comuni sul web: una prima indagine condotta su lessico e testualità". In L*ingue e culture dei media*, 8, 2, 2024: 43-44.

Mercatali, Pietro, Romano, Francesco, *Partecipazione amministrativa e strumenti tecnologici, in Amministrazione italiana*, Numero annuale 2013: 155-177.

Mercatali, Pietro, Legimatica e redazione delle leggi, Firenze: ITTIG-CNR, 1995, online su https://www.ittig.cnr.it/Ricerca/Testi/mercatali1995.htm#10

Romano, Francesco, Venturi, Giulia, "Metodi e strumenti per la modernizzazione della lingua delle istituzioni". In *Migranti Leggi Contratti: verso la chiarezza* (a cura di Annarita Miglietta), CALUMET – intercultural law and humanities review, 2025: 105-121.

Romano, Francesco, "Linguaggio giuridico e giovani: l'esperienza dei laboratori di progettazione di contenuti semplificati per i testi online della PA". In *Italiano LinguaDue*, volume 14 (2-2022): 714–723.

Romano, Francesco, "Adozione, diffusione e citazione della «Guida alla redazione degli atti amministrativi. Regole e suggerimenti»: alcune esperienze". In *La qualità degli atti normativi e amministrativi*, a cura di S. Panizza, Pisa: PUP, 2016.Sabatini, Alma, *Il sessismo nella lingua italiana*, Roma: Presidenza del Consiglio dei Ministri, 2016.

Sabatini, Francesco, *Lezione di italiano*, Milano: Mondadori, 2016.

Salvi, Alessandro, Caldonazzo, Cristina, Fioravanti, Chiara, Romano, Francesco, Chiappelli, Tiziana, Mangani, Sabrina, Fabbri, Maria e Mugnai, Margherita, "Percorsi di cittadinanza interculturale e digitale con le nuove generazioni: il progetto #IOPARTECIPO". In *Rivista italiana di informatica e diritto*, n. 2/2019, CNR – Roma, 2019: 1-15.

Solito, Laura, "Comunicazione istituzionale". In *Le parole della comunicazione*, (a cura di) C. Sorrentino, Pisa: Pacini, 2018.

Spinosa, Pierluigi, "Norma tecnica: NIR-URN 1.1 (novembre 2001). Assegnazione dei nomi uniformi ai documenti giuridici". In *Informatica e diritto*, XXVII Annata, Vol. X, n. 1, 2001: 233-266.

Talamo, Sergio, Zarriello, Roberto, *Nuovo manuale di comunicazione pubblica: comunicare la Pubblica amministrazione: teorie, tecniche e buone pratiche digital e social*, Roma: Centro di documentazione

giornalistica, 2022.

Turchi, Fabrizio, Romano, Francesco, Progetto IS-LeGI per la didattica digitale e la formazione, selected papers della Conferenza GARR Conferenza GARR 23 "Saperi interconnessi", DOI: 10.26314/GARR-Conf23-proceedings 09, 2023: 61-63.

Vellutino, Daniela, *L'italiano istituzionale per la comunicazione pubblica*, Bologna: Il Mulino, 2018.

Zagrebelsky, Gustavo, *Imparare democrazia*, Torino: Einaudi, 2007.

L'intelligenza artificiale per semplificare i testi della pubblica amministrazione: strategie di *prompt engineering*

Daniele Fortis
(Dirigente ASST Santi Paolo e Carlo, Milano)

Abstract

Questo contributo presenta i risultati preliminari di un progetto di semplificazione di testi amministrativi elaborato in un contesto applicativo concreto – quello di un ente pubblico della sanità lombarda – mediante l'uso dell'intelligenza artificiale generativa. Muovendo dall'evidenza che la complessità del linguaggio burocratico ostacola la fruizione dei servizi sanitari, lo studio mira a identificare la tecnica di *prompt engineering* e il modello linguistico di grandi dimensioni più efficaci nel migliorare la chiarezza. Un *corpus* di testi amministrativi è stato semplificato utilizzando quattro tecniche di *prompting* (*zero-shot, few-shot, chain-of-thought e tree-of-thought*) e due modelli (GPT-5 e Gemini 2.5 Flash). La valutazione dei risultati è stata effettuata attraverso un approccio metodologico misto che integra indici di leggibilità e una valutazione esperta automatizzata basata sui criteri del *plain language*. Tutti i testi semplificati mostrano un miglioramento rispetto agli originali. Le formule di leggibilità forniscono risultati non del tutto concordi, mentre la valutazione automatizzata sembra individuare nel *prompt tree-of-thought* la tecnica più efficace. Non emergono differenze prestazionali significative tra i due modelli testati. I risultati confermano il potenziale degli LLM e del *prompt engineering* nella semplificazione del linguaggio amministrativo, pur evidenziando la necessità di future verifiche su lettori reali.

1. Introduzione

Questo contributo illustra lo studio preliminare di un progetto di semplificazione dei testi amministrativi tramite l'intelligenza artificiale generativa elaborato presso la ASST Santi Paolo e Carlo di Milano[1], dove ricopro l'incarico di dirigente dell'Ufficio Relazioni con il Pubblico (URP) e dei servizi di accoglienza.

1 ASST è l'acronimo di Azienda Socio-Sanitaria Territoriale: un ente pubblico del Servizio Sanitario Regionale della Lombardia, istituito con la Legge Regionale n. 23/2015, con il compito di erogare sia cure ospedaliere, sia servizi sociali e sanitari territoriali (ambulatori, consultori, centri vaccinali, case di comunità, assistenza domiciliare). La ASST Santi Paolo e Carlo ricomprende due ospedali (San Paolo e San Carlo Borromeo) e i servizi sociosanitari territoriali dell'area sud-ovest di Milano.

Il progetto nasce dalla constatazione che molti testi dell'ente destinati al pubblico sono scritti in un linguaggio complesso e distante dall'uso comune, connotato dai tratti anticomunicativi tipici dell'italiano burocratico. Non mi dilungherò sulle caratteristiche (lessicali, sintattiche, testuali, pragmatiche) tipiche di questa varietà linguistica, perché esaurientemente descritte da ampia letteratura (Fortis 2005, Lubello 2014, Cortelazzo 2021), né sulle ragioni etiche e civili (Piemontese 2023) o economico-organizzative (Kimble 2023), anch'esse diffusamente argomentate, che legittimano gli interventi di semplificazione. Mi limiterò a riportare un dato empirico: l'esperienza quotidiana nella gestione del rapporto con l'utenza evidenzia come la scarsa comprensibilità dei testi prodotti dall'ente spesso ostacoli un'agevole fruizione dei servizi sanitari e sociosanitari da parte dei cittadini e sia fonte di contrasti e tensioni relazionali. Dall'analisi dei reclami presentati all'URP, infatti, emerge che, alla base delle rimostranze, vi è spesso un'incomprensione linguistica, dovuta a testi – avvisi, istruzioni, procedure, regolamenti – formulati in modo ambiguo e poco trasparente.

Di qui l'iniziativa di revisionare i testi aziendali rivolti al pubblico per renderli più chiari e accessibili. La scelta di ricorrere alla *Automatic Text Simplification* (ATS) è stata dettata da due considerazioni: la prima è che la grande quantità di testi da riformulare avrebbe reso troppo onerosa una revisione svolta dal personale addetto alla comunicazione; la seconda, che le prestazioni dei modelli linguistici di grandi dimensioni (LLM) nel migliorare la chiarezza dei testi sono oggi pari, e talvolta superiori, a quelle dei revisori umani, come attestato da vari studi (fra i molti, Hasebe *et al.* 2025, Nozza e Atanasio 2023, Fiorentino e Russodivito 2025).

Fra le possibili modalità di utilizzo dell'intelligenza artificiale, il *prompt engineering* – ottimizzare le istruzioni testuali (*prompt*) fornite al modello per ottenere risposte precise e utili – è apparso il più idoneo, per le ragioni che illustrerò nel paragrafo successivo.

Occorreva stabilire, a questo punto, quale tecnica di *prompting* e quale LLM potessero offrire la maggior qualità ed efficacia nel semplificare i testi. E questa è la domanda di ricerca alla quale questo lavoro intende rispondere.

2. I possibili approcci nell'utilizzo dei LLM e la scelta del *prompt engineering*

Schematizzando, si possono individuare tre principali approcci per personalizzare e utilizzare gli LLM. Si tratta, certo, di una semplificazione, perché questi approcci, nella prassi, non sono mutuamente esclusivi; tuttavia, questa tripartizione è generalmente riconosciuta come corretta, perché copre tutte le esigenze, nel senso che qualsiasi problema si affronti con un modello di AI, la soluzione ricade in una di queste tre categorie (o in una loro combinazione):

1. *Prompt engineering*. È il metodo più immediato e accessibile. Consiste nel formulare domande o istruzioni ben strutturate per sfruttare al meglio le capacità del modello

e guidarlo verso il risultato desiderato[2]. Non si modifica il modello stesso e i suoi parametri interni, ma si raffina il modo in cui gli si "parla".

2. *RAG (Retrieval-Augmented Generation*). Il modello viene collegato a una base di conoscenza esterna (documenti, *database*, ecc.) dalla quale recupera informazioni pertinenti per generare risposte accurate e aggiornate. Anche in questo caso, non si modifica strutturalmente il modello, ma gli si forniscono informazioni rilevanti rispetto al compito da eseguire.
3. *Fine-tuning*. Si riaddestra il modello su nuovi dati, modificando i suoi parametri per specializzarlo e consentirgli di eseguire al meglio un *task* specifico. È il metodo che offre la massima personalizzazione del modello, ma è anche il più impegnativo, perché richiede risorse computazionali elevate e dati di alta qualità.

Per semplificare i testi amministrativi della ASST, fra i tre approcci descritti ho ritenuto che il più indicato fosse il primo. Il *prompt engineering* rappresenta, infatti, il metodo più agile e leggero. Non comportando addestramento del modello, ma sfruttando le capacità del modello pre-addestrato, non richiede competenze tecniche avanzate, non ha costi e non necessita dell'intervento dei Sistemi Informativi aziendali o di una ditta esterna. Per il compito da svolgere, d'altronde, non si vede la necessità di ricorrere alla RAG o al *fine-tuning*: gli LLM, infatti, hanno già dimostrato di saper semplificare testi efficacemente utilizzando le sole impostazioni predefinite. Inoltre, l'ASST, pur operando in un ambito settoriale come quello sanitario, produce una vasta gamma di testi amministrativi eterogenei, la cui semplificazione non rappresenta un compito così specialistico da richiedere un riaddestramento mirato del modello o la costruzione di un *dataset* personalizzato di testi amministrativi semplificati da integrare tramite sistemi di *retrieval*. Simili interventi, anzi, potrebbero risultare controproducenti, rischiando di limitare la flessibilità e le capacità generaliste del modello.

Il maggior vantaggio del *prompt engineering*, tuttavia, risiede nella sua flessibilità operativa: questo approccio, infatti, consente un controllo granulare sull'*output*, permettendo di variare le caratteristiche del testo risultante dalla semplificazione (stile, livello di formalità, complessità, lunghezza, destinatario ideale, formato, struttura) in modo molto preciso, semplicemente modificando l'*input*, ossia le istruzioni. Non servono modifiche infrastrutturali o nuovo *training*.

Un esempio per chiarire l'affermazione che precede. Consideriamo questo testo, che esibisce alcuni dei tratti tipici del burocratese (lessico di registro elevato e arcaizzante, nominalizzazioni, lunghezza del periodo):

> Avverso il giudizio della Commissione Medica, il soggetto interessato può presentare formale istanza di riesame entro 60 giorni dalla notifica, unitamente alla documentazione sanitaria integrativa e difensiva idonea a comprovare i presupposti per la revoca o modifica del provvedimento originario.

2 Una definizione più precisa di *prompt engineering* è offerta da Chen *et al.* (2025:2) "The systematic design and optimization of input prompts to guide the responses of LLMs, ensuring accuracy, relevance, and coherence in the generated output."

Ai fini di una semplificazione adeguata, occorre tener conto di variabili contestuali come il destinatario, il *medium*, lo scopo, il genere testuale. Se questo passaggio ha valenza puramente informativa, si rivolge all'utenza generale ed è destinato a comparire, ad esempio, nel sito web istituzionale dell'ente o nella carta dei servizi, può essere appropriato dare del tu al lettore. È stata quindi impartita tale istruzione nel *prompt* fornito al modello («dai del tu al destinatario»), che ha prodotto questa versione semplificata:

> Se non sei d'accordo con il giudizio della Commissione Medica, puoi chiedere un nuovo esame entro 60 giorni da quando ricevi la comunicazione ufficiale. Per farlo, devi allegare alla richiesta scritta eventuali documenti medici aggiuntivi che dimostrino perché la decisione dovrebbe essere cambiata o revocata.

Se invece il passaggio è da inserire in una lettera o un'email indirizzata a uno specifico cittadino, occorre certamente usare il lei. È bastato indicarlo nel *prompt*, per ottenere dal modello questo testo:

> Se non condivide la decisione della Commissione Medica, può richiedere una nuova valutazione entro 60 giorni dalla comunicazione ufficiale. Alla richiesta deve allegare la documentazione sanitaria aggiuntiva che può aiutare a dimostrare perché il provvedimento iniziale dovrebbe essere cambiato o revocato.

In un regolamento – testo con valore normativo, quindi dotato di generalità-astrattezza – è opportuno esprimersi in terza persona. Un'ulteriore variazione del *prompt* ha prodotto questo risultato:

> Se l'assistito non condivide il giudizio della Commissione Medica, può chiedere un nuovo esame entro 60 giorni dalla notifica. Alla richiesta va allegata la documentazione sanitaria aggiuntiva utile a dimostrare perché il provvedimento iniziale dovrebbe essere modificato o revocato.

Questi esempi mostrano come, intervenendo sul *prompt*, sia possibile modulare il testo di *output* con un alto livello di personalizzazione, e quindi ottenere una semplificazione coerente con le istanze dell'utilizzatore, che può pilotare il risultato della revisione per renderlo adeguato alle proprie esigenze contestuali.

3. I *prompt* sperimentati

Da pratica empirica, il *prompt engineering* si è evoluto in un ambito di ricerca ormai strutturato. La letteratura scientifica ha classificato i *prompt* in minuziose ed elaborate tassonomie (ad esempio, quelle proposte da Chang *et al.* 2024 e da Sahoo *et al.* 2024), dando vita a una proliferazione spesso non necessaria di categorie e sottocategorie, riconducibili, in realtà, a un numero ristretto di tipi principali. Ho quindi deciso di testare l'efficacia nel semplificare testi amministrativi delle seguenti 4 tecniche di *prompt*:

1. *Zero-shot*. È la strategia più semplice e diretta, il livello base di interazione col modello: gli si impartiscono direttive, anche dettagliate, ma senza fornire alcun esempio su come affrontare il compito. L'intelligenza artificiale deve risolvere il problema affidandosi unicamente alle conoscenze generali acquisite durante l'addestramento, senza alcun riferimento contestuale nel *prompt*.
2. *Few-shot*. Si forniscono al modello alcuni esempi dimostrativi (tipicamente da 2 a 5) dell'*output* atteso, cioè di come eseguire il compito. Gli si offre un contesto da cui può apprendere uno schema da replicare e generalizzare. Gli esempi illustrano il formato atteso, lo stile o la logica da seguire, permettendo al modello di imparare per imitazione, ma sempre entro il contesto della singola conversazione (senza cioè riaddestrarlo, come avverrebbe col *fine-tuning*).
3. *Chain-of-thought*. Si chiede al modello di non generare direttamente la risposta, ma di «pensare ad alta voce», ossia di articolare esplicitamente il suo ragionamento nei passaggi sequenziali che conducono alla risposta finale, mostrando il processo logico verso la soluzione. Si decompone un problema complesso, suddividendolo in sottoproblemi più semplici.
4. *Tree-of-thought* costituisce un'estensione del *prompt chain-of-thought*, ma, mentre quest'ultimo incoraggia un percorso logico lineare, il *tree-of-thought* chiede al modello di seguire un percorso ramificato, di esplorare più percorsi di ragionamento alternativi, confrontando e soppesando le diverse opzioni, scartando via via le più deboli e mantenendo le più valide.

Ho quindi elaborato 4 *prompt*, riportati in Appendice.

Il primo, del tipo *zero-shot*, impartisce al modello istruzioni abbastanza dettagliate, basate sulle linee guida per la chiarezza dei testi amministrativi, su come semplificare il testo sottopostogli, senza però fornire alcun esempio.

Il secondo, del tipo *few-shot*, è uguale al primo, dal quale differisce solo per l'aggiunta di 3 esempi di frasi semplificate.

Il terzo, del tipo *chain-of-thought*, fornisce sostanzialmente le stesse linee guida dei *prompt* precedenti, ma indicando al modello le fasi che deve seguire, in modo sequenziale, nell'applicarle.

Il quarto, del tipo *tree-of-thought,* configura uno scenario in cui tre esperti linguistici, ciascuno con priorità diverse – il primo persegue soprattutto la chiarezza, il secondo privilegia la concisione, il terzo si preoccupa di preservare la precisione giuridica – interagiscono fra loro, confrontano le proprie versioni e raggiungono un testo semplificato di compromesso fra le rispettive esigenze in conflitto.

In tutti i *prompt* si raccomanda di non riassumere il testo (cosa che gli LLM, in assenza di tale indicazione, tendono a fare) e di non aggiungere informazioni (per evitare le cosiddette "allucinazioni": contenuti linguisticamente plausibili ma falsi o inventati, generati dai modelli come conseguenza del proprio funzionamento probabilistico).

4. La costruzione del *corpus* e la sua semplificazione

Ho costruito un *corpus* composto da 6 testi, ciascuno con una lunghezza media di circa 600 *token*, rivolti all'utenza e finalizzati a illustrare le modalità di fruizione dei servizi. Trattandosi di testi di natura prettamente informativa, seppure incentrati su pratiche e procedure amministrative, la loro ricaduta giuridica è modesta ed è quindi minimo il rischio che la semplificazione produca conseguenze legali indesiderate[3].

Ciascun testo è stato semplificato con i 4 *prompt* elaborati e con 2 LLM: GPT-5 e Gemini 2.5 Flash, scelti perché i più diffusi. Sono state quindi ottenute 8 versioni semplificate di ciascun testo, per un totale di 48.

Occorreva, a questo punto, valutare tutti i testi prodotti, oltre ai 6 originali, sotto il profilo della chiarezza e mettere a confronto i risultati, per stabilire quale *prompt* e quale LLM avesse conseguito la *performance* migliore.

5. Valutare i risultati

Per misurare la chiarezza di un testo esistono vari metodi, qualitativi e quantitativi, soggettivi e oggettivi. I diversi studi che si sono proposti di valutare il miglioramento della chiarezza sono ricorsi a soluzioni eterogenee. I principali metodi adottati a questo scopo, con i rispettivi vantaggi e limiti, sono i seguenti:

- *Test di comprensione e usabilità su lettori reali*. Si seleziona un campione di lettori ritenuti rappresentativi del *target* ideale del testo, ossia dotati di caratteristiche sociodemografiche simili a quelle dei supposti destinatari del medesimo. Ai soggetti individuati sono somministrati quesiti (di vario genere: domande a risposta multipla o del tipo vero/falso, procedura *cloze*, ecc.) volti a misurare la loro capacità di comprendere il testo e di usarlo, cioè di agire sulla sua base. Questo metodo rappresenta il *gold standard* sotto il profilo dell'attendibilità, ma è estremamente impegnativo e dispendioso in termini di tempo e risorse, quindi difficilmente praticabile (sono infatti relativamente pochi gli studi che vi fanno ricorso).
- *Rilevazione della comprensibilità percepita*. Al campione sono somministrati quesiti del tipo: «Quanto chiara ti sembra questa frase?», «Quale fra queste 2 versioni ritieni più comprensibile?» Quesiti di questo genere sono relativamente semplici da elaborare, ma è evidente che non misurano direttamente la comprensibilità del testo, bensì la percezione – soggettiva e spesso fallace – che ne hanno gli esaminati (possono avere l'impressione che il testo sia chiaro, ma non capirlo realmente).

3 I testi risultanti dalla semplificazione sono stati comunque sottoposti al vaglio di alcuni dipendenti della ASST dotati di formazione giuridica, i quali non hanno riscontrato criticità di rilievo, limitandosi a segnalare alcune lievi imprecisioni lessicali. Ad esempio, dove il LLM aveva sostituito *revocare* con *annullare*, è stato osservato che i due verbi non sono equivalenti sotto il profilo del significato giuridico. Tale discrepanza, tuttavia, non sembra particolarmente problematica in un testo informativo.

- *Valutazione esperta.* Il testo viene valutato non da un campione dei lettori effettivi cui è destinato, ma da esperti di linguistica, comunicazione, scrittura e discipline affini. Di fatto, questi ultimi verificano, perlopiù, se il testo rispetta le linee guida per la scrittura chiara ampiamente condivise (un'analisi critica in Fortis 2004, una sintesi recente in Cortelazzo 2021, una rassegna ragionata della letteratura in Fiorentino e Ganfi 2024), che però costituiscono solo il (probabile) presupposto della comprensibilità, ma non la garantiscono, e non sempre gli esperti riescono a predire le difficoltà incontrate dai lettori reali. Inoltre, i valutatori esprimono un giudizio inevitabilmente soggettivo, il che spiega lo scarso accordo tra loro (*inter-rater reliability)* che spesso si riscontra (come attestato, ad esempio, in Tavosanis 2025: 114). Anche questo metodo di valutazione è oneroso, a causa della difficoltà di reperire i valutatori. Tale difficoltà non sussiste, ovviamente, nel caso che la valutazione sia affidata all'intelligenza artificiale, opzione sempre più utilizzata nella ricerca recente.
- *Formule di leggibilità.* Si tratta di strumenti matematici che stimano il grado di facilità/difficoltà di un testo, talvolta rapportandolo al livello di istruzione del lettore, usando parametri linguistici quantitativi e quindi misurabili (lunghezza media di parole e frasi o, nel caso delle formule più sofisticate, ulteriori aspetti lessicali e sintattici). Le formule presentano il vantaggio di essere facili e rapide da usare, in quanto automatizzate, ma il limite di considerare solo aspetti linguistico-formali superficiali e non altri fattori che incidono sulla comprensibilità (coerenza logico-semantica, struttura informativa del testo, conoscenze pregresse del lettore).

In Tab. 1 propongo una sintesi dei vantaggi e dei limiti dei metodi illustrati, sotto i profili dell'attendibilità e della praticabilità:

Metodo	**Attendibilità**	**Praticabilità**
Test di comprensione e usabilità su lettori reali	Alta	Bassa
Rilevazione della comprensibilità percepita	Media	Media
Valutazione esperta	Media	Bassa (se umana) Alta (se automatizzata)
Formule di leggibilità	Media	Alta

Tabella 1: Metodi di misurazione della chiarezza.

Nello studio qui presentato, per valutare il miglioramento della chiarezza ho scelto due approcci complementari: le formule di leggibilità e la valutazione esperta affidata a un LLM tramite un questionario automatizzato, due metodi che ho ritenuto realizzassero un buon compromesso fra affidabilità e praticabilità. Le formule, infatti, malgrado i limiti sopra evidenziati, di cui sono consapevole, tanto da averli analizzati dettagliatamente in un precedente lavoro (Fortis 2005), alquanto critico – forse troppo,

in retrospettiva – verso tali strumenti, forniscono comunque una diagnosi orientativa e probabilisticamente valida, che, unita ai vantaggi dell'oggettività e dell'estrema rapidità e facilità d'uso, ne giustifica l'impiego. La valutazione esperta formulata dall'AI, inoltre, consente di superare la difficoltà di reperire dei valutatori esperti e la loro umana soggettività, e per questo rappresenta una soluzione adottata da un numero crescente di studi.

5.1 La misurazione della leggibilità

Ciascuno dei 54 testi (6 originali e 48 versioni semplificate) è stato sottoposto a misurazione della leggibilità mediante due noti indici: Gulpease (Lucisano e Piemontese 1988) e READ-IT Globale (Dell'Orletta, 2011)[4]. Successivamente, è stata calcolata la media dei punteggi ottenuti da ciascun gruppo di testi semplificati con la stessa combinazione di modalità (ad es. i 4 semplificati con GPT-5 e un *prompt zero-shot*). I risultati sono esposti in Tab. 2.

	Tecnica di prompt	**READ-IT global**	**Gulpease**
Testi originali	-	87,5	43,6
Testi semplificati con GPT-5	Zero-shot	67,5	68,4
	Few-shot	72,7	66,6
	Chain-of-thought	82,9	59,2
	Tree-of-thought	58,3	62,5
Testi semplificati con Gemini 2.5 flash	Zero-shot	54,6	62,8
	Few-shot	56,0	60,3
	Chain-of-thought	60,2	61,4
	Tree-of-thought	59,4	63,8

Tabella 2: Indici di leggibilità dei testi originali e indici medi di quelli semplificati.

Come si può notare, tutti i testi semplificati riportano punteggi medi migliori (cioè più bassi secondo READ-IT, più alti secondo Gulpease) rispetto agli originali. Tutti, tranne uno, conseguono un indice Gulpease superiore a 60, valore adeguato a chi possiede almeno la licenza media e considerato la soglia ideale per la comunicazione pubblica e, in generale, per i testi destinati a un *target* generalista, di lettori non specialisti.

I punteggi, tuttavia, sono discordanti fra i due indici per quanto riguarda il *prompt* e il LLM più efficaci. Questi risultati, quindi, dimostrano l'efficacia della semplificazione,

4 Il punteggio assegnato da READ-IT è espresso in percentuale e oscilla tra 0 (minima complessità, massima leggibilità) e 100 (massima complessità, minima leggibilità), quindi più è alto, meno il testo è leggibile. Al contrario, per Gulpease 0 indica la massima difficoltà di lettura, 100 la massima facilità.

ma non forniscono indicazioni coerenti che permettano di rispondere alla domanda di ricerca alla base di questo studio.

5.2 Il questionario di valutazione

Il secondo metodo di valutazione è consistito nella somministrazione automatizzata di un questionario (Tab. 3), mutuato, con adattamenti, da Alves *et al.* (2025), mirato a rilevare la presenza, nei testi analizzati, delle caratteristiche ritenute costitutive della chiarezza, ossia, fondamentalmente, il rispetto dei principi consolidati del *plain language*.

N.	Quesito	Peso
1.	Il testo semplificato omette informazioni presenti nel testo originale?	5
2.	Il testo semplificato include informazioni non presenti nel testo originale?	5
3.	La semplificazione ha compromesso la comprensibilità del contenuto?	3
4.	Il testo semplificato è adeguato al livello di conoscenza del pubblico di destinazione?	3
5.	Il testo presenta le informazioni più importanti per prime e quelle restanti in un ordine logico?	1
6.	Il testo struttura le informazioni utilizzando elenchi o tabelle?	1
7.	Il testo trasmette tendenzialmente una sola idea per frase?	1
8.	Il testo usa frasi brevi (la maggioranza non oltre 25 parole)?	1
9.	Il testo usa una sintassi semplice, con poche subordinate?	1
10.	Il testo utilizza prevalentemente verbi di forma attiva?	1
11.	Il testo evita di nominalizzare i verbi?	1
12.	Il testo spiega i termini tecnici inevitabili?	1
13.	Il testo utilizza parole precise?	1
14.	Il testo è conciso ed evita parole superflue e ridondanti?	1
15.	Il testo scrive per esteso gli acronimi di uso non comune?	1

Tabella 3: Il questionario di valutazione utilizzato.

Al modello (in questo caso GPT-5) è stato chiesto di valutare ciascun testo sulla base del questionario, assegnando, per ogni quesito, un punteggio proporzionato all'importanza attribuita all'aspetto considerato. Ai primi due *item*, ad esempio, è stato riconosciuto un peso maggiore, perché riguardano la preservazione del contenuto informativo dei testi originali, aspetto da ritenersi più rilevante rispetto ai singoli

parametri linguistici, come l'uso della diatesi attiva dei verbi o la brevità delle frasi. Il modello ha fornito i punteggi richiesti argomentandoli in modo sorprendentemente plausibile, ossia adducendo convincenti ragioni a sostegno delle valutazioni espresse. Tali punteggi sono stati poi convertiti in centesimi, per maggior chiarezza. I risultati sono riportati in Tab. 4.

LLM	**Tecnica di prompt**	**Punteggio medio (/100)**	**Media dei punteggi**
GPT-5	Zero-shot	90,3	90,9
	Few-shot	89,6	
	Chain-of-thought	91,1	
	Tree-of-thought	92,6	
Gemini 2.5 Flash	Zero-shot	87,7	90,8
	Few-shot	91,8	
	Chain-of-thought	88,1	
	Tree-of-thought	95,6	

Tabella 4: Esiti del questionario di valutazione.

Il punteggio medio più alto (95,6) è stato conseguito dai testi semplificati con Gemini e con il *prompt* del tipo *tree-of-thought*, che quindi sembra essere il più efficace, considerando che ha riscosso il punteggio maggiore anche fra i testi riformulati con GPT. Quanto agli LLM, se si calcolano le medie dei punteggi ottenuti da ciascuno, queste risultano pressoché pari (90,9 per GPT, 90,8 per Gemini), quindi non sembrano esserci significative differenze fra i due modelli.

6. Conclusioni

Il primo dato – per nulla sorprendente, in realtà, alla luce della letteratura più recente – che emerge da questo lavoro è che i modelli linguistici di grandi dimensioni, anche in assenza di addestramenti specifici o del supporto di basi documentali esterne, sono strumenti efficaci per ridurre la complessità linguistica dei testi burocratici. Ciò è quanto risulta dal fatto che tutti i testi semplificati hanno conseguito punteggi medi migliori rispetto agli originali, tanto secondo le formule di leggibilità (Gulpease e READ-IT Globale) quanto secondo la valutazione esperta automatizzata basata su criteri di *plain language*.

Rispetto alla domanda di ricerca posta come premessa a questo studio – quale tecnica di *prompt engineering* e quale modello linguistico risultino più efficaci nella

semplificazione dei testi amministrativi – le evidenze emerse consentono di delineare una risposta.

Mentre sul piano della leggibilità formale gli esiti si sono rivelati discordanti, la valutazione esperta automatizzata tramite questionario ha premiato in modo più netto il *prompt* di tipo *tree-of-thought*. Ciò suggerisce che la simulazione di un processo decisionale complesso, in cui punti di vista differenti e potenzialmente conflittuali (chiarezza, concisione, precisione giuridica) sono esplicitamente esaminati e bilanciati, consente al modello di produrre testi che raggiungono un miglior equilibrio tra semplificazione linguistica e tutela dell'integrità informativa.

Per quanto riguarda il confronto tra modelli, non emergono differenze sostanziali tra GPT-5 e Gemini 2.5 Flash: le medie dei punteggi ottenuti risultano pressoché sovrapponibili. Questo suggerisce che, almeno per il compito analizzato, la scelta del modello è meno critica rispetto a quella della tecnica di *prompting*; pertanto, le pubbliche amministrazioni possono scegliere il LLM in base a criteri organizzativi, economici o di *policy*, senza che ciò incida significativamente sulla qualità dei risultati.

Lo studio presenta anche alcuni innegabili limiti che è opportuno esplicitare. Innanzitutto, la dimensione contenuta del *corpus* analizzato riduce la generalizzabilità dei risultati. Un'indagine su un *corpus* più ampio e diversificato potrebbe fornire risultati più attendibili. In secondo luogo, la ricerca non ha coinvolto una valutazione estrinseca, ossia un test su lettori reali, il cui *feedback* è irrinunciabile per misurare l'effettiva comprensibilità. Inoltre, la valutazione esperta affidata a un LLM, per quanto innovativa e pratica, merita ulteriori approfondimenti per comprovarne l'affidabilità. È quindi auspicabile che i miglioramenti nella leggibilità rilevati attraverso metriche e valutazione esperta automatizzata siano in futuro validati attraverso test di comprensione e usabilità su campioni rappresentativi di lettori, così da verificare in modo empirico l'effettiva ricaduta della semplificazione sulla capacità dei cittadini di comprendere e utilizzare le informazioni.

Bibliografia

Alves, Karine, Matheus Silva, Edney Santos, George Valença, and Kellyton Brito. "A Generative AI Approach for Creating and Validating Simplified Versions of Government Documents". *Conference on Digital Government Research* 26 (2025).

Chang, Kaiyan, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. "Efficient Prompting Methods for Large Language Models: A Survey". arXiv preprint. 2404.01077 (2024). https://arxiv.org/abs/2404.01077

Cortelazzo, Michele. *Il linguaggio amministrativo: principi e pratiche di modernizzazione*. Roma: Carocci, 2021.

Dell'Orletta, Felice, Simonetta Montemagni, and 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:*73-83. Stroudsburg, PA:

Association for Computational Linguistics, 2011.

Fiorentino, Giuliana, e Marco Russodivito. "Umano vs AI: alcune considerazioni sulla semplificazione manuale e automatica del corpus ItaIst*"*. In *Amministrazione attiva: Semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, 53-78. Firenze: Cesati, 2025.

Fiorentino, Giuliana, e Vittorio Ganfi. "Parametri per semplificare l'italiano istituzionale: revisione della letteratura." *Italiano LinguaDue* 16(1) (2024): 220-237.

Fortis, Daniele. "Il linguaggio amministrativo italiano". *Revista de Llengua i Dret* 43 (2005): 47-117.

Fortis, Daniele. "L'uso delle formule di leggibilità nella scrittura amministrativa". *Rivista italiana di comunicazione pubblica* 23 (2005): 34-74.

Fortis, Daniele. "Semplificazione del linguaggio amministrativo: validità e limiti delle linee guida". *Rivista italiana di comunicazione pubblica* 20 (2004): 48-83.

Gigliotti, Claudia, e Mariachiara Pascucci. "Assessing the Effectiveness of ChatGPT-3.5 and ChatGPT-4o in Simplifying Italian Institutional Texts." *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, 2(2) (2025).

Gigliotti, Claudia, Walter Paci, Giovanni Acerboni, Alessandro Panunzi, e Maria Roberta Peugini. "Valutazione di tecniche di prompt engineering per la semplificazione dell'italiano burocratico e professionale." In *Amministrazione attiva: Semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, 119-138. Firenze: Cesati, 2025.

Hasebe, Yoichiro, and Jae-Ho Lee. "Divergent Strategies in Text Simplification: A Comparative Analysis of AI and Human Approaches in Language Processing". *Intercultural Pragmatics* 22 (2025): 203-230.

Kimble, Joseph. *Writing for Dollars, Writing to Please: The Case for Plain Language in Business, Government, and Law*. 2nd ed. Durham, NC: Carolina Academic Press, 2023.

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

Lucisano, Pietro, e Maria Emanuela Piemontese. "Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana". *Scuola e Città* 3 (1988): 57–68.

Nozza, Debora, e Giuseppe Attanasio. "Is It Really That Simple? Prompting Large Language Models for Automatic Text Simplification in Italian". In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023),* a cura di Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini e Nicole Novielli, 322–333. Venice: CEUR Workshop Proceedings, 2023.

Paci, Walter, Lorenzo Gregori, Giovanni Acerboni, Alessandro Panunzi e Maria Roberta Perugini. (2024) 2025. "Exploiting ChatGPT to Simplify Italian Bureaucratic and Professional Texts". *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 1 (1) (2024)

Piemontese, Maria Emanuela (a cura di). *Il dovere costituzionale di farsi capire*. Roma: Carocci, 2023.

Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle e Philip Resnik. "The Prompt Report: A

Systematic Survey of Prompt Engineering Techniques". arXiv preprint. 2406.06608 (2024). https://arxiv.org/abs/2406.06608.

Tavosanis, Mirko. "Valutare il miglioramento della chiarezza eseguito da intelligenze artificiali generative". In *Amministrazione attiva: Semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, 103–118. Firenze: Cesati, 2025.

APPENDICE

Sei un **esperto linguistico** specializzato nella semplificazione del linguaggio amministrativo.

Il tuo compito è riscrivere un testo amministrativo in un linguaggio più semplice e chiaro, comprensibile a un **lettore non esperto** nella materia trattata e dotato di un livello di istruzione corrispondente alla **scuola media superiore**.

Mantieni inalterato il contenuto del testo. Non riassumerlo e non condensarlo. Non omettere e non aggiungere alcuna informazione.

Nel riformulare il testo segui queste **linee guida**, dove è possibile:
- Identifica l'informazione più importante per il lettore e mettila all'inizio del testo e poi via via quelle meno importanti.
- Usa liste o tabelle per presentare le informazioni in modo chiaro.
- Fai in modo che ogni frase esprima, di norma, una sola idea.
- Riduci la complessità sintattica e i legami di subordinazione.
- Fai in modo che le frasi non superino, di norma, le 20-25 parole.
- Dai alle frasi una struttura lineare (soggetto + verbo + complemento).
- Trasforma i verbi passivi in verbi attivi.
- Usa frasi affermative anziché negative.
- Evita le frasi impersonali e sostituiscile con frasi con un soggetto espresso.
- Evita la nominalizzazione, cioè preferisci i verbi ai sostantivi astratti che da essi derivano.
- Preferisci le congiunzioni e le preposizioni consistenti in una parola sola alle locuzioni congiuntive e preposizionali formate da più parole.
- Sostituisci le parole burocratiche, antiquate, forbite con sinonimi più comuni.
- Sii conciso: elimina le parole superflue e ridondanti.
- Sostituisci i termini tecnici con parole più comuni di significato equivalente. Se un termine tecnico non ha un sinonimo più comune, usalo ma spiegandolo.
- Spiega gli acronimi, a meno che non siano di uso molto comune.

Il **testo da semplificare** è questo:
«TESTO»

1 – Prompt zero-shot

Sei un **esperto linguistico** specializzato nella semplificazione del linguaggio amministrativo.

Il tuo compito è riscrivere un testo amministrativo in un linguaggio più semplice e chiaro, comprensibile a **un lettore non esperto** nella materia trattata e dotato di un livello di istruzione corrispondente alla **scuola media superiore**.

Mantieni inalterato il contenuto del testo. Non riassumerlo e non condensarlo. Non omettere e non aggiungere alcuna informazione.

Nel riformulare il testo segui queste **linee guida**, dove è possibile:
- Identifica l'informazione più importante per il lettore e mettila all'inizio del testo e poi via via quelle meno importanti.
- Usa liste o tabelle per presentare le informazioni in modo chiaro.
- Fai in modo che ogni frase esprima, di norma, una sola idea.
- Riduci la complessità sintattica e i legami di subordinazione.
- Fai in modo che le frasi non superino, di norma, le 20-25 parole.
- Dai alle frasi una struttura lineare (soggetto + verbo + complemento).
- Trasforma i verbi passivi in verbi attivi.
- Usa frasi affermative anziché negative.
- Evita le frasi impersonali e sostituiscile con frasi con un soggetto espresso.
- Evita le nominalizzazioni, cioè preferisci i verbi ai sostantivi astratti che da essi derivano.
- Preferisci le congiunzioni e le preposizioni consistenti in una parola sola alle locuzioni congiuntive e preposizionali formate da più parole.
- Sostituisci le parole burocratiche, antiquate, forbite con sinonimi più comuni.
- Sii conciso: elimina le parole superflue e ridondanti.
- Sostituisci i termini tecnici con parole più comuni di significato equivalente. Se un termine tecnico non ha un sinonimo più comune, usalo ma spiegandolo.
- Spiega gli acronimi, a meno che non siano di uso molto comune.

Esempi su come devi procedere:
1. Testo originale: «La prenotazione delle visite specialistiche e delle prestazioni diagnostico-strumentali può essere effettuata online.» Testo semplificato: «I pazienti possono prenotare online le visite e gli esami.»
2. Testo originale: «Si invitano i pazienti a comunicare tempestivamente eventuali cambi di residenza.» Testo semplificato: «Se dovesse cambiare residenza, la invitiamo a comunicarcelo rapidamente.»
3. Testo originale: «Al fine di consultare il proprio FSE è necessario essere in possesso della Tessera Nazionale dei Servizi o dello SPID». Testo semplificato: «Per consultare il proprio Fascicolo Sanitario Elettronico (FSE), i pazienti devono avere la Tessera Nazionale dei Servizi o lo SPID.»

Il **testo da semplificare** è questo:
«TESTO»

2 – Prompt few-shot

Sei un **esperto linguistico** specializzato nella semplificazione del linguaggio amministrativo.

Il tuo compito è riscrivere un testo amministrativo in un linguaggio più semplice e chiaro, comprensibile a **un lettore non esperto** nella materia trattata e dotato di un livello di istruzione corrispondente alla **scuola media superiore**.

Nel farlo, segui le seguenti fasi:

1. **ANALIZZA IL TESTO E IL SUO CONTESTO**
- Individua il tema generale del testo e i concetti chiave.
- Identifica l'obiettivo e il destinatario del testo.
- Identifica la gerarchia delle informazioni assegnando a ciascuna un grado di importanza.
- Individua gli elementi che rendono il testo difficile da comprendere a livello sintattico e lessicale.

2. **RIDEFINISCI LA STRUTTURA**
- Metti l'informazione più importante per il cittadino all'inizio.
- Disponi le informazioni in ordine logico.
- Raggruppa le informazioni correlate.
- Usa elenchi puntati per presentare liste o opzioni.
- Inserisci connettivi per guidare la lettura.

3. **SEMPLIFICA LA SINTASSI**
- Spezza le frasi lunghe e complesse in frasi più brevi e lineari.
- Limita le subordinate.
- Segui l'ordine soggetto-verbo-oggetto.
- Usa la forma attiva dei verbi invece di quella passiva, dove possibile.
- Sciogli le nominalizzazioni.

4. **REVISIONA IL LESSICO**
- Sostituisci le parole burocratiche o desuete con parole di uso comune.
- Sostituisci i tecnicismi con parole comuni di significato equivalente, se esistono.
- Se un tecnicismo non ha un sinonimo di uso comune, mantienilo ma spiegalo con parole comuni.
- Sostituisci gli acronimi con la loro forma estesa, a meno che l'acronimo non sia noto.
- Rimuovi le parole superflue e ridondanti.

5. **VERIFICA IL RISULTATO**
- Verifica che il testo semplificato sia chiaro, scorrevole e coerente.
- Controlla che non ci siano ambiguità.
- Verifica che la versione semplificata non abbia omesso, aggiunto o modificato informazioni importanti.
- Assicurati che il contenuto giuridico del testo sia stato preservato.

Il **testo da semplificare** è questo:
"TESTO"

3 – Prompt chain-of-thought

Immagina di avere un team di **3 esperti linguistici** specializzati nella semplificazione del linguaggio amministrativo.

Il loro compito è riscrivere un testo amministrativo in un linguaggio più semplice e chiaro, comprensibile a **un lettore non esperto** nella materia trattata e dotato di un livello di istruzione corrispondente alla **scuola media superiore**.

Il testo deve rimanere **inalterato nel contenuto** e fedele all'originale.

Il **primo esperto** privilegia la **chiarezza e la leggibilità** e produce una versione semplificata del testo adottando un approccio radicale nell'applicare le *plain language guidelines* consolidate, ad esempio:
- antepone le informazioni principali a quelle secondarie
- riduce le subordinate
- sostituisce le costruzioni impersonali con costruzioni personali, anche dando del tu al destinatario
- converte i verbi passivi in verbi attivi
- scioglie le nominalizzazioni
- sostituisce le parole tecniche o burocratiche con parole comuni

Il **secondo esperto** dà priorità alla **concisione** e produce una versione semplificata del testo incentrata sull'eliminazione del superfluo:
- elimina le informazioni non essenziali
- accorcia le frasi
- rimuove le ripetizioni e le ridondanze
- elimina i riferimenti normativi se non funzionali allo scopo e al destinatario del testo

Il **terzo esperto** si focalizza sul mantenimento della **precisione giuridica** e produce una versione semplificata del testo adottando un approccio conservativo:
- identifica i termini giuridici da preservare
- mantiene i riferimenti normativi necessari
- si assicura che non vi siano ambiguità interpretative
- si accerta che nessuna informazione giuridicamente rilevante sia omessa

I tre esperti si riuniscono e **confrontano le proprie versioni**. **Soppesano le varie opzioni** e valutano le esigenze in conflitto, ad esempio:
- il primo vuole sciogliere le nominalizzazioni, ma il secondo obietta che ciò renderebbe il testo più verboso e meno compatto
- il secondo vuole eliminare i connettivi, ma il primo obietta che ciò renderebbe oscuri i nessi logici fra le informazioni
- il primo vuole sostituire i tecnicismi giuridici con parole più comuni, ma il terzo sostiene che i termini proposti non sono giuridicamente rigorosi

Dopo varie interazioni i 3 esperti raggiungono un giusto compromesso ed elaborano un testo che bilancia chiarezza, concisione e precisione giuridica.

Il **testo da semplificare** è questo:
«TESTO»

4 – Prompt tree-of-thought

Large Language Model per la semplificazione del linguaggio amministrativo: una valutazione comparativa tra SEMPL-IT, ChatGPT 5.1, Claude Sonnet 4.6 e Gemini 3

Vittorio Ganfi
(Università degli Studi di Bergamo)

Abstract

Dopo aver richiamato il ruolo della semplificazione manuale e i suoi limiti di scalabilità, il lavoro discute la semplificazione mediata da *Large Language Models* (LLM), mettendo in evidenza la centralità del prompt e la necessità di revisione umana in domini sensibili. Viene presentata una versione aggiornata di SEMPL-IT, sistema specializzato per l'italiano amministrativo, e viene impostato un confronto con modelli generalisti (ChatGPT 5.1, Claude Sonnet 4.6 e Gemini 3). L'analisi combina metriche di leggibilità, indicatori lessicali/strutturali e misure di similarità (*edit distance* e *semantic similarity*). I risultati mostrano che SEMPL-IT modula la profondità della riscrittura in funzione del destinatario mantenendo un'elevata preservazione semantica; i modelli generalisti incrementano la leggibilità in misura variabile, con Gemini 3 che massimizza gli indici di leggibilità ma presenta una riduzione della similarità semantica. L'analisi qualitativa conferma che, in testi con valore giuridico-amministrativo, l'impiego di LLM richiede controlli redazionali per prevenire slittamenti di contenuto, omissioni e possibili allucinazioni.

1. Introduzione

Molti testi prodotti dalla pubblica amministrazione risultano, per un'ampia parte dei destinatari, difficili da interpretare. La difficoltà non è motivata soltanto dalla presenza di tecnicismi, ma anche da una combinazione di altri fattori: sintassi ipotattica, lessico desueto e una progettazione testuale che non sempre rende evidente la gerarchia delle informazioni. La letteratura sulla scrittura amministrativa e sulla chiarezza dei documenti pubblici ha descritto con ampiezza questi tratti e le loro conseguenze sulla leggibilità (Fioritto 1997, 2009, 2018; Piemontese 1999, 2023; Cortelazzo & Pellegrino 2003; Cortelazzo 2015, 2021). Per questo motivo, la semplificazione linguistica è stata proposta come strategia per favorire l'accesso all'informazione e per ridurre la distanza comunicativa tra istituzioni e cittadini.

Nel contesto attuale, l'impiego di Large Language Models (LLM) permette riscritture automatiche rapide e formalmente scorrevoli. Proprio la rapidità, tuttavia, impone

criteri di controllo stringenti in ambiti come quello amministrativo, dove una riformulazione imprecisa può produrre effetti interpretativi rilevanti. Il presente contributo intende mostrare alcune potenzialità e alcuni limiti correlati alla semplificazione attraverso gli LLM, e presenta due scopi principali: (i) la presentazione di una versione aggiornata di SEMPL-IT, sistema dedicato alla semplificazione dell'italiano amministrativo (Fiorentino et al. in stampa b), e (ii) il confronto con LLM generalisti, ovvero GPT 5.1 (OpenAI 2025), Claude Sonnet 4.6 (Claude 2026) e Gemini 3 (Laurent 2025). Nei seguenti paragrafi verranno individuate brevemente alcune caratteristiche della complessità amministrativa. Vengono, quindi, confrontate la semplificazione manuale e quella mediata dai LLM. Si descrive la struttura di SEMPL-IT, mostrando come è stata disegnata e realizzata la risorsa, e infine si confrontano le semplificazioni effettuate dai vari LLM, integrando metriche quantitative e osservazioni qualitative.

2. La complessità amministrativa: caratteristiche e ricadute

Come noto, una parte consistente della comunicazione istituzionale è considerata come distante e opaca dai cittadini (Lubello 2014, 2017). Le cause di questa oscurità risiedono spesso nella pratica redazionale contraddistinta da terminologia tecnico-burocratica impiegata in modo sistematico; periodi lunghi con molte subordinate; scarsa segmentazione e macrostrutture prive di segnali orientativi (titoli, paragrafi, elenchi). Parallelamente, l'ultimo decennio mostra un rinnovato investimento verso le buone pratiche nella scrittura istituzionale. Si possono citare, a titolo puramente esemplificativo, il Manuale di comunicazione istituzionale e internazionale curato da Bombi (2013); le raccomandazioni operative per una scrittura più chiara di Cortelazzo (2021) e il bilancio di Piemontese (2023) che soppesa l'effettivo recepimento delle buone prassi di scrittura a trent'anni dalla pubblicazione del fondamentale Codice di Stile (Cassese 1994). A questi lavori vanno aggiunte sia esperienze di riscrittura e formazione in ambito giuridico (Romano 2023), sia studi che documentano la collaborazione tra linguisti e giuristi e che motivano l'adozione di tecnologie innovative per migliorare i testi amministrativi (Romano e Venturi 2025). La richiesta di chiarezza nella comunicazione pubblica è, infatti, centrale per ogni cittadino (Piemontese 2013, Cassese 2023, Vellutino e De Martino 2025, Simonelli 2025) e diventa ancora più urgente per persone con competenza linguistica non nativa (Vedovelli 2013, Miglietta 2015), poiché le varietà burocratiche si dimostrano scarsamente accessibili ai non nativi. La necessità di aumentare la leggibilità dei testi prodotti dalla pubblica amministrazione deriva da un confronto con la legislazione europea, come indicano analisi che raffrontano i livelli di complessità nei contesti legislativi italiani e comunitari (Mori 2019).

Nonostante la significativa mole di contributi, la scarsa trasparenza dei testi amministrativi resta un problema irrisolto (Piemontese 2023). Tale fenomeno riduce la comprensibilità dei testi pubblici, produce una interazione meno consapevole con le istituzioni, e può erodere la fiducia verso le pubbliche amministrazioni. Sul piano

analitico, si possono distinguere due tendenze frequenti, che orientano le scelte stilistiche degli estensori dei testi pubblici: (a) tendenza offuscatrice: ovvero la preferenza di strutture meno trasparenti (es. *ovvero* invece di oppure); e (b) la tendenza incrementale: scelta di strutture sintattiche più complesse (es. ipotassi anziché paratassi). La descrizione analitica di questi tratti è stata formalizzata da Fiorentino e Ganfi (2024), che propongono 13 parametri morfosintattici e 8 lessicali per rappresentare la complessità amministrativa. Fra le criticità che ricorrono con maggiore frequenza si osservano: i periodi molto lunghi, specialmente quando la subordinazione è intensa; l'uso estensivo dei passivi che risultano sovente meno immediati delle corrispettive costruzioni attive; le strutture informative poco guidate, con segmentazione ridotta e assenza di marcatori di percorso (titoli, sottotitoli, elenchi ecc.); e il predominio di un lessico connotato da una bassa frequenza d'impiego. Alla difficoltà sintattica va, infatti, sommata quella lessicale. Tra i lessemi poco usuali impiegati nei testi amministrativi, a titolo esemplificativo, si possono ricordare anglismi quali *open data, digital tax, gig economy, flat tax, step-child adoption, privacy, prosumer, sharing economy, cloud computing, crowdfunding, empowerment, performance, open access, service tax, smart cities, turnover* (cfr. Bombi 2013, 2015).

L'insieme delle caratteristiche menzionate crea un effetto cumulativo, in quanto più livelli di complessità aumentano il carico cognitivo e rendono più probabile una ricostruzione incompleta dell'informazione. Chi non dispone di adeguate risorse linguistiche e culturali rischia, di conseguenza, una forma di marginalizzazione nell'accesso ai servizi e ai diritti veicolati dai testi (Lubello e Nobili 2020).

3. Semplificazione manuale: controllo, tempi e costi

Nella letteratura scientifica relativa alla scrittura istituzionale, la semplificazione manuale è considerata un dispositivo efficace. Consente, infatti, di governare in modo puntuale le scelte di riscrittura e di mantenere l'equilibrio fra la chiarezza e la precisione necessaria nei testi di matrice giuridica e amministrativa (Cioffi 2025). Tuttavia, il limite principale della riscrittura manuale è la scalabilità, poiché risulta particolarmente oneroso impiegarla in modo sistematico per grandi volumi di documenti. I costi significativi prodotti da un impiego estensivo della scrittura manuale derivano soprattutto da tre fattori. Primariamente, la riscrittura manuale necessita di tempi di lavorazione ingenti, richiedendo lettura integrale, diagnosi dei punti critici e persino più cicli di revisione (Ferrari e Piantanida 2025). Secondariamente, i testi istituzionali presuppongono una verifica finale di un esperto, che deve appurare la completezza informativa e la correttezza giuridica del testo riscritto (Monaco 2025). In ultimo, la riscrittura manuale richiede competenze metalinguistiche avanzate, implicando la capacità di riconoscere in modo sistematico le fonti della complessità e di intervenire coerentemente su queste ultime. Un recente contributo sorregge, sul piano empirico, queste considerazioni: Fiorentino e Russodivito (2025) riportano che due semplificatori esperti hanno impiegato in media 20 ore per semplificare manualmente un sottocorpus di ItaIst

(Fiorentino et al. 2024) composto da 619 unità testuali (33.297 token) ricavate da otto documenti. Benché il campione testuale considerato non sia estremamente ampio, l'ordine di grandezza mostra quanto risulti complessa l'estensione di questo metodo di semplificazione a tutti i testi della pubblica amministrazione, senza un aumento proporzionale, e pertanto estremamente ingente, delle risorse.

Operativamente, la semplificazione manuale può essere descritta come una procedura che consta di più passaggi, ciascuno dei quali non può venire ridotto a scelte meccaniche. Il processo può essere ripartito nelle seguenti fasi operative: (a) lettura attenta; (b) individuazione delle criticità sul piano lessicale, sintattico e testuale; (c) applicazione di regole e raccomandazioni derivate da linee guida e manuali (ad es. Cassese 1994; Fioritto 1997); (d) revisione finale per verificare coerenza degli interventi e mantenimento dei contenuti. Dato che ogni riscrittura può incidere sull'interpretazione, il redattore deve valutare singolarmente ogni fase di rielaborazione testuale. Nel complesso, il processo richiede una competenza procedurale, ovvero la capacità di applicare in modo coerente principi di semplificazione ai testi amministrativi. Questa combinazione di competenze alte, tempi lunghi e costi significativi spiega l'interesse per soluzioni automatiche o ibride. Nella seguente immagine vengono schematizzate le fasi messe in atto nella semplificazione manuale:

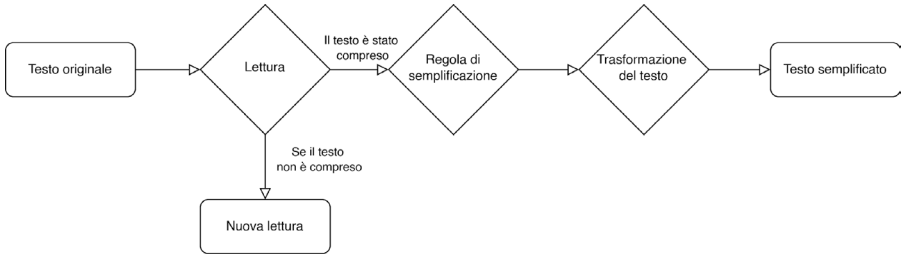

Figura 1: processo di semplificazione manuale (Ganfi e Russodivito 2025)

4. Semplificazione automatica con LLM: prompt e revisione

L'ampia produzione documentale delle pubbliche amministrazioni necessita, con ogni evidenza, di strumenti che riducano tempi e costi della riscrittura manuale. In questo ambito di impiego non possono essere trascurate le rilevanti potenzialità degli LLM, che permettono la generazione di testi contraddistinti da un livello di coerenza e fluidità elevato (Fiorentino e Tavosanis 2025). Va premesso al prosieguo dell'analisi condotta in questo paragrafo che l'uso degli LLM non elimina la necessità di presidio umano. La semplificazione automatica deve essere considerata un supporto al redattore e non una completa automazione del processo. Quest'ultima asserzione è particolarmente cogente in contesti dove le ricadute giuridiche e sociali di un eventuale errore della

macchina possono essere oltremodo rilevanti (Tavosanis 2024; Pascucci e Tavosanis 2024).

Il cambiamento di paradigma correlato all'impiego degli LLM non è meramente quantitativo (ovvero una minore esigenza di tempo e risorse), ma qualitativo: in quanto presuppone una riorganizzazione dell'intero processo di semplificazione e coinvolge competenze diverse da quelle impiegate per la semplificazione manuale (Ganfi e Russodivito 2025). L'utente interagisce con il sistema principalmente attraverso il prompt, che diventa un'interfaccia operativa capace di tradurre obiettivi comunicativi in istruzioni[1]. A differenza della riscrittura manuale, spesso iterativa e dipendente dall'esperienza del redattore (Fiorentino et al. in stampa a), la generazione di un testo semplificato con LLM avviene tipicamente in un singolo passaggio computazionale. La trasformazione non è ottenuta applicando regole esplicite in sequenza, ma emerge da un'azione globale che, attraverso l'impiego di un prompt, integra istruzioni e conoscenza implicita delle regole di semplificazione. Se non istruito con prompt adeguati, il modello può, quindi, produrre riscritture non allineate alle linee guida istituzionali italiane. Con l'uso degli LLM, la competenza richiesta all'utente viene rimodulata: è meno centrale la capacità di riscrivere e viene, invece, richiesta la capacità metalinguistica di tradurre in prompt chiari e controllati le esigenze della semplificazione. Come anticipato, anche in presenza di output all'apparenza soddisfacenti, rimane imprescindibile una revisione umana dei testi riscritti dai modelli linguistici per emendare omissioni, errori o riformulazioni fuorvianti. Nello schema riportato sotto viene rappresentato il processo di semplificazione automatica mediato dagli LLM:

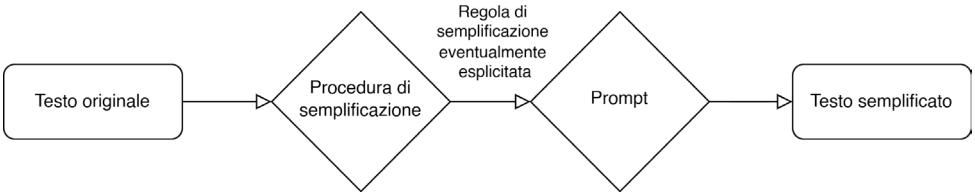

Figura 2: processo di semplificazione mediata dagli LLM (Ganfi e Russodivito 2025)

1 Il tema ha un rilievo non marginale nel dibattito, particolarmente acceso negli studi giuridici, relativo all'impiego dell'IA in ambito istituzionale (Monaco 2025).

5. SEMPL-IT e modelli generalisti: disegno di confronto

5.1 Prompting e granularità: zero-shot vs prompt chaining

Il confronto presentato in questo contributo paragona SEMPL-IT, un sistema specializzato per la semplificazione dell'italiano amministrativo (Fiorentino et al. in stampa b), e alcuni LLM generalisti, ovvero GPT 5.1, Claude Sonnet 4.6 e Gemini 3. In particolar modo, si intende mostrare la diversa granularità nelle semplificazioni correlata ad alcune strategie di prompting (Sahoo et al. 2024), inglobate nella struttura di SEMPL-IT. Nei seguenti paragrafi viene messo in luce che attraverso un prompt zero-shot il modello, ricevendo una sola istruzione generale, predilige una riscrittura globale dei testi semplificati. La scelta di un prompt chaining, di contro, produce una riscrittura guidata in cui gli interventi progressivi riguardano specifici aspetti linguistici. Come si vedrà, l'architettura di SEMPL-IT, inglobando queste due diverse strategie di prompting, può adattarsi a diverse esigenze degli utenti.

5.2 Dati e architettura di SEMPL-IT

SEMPL-IT è stato progettato per essere un *semplificatore testuale automatico* capace di adattare il grado della semplificazione alle esigenze degli utenti[2]. L'architettura del software prevede l'incorporazione di due modelli complementari, costruiti a partire da corpora paralleli generati seguendo strategie di semplificazione diverse. (a) SEMPL-IT esperti è disegnato per replicare trasformazioni granulari e guidate, preservando il valore giuridico e tratti testuali tipici del registro amministrativo. (b) SEMPL-IT utenti comuni è stato, invece, addestrato su dati ottenuti con prompt zero-shot, interviene in modo più incisivo, aumentando la leggibilità e modificando più profondamente la superficie del testo.

Nella versione più recente, SEMPL-IT integra, quindi, due modelli complementari che emulano strategie di prompt chaining e di zero-shot prompt per la semplificazione testuale. SEMPL-IT ricalca l'addestramento di modelli open-weight specializzati, in continuità con le pratiche di prompt design discusse in letteratura (Wei et al. 2022; Sahoo et al. 2024; Schulhoff et al. 2024; Amatriain 2024). Rappresenta, pertanto, un'evoluzione significativa rispetto alle precedenti versioni del sistema (Ganfi & Russodivito 2025). In particolare, le due diverse versioni del modello implementano diverse strategie di semplificazione: SEMPL-IT utenti comuni, replicando una strategia di zero-shot prompting, effettua una semplificazione più marcata, mentre SEMPL-IT esperti produce semplificazioni meno radicali, preservando maggiormente contenuto e forma degli originali. SEMPL-IT esperti adotta, infatti, un'architettura modulare che prevede modelli coordinati, ciascuno addestrato per una singola regola

2 SEMPL-IT è stato sviluppato dall'unità dell'Università del Molise coordinata da Giuliana Fiorentino e coinvolta nel progetto PRIN VerbACxSS.

di semplificazione (sintattica, lessicale o testuale). Tra le regole applicate progressivamente rientrano la disposizione dei costituenti in un ordine SVO, la trasformazione delle costruzioni passive in attive, la riduzione delle nominalizzazioni, la segmentazione dei periodi lunghi, la normalizzazione dei riferimenti normativi, la sostituzione di connettivi/espressioni burocratiche e la correzione di errori ortografici, grammaticali e di punteggiatura. L'uso di una serie di regole sequenziali ha lo scopo di intervenire sulla forma senza compromettere contenuto e valenza giuridica.

Il processo di sviluppo della risorsa[3] ha previsto diverse fasi di lavoro progressive: (1) sono state individuate le regole sulla base della letteratura scientifica; (2) è stato raccolto un primo sottocorpus di 24 documenti (circa 180.000 token) estratti dal corpus ItaIst[4]; (3) sono state create versioni semplificate del corpus attraverso GPT-4o usando prompt chaining e few-shot prompting (Sahoo et al. 2024; Schulhoff et al. 2024), ottenendo corpora paralleli distinti dalla regola di semplificazione applicata (Russodivito et al. 2024); (4) i corpora paralleli sono stati impiegati per l'addestramento e la validazione. Come base open-weight per la costituzione del modello è stato adottato Qwen2.5 (Qwen Team 2025) da 7B parametri[5], in ragione del bilanciamento tra capacità, sostenibilità computazionale e qualità dell'italiano generato. (5) Nella fase finale i modelli

3 Una più completa descrizione del processo che ha condotto allo sviluppo di SEMPL-IT è presentata in Fiorentino et al. in stampa b.

4 ItaIst è un corpus rappresentativo dell'italiano istituzionale che colleziona diverse tipologie testuali (carte dei servizi, bandi, atti di programmazione, accreditamento, testi sulla razionalizzazione delle partecipate). I documenti sono stati selezionati manualmente dai siti istituzionali di enti pubblici appartenenti a otto regioni, con un bilanciamento geografico tra nord, centro e sud (Basilicata, Calabria, Campania, Lazio, Lombardia, Molise, Toscana, Veneto). La versione del corpus utilizzata per l'addestramento di SEMPL-IT comprende 108 documenti. Per la versione finale del corpus si rimanda a Fiorentino et al. (2024), ItaIst corpus. https://huggingface.co/datasets/VerbACxSS/ItaIst.

5 Qwen2.5 appartiene a una generazione recente di modelli open-weight caratterizzati da un'architettura decoder-only moderna e da prestazioni competitive su una vasta gamma di compiti di elaborazione del linguaggio naturale, come documentato nel relativo technical report (Qwen Team 2025), e offre il vantaggio di poter essere eseguito, ispezionato e adattato anche in ambienti offline, aspetto cruciale per scenari applicativi istituzionali caratterizzati da vincoli di sicurezza, protezione dei dati e requisiti di sovranità tecnologica. La scelta della variante 7B consente inoltre di adottare in modo efficace tecniche di Parameter Efficient Fine-Tuning, in particolare approcci basati su QLoRA (Dettmers et al. 2023), che permettono di addestrare più modelli specializzati aggiornando solo una frazione limitata dei parametri, riducendo significativamente i costi computazionali ed energetici e rendendo praticabile l'addestramento di un insieme di modelli dedicati, ciascuno corrispondente a una specifica regola di semplificazione. L'intera pipeline di SEMPL-IT si ispira al paradigma della model distillation (Mansourian et al. 2025), in cui GPT-4o svolge il ruolo di teacher generando esempi di semplificazione utilizzati per addestrare modelli student più piccoli ed efficienti, implementati come adapter QLoRA innestati sul backbone Qwen2.5 e addestrati su infrastrutture GPU NVIDIA L40S mediante la libreria open source LLaMA-Factory (Fiorentino et al. in stampa b).

sono stati rifiniti con una procedura di fine-tuning. Nello schema seguente è riassunto il processo di addestramento di SEMPL-IT:

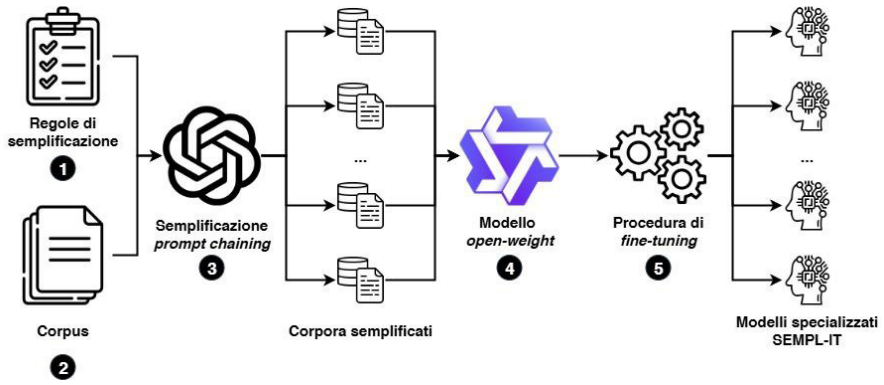

Figura 3: schema di addestramento di SEMPL-IT (Fiorentino et al. in stampa b)

Come mostra l'immagine 3, le regole sequenziali, individuate attraverso l'analisi della letteratura, hanno guidato la costruzione di più corpora paralleli che, in una ulteriore fase di sviluppo, sono stati usati per l'addestramento di Qwen. Dopo un'ulteriore fase di validazione del sistema, sono stati sviluppati i sottomodelli che permettono a SEMPL-IT esperti di applicare le regole sequenziali. Per lo sviluppo di SEMPL-IT utenti comuni è stato impiegato, invece, un dataset di addestramento ricavato da una versione di ItaIst semplificata con un prompt zero-shot, istruendo GPT-4o per ottenere una riformulazione più radicale. La struttura duplice di SEMPL-IT consente, quindi, all'utente di scegliere il livello di leggibilità più adatto al modello comunicativo di riferimento. Un redattore che scrive per interlocutori interni alla PA può preferire SEMPL-IT esperti per incrementare la leggibilità senza modificare eccessivamente struttura e valore giuridico dell'originale. Un cittadino o un utente esterno alla PA può optare per SEMPL-IT utenti comuni per ottenere una riscrittura più esplicita e facilmente accessibile.

I modelli sono consultabili tramite un sito web[6], provvisto di un'interfaccia facilmente accessibile. L'utente incolla il testo nell'apposita casella, seleziona la modalità (esperti o utenti comuni) e avvia la semplificazione con il pulsante «semplifica». Un esempio dell'interfaccia grafica è riportato di seguito.

6 L'ultima versione del modello rilasciata nel 2025 è accessibile utilizzando il seguente sito web https://sempl-it.unimol.it/home.

Figura 4: interfaccia di SEMPL-IT con la scelta tra utenti esperti e comuni.

5.3 Metriche e testo di confronto

Seguendo una procedura empirica[7] adottata in un altro studio (cfr. Ganfi in stampa), per valutare l'incremento di leggibilità sono stati impiegati due indici che misurano la complessità globale dei testi[8], ovvero gli indici di Gulpease (Lucisano e Piemontese 1988) e di Flesch-Vacca (Franchina e Vacca 1986)[9]. Sono state, inoltre, misurate alcune caratteristiche lessicali e strutturali che contraddistinguono i testi più complessi: per l'analisi vengono presi in considerazione i seguenti parametri:

1. *Percentuale di lemmi nel Nuovo vocabolario di base* (De Mauro e Chiari 2016). Il dato misura quanta parte del lessico appartiene a un repertorio di parole impiegate di frequente e maggiore disponibilità per un lettore medio. In generale, una percentuale più alta di lemmi appartenenti al vocabolario di base tende ad aumentare la comprensibilità, perché riduce la presenza di termini rari o specialistici.

7 Per l'analisi delle metriche è stata impiegata la piattaforma inglobata nel sito web **https://sempl-it.unimol.it/home** che permette di misurare il grado di complessità dei testi. In Ganfi in stampa sono introdotti alcuni aspetti preliminari discussi compiutamente in questo contributo.

8 Una ricognizione delle tecnologie impiegabili nella semplificazione dei testi amministrativi è presentata in Brunato e Venturi (2014).

9 L'indice di Gulpease (Lucisano & Piemontese, 1988) è progettato specificamente per l'italiano. Il punteggio aumenta all'aumentare della leggibilità. Quindi, testi caratterizzati da frasi più brevi e da parole mediamente meno lunghe tendono a ottenere valori più elevati. L'indice di Flesch–Vacca è un indice di leggibilità che stima la difficoltà di un testo a partire da parametri quantitativi quali il numero medio di sillabe per parola e il numero medio di parole per frase. Analogamente al Gulpease, anche in questo caso valori più alti corrispondono, in linea generale, a testi più accessibili. Pur condividendo la stessa direzione interpretativa, in quanto punteggi crescenti documentano una maggiore leggibilità, i due indicatori si differenziano per la base analitica, poiché l'indice di Gulpease è calibrato sull'italiano, mentre l'indice di Flesch–Vacca si fonda su una misurazione non specifica della complessità strutturale.

2. *Quantità di connettivi difficili*. Queste unità lessicali segnalano relazioni logiche tra proposizioni e segmenti testuali (causa, concessione, conseguenza, ecc.). I connettivi meno frequenti (ad esempio *nondimeno, laddove*) e quelli costituiti da più unità lessicali (ad esempio, *ai sensi di, a ragion di*) risultano meno trasparenti e possono aumentare lo sforzo interpretativo.
3. *Numero di frasi presenti nel documento*. Il parametro misura la segmentazione sintattica del testo. A parità di contenuto, un testo che colleziona un numero minore di frasi tende a presentare periodi più articolati e, pertanto, più complessi. Di contro, un numero maggiore di frasi si correla alla predilezione per la paratassi e per la linearità, producendo un incremento della leggibilità.
4. *Densità lessicale*. Il valore pesa la quota di parole piene (nomi, verbi, aggettivi, avverbi) rispetto al totale delle parole impiegate in un documento. Un testo provvisto di alta densità lessicale contiene molte unità informative concentrate e può, quindi, risultare più complesso da interpretare. Una minore densità lessicale si correla a una facilitazione della lettura.

Per stimare la fedeltà al testo di partenza sono state impiegate due diverse misure complementari:

1. *Edit distance* (Levenshtein, 1966). Questa metrica quantifica il numero minimo di operazioni (ad esempio, inserzioni, cancellazioni, sostituzioni) necessarie per trasformare una stringa nell'altra. Il valore, quindi, può essere impiegato per mostrare la prossimità esteriore di due testi.
2. *Semantic similarity* (Agirre et al., 2016). La metrica sfrutta rappresentazioni vettoriali distribuzionali per misurare la prossimità semantica di due testi anche quando differiscono formalmente.

Considerando l'insieme delle metriche e il novero delle caratteristiche linguistiche è possibile ricavare il grado di leggibilità e l'aderenza al testo originale dei documenti semplificati dai vari LLM. Nello specifico, si confronta in maniera articolata un testo originale, esemplificativo dei documenti prodotti dalle pubbliche amministrazioni, e le versioni generate da SEMPL-IT esperti, SEMPL-IT utenti comuni, GPT 5.1, Claude Sonnet 4.6 e Gemini 3. Tutti gli LLM ricevono come input un documento identico e per la valutazione dell'output si applicano gli stessi criteri di analisi. Inoltre, tutti i modelli generalisti sono interrogati attraverso il medesimo prompt zero-shot[10], in maniera da confrontare SEMPL-IT con un uso di un utente sprovvisto di nozioni di prompt design. La comparazione con gli LLM generalisti si rivela particolarmente proficua perché può rivelare limiti e potenzialità di modelli specializzati. Un modello specializzato, come SEMPL-IT, incorpora, infatti, una distinzione fra livelli di trasformazione,

10 Il prompt utilizzato in tutti gli LLM è «Semplifica questo testo in modo da renderlo più comprensibile al pubblico generale». Poiché il prompt può modulare fortemente la qualità della semplificazione (Fiorentino e Russodivito 2025), per ridurre la variabilità si è scelto un'istruzione essenziale. Ovviamente per interrogare le due versioni di SEMPL-IT non è stato impiegato alcun prompt, visto che l'interfaccia grafica del sistema non prevede l'uso di prompt espliciti.

che non è esplicitamente presente nei modelli generalisti sprovvisti di vincoli espliciti volti a preservare il valore giuridico del testo. Il testo originale consiste in una porzione della Carta dei Servizi dell'Assessorato del Territorio e dell'Ambiente della Regione Siciliana (28/05/2024), non presente nel dataset di addestramento di SEMPL-IT[11].

5.4 Risultati quantitativi

L'analisi che segue non mira esclusivamente a individuare il sistema che ottiene i migliori risultati nella semplificazione linguistica relativa all'italiano amministrativo, ma intende offrire uno spunto di riflessione sulle caratteristiche che differenziano due paradigmi nell'uso degli LLM, ovvero (i) la semplificazione controllata, basata su regole applicate in sequenza, paradigma rappresentato da SEMPL-IT esperti; e (ii) la semplificazione globale in zero-shot, nel campione indagato il paradigma è adottato sia da SEMPL-IT utenti comuni, sia dai modelli generalisti. Attraverso le metriche di similarità (*edit distance* e *semantic similarity*) si mette in mostra quanto i diversi paradigmi tendano a variare nell'allontanarsi dall'originale sia per la veste formale sia per il contenuto. L'insieme delle caratteristiche prese in considerazione permette, pertanto, di profilare un quadro articolato, per quanto non esaustivo, delle aspettative che gli utenti possono avere nel seguire i due diversi paradigmi e parallelamente delle esigenze cui rispondono le due diverse strategie di semplificazione. La tabella seguente riassume le principali metriche calcolate sul testo originale e sulle versioni semplificate.

	Originale	**SEMPL-IT esperti**	**SEMPL-IT utenti comuni**	**ChatGPT 5.1**	**Claude Sonnet 4.6**	**Gemini 3**
Gulpease	41	48	62	56	57	71
Flesch-Vacca	18	30	50	46	47	60
Vocabolario di base	290	293	311	284	183	210
Connettivi difficili	5	1 (-80 %)	0	1	1 (-80 %)	0
Numero frasi	14	23	36	26	16	32
Densità lessicale	57 %	56 %	52 %	52 %	53 %	49 %
Edit distance	/	358 (15 %)	44 %	63 %	65 %	68 %
Semantic similarity		100 %	96 %	98 %	90 %	89 %

Tabella 1: metriche di confronto nei vari sottogruppi

11 Per assicurare la confrontabilità, tutti i modelli sono applicati allo stesso estratto testuale (Carta dei Servizi dell'Assessorato del Territorio e dell'Ambiente della Regione Siciliana, 28/05/2024), non incluso nei dati di addestramento di SEMPL-IT, e vengono valutati con criteri identici. In appendice è riportato l'esempio di input e le riscritture prodotte.

L'analisi quantitativa conferma, in primo luogo, un incremento sistematico della leggibilità nelle riscritture rispetto all'originale, che presenta un indice Gulpease di 41 e un Flesch-Vacca di 18. La graduazione della semplificazione rispecchia il paradigma adottato nell'istruire gli LLM. SEMPL-IT esperti migliora in modo moderato (con un indice Gulpease di 48 e un Flesch-Vacca di 30), mentre SEMPL-IT utenti comuni mostra un incremento più netto degli indici di leggibilità, raggiungendo rispettivamente 62 e 50. Il dato è correlato alla predilezione per riscritture più didattiche[12]. Tra i modelli generalisti, ChatGPT 5.1 e Claude Sonnet 4.6 si collocano su valori intermedi e molto vicini, presentando un Gulpease di 56 e 57 e un Flesch-Vacca di 46 e di 47. Gemini 3, invece, è il modello che raggiunge la leggibilità più alta, in quanto mostra un Gulpease di 71 e un Flesch-Vacca di 60[13]. I dati sulla complessità generale sono sostenuti dagli altri parametri misurati: il numero di frasi cresce in tutte le versioni rispetto alle 14 dell'originale, ma con una distribuzione diversificata[14], e la segmentazione dei periodi sospinge l'aumento degli indici, soprattutto nei casi di SEMPL-IT utenti comuni e Gemini 3 in cui si registra l'incremento maggiore nel numero di frasi. In parallelo, i connettivi difficili diminuiscono drasticamente[15], mostrando che tutti gli LLM prediligono la sostituzione di connettivi tipici della prosa amministrativa con nessi più lineari. Tutti i modelli registrano un calo dell'originale densità lessicale del 57% a valori progressivamente più bassi[16]. La riduzione della densità lessicale si traduce nella distribuzione dell'informazione su più unità testuali e nell'incremento di elementi funzionali e istruzionali, tendenza particolarmente evidente nelle riscritture di Gemini 3. Sul piano lessicale, il conteggio assoluto delle occorrenze riconducibili al vocabolario di base cresce nelle due versioni di SEMPL-IT, passando dai 290 lessemi originali ai 293 del modello esperti ai 311 del modello utenti comuni, verosimilmente anche per effetto di esplicite istruzioni di dominio lessicale presenti nell'addestramento dei modelli. Il dato risulta inferiore, invece, nei generalisti che documentano tutti una riduzione rispetto all'originale[17]. Tuttavia, trattandosi di un dato non normalizzato e avendo i testi dimensioni diverse, la percentuale più bassa di vocabolario di base negli LLM generalisti può essere motivata con l'estensione più ridotta e con la maggiore schematicità degli output generati da questi sistemi.

12 Si consideri a titolo esemplificativo la seguente riscrittura di SEMPL-IT utenti comuni: «Questi atti possono essere consultati in due modi: 1. Online… 2. Per richiesta…».

13 Gemini 3 giunge a generare, infatti, trasformazioni che scompongono le procedure indicate nell'originale ed esposte analiticamente rivolgendosi direttamente al lettore/utente: «Puoi farlo in due modi»; «L'URP è il punto di contatto tra te e il Corpo Forestale».

14 Appaiono 23 frasi in SEMPL-IT esperti; 36 in SEMPL-IT utenti comuni; 26 in ChatGPT 5.1; 16 in Claude; 32 in Gemini

15 I connettivi difficili passano da 5 a 1 in SEMPL-IT esperti, ChatGPT 1 e Claude 1, mentre non ne appare nessuno nelle versioni semplificate con Gemini 0 e SEMPL-IT esperti.

16 Presentando i seguenti rispettivi valori: il 56% in SEMPL-IT esperti, il 52% in SEMPL-IT utenti comuni, il 52% in ChatGPT 5.1, il 53% in Claude e il 49% in Gemini 3.

17 I lessemi del vocabolario di base sono 284 in ChatGPT, 183 in Claude e 210 in Gemini 3.

Le misure di fedeltà dimostrano l'esistenza di una correlazione tra un più moderato incremento della leggibilità e preservazione semantica del contenuto originale. SEMPL-IT esperti, che, come visto, presenta i dati di leggibilità meno accentuati, mantiene la similarità semantica massima e una edit distance estremamente contenuta (358 operazioni; 15%). Questi dati collimano con il paradigma di semplificazione controllata impiegato da questo modello che sfrutta interventi locali e controllati. SEMPL-IT utenti comuni e ChatGPT 5.1 presentano similarità ancora piuttosto alta (rispettivamente il 96% e il 98%) pur presentando divergenze formali maggiori, con una edit distance rispettivamente di 44% e di 63%. Questi dati suggeriscono che i modelli introducono ristrutturazioni profonde che, comunque, parafrasano fedelmente i contenuti degli originali. Tuttavia, la considerazione congiunta delle metriche di similarità e del corpus indica che un'elevata similarità non esclude a fortiori scostamenti problematici dal senso originario. SEMPL-IT utenti comuni, nonostante presenti il 96% di similarità semantica, introduce un'inversione di ruolo tematico nell'esempio "Il Comando del Corpo Forestale ha il diritto di conoscere…" e una ridenominazione erronea dell'URP che nella sua versione diviene "Unità di Recepimento e di Pubblicazione". Simili fenomeni, evidenti nell'analisi qualitativa (§5.5), sfuggono alle metriche aggregative e possono condizionare le valutazioni meramente quantitative degli LLM. In Claude e Gemini la combinazione di edit distance alta (rispettivamente 65% e 68%) e similarità semantica più bassa (90% e 89%) si lega a un incremento di riscritture profonde che prevedono omissioni di porzioni testuali e aggiunte esplicative, come si mostrerà nel prossimo paragrafo.

I dati quantitativi analizzati in questa sezione descrivono con chiarezza l'interrelazione tra gli indici globali di complessità e le diverse caratteristiche lessicali e sintattiche associate alla leggibilità. Viene, inoltre, documentato che tutti i modelli producono un decremento della complessità rispetto al testo originale. Viene confermata la relazione di interdipendenza tra l'adozione di un paradigma di semplificazione controllata e un mantenimento più accentuato del significato originale. Tuttavia, la mera analisi delle metriche non assicura la fedeltà dei contenuti. La valutazione delle allucinazioni può, pertanto, sfuggire all'analisi quantitativa, perché divergenze semanticamente decisive possono coesistere con profili metrici apparentemente eccellenti.

5.5 Osservazioni qualitative

L'analisi qualitativa si rivela indispensabile per interpretare coerentemente le differenze emerse dalle metriche di leggibilità e, soprattutto, per rilevare fenomeni non riducibili a parametri quantitativi. Nel corpus considerato, un primo ambito di intervento degli LLM riguarda la segmentazione e la conseguente ricomposizione informativa dei contenuti. Ad esempio, il periodo dell'originale riguardante il diritto di accesso presenta una struttura fortemente ipotattica e ad alta densità di incassamento[18]. Nelle riscritture

18 «Il diritto di conoscere atti, documenti o dati… può essere esercitato… o, quando questi

SEMPL-IT esperti il periodo viene sciolto in unità più brevi e autonome, operazione che riduce la complessità di elaborazione in lettura e risulta coerente con l'innalzamento degli indici di leggibilità. La stessa tendenza è accentuata in SEMPL-IT utenti comuni, dove la segmentazione si accompagna a una riorganizzazione macro-strutturale tramite enumerazione procedurale[19]. Analogamente, ChatGPT 5.1 riorganizza i contenuti attraverso un'impaginazione a titoli e l'uso di liste. Claude Sonnet 4.6 e Gemini 3 selezionano una ristrutturazione testuale ancora più radicale, in quanto il primo scompone il periodo in una serie di procedure[20], mentre il secondo distribuisce le informazioni in una guida numerata[21]. La leggibilità in questi casi non deriva solo dalla semplificazione micro-sintattica, ma anche dall'introduzione di indici paratestuali (titoli, numerazione, liste) che incrementano la trasparenza discorsiva senza essere necessariamente catturati dagli indici quantitativi.

Un secondo ambito di semplificazione, strettamente connesso al primo, riguarda l'esplicitazione sintattico-referenziale e il trattamento delle catene anaforiche. Il testo di partenza utilizza pronomi e deittici testuali ("*gli stessi*", "*questi*") e, nell'elenco, mantiene un soggetto sottinteso ("*esercita*…"). Le riscritture degli LLM tendono, invece, a preferire i riferenti espliciti, distribuendo l'informazione su più unità e riducendo le riprese anaforiche pronominali o implicite. In SEMPL-IT esperti questo fenomeno è evidente nella ripetizione del soggetto pieno nelle voci puntate ("*il Corpo Forestale esercita*…"; "*il Corpo Forestale tutela*…"), laddove l'originale, come premesso, ricorre a una serie verbale con soggetto implicito. Anche nelle altre porzioni testuali l'esplicitazione aumenta, come nella seguente frase di Gemini 3: "*Sede: L'ufficio principale si trova a Palermo*" in cui trasforma l'originale ("*L'URP_ Area1 "Affari generali"… ha sede a Palermo*") in una predicazione più esplicita. L'orientamento al destinatario diventa particolarmente marcato in Gemini 3, che impiega la seconda persona per riferirsi direttamente al lettore, usando espressioni direttive ("*Puoi farlo in due modi:*"; "*L'URP è il punto di contatto tra te e il Corpo Forestale*"), con un effetto pragmatico di esplicitare le procedure originali. La semplificazione, in questo caso, incrementa l'usabilità, ma sposta il documento verso un registro divulgativo e, soprattutto, altera la cornice testuale istituzionale. Un simile fenomeno si apprezza anche nella riscrittura "*per aiutarci a migliorare i nostri servizi*", dove compare un riferimento al centro deittico riconducibile alla prima persona plurale assente nel testo di partenza.

Sul piano del lessico e della connessione logico-testuale, le riscritture convergono verso strategie di riduzione dei tecnicismi e del condensamento informativo. L'originale convoglia in una singola frase la descrizione della procedura e le condizioni[22] nei testi

non sono pubblicati, con la richiesta di accesso così come previsto dalle vigenti norme…».

19 «Questi atti possono essere consultati in due modi: 1. Online… 2. Per richiesta.»

20 «Come accedere agli atti pubblici»; «A cosa serve l'URP».

21 «*1. Come consultare documenti e atti*», «*2. L'Ufficio Relazioni con il Pubblico (URP)*», «*3. Di cosa si occupa il Corpo Forestale?*»

22 «…può essere esercitato attraverso… o, quando… non sono pubblicati, con la richiesta…

semplificati tali relazioni vengono spesso riscritte in periodi che distendono il contenuto informativo, impiegando periodi ipotetici lineari in cui le condizioni appaiono nelle protasi[23]. In parallelo, si osserva un abbassamento del registro lessicale, come per l'"*inoltrati*" (originale: "*possono essere inoltrati all'URP*"), che diventa "*inviati direttamente*" (ChatGPT 5.1) o "*possono inviare…*" (SEMPL-IT utenti comuni). Si possono evidenziare anche riformulazioni delle nominalizzazioni originali che vengono riscritte con verbi pieni, come in "*visualizzazione*", che diviene "*consultate*" (ChatGPT 5.1), "*visitando il sito web ufficiale*" (Claude Sonnet 4.6). Queste scelte contribuiscono alla riduzione della densità lessicale e a rendere il testo più accessibile a chi ha scarsa dimestichezza con la prassi testuale della pubblica amministrazione.

Sul versante della fedeltà semantica, l'analisi qualitativa mette in luce criticità non rilevabili attraverso le metriche. Nel corpus emergono infatti tre tipologie di divergenza particolarmente rilevanti in un dominio amministrativo: (i) slittamenti di ruolo tematico, (ii) allucinazioni o aggiunte non ricavabili dagli originali, e (iii) omissioni di elementi giuridicamente salienti. Un caso paradigmatico di slittamenti dei ruoli si osserva nella seguente frase da SEMPL-IT utenti comuni: "*Il Comando del Corpo Forestale ha il diritto di conoscere gli atti, i documenti o i dati che detiene*". La frase rovescia la struttura argomentale del testo di partenza ("*Il diritto di conoscere… può essere esercitato…*"), attribuendo il diritto all'ente anziché ai destinatari (cittadini/utenti). Si tratta di una modifica locale, del tutto coerente sul piano sintattico, ma semanticamente problematica in quanto altera il detentore del diritto e, quindi, il significato dell'intero segmento. Il medesimo SEMPL-IT utenti comuni produce un'ulteriore riscrittura problematica, riconducibile a una vera e propria allucinazione terminologica. L'acronimo URP viene reinterpretato come "*Unità di Recepimento e di Pubblicazione (URP)*", mentre il testo di partenza esplicita chiaramente "*Ufficio Relazioni con il Pubblico - URP*". In un testo istituzionale, una simile ridenominazione non consiste in un errore marginale, poiché incide sulla riconoscibilità dell'ufficio e sulla sua collocazione nell'ecosistema amministrativo.

Tra le aggiunte non supportate, l'esempio più evidente è in Gemini 3, che specifica: "*Molti documenti sono già consultabili online nella sezione "Amministrazione Trasparente"*". La sezione *Amministrazione Trasparente* non compare nel testo di partenza e, pur essendo un'inferenza verosimile nel contesto della trasparenza amministrativa, costituisce informazione aggiunta che richiederebbe verifica. L'utente potrebbe, infatti, essere indotto a cercare una sezione del sito non presente o denominata diversamente. Il ricorso all'allucinazione, ovvero a un arricchimento informativo plausibile ma fuorviante, è particolarmente problematico, in quanto tende a passare inosservato in assenza di un

così come previsto…»

23 Ad esempio, ChatGPT 5.1 riscrive in «Se i documenti non sono pubblicati online, è possibile richiederli…», Claude Sonnet 4.6 in «presentando una richiesta formale… se il documento non è online», e SEMPL-IT utenti comuni opta per «se gli atti non sono pubblicati online, gli utenti possono richiederne l'accesso».

controllo umano. Senza un'attenta revisione, i testi semplificati dagli LLM possono produrre effetti pratici, come disorientamento, o errori procedurali, non trascurabili.

Infine, in merito alle omissioni, Claude Sonnet 4.6 semplifica la cornice normativa fino a neutralizzarla. Come si vede dalla riscrittura "*presentando una richiesta formale di accesso, secondo quanto previsto dalla legge*" i riferimenti puntuali a "*D.lgs. 33/2013*" e "*legge 241/1990*" presenti nel testo di partenza vengono cancellati. In un genere testuale in cui i richiami normativi svolgono una funzione di ancoraggio alla legislazione vigente e, quindi, di legittimazione, una simile omissione riduce la precisione giuridica del documento, pur migliorandone l'immediatezza. Anche in SEMPL-IT esperti, che come mostrato in precedenza impiega un paradigma di semplificazione controllata, affiorano segnali di instabilità semantica, dovuta a una errata rimodulazione dei ruoli tematici. La seguente frase "*Quando non sono pubblicati, l'ente riceve l'accesso*" suggerisce una lettura erronea, in quanto a ricevere l'accesso ai contenuti è l'utente e non l'ente. Anche in questo caso la semplificazione ha introdotto delle deviazioni nella rappresentazione delle procedure descritte nel documento originale.

Nel complesso, i dati indicano che le strategie più efficaci per incrementare la leggibilità dei testi amministrativi (ad esempio, segmentazione, uso di liste e titoli, esplicitazione dei referenti) possono correlarsi a un'alta fedeltà rispetto ai contenuti originali, benché non la garantiscano. I modelli più controllati (nel campione, SEMPL-IT esperti e, spesso, ChatGPT 5.1) tendono a mantenere meglio i contenuti e l'assetto informativo del testo. Gli LLM che massimizzano l'accessibilità attraverso riscritture più radicali, riorganizzando il materiale in una struttura a guida che esplicita le procedure originali (come nel testo di Gemini 3) o sintetizzando il contenuto (Claude Sonnet 4.6) sono maggiormente esposti a omissioni e aggiunte non previste nell'originale. Al contempo, anche quando la semplificazione appare fluida, restano possibili slittamenti localizzati ma semanticamente rilevanti, come nei casi dell'assegnazione del diritto al Comando invece che all'utente o della errata ricostruzione dell'acronimo *URP*. Simili esiti corroborano l'esigenza, metodologicamente centrale nei testi amministrativi, di una revisione umana sistematica, orientata non solo alla correttezza linguistica ma anche alla preservazione dei ruoli, dei riferimenti normativi e dei presupposti pragmatici del testo di partenza.

6. Conclusioni

Il contributo ha discusso, sfruttando metriche quantitative e analisi qualitativa, opportunità e rischi dell'uso di LLM per semplificare l'italiano amministrativo. Ci si è soffermati, dopo una rapida descrizione della risorsa, soprattutto sul confronto tra SEMPL-IT e altri LLM generalisti. La progettazione di SEMPL-IT che ingloba paradigmi di semplificazione diversi permette di calibrare la riscrittura su differenti destinatari, producendo interventi più controllati nella modalità esperti, e una semplificazione più profonda nella modalità utenti comuni. Il confronto fra modelli generalisti (ChatGPT

5.1, Claude Sonnet 4.6, Gemini 3), interrogati con prompt zero-shot, mostra che la leggibilità dei testi amministrativi semplificati con LLM può crescere sensibilmente anche fuori da sistemi specializzati e anche senza l'adozione di strategie di prompting design evolute. Tuttavia, i risultati delle analisi qualitative e quantitative indicano una correlazione tra l'incremento della leggibilità e l'allontanamento dai contenuti originali. Nello specifico, Gemini 3 raggiunge gli indici di leggibilità più alti e, al contempo, realizza una riduzione della similarità semantica; ChatGPT 5.1 e Claude Sonnet 4.6 si collocano su valori intermedi, con buona leggibilità e, nel caso di ChatGPT 5.1, una fedeltà più elevata. Nel complesso, il confronto con gli LLM generalisti mostra che SEMPL-IT, malgrado sia stato sviluppato ricorrendo a risorse enormemente più modeste di quelle impiegate per le controparti commerciali, risulta essere competitivo. SEMPL-IT, infatti, da una parte combina incremento della leggibilità e alta preservazione semantica e, dall'altra, ingloba due diverse strategie di semplificazione nelle due versioni esperti e utenti comuni. Il contributo, infine, ha riaffermato la necessità di affiancare sempre un'accurata revisione umana alla semplificazione automatica dei testi.

Riferimenti bibliografici

Agirre, Eneko, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. "SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation". *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, (2016): 497–511.

Amatriain, Xavier. "Prompt Design and Engineering: Introduction and Advanced Methods". *arXiv preprint* arXiv:2401.14423, (2024).

Bombi, Raffaella. "Anglicismi e burocratese: felice convivenza o relazioni pericolose?". In *Manuale di comunicazione istituzionale e internazionale*, a cura di Raffaella Bombi, (2013): 69–86.

Bombi, Raffaella, a cura di. *Manuale di comunicazione istituzionale e internazionale*. Roma: Il Calamo, 2013.

Bombi, Raffaella. "Possono essere accettabili alcuni anglicismi?". In *Quale comunicazione tra Stato e cittadini oggi? Per un nuovo manuale di comunicazione istituzionale e internazionale*, a cura di Raffaella Bombi, (2015): 55–72.

Brunato, Dominique, e Giulia Venturi. "Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici". In *Diritto, linguaggio e tecnologie dell'informazione*. *Informatica e Diritto* 1, (2014): 111–142.

Cassese, Sabino, a cura di. *Codice di stile delle comunicazioni scritte ad uso delle amministrazioni pubbliche*. Roma: Istituto Poligrafico e Zecca dello Stato, 1994.

Cassese, Sabino. "Prefazione". In *Il dovere costituzionale di farsi capire*, a cura di Maria Emanuela Piemontese, (2023): 7–11.

Cioffi, Alessandro. "Scrittura e semplificazione nel diritto amministrativo. Un'introduzione". In *Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, (2025): 273–286.

Claude Team. *Technical Report of Claude 4.6 Sonnet*. 2026.

Cortelazzo, Michele A. "La semplificazione dei testi amministrativi: le buone pratiche". In *Quale comunicazione tra Stato e cittadini oggi?*, a cura di Raffaella Bombi, (2015): 93–110.

Cortelazzo, Michele A. *Il linguaggio amministrativo*. Roma: Carocci, 2021.

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

De Mauro, Tullio, e Isabella Chiari. "Il nuovo vocabolario di base della lingua italiana". *Internazionale*, (23 dicembre 2016).

Dell'Orletta, Felice, Simonetta Montemagni, e Giulia Venturi. "READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification". *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, (2011): 73–83.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, e Luke Zettlemoyer. "QLoRA: Efficient Finetuning of Quantized LLMs". *arXiv preprint* arXiv:2305.14314, (2023).

Ferrari, Angela, e Giorgia Piantanida. "Il prezzo della semplificazione linguistica usualmente intesa. Riflessioni in prospettiva teorica". In *Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, (2025): 171–184.

Fioravanti, Chiara. "Semplificare i moduli della PA: Approcci e metodi a partire dall'utenza dei cittadini stranieri". *CALUMET. Intercultural Law and Humanities Review* 22, (2025): 55–72.

Fiorentino, Giuliana, Marco Russodivito, Vittorio Ganfi, e Rocco Oliveto. "Validazione e confronto tra semplificazione automatica e semplificazione manuale di testi in italiano istituzionale ai fini dell'efficacia comunicativa". *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, in stampa.

Fiorentino, Giuliana, Marco Russodivito, Vittorio Ganfi, e Rocco Oliveto. "SEMPL-IT: il primo LLM per la semplificazione dell'italiano amministrativo". *Atti del XXV Congresso Internazionale dell'Associazione Italiana di Linguistica Applicata*, in stampa.

Fiorentino, Giuliana, e Vittorio Ganfi. "Parametri per semplificare l'italiano istituzionale: revisione della letteratura". *Italiano LinguaDue* 16(1), (2024): 220–237.

Fiorentino, Giuliana, e Marco Russodivito. "Umano vs AI: alcune considerazioni sulla semplificazione manuale e automatica dal corpus ItaIst". In *Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, (2025): 53–78.

Fiorentino, Giuliana, e Mirko Tavosanis. "Chiaro, sintetico e brillante: l'italiano dei testi redatti per l'AI funziona?". *LId'O. Lingua italiana d'oggi* 21, (2024): 37–65.

Fioritto, Alfredo. *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Roma: Presidenza del Consiglio dei Ministri, 1997.

Fioritto, Alfredo. *Manuale di stile dei documenti amministrativi*. Bologna: Il Mulino, 2009.

Fioritto, Alfredo. "Linguaggio amministrativo e cittadinanza". *Italiano LinguaDue* 10(1), (2018).

Franchina, Vincenzo, e Roberto Vacca. "Adaptation of Flesch Readability Index on a Bilingual Text Written by the Same Author both in Italian and English Languages". *Linguaggi* 3, (1986): 47–49.

Ganfi, Vittorio. "LLM e la semplificazione del linguaggio amministrativo: un confronto tra

SEMPL-IT e GPT 5.1". In stampa. Firenze: Cesati.

Ganfi, Vittorio, e Marco Russodivito. "SEMPL-IT: un modello di intelligenza artificiale per la semplificazione dell'italiano". In *Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, (2025): 79–102.

Giacalone Ramat, Anna. *Verso l'italiano. Percorsi e strategie di acquisizione*. Roma: Carocci, 2003.

Laurent, Antoine. "Google TPUs Explained: Architecture & Performance for Gemini 3". 2025.

Levenshtein, Vladimir I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals". *Soviet Physics Doklady* 10(8), (1966): 707–710.

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

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

Lubello, Sergio, e Cecilia Nobili. "150 anni di scrittura 'per' i cittadini: il destinatario dimenticato". In *Pragmatica storica dell'italiano: modelli e usi comunicativi del passato*, a cura di Giovanni Alfieri, Gaetano Alfonzetti, Daniela Motta e Rosaria Sardo, (2020): 209–216.

Lucisano, Pietro, e Maria Emanuela Piemontese. "Gulpease: una formula per la predizione della leggibilità di testi in lingua italiana". *Scuola e città*, (1988): 110–124.

Miglietta, Annarita. "L'immigrato, l'italiano e il burocratese". *Lingue Linguaggi* 16, (2015): 463–483.

Monaco, Teresa. "Semplificazione e intelligenza artificiale, quando la macchina 'apre' la fiducia". In *Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, (2025): 355–370.

Mori, Laura. "Complessità sintattica e leggibilità. Un monitoraggio linguistico per la valutazione dell'accessibilità dei testi legislativi europei e italiani". *Studi Italiani di Linguistica Teorica e Applicata* XLVIII(3), (2019): 627–657.

Pascucci, Marco, e Mirko Tavosanis. "Confronto tra diversi tipi di valutazione del miglioramento della chiarezza di testi amministrativi in lingua italiana". *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, (2024).

Piemontese, Maria Emanuela. *Capire e farsi capire*. Roma: Tecnodid, 1996.

Piemontese, Maria Emanuela. "Aspetti linguistici della semplificazione e leggibilità di testi di interesse pubblico e istituzionale". In *Manuale di comunicazione istituzionale e internazionale*, a cura di Raffaella Bombi, (2013): 237–266.

Piemontese, Maria Emanuela. *Il dovere costituzionale di farsi capire. A trent'anni dal Codice di stile*. Roma: Carocci, 2023.

Qwen Team. "Qwen2.5 Technical Report". *arXiv preprint* arXiv:2412.15115, (2025).

Romano, Francesca. "Linguaggio giuridico e giovani: l'esperienza dei laboratori di progettazione di contenuti semplificati per i testi online della PA". *Italiano LinguaDue* 14, (2023): 714–723.

Romano, Francesca, e Giulia Venturi. "Metodi e strumenti per la modernizzazione della lingua delle istituzioni". *CALUMET. Intercultural Law and Humanities Review* 22, (2025): 105–121.

Russodivito, Marco, Vittorio Ganfi, Giuliana Fiorentino, e Rocco Oliveto. "AI vs. Human Effectiveness of LLMs in Simplifying Italian Administrative Documents". *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, (2024): 1–12.

Sahoo, Pranav, Ayush Singh, Abhishek Jha, Ashutosh Kumar Jha, Souvik Gupta, e Shashank Mishra. "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications". *arXiv preprint* arXiv:2402.07927, (2024).

Schulhoff, Sander, Guilherme Ilharco, Anaeze Wong, Yu Zhang, Hyojin Kim, Jeremy He, Yifei Jiang, e Marco Tulio Ribeiro. "The Prompt Report: A Systematic Survey of Prompting Techniques". *arXiv preprint* arXiv:2406.06608, (2024).

Simonelli, Maria Ausilia. "La semplificazione della scrittura amministrativa: motivazioni ed efficacia comunicativa". In *Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, (2025): 335–342.

Vedovelli, Massimo. "(Il-)leggibilità dei testi italiani di comunicazione istituzionale e politica linguistica". In *Manuale di comunicazione istituzionale e internazionale*, a cura di Raffaella Bombi, (2013): 305–326.

Vellutino, Daniele, e Maria De Martino. "Dalla semplificazione alla leggibilità. Riflessioni linguistiche e cognitive sulla comprensione dei testi istituzionali". In *Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa*, a cura di Giuliana Fiorentino, Alessandro Cioffi e Maria Ausilia Simonelli, (2025): 227–240.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, e Denny Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". *Advances in Neural Information Processing Systems* 35, (2022).

Xu, Zijian, Samyak Jain, e Mohan S. Kankanhalli. "Hallucination is Inevitable: An Innate Limitation of Large Language Models". *arXiv preprint* arXiv:2401.11817, (2024).

Sitografia

Giuliana Fiorentino, Vittorio Ganfi, Rocco Oliveto, Marco Russodivito, Alessandro Cioffi, Maria Ausilia Simonelli, Ludovico Di Benedetto (2024), ItaIst corpus. https://huggingface.co/datasets/VerbACxSS/ItaIst. Ultima consultazione: 09.01.2026.

Giuliana Fiorentino, Vittorio Ganfi, Rocco Oliveto, Marco Russodivito, Alessandro Cioffi, Maria Ausilia Simonelli, Ludovico Di Benedetto (2025), SEMPL-IT. https://sempl-it.unimol.it/home. Ultima consultazione: 09.01.2026.

Appendice

Estratto di circa 4.000 caratteri dalla Carta dei Servizi dell'Assessorato del Territorio e dell'Ambiente della Regione Sicilia del 28-05-2024 (https://www.regione.sicilia.it/sites/default/files/2024-05/Carta%20dei%20Servizi%20Comando%20Corpo%20Forestale%202024.pdf).

Originale
Diritto di accesso agli atti
Il diritto di conoscere atti, documenti o dati formati o detenuti dal Comando del Corpo Forestale può essere esercitato attraverso la visualizzazione degli stessi nel sito web del

Comando o, quando questi non sono pubblicati, con la richiesta di accesso così come previsto dalle vigenti norme, quali il D.lgs. 33/2013 "Riordino della disciplina riguardante il diritto di accesso civico e gli obblighi di pubblicità e diffusione di informazione da parte delle pubbliche amministrazioni" e la legge 241/1990 "Nuove norme in materia di procedimento amministrativo e di diritto di accesso agli atti amministrativi".

URP: ruoli e funzioni - finalità dell'Ufficio
L'Ufficio Relazioni con il Pubblico - URP – è un luogo aperto ed accessibile organizzato per ricevere gli utenti (cittadini, associazioni, imprese, enti locali, altri Enti) facilitando l'accesso ad informazioni e procedure; se la domanda è di natura molto specialistica, l'utente riceverà indicazioni precise per rivolgersi alla struttura, al dirigente responsabile, al responsabile del procedimento che meglio può accompagnare il cittadino nel percorso di ricerca e approfondimento e assicurare risposte esaurienti.

L'URP - Area 1 "Affari generali" del Comando Corpo Forestale ha sede a Palermo.

La raccolta di suggerimenti e reclami ha lo scopo di favorire il miglioramento dei servizi forniti all'utenza dalle strutture del Comando.

Richieste di informazioni, segnalazioni ed eventuali reclami possono essere inoltrati all'URP del Comando del Corpo Forestale.

Competenze e attività del CFRS
Il Corpo Forestale della Regione Siciliana (CFRS) è stato istituito con legge regionale 5 aprile 1972 n. 24 per svolgere, nell'ambito del territorio regionale, le funzioni e i compiti attribuiti in campo nazionale dall'ex Corpo Forestale dello Stato, oggi Carabinieri Forestali.

La presenza della "Forestale" in Sicilia può farsi risalire al Codice Forestale che Francesco I di Borbone estese al Regno di Sicilia il 26 marzo 1827. Il Corpo Forestale della Regione Siciliana (CFRS) è la struttura operativa di riferimento per la prevenzione e la lotta agli incendi boschivi, prende parte all'organizzazione e allo svolgimento delle attività di protezione civile e, attraverso una capillare attività di sorveglianza, controllo e repressione tutela il patrimonio faunistico e naturalistico della Regione.

In particolare:

• esercita in modo sistematico e continuativo attività di prevenzione e lotta contro gli incendi dei boschi e della vegetazione, a livello regionale con il coordinamento del Servizio Antincendi Boschivi (SAB) e a livello provinciale attraverso l'attività degli Ispettorati Ripartimentali delle Foreste.

• esercita attività finalizzate alla prevenzione, sorveglianza e soccorso in ambienti montani attraverso i Distaccamenti forestali di protezione civile "Madonie" ed "Etna".

• tutela i pascoli e, attraverso gli Ispettorati Ripartimentali delle Foreste (SIRF), esercita azione di tutela e vigilanza su terreni sottoposti a vincoli idrogeologici e su parchi, aree naturali protette e riserve naturali;

• esercita attività di contrasto nei casi di illecito smaltimento dei rifiuti e immissione di agenti inquinanti nell'ambiente naturale, bracconaggio, taglio abusivo di piante o boschi, speculazione edilizia, irregolare sfruttamento delle cave;

• esercita la prevenzione e la repressione delle frodi nel campo agroalimentare, vigila sugli Organismi che operano nell'ambito delle produzioni di qualità, accerta le violazioni in materia di contrasto alle contraffazioni agroalimentari, etichettature e tracciabilità dei prodotti;

• svolge servizi di vigilanza nel corso di manifestazioni pubbliche, politiche e sportive;

• concorre, assieme alle altre forze di polizia, all'espletamento dei servizi di ordine e sicurezza pubblica, in particolar modo in località situate in zone rurali e montane nelle quali operano prevalentemente i Distaccamenti forestali.

Originale RIDOTTO

Diritto di accesso agli atti Il diritto di conoscere atti, documenti o dati formati o detenuti dal Comando del Corpo Forestale può essere esercitato attraverso la visualizzazione degli stessi nel sito web del Comando o, quando questi non sono pubblicati, con la richiesta di accesso così come previsto dalle vigenti norme, quali il D.lgs. 33/2013 "Riordino della disciplina riguardante il diritto di accesso civico e gli obblighi di pubblicità e diffusione di informazione da parte delle pubbliche amministrazioni" e la legge 241/1990 "Nuove norme in materia di procedimento amministrativo e di diritto di accesso agli atti amministrativi". L'Ufficio Relazioni con il Pubblico - URP – è un luogo aperto ed accessibile organizzato per ricevere gli utenti (cittadini, associazioni, imprese, enti locali, altri Enti) facilitando l'accesso ad informazioni e procedure; se la domanda è di natura molto specialistica, l'utente riceverà indicazioni precise per rivolgersi alla struttura, al dirigente responsabile, al responsabile del procedimento che meglio può accompagnare il cittadino nel percorso di ricerca e approfondimento e assicurare risposte esaurienti. L'URP - Area 1 "Affari generali" del Comando Corpo Forestale ha sede a Palermo. La raccolta di suggerimenti e reclami ha lo scopo di favorire il miglioramento dei servizi forniti all'utenza dalle strutture del Comando. Richieste di informazioni, segnalazioni ed eventuali reclami possono essere inoltrati all'URP del Comando del Corpo Forestale. Il Corpo Forestale della Regione Siciliana (CFRS) è la struttura operativa di riferimento per la prevenzione e la lotta agli incendi boschivi, prende parte all'organizzazione e allo svolgimento delle attività di protezione civile e, attraverso una capillare attività di sorveglianza, controllo e repressione tutela il patrimonio faunistico e naturalistico della Regione. In particolare: • esercita in modo sistematico e continuativo attività di prevenzione e lotta contro gli incendi dei boschi e della vegetazione, a livello regionale con il coordinamento del Servizio Antincendi Boschivi (SAB) e a livello provinciale attraverso l'attività degli Ispettorati Ripartimentali delle Foreste. • esercita attività finalizzate alla prevenzione, sorveglianza e soccorso in ambienti montani attraverso i Distaccamenti forestali di protezione civile "Madonie" ed "Etna". • tutela i pascoli e, attraverso gli Ispettorati Ripartimentali

delle Foreste (SIRF), esercita azione di tutela e vigilanza su terreni sottoposti a vincoli idrogeologici e su parchi, aree naturali protette e riserve naturali; • esercita attività di contrasto nei casi di illecito smaltimento dei rifiuti e immissione di agenti inquinanti nell'ambiente naturale, bracconaggio, taglio abusivo di piante o boschi, speculazione edilizia, irregolare sfruttamento delle cave;

Semplificato SEMPL-IT (esperti)

Diritto di accesso agli atti Il diritto di conoscere atti, documenti o dati detenuti dal Comando del Corpo Forestale può essere esercitato visualizzando gli stessi sul sito web del Comando. Quando non sono pubblicati, l'ente riceve l'accesso come previsto dalle norme vigenti. Tra queste, si annoverano il D.Lgs. 33/2013 "Riordino della disciplina riguardante il diritto di accesso civico e gli obblighi di pubblicità e diffusione di informazione da parte delle pubbliche amministrazioni" e la L. 241/1990 "Nuove norme in materia di procedimento amministrativo e di diritto di accesso agli atti amministrativi". L'Ufficio Relazioni con il Pubblico - URP – è un luogo aperto e accessibile per ricevere gli utenti (cittadini, associazioni, imprese, enti locali, altri Enti). Questo servizio facilita l'accesso ad informazioni e procedure. Se la richiesta è molto specialistica, l'utente riceverà indicazioni precise per rivolgersi alla struttura, al dirigente responsabile, al responsabile del procedimento che meglio può accompagnare il cittadino nel percorso di ricerca e approfondimento. L'ente assicura risposte esaurienti. L'URP - Area 1 "Affari generali" del Comando Corpo Forestale si trova a Palermo. La raccolta di suggerimenti e reclami ha lo scopo di migliorare i servizi forniti all'utenza dalle strutture del Comando. L'ente riceve richieste di informazioni, segnalazioni ed eventuali reclami inviandole all'URP del Comando del Corpo Forestale. Il Corpo Forestale della Regione Siciliana (CFRS) esercita la struttura operativa di riferimento per prevenire e contrastare gli incendi boschivi. Partecipa all'organizzazione e allo svolgimento delle attività di protezione civile. Con una capillare attività di sorveglianza, controllo e repressione, l'ente tutela il patrimonio faunistico e naturalistico della Regione. In particolare: • il Corpo Forestale esercita in modo sistematico e continuativo attività di prevenzione e lotta contro gli incendi dei boschi e della vegetazione. A livello regionale, opera con il coordinamento del Servizio Antincendi Boschivi (SAB). A livello provinciale, tramite l'attività degli Ispettorati Ripartimentali delle Foreste. • il Corpo Forestale esercita attività finalizzate alla prevenzione, sorveglianza e soccorso in ambienti montani tramite i Distaccamenti forestali di protezione civile "Madonie" ed "Etna". • il Corpo Forestale tutela i pascoli e, con gli Ispettorati Ripartimentali delle Foreste (SIRF), esercita azioni di tutela e vigilanza su terreni sottoposti a vincoli idrogeologici e su parchi, aree naturali protette e riserve naturali. • il Corpo Forestale esercita attività di contrasto nei casi di illecito smaltimento dei rifiuti e immissione di agenti inquinanti nell'ambiente naturale. Bracconaggio, taglio abusivo di piante o boschi, speculazione edilizia, irregolare sfruttamento delle cave sono altresì oggetto di interventi.

Semplificato SEMPL-IT utenti comuni
Il Comando del Corpo Forestale ha il diritto di conoscere gli atti, i documenti o i dati che detiene. Questi atti possono essere consultati in due modi: 1. Online: il Comando pubblica gli atti sul proprio sito web. 2. Per richiesta: se gli atti non sono pubblicati online, gli utenti possono richiederne l'accesso seguendo le norme vigenti. Le norme principali che regolano questo diritto sono: - Il Decreto Legislativo 33/2013, che disciplina il diritto di accesso civico e gli obblighi di pubblicità delle amministrazioni pubbliche. - La Legge 241/1990, che regola il diritto di accesso agli atti amministrativi. L'Unità di Recepimento e di Pubblicazione (URP) del Comando del Corpo Forestale riceve gli utenti. Questi utenti possono essere cittadini, associazioni, imprese, enti locali o altri enti. L'URP aiuta gli utenti a trovare informazioni e a seguire le procedure. Se la richiesta è molto complessa, l'URP fornisce indicazioni per rivolgersi alla struttura più adatta, al dirigente responsabile o al responsabile del procedimento. Questo garantisce risposte complete e precise. L'URP, che si trova a Palermo, è l'Area 1 "Affari generali" del Comando del Corpo Forestale. La raccolta di suggerimenti e reclami ha lo scopo di migliorare i servizi offerti agli utenti. Gli utenti possono inviare richieste di informazioni, segnalazioni o reclami all'URP del Comando del Corpo Forestale.

Il Corpo Forestale della Regione Siciliana (CFRS) è la struttura principale che si occupa di prevenire e contrastare gli incendi boschivi. Questo corpo opera anche per la protezione civile, organizzando e svolgendo attività di soccorso e prevenzione in ambienti montani.

- Il CFRS opera in modo continuo e sistematico per prevenire e contrastare gli incendi nei boschi e nella vegetazione. A livello regionale, il CFRS collabora con il Servizio Antincendi Boschivi (SAB). A livello provinciale, il CFRS opera con gli Ispettorati Ripartimentali delle Foreste.

- Il CFRS esegue attività di prevenzione, sorveglianza e soccorso in ambienti montani. Queste attività sono svolte con i Distaccamenti forestali di protezione civile "Madonie" e "Etna".

- Il CFRS si occupa della tutela dei pascoli e, insieme agli Ispettorati Ripartimentali delle Foreste (SIRF), esercita azione di tutela e vigilanza su terreni sottoposti a vincoli idrogeologici. Inoltre, il CFRS opera in parchi, aree naturali protette e riserve naturali.

- Il CFRS si occupa anche di contrastare illeciti come lo smaltimento di rifiuti e l'immissione di agenti inquinanti nell'ambiente. Inoltre, il CFRS contrasta il bracconaggio, il taglio abusivo di piante o boschi, la speculazione edilizia e l'irregolare sfruttamento delle cave.

Semplificato ChatGPT 5.1 Prompt "Semplifica questo testo in modo da renderlo più comprensibile al pubblico generale"
Diritto di accesso agli atti
I cittadini hanno il diritto di conoscere atti, documenti e dati prodotti o conservati dal Comando del Corpo Forestale. Queste informazioni possono essere consultate direttamente sul sito web del Comando. Se i documenti non sono pubblicati online,

è possibile richiederli presentando una domanda di accesso, secondo quanto previsto dalla legge (in particolare il D.lgs. 33/2013 e la legge 241/1990).

URP – Ufficio Relazioni con il Pubblico: ruolo e funzioni

L'Ufficio Relazioni con il Pubblico (URP) è un ufficio aperto a tutti che aiuta cittadini, associazioni, imprese ed enti a ottenere informazioni e a orientarsi tra i servizi e le procedure del Comando.

Quando una richiesta è molto specifica o tecnica, l'URP indirizza l'utente all'ufficio o al responsabile più competente, in modo da garantire risposte chiare e complete.

L'URP – Area 1 "Affari Generali" del Comando del Corpo Forestale ha sede a Palermo.

L'ufficio raccoglie anche suggerimenti e reclami, con l'obiettivo di migliorare i servizi offerti.

Richieste di informazioni, segnalazioni e reclami possono essere inviati direttamente all'URP del Comando del Corpo Forestale.

Competenze e attività del Corpo Forestale della Regione Siciliana (CFRS).

Il Corpo Forestale della Regione Siciliana (CFRS) è stato istituito nel 1972 e svolge, nel territorio regionale, le funzioni che a livello nazionale sono oggi affidate ai Carabinieri Forestali.

La presenza della Forestale in Sicilia ha origini storiche antiche, risalenti al 1827.

Il CFRS è il principale punto di riferimento per:

la prevenzione e la lotta agli incendi boschivi;

la protezione civile;

la tutela dell'ambiente, della fauna e del patrimonio naturale della Regione.

In particolare, il CFRS:

previene e combatte gli incendi nei boschi e nella vegetazione, a livello regionale e provinciale;

svolge attività di prevenzione, sorveglianza e soccorso in zone montane, come l'Etna e le Madonie;

tutela i pascoli, i terreni soggetti a vincoli idrogeologici, i parchi, le riserve naturali e le aree protette;

contrasta reati ambientali come l'abbandono illegale dei rifiuti, l'inquinamento, il bracconaggio, il taglio abusivo di alberi, l'abusivismo edilizio e lo sfruttamento irregolare delle cave;

previene e reprime le frodi nel settore agroalimentare, controllando qualità, etichettatura e tracciabilità dei prodotti;

svolge servizi di vigilanza durante manifestazioni pubbliche, politiche e sportive;

collabora con le altre forze di polizia per garantire l'ordine e la sicurezza pubblica, soprattutto nelle aree rurali e montane.

Semplificato Claude Sonnet 4.6 – Prompt "Semplifica questo testo in modo da renderlo più comprensibile al pubblico generale"
Come accedere agli atti pubblici

Chiunque può consultare documenti, dati e atti del Corpo Forestale in due modi:

- visitando il sito web ufficiale, dove molti documenti sono già disponibili;
- presentando una richiesta formale di accesso, secondo quanto previsto dalla legge, se il documento non è online.

A cosa serve l'URP (Ufficio Relazioni con il Pubblico)
L'URP è lo sportello di riferimento per cittadini, associazioni, imprese ed enti che hanno bisogno di informazioni o assistenza. Se la richiesta è complessa o molto tecnica, il personale dell'URP indica a chi rivolgersi all'interno del Comando per ottenere una risposta completa.

La sede dell'URP si trova a Palermo.

Attraverso l'URP è anche possibile inviare suggerimenti, segnalazioni o reclami per contribuire a migliorare i servizi offerti.

Cosa fa il Corpo Forestale della Regione Siciliana (CFRS)
Il CFRS è l'ente regionale che si occupa di proteggere boschi, natura e territorio. In particolare:

- Combatte gli incendi boschivi, coordinando le attività di prevenzione e spegnimento su tutto il territorio regionale e provinciale.
- Interviene in montagna per attività di sorveglianza e soccorso, grazie ai distaccamenti specializzati delle Madonie e dell'Etna.
- Protegge pascoli, parchi e aree naturali, vigilando anche sui terreni soggetti a vincoli idrogeologici e sulle riserve naturali.
- Contrasta i reati ambientali, come lo smaltimento illegale di rifiuti, l'inquinamento, il bracconaggio, il taglio abusivo di alberi e le costruzioni o cave irregolari.

Semplificato Gemini 3 – Prompt "Semplifica questo testo in modo da renderlo più comprensibile al pubblico generale"
Guida ai Servizi e alle Funzioni del Corpo Forestale (Regione Siciliana)
Il Corpo Forestale della Regione Siciliana (CFRS) si occupa di proteggere la natura, prevenire gli incendi e garantire che le leggi ambientali siano rispettate.

1. Come consultare documenti e atti (Diritto di Accesso)

Ogni cittadino ha il diritto di conoscere i documenti e i dati in possesso del Comando. Puoi farlo in due modi:

- Sito Web: Molti documenti sono già consultabili online nella sezione «Amministrazione trasparente».
- Richiesta diretta: Se i documenti non sono online, puoi richiederli ufficialmente (secondo le leggi 241/90 e il D.lgs. 33/2013).

2. L'Ufficio Relazioni con il Pubblico (URP)

L'URP è il punto di contatto tra te e il Corpo Forestale. Serve a facilitare l'accesso alle informazioni e alle procedure.

Sede: L'ufficio principale si trova a Palermo.

Cosa fa: Ti aiuta a trovare le informazioni che cerchi. Se la tua domanda è molto tecnica, l'URP ti indicherà l'ufficio o il dirigente specifico a cui rivolgerti.

Suggerimenti e Reclami: Puoi inviare segnalazioni o reclami all'URP per aiutarci a migliorare i nostri servizi.

3. Di cosa si occupa il Corpo Forestale?

Il Corpo Forestale interviene su più fronti per difendere il territorio siciliano:

Lotta agli incendi: È la struttura principale che organizza e coordina lo spegnimento degli incendi nei boschi e nelle campagne.

Soccorso e Montagna: Grazie ai distaccamenti "Madonie" ed "Etna", svolge attività di sorveglianza e soccorso in ambienti montani.

Tutela dell'ambiente: Controlla parchi, riserve naturali e aree protette, verificando il rispetto dei vincoli del terreno.

Contrasto agli illeciti: Combatte i crimini ambientali come:

Bracconaggio e tagli abusivi di alberi.

Smaltimento illegale di rifiuti e inquinamento.

Abusi edilizi e gestione irregolare delle cave.

Un regolamento TARI, un contratto di lavoro, un avviso sanitario: testi che incidono sui diritti e sui doveri quotidiani non possono restare un labirinto di formule, tecnicismi e periodi interminabili. La chiarezza è una condizione della cittadinanza democratica.

Gli atti del Convegno "Linguaggio, complessità e comprensione: strumenti e strategie per la semplificazione" riuniscono linguisti, giuristi e informatici per misurare e affrontare l'opacità della scrittura amministrativa. Dai *corpora* di regolamenti comunali alle comunicazioni tributarie, dai CCNL ai documenti della sanità pubblica, i contributi mostrano dove nasce la difficoltà: densità lessicale, nominalizzazioni, sintassi complessa, formule cristallizzate e informazioni mal gerarchizzate.

Ampio spazio è dedicato al Natural Language Processing, agli indici di leggibilità come Gulpease e READ-IT, alla profilazione linguistica e alle riscritture con Large Language Models. Ma l'automazione non basta: nei testi con valore prescrittivo, la fedeltà semantica e la revisione umana restano indispensabili, poiché semplificare non significa impoverire il diritto, ma organizzare la complessità in modo che ogni cittadino possa capire, decidere e partecipare.

www.ledizioni.it

PDF IN OPEN ACCESS