Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici

Dominique Brunato e Giulia Venturi
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) – Pisa
Italian Natural Language Processing Lab (ItaliaNLP Lab) – www.italianlp.it
dominique.brunato@ilc.cnr.it , giulia.venturi@ilc.cnr.it

1. Introduzione

La scarsa accessibilità della *lingua del diritto* nelle sue diverse declinazioni è stata spesso oggetto di critiche mosse da prospettive diverse (quella dei giuristi, dei linguisti, dei comunicatori pubblici, quella soprattutto dei cittadini). Se alcuni risultati concreti verso la semplificazione della vasta tipologia di documenti che possono essere considerati *testi giuridici*[1] sono stati raggiunti in alcuni paesi europei già da molti anni[2], in Italia il dibattito in materia di miglioramento dell'accessibilità di testi giuridici sembra non avere via di uscita. Da un lato, infatti, è stato fatto notare che i difetti della legislazione italiana dipendono «dalla circostanza che quest'ultima a conti fatti è una succursale del linguaggio burocratico, sia perché i disegni di legge vengono concepiti non di rado negli uffici legislativi dei ministeri, sia perché la legge stessa … si è ormai amministrativizzata, nel senso che regola questioni minute e di dettaglio, un tempo ascritte al dominio pressoché esclusivo dell'atto amministrativo»[3]. Dall'altro, si è sottolineato che, nel ricalcare il linguaggio della legge, a cui fortemente si ispira, «il linguaggio burocratico esibisce una qualità quasi sempre assai inferiore a quello legislativo, di cui rappresenta una sorta di "parente povero", se non addirittura di "caricatura"»[4].

Di fatto, se, per effetto di cattive scelte stilistiche, le norme risultano oscure e difficilmente accessibili ai non specialisti, ancora più forti sono i disagi che arrecano alla scrittura burocratico-amministrativa. Dal momento che l'attività delle amministrazioni pubbliche investe direttamente la vita di tutti i membri di una comunità nazionale, i testi amministrativi, pur dovendo mantenere i requisiti di legittimità e formalità, hanno anche una forte componente comunicativa che non può essere tralasciata e che deve tradursi in un linguaggio che sia adeguato alle competenze del cittadino

[1] Cfr. B. MORTARA GARAVELLI, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, Torino, Einaudi, 2001, 19-34 pp.; A. FIORITTO, *Manuale di stile dei documenti amministrativi*, Bologna, Il Mulino, 2009, 29-42 pp.
[2] Un esempio su tutti è quello della Svezia, che già dai primi anni Settanta ha istituito la figura del "tutore della lingua", un esperto di questioni linguistiche e aspetti della semplificazione che collabora alle riunioni del Consiglio dei Ministri, insieme ai tecnici e agli esperti della materia da legiferare, con il compito di vigilare sulla comprensibilità dei testi e delle proposte di legge finali.
[3] M. AINIS, *La legge oscura. Come e perché non funziona*, Roma-Bari, Laterza, 1997, 189-190 pp.
[4] D. FORTIS, *Il linguaggio amministrativo italiano*, in "Revista de Liengua i dret", n. 43, pp. 47-116, 2005, 55 p.

medio. Quando invece la testualità burocratica si uniforma ciecamente a quella legislativa, perde di vista la sua funzione e il suo destinatario.

È questo il motivo per cui negli ultimi vent'anni si sono susseguiti a livello nazionale e internazionale numerosi progetti dedicati alla semplificazione linguistica della *lingua del diritto* nelle sue molteplici varietà. Nell'ambito di queste iniziative, un ruolo di primo piano è stato riconosciuto alle tecnologie informatiche come strumenti di verifica della qualità e dell'accessibilità dei testi giuridici. In particolare, un'attenzione specifica è stata rivolta al miglioramento in termini di efficacia comunicativa della prosa burocratica: «E' questo infatti l'uso specialistico in campo legale che ha più dirette conseguenze sull'uso comune, dal momento che la maggior parte delle persone viene a contatto con la lingua della legge attraverso le sue incarnazioni amministrative e burocratiche»[5].

Nell'era infatti della digitalizzazione e degli Open Data, migliorare l'accesso all'informazione contenuta in grandi quantità soprattutto di testi scritti sta diventato una questione fondamentale, come ricordato anche dalle *Web Content Accessibility Guidelines*[6] proposte dalla *Web Accessibility Initiative* e dalla *Europe 2020: Europe's growth strategy*[7], con la quale l'Unione europea ha posto tra i suoi obiettivi più ambiziosi quello di provvedere ad una «smart, sustainable and inclusive growth» grazie alla quale l'informazione sia facilmente accessibile per un ampio ed eterogeneo gruppo di cittadini, compresi individui con difficoltà di lettura dovute ad un basso livello di scolarizzazione, al fatto che i testi in questione non sono scritti nella loro lingua madre o ancora ad alcune specifiche disabilità linguistiche. Lo scopo è quello di mettere a punto metodi per ridurre le inutili complessità create dalle burocrazie di tutta Europa, complessità che non fanno altro che rallentare lo sviluppo nazionale e internazionale. Al contrario, attività di semplificazione burocratico-legislativa consentirebbero agli Stati, e ai privati, di risparmiare tempo, impegno e risorse finanziarie.

Il presente contributo si colloca in questo orizzonte di attività. Esso prende le mosse dai risultati ottenuti nell'ambito del filone di ricerche avviato negli ultimi anni e attivo a livello internazionale nel quale analisi linguistiche generate da strumenti di Trattamento Automatico del Linguaggio sono oggi usate per misurare il livello di leggibilità di un testo come passo preliminare alla sua semplificazione. A nostra conoscenza, tale studio rappresenta il primo tentativo volto a mostrare come tecnologie linguistico-computazionali allo stato dell'arte per la lingua italiana incomincino ad essere oggi mature per essere usate non solo come ausilio per definire la leggibilità di testi giuridici ma anche come guida per una stesura semplificata di tali testi. A questo scopo sarà illustrato READ-IT[8], il primo e al momento unico strumento di valutazione della leggibilità oggi esistente per la lingua italiana basato su strumenti di Trattamento Automatico del Linguaggio. Saranno dunque presentati alcuni esperimenti condotti utilizzando READ-IT nell'analisi di un testo legislativo 'speciale' come la *Costituzione italiana* (cf. § 3.3) e nell'analisi di documenti amministrativi (cf. § 4). In quest'ultimo esperimento, l'obiettivo sarà quello di dimostrare come l'approccio qui descritto

[5] B. MORTARA GARAVELLI, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, *cit.*, 7-8 pp.
[6] http://www.w3.org/TR/WCAG20/
[7] http://ec.europa.eu/europe2020/pdf/europe_2020_explained.pdf
[8] F. DELL'ORLETTA, S. MONTEMAGNI, G. 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)", Edimburgo, UK, 30 luglio 2011, 2011, pp. 73-83.

sia un valido ausilio alla semplificazione guidata di un testo giuridico. È stato scelto come esempio il caso della prosa burocratico-amministrativa perché, come già messo in luce, l'accessibilità a tali documenti rappresenta un elemento chiave della comunicazione istituzioni-cittadini. La comprensibilità di questi testi è condizione imprescindibile per una loro fruizione corretta e va ben al di là dell'essere un semplice atto di cortesia che l'amministrazione rivolge ai propri utenti, per configurarsi, piuttosto, come estensione del «diritto all'informazione» tutelato dall'articolo 21 della Costituzione[9], nonché come strumento per dare attuazione al più generale principio di trasparenza dell'attività amministrativa sancito dalla legge n. 241 del 7 agosto 1990 (*Nuove Norme in materia di procedimento amministrativo e di diritto di accesso ai documenti pubblici*). Sono queste considerazioni che, a partire dai primi anni Novanta, hanno animato quel «moto di riforma del linguaggio amministrativo, che si prefigge di renderlo più chiaro e accessibile ai cittadini»[10].

2. La misura della leggibilità di testi giuridici

È infatti agli inizi degli anni '90 che inizia a nascere in Italia la consapevolezza per una maggiore attenzione alla redazione di testi giuridici scritti in una lingua chiara e comprensibile. Sulla scia di quanto stava avvenendo sul piano internazionale, anche in Italia «lo sviluppo del dibattito sulla tecnica legislativa ha messo in luce come anche il legislatore, nei momenti in cui crea la norma, debba tener conto del modo in cui viene espressa e ricevuta»[11]. È questo il contesto in cui sono avviate in quegli anni le prime iniziative istituzionali finalizzate alla stesura di manuali e codici di regole e suggerimenti per la redazione di atti normativi e amministrativi[12] e prendono anche avvio le prime attività finalizzate a mettere a punto metodi e tecniche informatici in grado di offrire «una serie di strumenti che vanno dai semplici editori di testi con correttori ortografici, ai controlli di leggibilità [...] e alle tecniche di disambiguazione appoggiate su approcci di intelligenza artificiale»[13].

---

[9] M.E. PIEMONTESE, *Il linguaggio della pubblica amministrazione nell'Italia d'oggi. Aspetti problematici della semplificazione linguistica*, in G. Alfieri, A. Cassola (a cura di), "La *Lingua d'Italia*. Usi pubblici e istituzionali", Atti del XXlX Congresso Internazionale di Studi della SLI (Malta, 3-5 novembre 1998), Roma, Bulzoni, 1999, pp. 269-292, 269 p.

[10] Come osservato da D. FORTIS, *Il linguaggio amministrativo italiano*, *cit.*, 89 p., nonostante tale legge «non tratti esplicitamente del linguaggio, l'esigenza di una scrittura amministrativa più chiara costituisce un suo corollario ed è implicita nel suo spirito: consentire ai cittadini di accedere a documenti che comunque non riuscirebbe a comprendere sarebbe infatti un controsenso, che vanificherebbe, di fatto, tale diritto».

[11] La citazione è tratta dal contributo di Tullio De Mauro a E. ZUANELLI (a cura di), *Il diritto all'informazione in Italia, Roma*, Presidenza del Consiglio dei Ministri. Dipartimento per l'informazione e l'editoria, 1990, 219, p.

[12] Vedi M.E. PIEMONTESE, *Il linguaggio della pubblica amministrazione nell'Italia d'oggi. Aspetti problematici della semplificazione linguistica*, *cit.*, 270-271 pp. per una rassegna delle tappe più significative fino alla fine degli anni '90 «segnate dall'apparato statale nella direzione della dichiarazione e affermazione del principio di semplificare i testi normativi, amministrativi ecc.». Per una rassegna bibliografica aggiornata dei manuali sino ad oggi redatti a livello nazionale e regionale vedi http://www.maldura.unipd.it/buro/; P. MERCATALI, F. ROMANO, *I documenti dello stato digitale. Regole e tecniche per la semplificazione*, Edizioni Studio AD.ES, collana d'informatica giuridica, vol. 2, 2013. Ad oggi, il riferimento più attuale è la *Guida alla redazione degli atti amministrativi. Regole e suggerimenti*, redatta dai ricercatori dell'Istituto di Teorie e Tecniche dell'informazione giuridica (ITTIG), in collaborazione con l'Accademia della Crusca. La *Guida* è navigabile e scaricabile alla pagina http://www.pacto.it/content/view/416/48/

[13] G. TADDEI ELMI, *Dalla Legistica alla Legimatica*. In: C. Biagioli, P. Mercatali, G. Sartor (a cura di), Legimatica. Informatica per legiferare, Napoli, ESI, 1995, pp. 267-273, 271 p. Per una rassegna dei primi sistemi si rimanda a C. BIAGIOLI, *Legimatica: verso una seconda generazione*, in C. Biagioli, P. Mercatali, G. Sartor (a cura di), "Legimatica. Informatica per legiferare", Napoli, ESI, 1995, pp. 75-91; P. MERCATALI, *Dodici anni di legimatica. Da una parola a una disciplina*, in "Iter Legis", vol. 6, 2004, pp. 97-114.

Ed è allora che iniziano a diffondersi in Italia le primissime attività volte all'applicazione di metodi quantitativi per la misurazione della leggibilità di testi giuridici. Tra i casi più significativi è d'interesse qui ricordare gli esperimenti condotti congiuntamente dai ricercatori dell'allora Istituto di Documentazione Giuridica del CNR di Firenze e dai ricercatori dell'Istituto di Linguistica Computazionale del CNR di Pisa, finalizzati allo sviluppo di «un software completo ed articolato, che permetta il controllo della correttezza, leggibilità e coerenza linguistica di un testo giuridico»[14]. L'obiettivo era quello di «stabilire dei paradigmi di comportamento linguistico», quali la distribuzione del lessico nel testo, in grado di «offrire una misurazione globale della complessità sintattico-semantica di un testo giuridico»[15], superando in questo modo i limiti delle formule di leggibilità del testo (come la formula Flesch) sino a quel momento ampiamente utilizzate soprattutto nel contesto nord-americano del movimento *Plain Language*.[16]

Per la lingua italiana, numerosi studi sul calcolo della leggibilità di testi giuridici sono stati condotti utilizzando l'unico indice allora esistente specificatamente messo a punto per l'italiano, l'indice Gulpease[17]. Come ricorda Bice Mortara Garavelli[18], prima che allo studio dei testi normativi, l'attenzione dei linguisti si è concentrata sul linguaggio dell'amministrazione e della burocrazia. Ne sono una testimonianza i contributi di Emanuela Piemontese concentrati soprattutto (ma non solo) sull'analisi dell'universo composito dei documenti della pubblica amministrazione[19].

La metodologia per il calcolo della leggibilità di un testo basata sull'uso di strumenti di Trattamento Automatico del Linguaggio allo stato dell'arte descritta in questo contributo si inserisce appunto in questo filone di ricerche. Si intende qui mettere in luce come gli strumenti di analisi automatica del testo allora a disposizione non permettessero di individuare in modo automatico l'ampia gamma di

[14] C. BIAGIOLI, G. BIANUCCI, P. MERCATALI, D. TISCORNIA, *Introduzione. L'analisi automatica dei testi giuridici e politici*, in P. Mercatali (a cura di), "Computer e linguaggi settoriali. Analisi automatica di testi giuridici e politici", Milano, Franco Angeli, 1988, pp. 15-27, 24 p.

[15] C. BIAGIOLI, P. MERCATALI, D. TISCORNIA, *Le formule per l'analisi automatica della leggibilità: la formula di Flesch per il controllo di documenti giuridici*, in P. Mercatali (a cura di), "Computer e linguaggi settoriali. Analisi automatica di testi giuridici e politici", Milano, Franco Angeli, 1988, pp 45-99, 49 p.

[16] Nonostante il grande successo delle formule di leggibilità, anche negli USA c'è chi negli stessi anni '80 discute in modo critico la possibilità di poter definire il livello di leggibilità e comprensibilità di un testo giuridico facendo affidamento unicamente su caratteri generali e formali del testo. È il caso, ad esempio, di V.R. CHARROW, J.A. CRANDALL, R.P. CHARROW, *Characteristics and Functions of Legal Language*, in R. Kittredge, J. Lehrberger (a cura di), "Sublanguage: Studies of Language in Restricted Semantic Domains", deGruyter, Berlin, 1982, pp. 177-190, in cui viene riportata la situazione assurda di «simplifying tax forms to an 8th-grade level, as measured by a readability formula, and then finding, as one would expect, that 8th graders cannot fill one out, or even understand it». L'obiettivo di Charrow e colleghi era infatti quello di denunciare il fatto che tale indicatore fosse fondato «in misapprehension that the number of syllables per word and the number of words per sentence are accurate indicators of the comprehensibility of a document».

[17] P. LUCISANO, M.E. PIEMONTESE, *Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana*, in "Scuola e Città", vol. 3, 1988, pp. 57-68.

[18] B. MORTARA GARAVELLI, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, *cit.*

[19] M.E. PIEMONTESE, *Capire e farsi capire. Teorie e tecniche della scrittura controllata*, Napoli, Tecnodid, 1996, 123-193 pp.; M.E. PIEMONTESE, *Il linguaggio della pubblica amministrazione nell'Italia d'oggi. Aspetti problematici della semplificazione linguistica*, *cit.*; M.E. PIEMONTESE, *Leggibilità e comprensibilità delle leggi italiane. Alcune osservazioni quantitative e qualitative*, in D. Veronesi (a cura di), "Linguistica giuridica italiana e tedesca: obiettivi, approcci, risultati", atti del Convegno di studi (Bolzano, 1-3 ottobre 1998), Unipress, Padova, 2000, pp. 103-117; M.E. PIEMONTESE, *Leggibilità e comprensibilità dei testi delle pubbliche amministrazioni: problemi risolti e problemi da risolvere*, in S. Covino (a cura di), "La scrittura professionale. Ricerca, prassi, insegnamento", atti del l Convegno di studi (Perugia, Università per stranieri, 23-25 ottobre 2000), 2001, Firenze, Olschki, pp. 119-130; M.E. PIEMONTESE, M.T. TIRABOSCHI, *Leggibilità e comprensibilità dei testi della pubblica amministrazione. Strumenti e metodologie di ricerca al servizio del diritto a capire testi di rilievo pubblico*, in. E. Zuanelli (a cura di), "Il diritto all'informazione in Italia", Roma, Presidenza del Consiglio dei Ministri. Dipartimento per l'informazione e l'editoria, 1990, pp. 225-246.

caratteristiche linguistiche correlate al diverso livello di leggibilità di un testo (giuridico o meno). Al contrario oggi, come discusso nel paragrafi che seguono, gli strumenti di annotazione linguistica automatica del testo possono essere considerati un punto di partenza affidabile per ricavare utili indicatori del grado di leggibilità di un testo (anche giuridico) a partire dalle principali caratteristiche linguistiche in esso rintracciate. L'obiettivo è quello di suggerire come un tale approccio sia il primo passo per arrivare a definire un indice di qualità (linguistica) redazionale di testi giuridici.

L'approccio si inserisce inoltre nel recente filone di ricerche attivo a livello internazionale e rivolto all'uso di strumenti di Trattamento Automatico del Linguaggio, o in generale di metodi per catturare aspetti del profilo linguistico, con l'obiettivo comune di misurare il livello di accessibilità di un testo giuridico. È il caso, ad esempio, della *Readability Research Platform*[20] sviluppata presso la Research School of Computer Science dell'Australian National University. La piattaforma, basata su strumenti di annotazione linguistica del testo, è in grado di tracciare un profilo di alcune caratteristiche di un testo legislativo, caratteristiche non solo di base (es. lunghezza delle frasi) ma anche relative a livelli più avanzati di descrizione linguistica (es. la distribuzione di costituenti morfo-sintattici). Il profilo linguistico così tratteggiato si configura come il punto di partenza per calcolare il livello di leggibilità di un testo legislativo. Facendo affidamento su caratteristiche più di base (come ad esempio il livello di entropia concettuale, calcolata come il numero di parole 'diverse' che ricorrono in un testo), Katz & Bommarito[21] hanno recentemente proposto un *framework* empirico per arrivare a misurare il livello di complessità di un testo giuridico definito come l'unione della struttura (la struttura delle parti del testo, come capitoli, articoli ecc…), della lingua e del grado di interdipendenza (dato dal numero di citazioni intratestuali) di un testo.

3. READ-IT: uno strumento automatico per l'analisi della leggibilità di un testo

Gli ultimi anni hanno visto il progressivo affermarsi a livello internazionale del ricorso a tecnologie linguistico-computazionali per la misurazione automatica della leggibilità di un testo. A differenza dei metodi sino ad oggi adottati, come ad esempio la formula Flesch-Kincaid[22], utilizzata per la lingua inglese, o l'indice Gulpease per la lingua italiana, questa seconda generazione di misuratori di leggibilità non fa affidamento unicamente su caratteristiche generali e formali del testo, quali la lunghezza della frase e la lunghezza delle parole. L'utilizzo di strumenti di annotazione linguistica automatica permette infatti di definire la leggibilità di un testo sulla base di parametri linguistici più complessi e che fino ad ora sembravano essere inattingibili se non attraverso un accurato lavoro manuale. Tali parametri spaziano tra i vari livelli di analisi linguistica e sono rintracciati in modo automatico a partire dall'output del processo di annotazione automatica del testo.

[20] M. CURTOTTI, E. MCCREATH, *A Right to Access Implies A Right to Know: An Open Online Platform for Research on the Readability of Law*, in "Journal of Open Access to Law", vol. 1(1), 2013, pp. 1-56; la piattaforma è consultabile al sito http://buttle.anu.edu.au/readability/readability.wsgi

[21] D.M. KATZ, M.J. BOMMARITO, *Measuring the Complexity of the Law: The United States Code*, (August 1, 2013), 2014, disponibile all'indirizzo http://ssrn.com/abstract=2307352

[22] J. P. KINCAID, R. LIEUTENANT, R.P. FISHBURNE, R. L. ROGERS, B.S. CHISSOM, *Derivation of new readability formulas for Navy enlisted personnel*, Research Branch Report, Millington, TN: Chief of Naval Training, 1975, pp. 8-75.

Per quanto riguarda la lingua italiana, il primo e al momento unico strumento sviluppato che si basa su questi presupposti è rappresentato da READ-IT[23] sviluppato dall'*Italian Natural Language Processing Laboratory* (ItaliaNLP Lab)[24] dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa[25] e concepito per fornire anche un supporto alla redazione semplificata di un testo attraverso l'identificazione dei suoi luoghi di complessità. READ-IT implementa un indice di leggibilità "avanzato" basato su analisi linguistica multi-livello del testo condotta con strumenti che rappresentano lo stato dell'arte per il trattamento automatico della lingua italiana. READ-IT, sulla base dei risultati del monitoraggio di una serie di caratteristiche linguistiche rintracciate in un corpus a partire dall'output di strumenti di annotazione linguistica automatica, permette di calcolare la leggibilità dei testi di cui il corpus è composto classificandoli come testi di *facile* o *difficile* lettura. La classificazione è realizzata da un classificatore statistico che associa i testi in input (linguisticamente annotati) a due classi di lettura definite a priori. Si tratta di classi formate da testi tratti dal corpus *Due Parole*[26], un giornale scritto con una lingua giornalistica volutamente semplificata per essere compresa da persone con un basso livello di scolarizzazione o con disabilità cognitive, considerati testi di *facile* lettura, e dal corpus *La Repubblica*, porzione del corpus CLIC-ILC[27], considerati testi di *difficile* lettura. L'appartenenza ad una delle due classi è stabilita sulla base del grado di similarità tra la distribuzione di alcune delle caratteristiche linguistiche monitorate. Ad esempio, testi con valori di ricchezza lessicale, lunghezza delle relazioni di dipendenza, lunghezza di sequenze di complementi preposizionali modificatori di teste nominali, ecc... più vicini ai valori di monitoraggio linguistico di *Due Parole* sono classificati come testi di facile lettura rispetto a testi che mostrano valori più simili a quelli di *La Repubblica*.

Un tratto caratterizzante di READ-IT, innovativo rispetto alla letteratura internazionale in materia, consiste in una valutazione della leggibilità articolata su due livelli: il documento e la singola frase. La valutazione rispetto alla frase rappresenta un'importante novità dell'approccio sottostante a READ-IT: attraverso l'identificazione dei luoghi di complessità del testo (individuati a livello della singola frase) che necessitano di revisione e semplificazione, lo strumento risulta essere un utile ausilio per la semplificazione del testo[28].

Ampiamente sperimentato su diverse tipologie di testi[29], 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 della sua azione formativa; operatore di call center-utente, per fornire un supporto alla redazione dei testi usati nei call centers migliorando i processi di comunicazione con l'utente; medico-paziente, per assistere la

[23] F. DELL'ORLETTA, S. MONTEMAGNI, G. VENTURI, *READ-IT: assessing readability of Italian texts with a view to text simplification, cit.*
[24] www.italianlp.it
[25] Una demo on-line di READ-IT è disponibile alla pagina http://www.italianlp.it/demo/read-it/
[26] http://www.dueparole.it/default_.asp
[27] R. MARINELLI, L. BIAGINI, R. BINDI, S. GOGGI, M. MONACHINI, P. ORSOLINI, E. PICCHI, S. ROSSI, N. CALZOLARI, A. ZAMPOLLI, *The Italian PAROLE corpus: an overview*, in A. Zampolli et al. (a cura di), "Computational Linguistics in Pisa", XVI-XVII(1), Pisa-Roma, IEPI, 2003, pp. 401-421.
[28] F. DELL'ORLETTA, M. WIELING, G. VENTURI, A. CIMINO, S. MONTEMAGNI, *Assessing the Readability of Sentences: Which Corpora and Features?*, in "Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)", 26 June, Baltimore, Maryland, Association for Computational Linguistics, 2014, pp. 163-173.
[29] F. DELL'ORLETTA, S. MONTEMAGNI, G. VENTURI, *Assessing Document and Sentence Readability in Less Resourced Languages and across Textual Genres*, in Thomas François and Delphine Bernhard (a cura di), "International Journal of Applied Linguistics (ITL)", Special Issue on Readability and Text Simplification, 2014, (in corso di stampa).

redazione di consensi informati semplici e leggibili. In questo contributo l'intento è quello di dimostrare come READ-IT possa essere usato con successo per calcolare il livello di leggibilità di testi giuridici e per valutare l'efficacia della comunicazione legislatore e/o amministratore-cittadino, allo scopo di semplificare e migliorare i processi di comunicazione tra istituzioni e cittadini.

In quanto segue, saranno prima descritti gli strumenti di Trattamento Automatico del Linguaggio su cui si basa READ-IT (§ 3.1); sarà poi introdotta la metodologia di monitoraggio linguistico alla base del calcolo della leggibilità (§ 3.2); nel Paragrafo 3.3. sarà infine presentato un esempio dell'output di READ-IT.

3.1. Gli strumenti di Trattamento Automatico del Linguaggio

Gli strumenti di Trattamento Automatico del Linguaggio operando in successione, permettono di rendere progressivamente esplicita l'informazione linguistica contenuta in un testo. Per ogni livello di descrizione linguistica uno specifico componente di analisi identifica in modo automatico la struttura del testo, utilizzando come input il risultato prodotto dal componente precedente. L'identificazione della struttura linguistica del testo, o annotazione, avviene tipicamente in modo incrementale, attraverso analisi linguistiche a livelli di complessità crescente: "tokenizzazione", ovvero segmentazione del testo in parole ortografiche (o *tokens*); analisi morfo-sintattica e lemmatizzazione del testo *tokenizzato*; analisi della struttura sintattica della frase in termini di relazioni di dipendenza.

La Tabella 1 mostra un esempio del risultato del processo incrementale di annotazione linguistica del seguente periodo:

Le disposizioni di cui alla presente lettera si applicano anche nei confronti degli altri organi tenuti all'adozione di strumenti urbanistici.

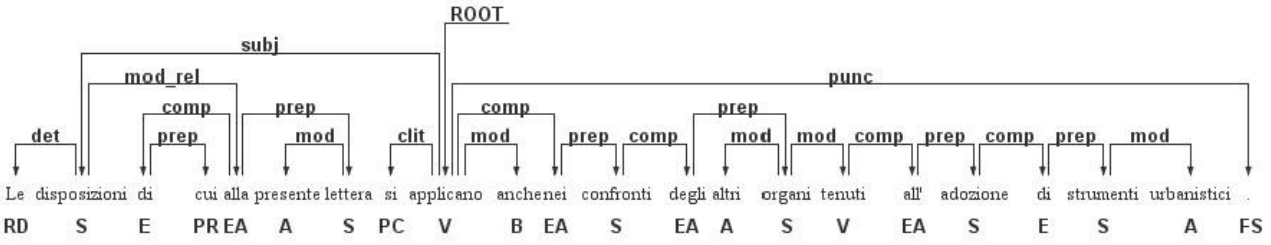
		Lemmatizzazione	Annotazione morfo-sintattica			Annotazione sintattica	
Id	Forma	Lemma	CPoS	FPoS	Tratti morfologici	Testa sintattica	Relazione
1	Le	il	R	RD	num=p|gen=f	2	det
2	disposizioni	disposizione	S	S	num=p|gen=f	9	subj
3	di	di	E	E	-	5	comp
4	cui	cui	P	PR	num=n|gen=n	3	prep
5	alla	a	E	EA	num=s|gen=f	2	mod_rel
6	presente	presente	A	A	num=n|gen=n	7	mod
7	lettera	lettera	S	S	num=s|gen=f	5	prep
8	si	si	P	PC	num=n|per=3|gen=n	9	clit
9	applicano	applicare	V	V	num=p|per=3|mod=i|ten=p	0	ROOT
10	anche	anche	B	B	-	9	mod
11	nei	in	E	EA	num=p|gen=m	9	comp
12	confronti	confronto	S	S	num=p|gen=m	11	mod
13	degli	di	E	EA	num=p|gen=m	12	prep
14	altri	altro	A	A	num=p|gen=m	15	mod
15	organi	organo	S	S	num=p|gen=m	13	prep
16	tenuti	tenere	V	V	num=p|mod=p|gen=m	15	mod
17	all'	a	A	EA	num=s|gen=n	16	comp
18	adozione	adozione	S	S	num=s|gen=f	17	prep
19	di	di	E	A	-	18	comp

		Lemmatizzazione	Annotazione morfo-sintattica			Annotazione sintattica	
Id	Forma	Lemma	CPoS	FPoS	Tratti morfologici	Testa sintattica	Relazione
20	strumenti	strumento	S	S	num=p|gen=m	19	prep
21	urbanistici	urbanistico	A	A	num=p|gen=m	20	mod
22	.	.	F	FS	-	9	punc

Tabella 1: Un esempio di annotazione linguistica.

Innanzitutto, il periodo è stato individuato grazie alla fase di segmentazione in periodi di una direttiva comunitaria in materia ambientale. Durante la successiva fase di tokenizzazione, all'interno del periodo sono stati riconosciuti i tokens corrispondenti alle singole forme (colonna *Forma*), identificate univocamente da un numero progressivo (colonna *Id*). La fase di disambiguazione morfo-sintattica ha permesso di associare ad ogni token individuato *i)* la corretta categoria morfo-sintattica (colonna *CPoS* e *FPoS*)[30] che il token ha nel contesto specifico, *ii)* i relativi tratti morfologici (colonna *Tratti morfologici*) e *iii)* il lemma corrispondente (colonnna *Lemma*). Ad esempio, la forma 'disposizioni' (Id=2) viene ricondotta al lemma 'disposizione', viene annotato con la categoria sostantivo (S) e viene inoltre riconosciuto che si tratta di una forma plurale (num=p) e femminile (gen=f).

Il risultato dell'annotazione sintattica riportato nelle colonne *Testa sintattica* e *Relazione* della Tabella 1 permette inoltre di stabilire che, ad esempio, il sostantivo 'disposizioni' è il soggetto (subj) del verbo 'applicano', il quale costituisce la testa sintattica della relazione. Questa informazione è riportata nella colonna *Testa* dove è infatti segnalato che la testa sintattica del dipendente 'disposizioni' ha Id=9, l'Id cioè del token 'applicano'. In questo caso 'applicano' ha testa sintattica 0 dal momento che rappresenta il verbo della frase principale, radice (root) dell'albero sintattico dell'intero periodo. La fase di annotazione sintattica a dipendenze permette dunque di fornire una descrizione esplicita dell'intero albero sintattico del periodo analizzato, sotto forma di relazioni di dipendenza che legano i tokens che lo compongono. L'informazione può inoltre essere graficamente visualizzata, come mostra la Figura 1 che riporta la struttura sintattica della frase annotata, rappresentata come una serie di nodi lessicali (i singoli tokens), messi in collegamento da archi di dipendenza a loro volta etichettati con il nome del tipo di relazione di dipendenza (gli archi e le etichette graficamente rappresentati).



[30] Per ogni token viene riconosciuta la categoria morfo-sintattica generale (CPoS) e eventuali sottocategorie (FPoS). Ad esempio, alla forma (token) 'alla' viene associata la categoria preposizione (E) e viene ulteriormente specificato che si tratta di una preposizione articolata (EA). Allo stesso modo, il token '.' viene annotato come un segno di punteggiatura (F) di fine periodo (FS).

Figura 1: Un esempio di rappresentazione grafica dell'annotazione sintattica a dipendenze.

In questo studio è stata utilizzata *LinguA* (*Linguistic Annotation pipeline*), una catena di strumenti statistici di Trattamento Automatico del Linguaggio sviluppati in modo congiunto dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa[31]. Tali strumenti rappresentano lo stato dell'arte per la lingua italiana essendo risultati gli strumenti più precisi e affidabili nell'ambito della campagna di valutazione di strumenti per l'analisi automatica dell'italiano, EVALITA-2009[32]. In particolare, il modulo di annotazione morfo-sintattica ha dimostrato un'accuratezza del 96,34%[33] nell'identificazione simultanea della categoria grammaticale e dei tratti morfologici associati. Per quanto riguarda l'analisi a dipendenze, il modulo di annotazione sintattica a dipendenze realizzato dal parser *DeSR*[34] raggiunge livelli di LAS[35] e UAS[36] in linea con lo stato dell'arte dell'analisi a dipendenze, pari a 83,38% e 87,71% rispettivamente.

Come sottolineato per la prima volta da Gildea[37], gli strumenti di Trattamento Automatico del Linguaggio hanno una drastica diminuzione di accuratezza quando sono impiegati nell'analisi di tipologie di testi rappresentativi di un dominio diverso da quello sui quali gli strumenti sono stati sviluppati. Si tratta della questione nota come *Domain Adaptation*, attività di ricerca volta a definire metodologie di adattamento degli strumenti all'analisi di testi che appartengono a un dominio diverso da quello rispetto al quale sono stati sviluppati. Il dominio giuridico non rappresenta un'eccezione, come recentemente dimostrato dal "Domain Adaptation Track"[38] di EVALITA-2012[39] e dal "First Shared Task on Dependency Parsing of Legal Text"[40] dell'edizione 2012 del workshop "Semantic Processing of Legal Texts" (SPLeT-2012)[41], dove sono state messe a punto

[31] Una demo di *LinguA* è disponibile alla pagina http://linguistic-annotation-tool.italianlp.it/
[32] http://www.evalita.it/2009
[33] Il modulo di annotazione morfosintattica è descritto da F. DELL'ORLETTA, *Ensemble system for Part-of-Speech tagging*, in "Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)", Reggio Emilia, 2009, disponibile alla pagina http://www.evalita.it/sites/evalita.fbk.eu/files/proceedings2009/PoSTagging/POS_ILC.pdf. L'accuratezza è calcolata come il rapporto tra il numero di tokens classificati correttamente e il numero totale di tokens analizzati.
[34] G. ATTARDI, F. DELL'ORLETTA, M. SIMI, J. TURIAN, *Accurate Dependency Parsing with a Stacked Multilayer Perceptron*, in "Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)", Reggio Emilia, 2009, disponibile alla pagina http://www.evalita.it/sites/evalita.fbk.eu/files/proceedings2009/Parsing/Dependency/DEP_PARS_UNIPI_UNI_MONTREAL.pdf
[35] LAS (*Labelled Attachment Score*) è una metrica che indica la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda sia la testa sintattica sia la dipendenza che le lega.
[36] UAS (*Unlabelled Attachment Score*) è una metrica che indica la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda l'identificazione della testa sintattica.
[37] D. GILDEA, *Corpus variation and parser performance*, in "Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)", Pittsburgh, PA, 2001, pp. 167-202.
[38] F. DELL'ORLETTA, S. MARCHI, S. MONTEMAGNI, G. VENTURI, T. AGNOLONI, E. FRANCESCONI, *Domain Adaptation for Dependency Parsing at Evalita 2011*, in Magnini B., Cutugno F., Falcone M., Pianta E. (a cura di), "Evaluation of Natural Language and Speech Tool for Italian", LNCS–LNAI, Vol. 7689, Springer-Verlag Berlin Heidelberg, 2013, pp. 58-69.
[39] http://www.italianlp.it/software/evalita-2011-domain-adaptation-for-dependency-parsing/
[40] F. DELL'ORLETTA, S. MARCHI, S. MONTEMAGNI, B. PLANK, G. VENTURI, *The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts*, in "Proceedings of the LREC 2012 4th Workshop on Semantic Processing of Legal Texts (SPLeT 2012)", Istanbul, Turkey, 27 May, 2012, pp. 42.
[41] http://www.italianlp.it/software/first-shared-task-on-dependency-parsing-of-legal-texts-at-splet-2012/

una serie di metodologie per adattare strumenti di Trattamento Automatico del Linguaggio sviluppati per l'analisi di testi giornalistici, considerati rappresentativi della lingua comune, ai testi giuridici sia italiani sia inglesi. Particolare interesse è stato dedicato all'analisi di quali aspetti influiscono di più sul grado di accuratezza dell'annotazione sintattica. Una tale attenzione è legata al fatto che questo livello costituisce il punto di partenza per numerose applicazioni pratiche, quali ad esempio l'estrazione automatica di informazione, la traduzione automatica, il Question Answering, ecc...

Per superare questo problema, gli strumenti statistici di Trattamento Automatico del Linguaggio usati in questo lavoro sono stati adattati unendo due *training* corpora rappresentativi di due diversi domini: la treebank ISST-TANL[42], composta da articoli di giornale considerati rappresentativi della lingua comune, e il corpus TEMIS[43], una collezione di testi legislativi e amministrativi italiani annotata fino a livello sintattico, rappresentativi del dominio giuridico. Questo ha permesso di mantenere l'accuratezza degli strumenti allo stato dell'arte.

3.2. Il monitoraggio linguistico

Come precedentemente introdotto nel Paragrafo 3, il calcolo della leggibilità operato da READ-IT si basa sui risultati del monitoraggio di una serie di caratteristiche linguistiche rintracciate in un corpus a partire dall'output dei diversi livelli di annotazione linguistica: lemmatizzazione, annotazione morfo-sintattica e annotazione sintattica a dipendenze. Grazie a tale metodologia di monitoraggio, il profilo linguistico di un testo è ricostruito sulla base della distribuzione di tratti linguistici che spaziano tra diversi livelli di descrizione linguistica: lessicale, morfo-sintattico e sintattico. La Tabella 2 riporta alcuni dei tratti monitorati.

Prendendo le mosse da una più generale metodologia di monitoraggio della lingua italiana nelle sue varietà diamesiche, diastratiche, diafasiche introdotta per la prima volta da Dell'Orletta e Montemagni[44] e Montemagni[45] ai quali si rinvia per una descrizione dettagliata, tale metodo è stato sperimentato su varie tipologie di testi, come ad esempio le produzioni scritte e i materiali didattici offerti nella scuola primaria e secondaria allo scopo di monitorare le competenze linguistiche di apprendenti l'italiano come L2[46]. Per quanto riguarda il dominio giuridico, un confronto tra il

[42] F. DELL'ORLETTA, S. MARCHI, S. MONTEMAGNI, B. PLANK, G. VENTURI, *The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts*, in "Proceedings of the LREC 2012 4th Workshop on Semantic Processing of Legal Texts (SPLeT 2012)", Istanbul, Turkey, 27 May, 2012, pp. 42.

[43] G. VENTURI, *Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts*, in "Proceedings of the LREC 2012 4th Workshop on Semantic Processing of Legal Texts", Istanbul, Turkey, 27 May, 2012, pp. 1-12.

[44] F. DELL'ORLETTA, S. MONTEMAGNI, *Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scolastico*, in "Atti del XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI 2010)", 27-29 settembre, Viterbo, 2012.

[45] S. MONTEMAGNI, *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*, in "Studi Italiani di Linguistica Teorica e Applicata (SILTA)", Anno XLII, Numero 1, 2013, pp. 145-172.

[46] F. DELL'ORLETTA, S. MONTEMAGNI, E.M. VECCHI, G. VENTURI, *Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria*, in Giovanni Carlo Bruno, Immacolata Caruso, Manuela Sanna, Immacolata Vellecco (a cura di), "Percorsi migranti: uomini, diritto, lavoro, linguaggi", Milano, McGraw-Hill, 2011, pp. 319-336.

profilo linguistico di diversi tipi di testi (atti legislativi, atti amministrativi e sentenze) rappresentativi di diverse varietà della lingua del diritto è descritto in Venturi[47].

Tabella 2: Alcune delle caratteristiche considerate in fase di monitoraggio linguistico da READ-IT .

Tipo di caratteristica	Livello di annotazione linguistica	Caratteristica
Di base	Divisione in frasi	Lunghezza media dei periodi e delle parole
Lessicale	Lemmatizzazione e annotazione morfo-sintattica	Percentuale di lemmi appartenenti al *Vocabolario di Base* del *Grande dizionario italiano dell'uso* (De Mauro, 2000)
		Distribuzione dei lemmi rispetto ai repertori di uso (Fondamentale, Alto uso, Alta disponibilità)
Morfo-sintattico	Annotazione morfo-sintattica	Distribuzione delle categorie morfo-sintattiche
		Densità lessicale
Sintattico	Annotazione sintattica a dipendenze	Distribuzione dei vari tipi di relazioni di dipendenza
		Arità verbale
		Caratteristiche relative alla struttura dell'albero sintattico analizzato: - altezza media dell'intero albero, - lunghezza media della più lunga relazione di dipendenza
		Caratteristiche relative all'uso della subordinazione: - distribuzione di frasi principali vs. subordinate, - lunghezza media di sequenze consecutive di subordinate
		Caratteristiche relative alla modificazione nominale: - lunghezza media dei complementi preposizionali dipendenti in sequenza da un nome

3.3. Un esempio: la *Costituzione italiana* del 1947

Come esempio di output di READ-IT, abbiamo scelto di riportare i risultati di un esperimento condotto sulla *Costituzione italiana* nella sua versione originaria del 1947. La scelta è motivata dall'intenzione di verificare l'uso della metodologia di valutazione della leggibilità qui proposta su un tipo di testo legislativo a lungo studiato sia da linguisti, sia da giuristi, sia da esperti di informatica giuridica. Tali lavori hanno dimostrato come la nostra Costituzione sia caratterizzata da una prosa che si distingue per una «scorrevolezza e relativa facilità di lettura della nostra Carta

[47] G. VENTURI, *Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach*, in "The International Journal of Speech, Language and the Law (IJSLL)", 2014, (in corso di stampa).

fondamentale in confronto alla grande maggioranza dei testi normativi italiani»[48] a dimostrazione di uno «straordinario impegno dei *Costituenti*» e di un «non comune impegno linguistico»[49].

Come illustrato nella Figura 2, l'interfaccia di READ-IT permette di copiare e incollare nella scheda *Testo da analizzare* il testo di cui si intende calcolare il livello di leggibilità.

Figura 2: Il testo della *Costituzione Italiana* del 1947 da analizzare.

Una volta che gli strumenti di Trattamento Automatico del Linguaggio hanno annotato linguisticamente il testo in input, è possibile visualizzare il risultato del calcolo della leggibilità nella scheda *Analisi globale della leggibilità*, come si può vedere nella Figura 3[50]. Oltre al calcolo del valore di Gulpease, READ-IT conduce la valutazione globale della leggibilità del testo rispetto a quattro diversi indici calcolati sulla base di quattro diverse configurazioni di caratteristiche del testo:

- *Dylan BASE*: in questo modello, le caratteristiche considerate sono quelle usate nelle misure tradizionali della leggibilità di un testo (ovvero la lunghezza della frase e la lunghezza delle parole);

[48] B. MORTARA GARAVELLI, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, *cit.*, 1 p.
[49] T. DE MAURO, *Introduzione. Il linguaggio della Costituzione. In: Costituzione della Repubblica Italiana (1947)*, Torino, UTET, 2006, pp. vii-xxxii.
[50] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura3.jpg

- *Dylan LESSICALE*: questo modello si focalizza sulle caratteristiche lessicali del testo (ovvero la composizione del vocabolario e la sua ricchezza lessicale);
- *Dylan SINTATTICO*: questo modello si basa su informazione di tipo grammaticale, ovvero sulla combinazione di tratti morfo-sintattici e sintattici;
- *Dylan GLOBALE*: si tratta di un modello basato sulla combinazione di tutti i tratti considerati dagli altri modelli.

Per ciascun modello, la percentuale esprime il livello di difficoltà, ovvero si riferisce alla probabilità di appartenenza del testo in esame alla classe dei testi di *difficile* leggibilità[51]: la barra a fianco esprime visivamente questo valore, dove il rosso rappresenta la probabilità di appartenenza alla classe dei testi difficili e il verde a quelli di facile lettura.

Nel caso specifico, la *Costituzione italiana* ha un valore di difficoltà di lettura dell'99,4% dato dal modello *Dylan GLOBALE*, risulta dunque un testo di difficile lettura. A differenza del punteggio di leggibilità dato dall'indice Gulpease, pari a 54,9[52] (riportato nell'interfaccia di READ-IT), READ-IT fornisce un punteggio diverso a seconda del modello di calcolo della leggibilità considerato. Rispetto, ad esempio, al modello *Dylan BASE*, la *Costituzione* si rivela semplice, con un livello di difficoltà del 21,9%. Così come anche sulla base del modello *Dylan SINTATTICO*, che tiene in considerazione le caratteristiche sintattiche, il testo risulta meno complesso (46,9%) rispetto al modello globale basato sull'intero insieme di caratteristiche linguistiche.

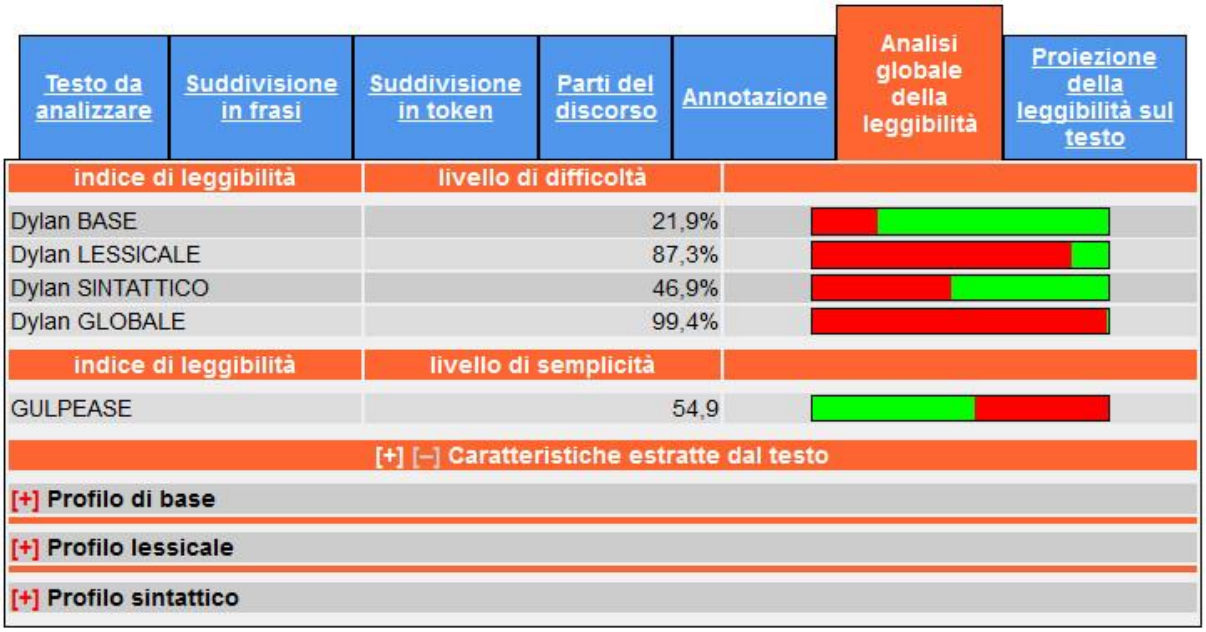


Figura 3: Risultato del calcolo globale della leggibilità della *Costituzione italiana*.

Un'analisi completa di tali differenze può essere condotta tenendo in considerazione le caratteristiche catturate da READ-IT in fase di monitoraggio linguistico del testo. Come si può vedere nella Figura 3, la sezione dell'interfaccia *Caratteristiche estratte dal testo* riporta i risultati

[51] I punteggi di leggibilità di READ-IT vanno dunque da 0 a 100: più il valore percentuale è basso, più il testo in esame è semplice.
[52] In base alla scala di leggibilità di Gulpease, un testo con punteggio di pari a 54,9 è un testo di difficile lettura per chi ha la licenza media. Rispetto a questo indice, la *Costituzione* risulta dunque un testo di media difficoltà di lettura.

del monitoraggio di un sottoinsieme (selezionato come significativo) delle caratteristiche linguistiche utilizzate da READ-IT nella misurazione della leggibilità. Se consideriamo, ad esempio, i tratti considerati nella ricostruzione del *Profilo di base* del testo analizzato (vedi Figura 4[53]), notiamo che la *Costituzione* contiene frasi con una lunghezza media pari a circa 16 tokens (15,8) per frase, una lunghezza che si avvicina di più a quella dei testi di facile lettura (che contengono frasi con una lunghezza media pari a 19 tokens) che non a quella dei testi di difficile lettura (con lunghezza media di frasi pari a 27 tokens) [54]. Nell'intera sezione, per ogni caratteristica riportata, oltre al valore numerico, viene fornita una rappresentazione grafica che mette a confronto il dato relativo al testo oggetto dell'analisi (corrispondente alla barra azzurra) con la corrispondente informazione rilevata nei corpora di riferimento di facile (barra verde) e difficile (barra rossa) lettura. Il rettangolino a fianco fornisce una classificazione semantica del dato rilevato in relazione al testo oggetto dell'analisi.

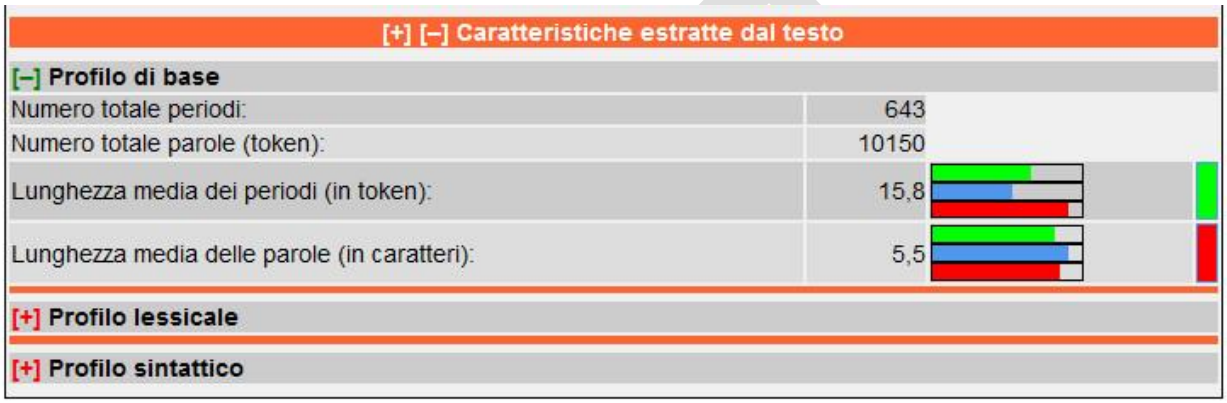


Figura 4: Caratteristiche linguistiche del *Profilo di base* della *Costituzione italiana*.

Il punteggio di leggibilità ottenuto sulla base del modello *Dylan LESSICALE* (87,3%), che tiene in considerazione le informazioni lessicali rintracciate all'interno del testo considerato, ci restituisce una *Costituzione* lessicalmente più vicina a *La Repubblica* che a *Due Parole*. Ciò è confermato dai risultati del monitoraggio linguistico di questo livello di analisi riportati nel *Profilo lessicale* (vedi Figura 5[55]). Rispetto, ad esempio, alla distribuzione di lemmi che appartengono al *Vocabolario di Base* (VdB), la *Costituzione*, con una percentuale pari al 58,2, ha valori più simili a quelli dei testi degli articoli di *La Repubblica* (nei quali la percentuale di lemmi del VdB è del 67,3) che a quelli di *Due Parole* (74,8%). Sebbene il modello lessicale di READ-IT indichi che la *Costituzione* si avvicina di più al polo qui considerato di difficile lettura, tuttavia dai risultati del monitoraggio emerge che la *Costituzione* contiene una percentuale di *Lessico Fondamentale* pari a quasi il 70%; ciò è indicativo di un testo pensato per essere leggibile ad un ampio pubblico di lettori.

[53] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura4.jpg

[54] Per visualizzare nell'interfaccia web i valori dei testi di facile e difficile lettura di riferimento è sufficiente passare con il cursore sulla barra verde o rossa.

[55] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura5.jpg

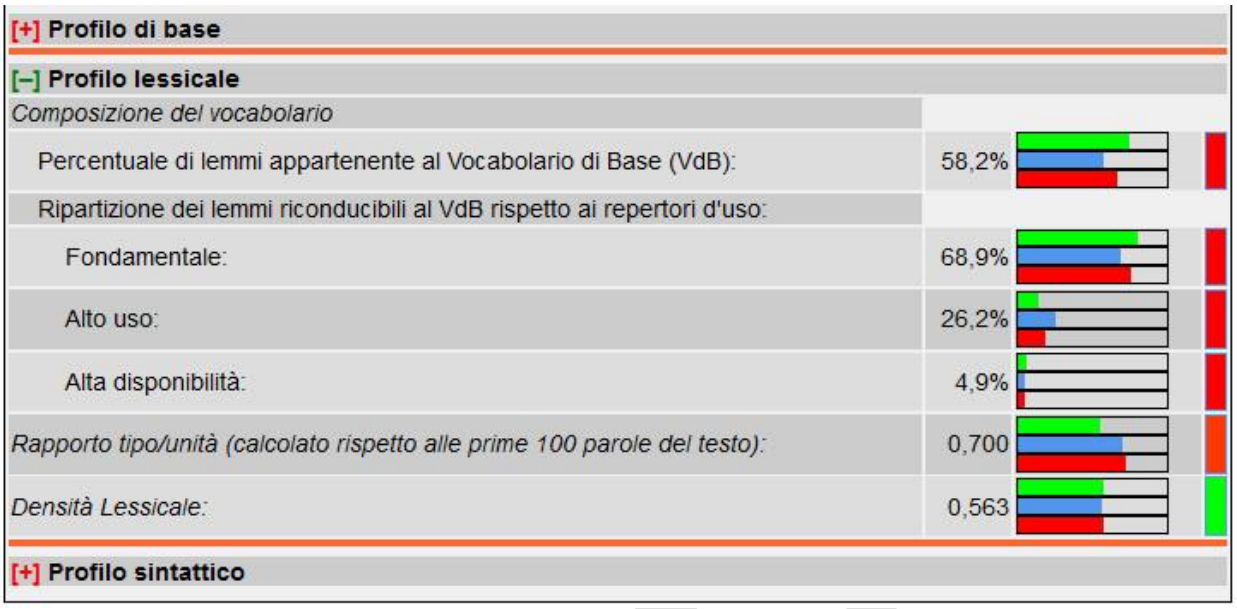


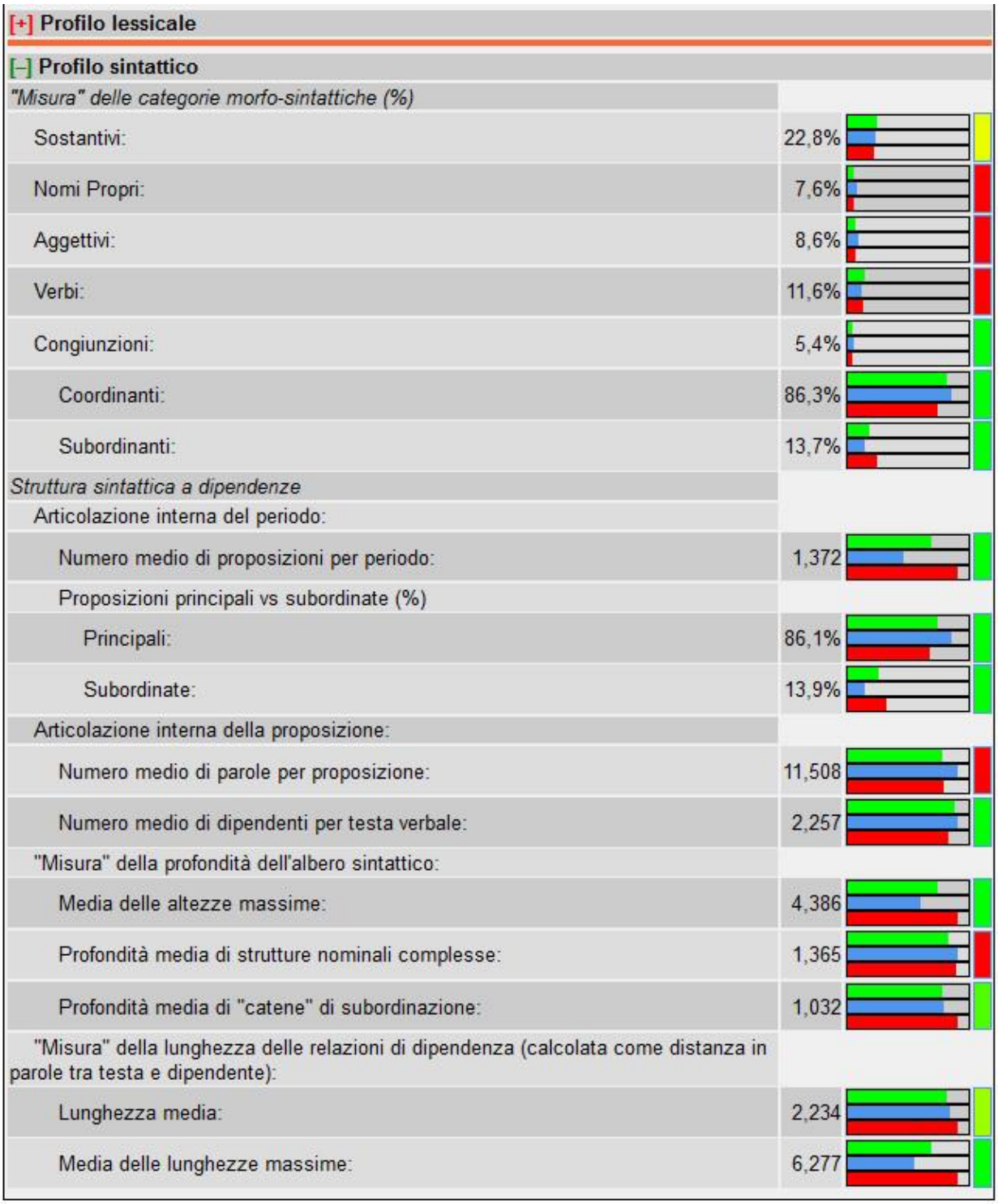
Figura 5: Caratteristiche linguistiche del *Profilo lessicale* della *Costituzione italiana*.

Sebbene tali risultati siano in contraddizione con quanto sino ad oggi fatto osservare circa la chiarezza e semplicità della *Costituzione italiana*, tuttavia, analizzando il livello di leggibilità del testo rispetto al suo profilo sintattico troviamo conferma del «non comune impegno linguistico»[56] dei padri costituenti verso la redazione di un testo leggibile e comprensibile. Se consideriamo infatti i tratti tenuti in considerazione nella ricostruzione del *Profilo sintattico* (vedi Figura 6[57]), notiamo che la *Costituzione* ha caratteristiche sintattiche più simili ai testi di semplice lettura, indicatori di semplicità che influiscono poi sul calcolo della leggibilità rispetto al modello *Dylan SINTATTICO*.

Per alcune caratteristiche sintattiche il testo mostra addirittura un comportamento riconducibile ad una facilità di lettura ancora maggiore dei testi qui considerati di facile lettura. È il caso, ad esempio, della media delle altezze massime degli alberi sintattici (*Media delle altezze massime* nella Figura 6); mentre i testi di facile lettura (la barra verde di riferimento) hanno valori medi pari a 5,292, la *Costituzione* ha un valore pari a 4,386. Le stesse differenze si ritrovano anche per quanto riguarda la lunghezza media delle relazioni di dipendenza sintattica (*Media delle lunghezze massime*); mentre i testi di facile lettura hanno lunghezze medie pari a 7,929, la *Costituzione* contiene frasi con relazioni di dipendenza lunghe in media 6,277.

[56] T. DE MAURO, *Introduzione. Il linguaggio della Costituzione*, *cit.*

[57] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura6.jpg



[+] Profilo lessicale

[–] Profilo sintattico

"Misura" delle categorie morfo-sintattiche (%)	
Sostantivi:	22,8%
Nomi Propri:	7,6%
Aggettivi:	8,6%
Verbi:	11,6%
Congiunzioni:	5,4%
Coordinanti:	86,3%
Subordinanti:	13,7%
Struttura sintattica a dipendenze	
Articolazione interna del periodo:	
Numero medio di proposizioni per periodo:	1,372
Proposizioni principali vs subordinate (%)	
Principali:	86,1%
Subordinate:	13,9%
Articolazione interna della proposizione:	
Numero medio di parole per proposizione:	11,508
Numero medio di dipendenti per testa verbale:	2,257
"Misura" della profondità dell'albero sintattico:	
Media delle altezze massime:	4,386
Profondità media di strutture nominali complesse:	1,365
Profondità media di "catene" di subordinazione:	1,032
"Misura" della lunghezza delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente):	
Lunghezza media:	2,234
Media delle lunghezze massime:	6,277

Figura 6: Caratteristiche linguistiche del *Profilo sintattico* dei primi 12 articoli della Costituzione.

Come precedentemente fatto notare, un tratto caratterizzante READ-IT, innovativo rispetto alla letteratura internazionale in materia, consiste in una valutazione della leggibilità articolata su due livelli: il documento e la singola frase. La valutazione rispetto alla frase è stata esplicitamente concepita per fornire un supporto al redattore del testo e guidarlo nel processo di revisione e semplificazione. I risultati di questo livello più granulare di calcolo della leggibilità sono contenuti nella scheda *Proiezione della leggibilità sul testo* dove è possibile identificare le frasi che

necessitano di revisione. Come si può vedere nella Figura 7[58], per ogni frase viene riportato il livello di difficoltà base (*base*), lessicale (*less.*), sintattico (*sint.*) e globale (*glob.*) in colonne distinte, livello calcolato dai corrispondenti modelli di analisi della leggibilità. Il livello di difficoltà è rappresentato cromaticamente mediante colori che vanno dal verde (frase leggibile) al rosso (frase particolarmente difficile): il rosso, così come sfumature giallo-arancioni, marcano frasi che necessitano di revisione.

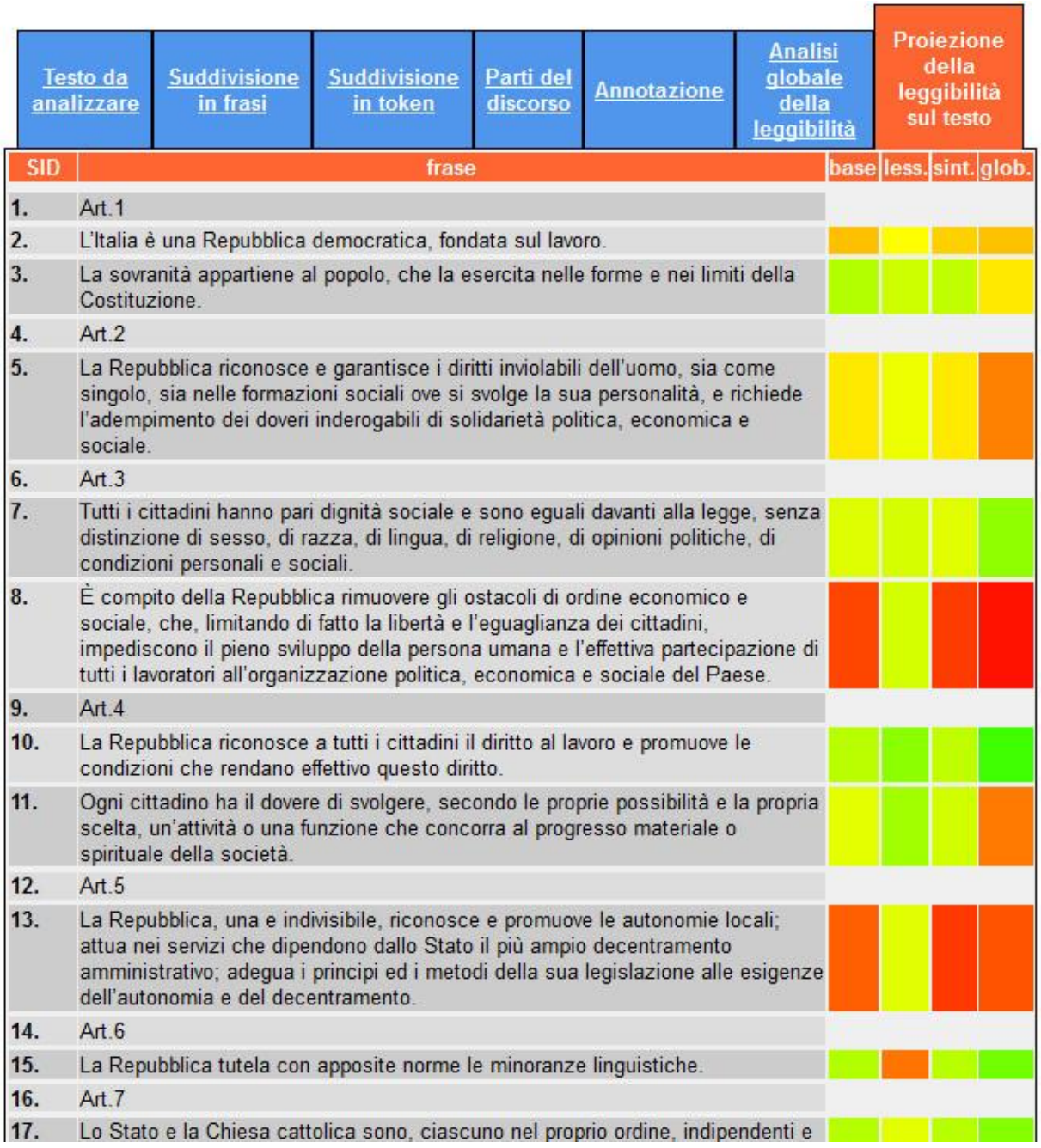


Figura 7: Leggibilità delle singole frasi contenute nella *Costituzione italiana*.

[58] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura7.jpg

Mentre dunque una frase come la terza è mediamente semplice, con un punteggio di leggibilità pari a 54,4%, l'ottava frase è decisamente più complessa con un punteggio di 96,5%[59]. Questo è un segnale per il redattore che viene così invitato a riformulare la frase facendo uso di strutture sintattiche più semplici, ad esempio evitando l'uso di una subordinata implicita espressa con il gerundio ('limitando di fatto la libertà e l'eguaglianza dei cittadini') incassata all'interno di una subordinata relativa ('che, limitando di fatto la libertà e l'eguaglianza dei cittadini, impediscono il pieno sviluppo della persona umana e …').

4. READ-IT e i testi della pubblica amministrazione: un esempio di applicazione

Obiettivo di questo paragrafo è mostrare come le indicazioni offerte da READ-IT possano concretizzarsi in uno strumento redazionale efficace per migliorare la qualità di testi amministrativi, soprattutto di quelli rivolti alla collettività, e dunque favorire una comunicazione istituzionale più efficiente.

A titolo di esempio, è stato selezionato un testo tratto da un corpus più ampio di documenti amministrativi[60] che, sebbene esemplificativi di tipologie testuali diverse (es. lettere, moduli, informative), sono accomunati dall'avere come principale destinatario il cittadino comune. Come ricordato nell'introduzione, sono proprio questi testi a richiedere un'attenzione particolare, dal momento che troppo spesso l'amministrazione comunica al cittadino adottando uno stile improprio e un linguaggio oscuro e poco accessibile, meglio conosciuto come *burocratese*.

Nell'ambito delle numerose iniziative in favore della semplificazione del linguaggio amministrativo, un indice di leggibilità qualitativamente "avanzato" può offrire un contributo significativo: esso infatti può rappresentare un ausilio per il redattore di testi di rilievo pubblico per verificare l'aderenza del proprio scritto alle linee guida della semplificazione – di cui la *Guida alla redazione degli atti amministrativi. Regole e suggerimenti* (cf. § 2) costituisce ad oggi la sintesi più aggiornata – e, in caso negativo, riadattarlo in questa direzione.

Il testo scelto riflette inoltre un'altra particolarità del corpus da cui proviene, che si configura come una sorta di «corpus parallelo monolingue». Ciò significa che ciascun testo della collezione si compone tanto della sua versione originale, prodotta internamente dalle singole amministrazioni, tanto della relativa versione semplificata, frutto di un lavoro di riscrittura[61] ad opera di linguisti,

[59] Per visualizzare nell'interfaccia web i valori di leggibilità relativi ad ogni livello è sufficiente passare con il cursore sulle colonne colorate corrispondenti.

[60] Il corpus è stato raccolto nell'ambito di uno studio più esteso volto ad analizzare le peculiarità della prosa burocratica e la leggibilità dei testi amministrativi da una prospettiva linguistico-computazionale. Per un approfondimento si rimanda a D. BRUNATO, *Complessità necessaria o stereotipi del burocratese? Un'indagine sulla leggibilità del linguaggio amministrativo da una prospettiva linguistico-computazionale*, 1° volume di atti del XIII Congresso SILFI, 2014, (in corso di stampa).

[61] I testi qui considerati, così come le relative riscritture, provengono da esercitazioni sulle tecniche di scrittura professionale tenutesi nell'ambito di corsi universitari e di corsi di aggiornamento indirizzati ai dipendenti pubblici, organizzati dal Dipartimento di Linguistica dell'Università di Padova, sotto la supervisione del linguista Prof. Michele Cortelazzo. Alcuni di questi testi, parte del cosiddetto corpus TACS (testi amministrativi chiari e semplici) sono consultabili all'indirizzo: http://www.maldura.unipd.it/buro/index.html

ispirato a quei principi di «chiarezza, semplicità e sinteticità»[62] che ritroviamo esplicitati nei diversi contributi alla semplificazione del linguaggio amministrativo.

In virtù di questa caratterizzazione, si può assumere che le due versioni del testo – quella autentica e quella adattata dal linguista – rappresentino rispettivamente gli estremi opposti verso cui può tendere la scrittura amministrativa. Pertanto, l'interesse è qui mostrare come una serie di parametri linguistici estratti automaticamente dal testo, e caratterizzanti la nozione di leggibilità a diversi livelli di rappresentazione, consenta non solo di discriminare in maniera automatica tra testi originali e testi riscritti, bensì di fornire anche un'indicazione qualitativa sulla natura della semplificazione, grazie alla capacità di intercettare quelle strutture lessicali, morfo-sintattiche e sintattiche su cui il linguista è intervenuto allo scopo di rendere il testo originale più comprensibile. La possibilità di riconoscere i luoghi di complessità del testo in maniera così granulare è il presupposto all'uso di READ-IT come strumento per la semplificazione semiautomatica del testo stesso.

L'esempio scelto contiene il testo di una lettera inviata da un' amministrazione comunale ad un privato cittadino, in cui viene comunicata la necessità di richiedere un sopralluogo tecnico come condizione preliminare per dichiarare la condizione di inabitabilità del proprio immobile. Di seguito si riportano il testo originale (i) e la sua versione riscritta (ii).

(i)

> *A seguito della dichiarazione sostitutiva dell'atto notorio di cui alla L. 15/68 presentata dalla S.V. il 25.06.1998, siamo a comunicare che l'atto è stato trasmesso per i controlli di competenza all'Ufficio Tecnico Comunale, che, con nota n. 4007 del 19.10.1998, ha precisato di non aver rilasciato dichiarazione di inabitabilità o inagibilità per l'immobile in oggetto specificato.*
>
> *Si precisa che i proprietari degli immobili non hanno alcun titolo a dichiarare lo stato di inabitabilità - inagibilità di un fabbricato; le norme in materia stabiliscono infatti che la suddetta dichiarazione è rilasciata dal Sindaco (art. 4 D.P.R. 423/94, art. 222 del R.D. 1264/34, art. 38 L. 142/90).*
>
> *In base a quanto specificato, le dichiarazioni sostitutive dell'atto di notorietà sono valide nel caso in cui già preesista un provvedimento di inabitabilità - inagibilità, che dovrà essere prodotto allo scrivente ufficio.*
>
> *Nel caso in cui la S.V. sia sprovvista di tale provvedimento, La invitiamo a richiedere, con la massima urgenza, un sopralluogo dell'Ufficio Tecnico Comunale (Settore Edilizia Privata - via fra' P. Sarpi, 2 - Telefono 8704707).*
>
> *Si fa presente che le mendaci dichiarazioni in atti pubblici e l'occupazione di immobili dichiarati inabitabili sono sanzionate penalmente.*
>
> *Rammentiamo infine che per inabitabilità/inagibilità sopravvenuta di un edificio è prevista la presentazione della denuncia di variazione ICI, ai sensi dell'art. 10, comma 4, del Decreto Legislativo 504/92.*
>
> *Per ulteriori informazioni, si invita a presentarsi agli sportelli di questo Ufficio, in Prato della Valle n. 98/99 o a telefonare allo 049/8205820-1*
>
> *Distinti saluti.*

[62] Direttiva del Ministro per la Funzione pubblica dell'8 maggio 2002 («Direttiva sulla semplificazione del linguaggio dei testi amministrativi»), art. 8.

(ii)

> *Egregio Signore,*
>
> *con la dichiarazione sostitutiva dell'atto notorio, il 25.6.1998 Lei ha dichiarato l'inabitabilità o l'inagibilità dell'immobile di via Roma 1. L'Ufficio Tecnico Comunale ci ha però precisato di non aver rilasciato nessuna dichiarazione di inabitabilità o inagibilità per quell'immobile.*
>
> *La dichiarazione sostitutiva dell'atto notorio può essere presentata dal proprietario solo quando esiste una dichiarazione di inabitabilità o inagibilità rilasciata dal Sindaco.*
>
> *La invitiamo pertanto a portare nei nostri uffici tale provvedimento. Se ne è sprovvisto, richieda al più presto un sopralluogo all'Ufficio Tecnico Comunale (via fra' P. Sarpi, 2 - tel. 049 8704707).*
>
> *Le ricordiamo che la legge punisce chi rilascia false dichiarazioni o il proprietario di immobili che vengono utilizzati dopo essere stati dichiarati inabitabili o inagibili.*
>
> *Le ricordiamo inoltre che, quando un immobile viene dichiarato inagibile o inabitabile, bisogna presentare la denuncia di variazione I.C.I. prevista dall'art. 10, comma 4, del Decreto Legislativo 504/92.*
>
> *Per ulteriori informazioni, Lei si può rivolgere all'Ufficio I.C.I. (Prato della Valle n. 98/99, tel. 049 8205820-1).*
>
> *Distinti saluti.*

Come si può notare, la "traduzione" dal *burocratese* ad una prosa più agile ed efficace sul piano comunicativo, ma comunque fedele alle intenzioni dell'emittente, si è concretizzata in interventi a diversi livelli della struttura linguistica. Sul piano lessicale, ad esempio, parole e locuzioni di uso tipicamente burocratico sono state sostituite da sinonimi più familiari, che pur ne preservano il significato di partenza (es. *S.V.* con *lei*; *l'immobile in oggetto* con *quell'immobile; sanzionare* con *punire; nel caso in cui* con *se).* Ovviamente, le modifiche non hanno agito invece sui tecnicismi e sui termini di dominio, quando necessari a mantenere la correttezza del messaggio originale (es. *I.C.I.*, *atto notorio*, *inagibile*, *inabitabile*).

Ben più marcate sono invece le trasformazioni sul piano sintattico che non solo portano ad una riduzione della lunghezza media della frase[63], ma soprattutto producono un testo dal tono comunicativo meno distaccato, chiariscono i soggetti delle azioni indicate ed esplicitano alcuni informazioni che nella formulazione originale venivano lasciate sottese. In questa direzione vanno, ad esempio, la sostituzione delle forme impersonali del verbo con la forma personale (es. *Si fa presente* vs *Le ricordiamo; Si invita a presentarsi* vs *Lei può rivolgersi),* la trasformazione del passivo in attivo (*le mendaci dichiarazioni* [...] *sono sanzionate penalmente* vs *la legge punisce chi rilascia false dichiarazioni),* lo scioglimento delle nominalizzazioni in particolari contesti sintattici

[63] A proposito del linguaggio burocratico, Cortelazzo e Viale fanno notare che «l'unione tra complessità lessicale e complessità morfologica genera 'ipertrofia': la lingua burocratica utilizza più parole di quella comune per dire le stesse cose». La citazione è contenuta in: M. CORTELAZZO e M. VIALE, *Storia del linguaggio politico, giuridico e amministrativo nella Romània: italiano / Geschichte der Sprache der Politik, des Rechts und der Verwaltung in der Romània: Italienisch*, in: Gerhard Ernst, Martin-Dietrich Gleßgen, Christian Schmitt und Wolfgang Schweickard (Hg.), Romanische Sprachgeschichte. Ein internationales Handbuch zur Geschichte der romanischen Sprachen, 2. Teilband / Histoire linguistique de la Romània. Manuel international d'histoire linguistique de la Romània, Tome 2, Berlin – New York, Walter de Gruyter Verlag, 2006, pp. 2112-2123, 2118 p.

(es. *per inabitabilità/inagibilità sopravvenuta* vs *quando un immobile viene dichiarato inagibile o inabitabile)*[64].

Abbiamo dunque sottoposto i due testi all'analisi in READ-IT, di cui mostriamo innanzitutto l'output rispetto all'analisi globale della leggibilità (Figure 8[65] e 9[66]).

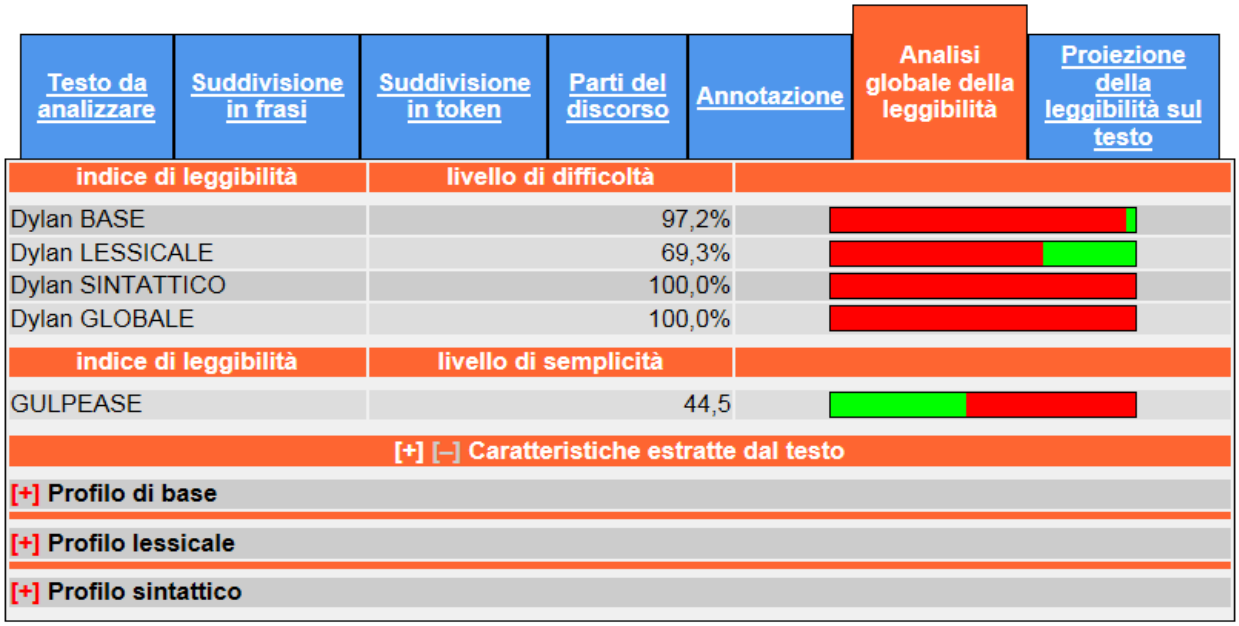


Figura 8: Risultato del calcolo globale della leggibilità sul testo originale

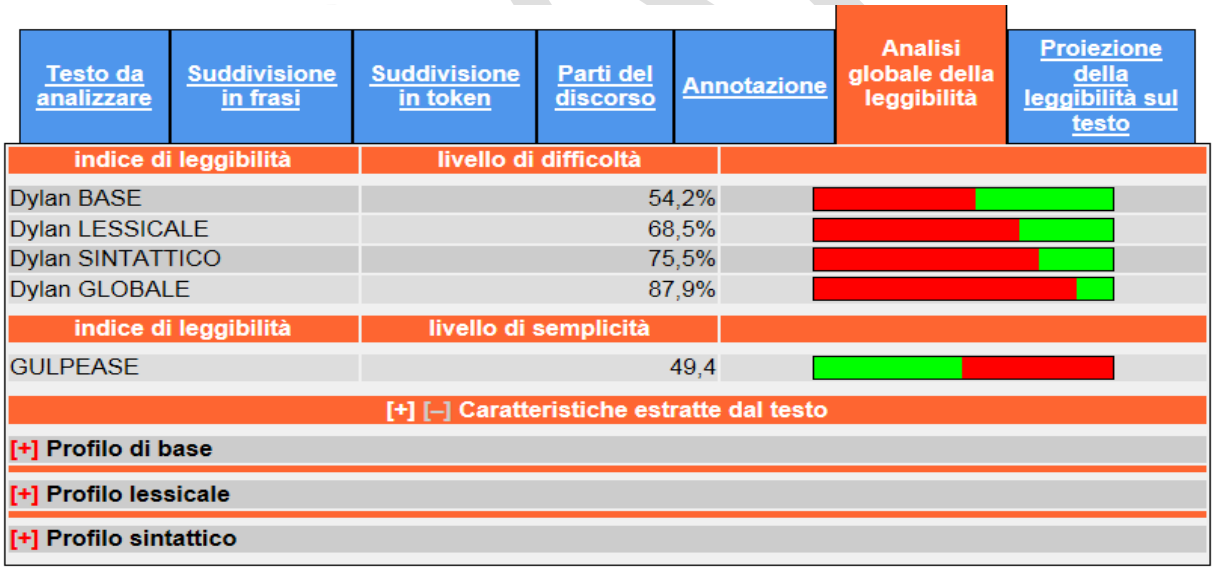


Figura 9: Risultato del calcolo globale della leggibilità sul testo semplificato

[64] Ulteriori commenti sugli esiti della semplificazione sono disponibili navigando la pagina http://www.maldura.unipd.it/buro/ e seguendo i links TACS/raccolta/richiesta di una dichiarazione di inabitabilità.

[65] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura8.jpg

[66] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura9.jpg

Si può chiaramente osservare che il risultato della semplificazione viene intercettato da tutti i modelli di analisi della leggibilità disponibili in READ-IT. In particolare, se il modello *Dylan base*, ispirato alle formule di leggibilità tradizionali (cf. § 3.3), assegna alla versione riscritta un punteggio di leggibilità quasi raddoppiato rispetto al testo originale, ben più interessante è valutare in maniera comparativa i risultati ottenuti dai modelli basati sul computo di parametri linguistici più sofisticati. Questi dati rispecchiano quanto osservato nell'analisi qualitativa, ovvero come a rendere il testo originale poco accessibile al lettore sia soprattutto la sua costruzione sintattica, più che quella lessicale. Se infatti il punteggio riportato dal modello *Dylan Lessicale* risulta pressoché invariato (69,3% per il testo originale e 68,5% per quello semplificato), la diminuzione dell'indice di difficoltà sintattica tra le due versioni è pari invece a quasi 25 punti percentuali. Tale risultato suggerisce che le caratteristiche linguistiche contemplate dal modello *Dylan Sintattico* sono effettivamente buone spie per tradurre in una metrica computazionale tipologie diverse di interventi di semplificazione sintattica, quali ad esempio lo scioglimento delle nominalizzazioni o la riduzione dei fenomeni di marcatezza (es. frasi passive o impersonali). A questo proposito, il confronto più dettagliato delle caratteristiche sintattiche monitorate dal relativo modello, disponibile nella sezione dedicata al *Profilo sintattico* (Figure 10[67] e 11[68]), è piuttosto significativo.

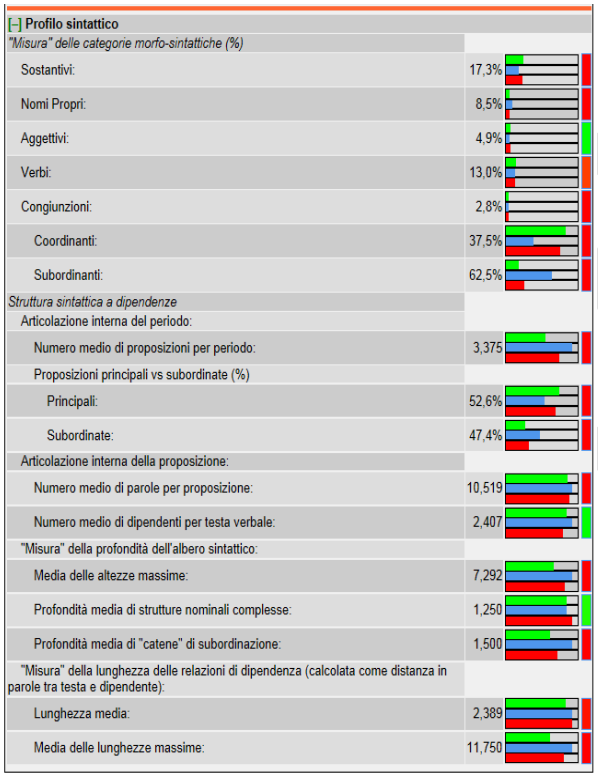


Figura 10: Caratteristiche linguistiche del *Profilo Sintattico* del testo originale

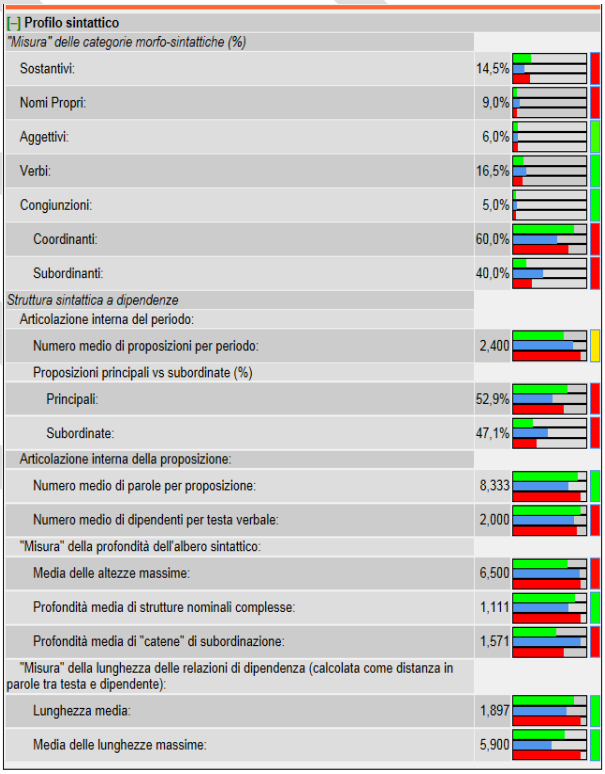


Figura 11: Caratteristiche linguistiche del *Profilo Sintattico* del testo semplificato

Dall'analisi comparativa osserviamo, ad esempio, che la versione semplificata riporta valori più bassi rispetto a tutte quelle caratteristiche linguistiche che descrivono la struttura dell'albero sintattico della frase, in termini sia di profondità che di lunghezza delle relazioni di dipendenza. È il

[67] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura10.jpg

[68] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura11.jpg

caso della *lunghezza media delle relazioni di dipendenza* (penultima riga nelle figure 10 e 11), che diminuisce di più di 0,4 punti (da 2,389 nel testo originale a 1,897 nel testo semplificato), e soprattutto della *media delle lunghezze massime* (ultima riga delle stesse figure), pari a 11,75 nel testo originale e a 5,90 in quello semplificato. La presenza di dipendenze sintattiche più lunghe, che viene calcolata come numero di parole che separano il costituente testa della relazione dal relativo dipendente, si può incrociare con il dato relativo alla *profondità media delle strutture nominali complesse*, pari a 1,25 nel testo originale e a 1,11 in quello semplificato: quest'ultima variazione, sebbene non particolarmente accentuata, è comunque molto interessante, dal momento che la presenza di lunghe catene di modificatori del nome è un marcatore riconosciuto del linguaggio giuridico, segnalato come «fonte di oscurità e difficoltà interpretative» in questa tipologia di testi[69]. Pertanto, l'attestazione di valori più alti rispetto a questo parametro riflette, nel testo di partenza, una certa tendenza del dipendente della pubblica amministrazione ad assumere uno stile comunicativo che ricalca il modo in cui sono scritte le leggi. Di contro, la corrispondente diminuzione nella versione semplificata segnala che gli autori della riscrittura hanno ritenuto necessario intervenire su questo tipo di costrutti nominali, soprattutto laddove gli stessi intervengono spezzando la continuità frasale (ad esempio perché ricorrono tra un link di dipendenza sintattica soggetto-verbo) ed è indicativo che l'esito della semplificazione restituisca un testo che è addirittura più semplice dei testi di facile lettura rispetto a questi parametri.

La portata applicativa delle indicazioni fornite da READ-IT diventa ancor più interessante nel momento in cui l'attenzione si sposta dall'analisi della leggibilità del testo a quella delle singole frasi (Figura 12[70] e Figura 13[71]).

Testo da analizzare	Suddivisione in frasi	Suddivisione in token	Parti del discorso	Annotazione	Analisi globale della leggibilità	Proiezione della leggibilità sul testo			
SID	frase					base	less.	sint.	glob.
1.	A seguito della dichiarazione sostitutiva dell'atto notorio di cui alla L. 15/68 presentata dalla S.V. il 25.06.1998, siamo a comunicare che l'atto è stato trasmesso per i controlli di competenza all'Ufficio Tecnico Comunale, che, con nota n. 4007 del 19.10.1998, ha precisato di non aver rilasciato dichiarazione di inabitabilità o inagibilità per l'immobile in oggetto specificato.								
2.	Si precisa che i proprietari degli immobili non hanno alcun titolo a dichiarare lo stato di inabitabilità - inagibilità di un fabbricato; le norme in materia stabiliscono infatti che la suddetta dichiarazione è rilasciata dal Sindaco (art. 4 D.P.R. 423/94, art. 222 del R.D. 1264/34, art. 38 L. 142/90).								
3.	In base a quanto specificato, le dichiarazioni sostitutive dell'atto di notorietà sono valide nel caso in cui già preesista un provvedimento di inabitabilità - inagibilità, che dovrà essere prodotto allo scrivente ufficio.								
4.	Nel caso in cui la S.V. sia sprovvista di tale provvedimento, La invitiamo a richiedere, con la massima urgenza, un sopralluogo dell'Ufficio Tecnico Comunale (Settore Edilizia Privata - via fra' P. Sarpi, 2 - Telefono 8704707).								
5.	Si fa presente che le mendaci dichiarazioni in atti pubblici e l'occupazione di immobili dichiarati inabitabili sono sanzionate penalmente.								
6.	Rammentiamo infine che per inabitabilità/inagibilità sopravvenuta di un edificio è prevista la presentazione della denuncia di variazione ICI, ai sensi dell'art. 10, comma 4, del Decreto Legislativo 504/92.								
7.	Per ulteriori informazioni, si invita a presentarsi agli sportelli di questo Ufficio, in Prato della Valle n. 98/99 o a telefonare allo 049/8205820-1								
8.	Distinti saluti.								

Figura 12: Leggibilità delle singole frasi contenute nel testo originale.

[69] Osserva B. MORTARA GARAVELLI, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani, cit.*, 173-175 pp., che i testi giuridici si caratterizzano per l'uso di sintagmi nominali a volte anche molto lunghi, ricchi di nominalizzazioni che si presentano come «grappoli di astrazioni concatenate in "complementi del nome"», ossia di «parafrasi riduttive, che contraggono in un nome (astratto) gli elementi di una proposizione (verbo + argomenti del verbo)».

[70] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura12.jpg

[71] La versione a colori dell'immagine è disponibile alla pagina http://www.italianlp.it/wp-content/uploads/downloads/figure_READ-IT/figura13.jpg

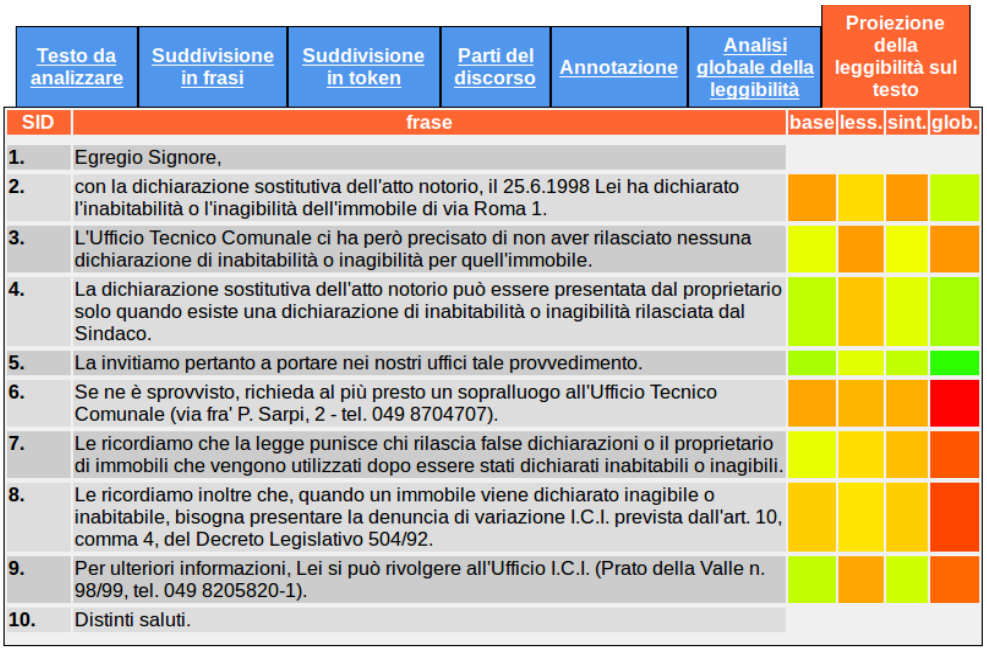


Figura 13: Leggibilità delle singole frasi contenute nel testo semplificato.

A questo livello, la capacità di intercettare le differenze tra le due versioni si traduce nell'attribuzione di punteggi di complessità mediamente più elevati alle frasi del testo originale. Ancora una volta è interessante osservare come le frasi originali risultino più difficili non solo in virtù della maggior lunghezza, aspetto che incide sulla leggibilità del modello base (colonna *base*), ma anche di una costruzione sintattica complessa (colonna *sint.*), come ben esemplifica a livello cromatico il colore rosso che ottiene la prima frase del testo originale. Al contrario, la revisione non abbassa di molto la difficoltà sul piano lessicale (colonna *less.*), segnale della presenza di alcune peculiarità del testo burocratico che, seppur poco ricorrenti nella lingua comune, non sempre possono essere semplificate.

È innegabile infatti che la lingua burocratico-amministrativa, definita «lingua settoriale non specialistica»[72], presenti alcuni aspetti di complessità "ineliminabile" che interessano soprattutto il repertorio terminologico; ciò è dovuto non tanto alla presenza di un lessico burocratico connaturato, quanto all'eterogeneità delle materie che l'amministrazione si trova a disciplinare, che richiedono spesso l'uso di sottocodici propri del settore oggetto di trattazione (es. sanità, edilizia, giurisprudenza). È a questo livello che la specializzazione di un indice di leggibilità "avanzato" diventa importante al fine di discriminare tra l'inutile *burocratese* e il necessario lessico di dominio.

5. Conclusioni e sviluppi futuri

In questo contributo abbiamo illustrato una metodologia per il calcolo della leggibilità di un testo basata su strumenti di Trattamento Automatico del Linguaggio ed espressamente rivolta alla sua semplificazione. Tale metodologia, già sperimentata con successo su diverse tipologie di testi, si è rivelata affidabile anche nel caso dei testi giuridici. A nostra conoscenza, tale studio rappresenta il

[72] A.A. SOBRERO, *Lingue speciali*, in: Introduzione all'italiano contemporaneo. La variazione e gli usi, (a cura di) A.A Sobrero, Roma - Bari, pp. 237-277, 237 p.

primo tentativo volto a mostrare come tecnologie linguistico-computazionali allo stato dell'arte per la lingua italiana incomincino ad essere oggi mature per essere usate non solo come ausilio per definire la leggibilità di testi giuridici ma anche come guida per una loro stesura semplificata.

Sebbene il metodo di monitoraggio linguistico e calcolo della leggibilità qui adottato sia stato progettato per l'analisi di testi rappresentativi della lingua comune, gli esperimenti condotti hanno dimostrato come la metodologia seguita riesca tuttavia ad intercettare le difficoltà della *lingua del diritto*, mettendone in luce gli specifici luoghi di complessità. In particolare, l'approccio si è rivelato un punto di partenza affidabile all'interno di un processo di semplificazione assistita di documenti amministrativi, processo tanto più centrale in una reale società inclusiva dove il rapporto istituzioni-cittadini dovrebbe essere al centro della vita democratica di uno Stato.

Grazie all'innovativa possibilità offerta da READ-IT di valutare la leggibilità non solo di un intero documento, ma anche di ogni singola frase in esso contenuta, è stato possibile delineare diverse strategie di semplificazione del testo, strategie specifiche per ogni livello di analisi linguistica considerato. Ad oggi, il metodo qui presentato è in grado di intercettare marcatori che incidono soprattutto sul livello di complessità sintattica, guidandone in questo modo la semplificazione. Al contrario, è soprattutto a livello lessicale che la specializzazione dell'approccio al calcolo della leggibilità si è dimostrata necessaria. Una possibile soluzione potrebbe ad esempio riguardare la specializzazione delle risorse lessicali di riferimento. Oltre a verificare la distribuzione di lemmi appartenenti al *Vocabolario di Base* della lingua italiana, la composizione interna del vocabolario del testo in esame potrebbe essere confrontata con repertori terminologici di dominio. L'obiettivo è quello di non penalizzare indiscriminatamente l'uso di terminologia specialistica all'interno di documenti amministrativi che, seppur meno rappresentata nella lingua comune, è funzionale alla loro interpretazione corretta.

Tra le possibili direzioni di ricerca offerte dalla metodologia di analisi della leggibilità e semplificazione del testo qui presentata vi è la sua applicazione su altre tipologie di documenti rappresentativi della lingua del diritto. Come precedentemente messo in luce da Venturi[73], testi che appartengono a varietà diverse del linguaggio giuridico non solo mostrano profili linguistici diversi tra di loro ma differiscono anche in modo diverso rispetto a testi rappresentativi della lingua comune. È questo il motivo per cui stiamo al momento sperimentando *i)* se i tratti linguistici che il processo di monitoraggio ha rivelato essere particolarmente caratterizzanti ad esempio testi giurisprudenziali come le sentenze sono catturate da READ-IT, *ii)* in che misura essi influiscano nella valutazione della leggibilità di questi documenti e *iii)* se sono necessarie specializzazioni che permettano di non tralasciare aspetti specifici di diverse varietà della lingua del diritto.

La validazione dei risultati del calcolo della leggibilità rispetto a test di comprensione sottoposti a soggetti umani è infine tra le attività di ricerca che abbiamo intenzione di portare avanti. La metodologia di definizione di un indice di leggibilità qui descritta parte dall'idea esposta da Piemontese[74] che «qualunque sia il tipo di testo, è possibile stabilire, in rapporto al destinatario, un punto critico di leggibilità» e che dunque si tratta di «stabilire a quali condizioni, ed entro quali bande di oscillazione, un testo può essere definito, di volta in volta, di facile lettura». È questo il

[73] G. VENTURI, *Investigating legal language peculiarities across different types of Italian legal texts: an NLP-based approach*, cit.
[74] M.E. PIEMONTESE, *Capire e farsi capire. Teorie e tecniche della scrittura controllata*, *cit.*, 128 p.

motivo per cui crediamo che complemento naturale di questo studio sia una fase di validazione empirica sia del punteggio di leggibilità fornito da READ-IT sia del risultato del processo di semplificazione del testo guidato dai suggerimenti forniti dallo strumento. Se un tale approccio è valido in generale per qualsiasi tipo di testo, il riscontro di indici automatici con giudizi umani di comprensione del testo è tanto più importante nel caso di documenti come quelli giuridici al centro di un'efficace ed efficiente comunicazione istituzioni-cittadino e per questo rivolti ad un pubblico quanto mai eterogeneo rispetto alle competenze linguistiche.