Dominique Brunato
(Università degli studi di Siena / Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR)[1])

Complessità necessaria o stereotipi del "burocratese"? Un'indagine sulla leggibilità del linguaggio amministrativo da una prospettiva linguistico-computazionale

1. Introduzione

La complessità del linguaggio amministrativo è un tema che suscita l'interesse tanto dello studioso della lingua nelle sue varietà d'uso, quanto di chi si occupa di comunicazione pubblica; quasi certamente risulta familiare anche alla maggior parte dei cittadini, chiamati nel loro agire quotidiano a confrontarsi con gli scritti delle pubbliche amministrazioni, siano essi in forma di richieste di documenti personali, moduli per l'iscrizione scolastica, bandi di concorso e così via.

Numerosi in letteratura sono gli studi condotti con gli strumenti e i metodi dell'indagine linguistica tradizionale che hanno descritto i tratti tipici del "burocratese", tanto rispetto alle scelte lessicali e sintattiche, quanto sul piano dell'organizzazione dei contenuti testuali[2]. In molti casi, le competenze del linguista sono state affiancate da quelle dell'esperto di dominio (giurista, funzionario amministrativo), allo scopo di proporre suggerimenti e indicazioni concrete per la semplificazione di questo linguaggio[3], un'esigenza che negli ultimi vent'anni è divenuta a più riprese materia d'interesse anche in sede legislativa[4]. Nell'ambito di queste iniziative è stato spesso ribadito che la semplificazione non può né deve risultare in una banalizzazione del testo stesso; il linguaggio

[1] La metodologia di monitoraggio del profilo linguistico del testo applicata in questo caso di studio costituisce una delle linee di ricerca del laboratorio ItaliaNLP Lab dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR). Un ringraziamento particolare va ai membri del laboratorio, in particolare Simonetta Montemagni, Giulia Venturi e Felice Dell'Orletta per gli spunti di riflessione e il supporto applicativo nella costituzione dei corpora e nell'estrazione dei dati. Inoltre, ringrazio il professor Michele Cortelazzo che ha gentilmente reso disponibile la versione elettronica di gran parte dei testi amministrativi selezionati per il corpus qui presentato.

[2] Cfr, fra gli altri, Berruto (1987: 163-166); Fortis (2005: 57-88); Serianni (2003: 123-139); Beccaria (1988: 180-192); Raso (2005).

[3] Tra i numerosi contributi alla semplificazione del linguaggio amministrativo, ispirati al modello del Plain Language di derivazione anglosassone, citiamo qui solo alcuni: la *Guida alla redazione degli atti amministrativi. Regole e suggerimenti*, il più recente manuale sull'argomento, risultato di un progetto di ricerca promosso dall'Istituto di Teorie e Tecniche dell'Informazione Giuridica (ITTIG) del CNR e l'Accademia della Crusca; la *Guida* è disponibile in rete alla pagina web: <http://www.pacto.it/content/view/416/48/>; il *Manuale di scrittura amministrativa*, a cura di Franceschini e Gigli (2003) e la *Guida alla scrittura istituzionale* di Cortelazzo e Pellegrino (2003). Di quest'ultima è stata offerta una trattazione sintetica, espressa in forma di «trenta regole» da intendersi come suggerimenti operativi per una comunicazione istituzionale più chiara ed efficace. Le regole sono consultabili alla pagina web curata dal Dipartimento di Linguistica dell'Università degli studi di Padova: <http://www.maldura.unipd.it/buro/trentaregole.html>.

[4] Per una ricostruzione delle tappe più significative del processo della semplificazione in sede legislativa, si vedano tra gli altri: Nelli (2009: 901-914); Piemontese (1999: 270-71); Fortis, cit., pp. 89-97.

amministrativo, infatti, presenta per sua natura delle caratteristiche di complessità rispetto alla lingua comune, che non sempre possono essere semplificate e che derivano dalla complessità stessa dell'attività della pubblica amministrazione, cfr. Fioritto (1997: 69). È il caso, ad esempio, di una certa specializzazione lessicale dovuta non tanto alla presenza di un lessico burocratico intrinseco, quanto alla necessità di affrontare materie diverse, che spaziano dalla sanità, alla scuola, all'edilizia, e che richiedono l'uso di tecnicismi di dominio non facilmente traducibili nel linguaggio ordinario. Al contrario – e sempre limitandosi alla dimensione del lessico – l'uso di tecnicismi collaterali, il ricorso a termini arcaici, l'abuso di derivati nominali e nominalizzazioni, caratterizzano quello stile comunicativo oscuro, farraginoso e poco comprensibile ancora diffuso in molte amministrazioni italiane, per cui Italo Calvino in un noto articolo del 1965 coniò il termine di «antilingua».

Obiettivo di questo contributo è mostrare come una metodologia di monitoraggio del profilo linguistico di un testo fondata sull'uso di tecnologie avanzate per il Trattamento Automatico del Linguaggio (TAL)[5] possa contribuire a raffinare la definizione di complessità linguistica rispetto al genere testuale indagato, permettendo di discriminare in maniera automatica le caratteristiche di complessità "necessaria" da quelli che invece appaiono come inutili artifici del "burocratese".

A questo scopo è stato selezionato un 'corpus parallelo monolingue' di documenti amministrativi italiani, ovvero internamente ripartito in due sotto-corpora: il primo costituito dai testi nella loro versione originale, il secondo dalle relative riscritture, semplificate da esperti linguisti. I testi di entrambe le collezioni sono stati arricchiti con informazione linguistica a diversi livelli di descrizione formale (in particolare, lessico, morfo-sintassi e sintassi), grazie ad una piattaforma di strumenti di annotazione automatica del testo allo stato dell'arte per la lingua italiana. L'annotazione ha costituito la premessa all'applicazione della metodologia di monitoraggio linguistico, nella quale i corpora così trattati sono stati analizzati in una duplice ottica contrastiva, tanto rispetto alla distinzione interna al corpus amministrativo (testi originali vs semplificati), tanto rispetto ad altri corpora ugualmente annotati, selezionati come rappresentativi di altri generi e varietà testuali, andando a monitorare - per ognuno - la distribuzione di una serie di caratteristiche linguistiche estratte automaticamente dal testo e identificative di tendenze del lessico, della struttura morfo-sintattica e di quella sintattica; l'assunzione seguita, mutuata dagli studi classici di tipo corpus-based sulla *register variation*, è infatti quella che «linguistic features from all levels function together as underlying dimensions of variation, with each dimension defining a different set of linguistic relations among registers», cfr. Biber (1993: 219). Nella scelta delle caratteristiche oggetto di monitoraggio, un'attenzione particolare è stata dedicata alla valutazione di quei parametri linguistici già risultati predittivi del livello di leggibilità di testi di ambito giornalistico, secondo

[5] [cfr. Montemagni (2013), per un approfondimento della metodologia e i riferimenti lì riportati per la sua applicazione ad alcuni casi di studio.]

l'approccio sottostante a READ-IT[6], che costituisce ad oggi il primo strumento "avanzato" per la valutazione automatica della leggibilità dei testi disponibile per la lingua italiana.

La valutazione automatica della leggibilità di un testo rappresenta infatti una delle principali realizzazioni applicative della metodologia di monitoraggio fondata su caratteristiche rintracciabili dall'annotazione automatica multi-livello del testo ed è particolarmente rilevante rispetto alla tematica affrontata da questo articolo. È la stessa *Guida alla Redazione degli Atti Amministrativi. Regole e suggerimenti*, (cfr. nota 3), ad invitare il redattore dell'atto amministrativo a «costruire le frasi tenendo conto dei requisiti di leggibilità secondo gli indici correnti» (ivi, p.16). Tuttavia, gli indici di leggibilità tradizionali, quali Gulpease per la lingua italiana (Piemontese e Lucisano, 1988), fanno affidamento unicamente a caratteristiche formali del testo, tipicamente la lunghezza della frase e della parola, il cui potere predittivo rispetto alla difficoltà di lettura posta da uno scritto è molto limitato, soprattutto quando si affrontano testi appartenenti al dominio dei linguaggi settoriali, quale la prosa burocratico-amministrativa[7]. Al contrario, gli indici di leggibilità di nuova generazione come READ-IT, sfruttando la crescente affidabilità delle analisi linguistiche generate dagli strumenti di TAL, sono in grado di intercettare parametri di complessità linguistica molto più raffinati, definiti tanto sulla base della letteratura linguistica e psicolinguistica, quanto sull'evidenza emersa dal monitoraggio di corpora accuratamente selezionati come rappresentativi di generi e varietà di lingua diversi.

Il presente contributo si articola come segue: il paragrafo successivo (§ 2) introduce brevemente alcune nozioni sulle tecnologie linguistico-computazionali utilizzate per l'annotazione automatica dei corpora analizzati, la cui descrizione è contenuta nel paragrafo 3; segue la presentazione delle caratteristiche linguistiche monitorate comparativamente nei corpora (§ 4); il paragrafo 5 illustra le distribuzioni quantitative più significative riportate dai corpora annotati rispetto alle caratteristiche linguistiche selezionate.

2. Le tecnologie linguistico-computazionali per l'annotazione automatica del testo

Come anticipato nell'introduzione, il prerequisito alla metodologia di monitoraggio è l'annotazione linguistica del testo effettuata con strumenti di Trattamento Automatico del Linguaggio. L'annotazione automatica del testo permette di identificare la struttura linguistica sottostante al testo e di renderla progressivamente esplicita. Il processo di annotazione avviene tipicamente in maniera incrementale ed è realizzato da una serie di moduli distinti che, operando in successione,

[6] Dell'Orletta / Montemagni / Venturi (2011a).

[7] Nella definizione di Sobrero (1993: 237) il linguaggio amministrativo è ritenuto una «lingua settoriale non specialistica».

generano analisi linguistiche progressivamente più complesse per il tipo di informazione estratta dal testo (Montemagni, 2013). Ai livelli più superficiali di analisi, il testo viene distinto in periodi (operazione di "sentence splitting") e "tokenizzato", ossia segmentato in parole ortografiche (o *tokens*); seguono l'analisi morfo-sintattica e la lemmatizzazione del testo precedentemente *tokenizzato* e infine l'analisi della struttura sintattica della frase in termini di relazioni di dipendenza.

Un esempio del risultato del processo incrementale di annotazione linguistica di un periodo, estratto dal corpus burocratico indagato, è visibile in Tabella 1:

Si fa presente che le mendaci dichiarazioni in atti pubblici e l'occupazione di immobili dichiarati inabitabili sono sanzionate penalmente.

		Lemmatizzazione	**Annotazione morfo-sintattica**			**Annotazione sintattica**	
Id	*Forma*	*Lemma*	*CPos*	*FPos*	*Tratti morfologici*	*Testa Sintattica*	*Tipo di relazione*
1	Si	si	P	PC	num=n|per=3|gen=n	2	clit
2	fa	fare	V	V	num=s|per=3|mod=i |ten=p	0	ROOT
3	presente	presente	A	A	num=s|gen=n	2	pred
4	che	che	C	CS	--	2	arg
5	le	il	R	RD	num=p|gen=f	7	det
6	mendaci	mendace	A	A	num=p|gen=n	7	mod
7	dichiarazioni	dichiarazione	S	S	num=p|gen=f	19	subj_pass
8	in	in	E	E	_	7	comp
9	atti	atto	S	S	num=p|gen=m	8	prep
10	pubblici	pubblico	A	A	num=p|gen=m	9	mod
11	e	e	C	CC	--	7	con
12	l'	il	R	RD	num=s|gen=n	13	det
13	occupazione	occupazione	S	S	num=s|gen=f	19	subj_pass
14	di	di	E	E	--	13	comp
15	immobili	immobile	S	S	num=p|gen=m	14	prep
16	dichiarati	dichiarato	A	A	num=p|gen=m	15	mod
17	inabitabili	inabitabile	A	A	num=p|gen=n	15	mod
18	sono	essere	V	VA	num=p|per=3|mod=i|ten=p	19	aux
19	sanzionate	sanzionare	V	V	num=p|mod=p|gen=f	4	sub
20	penalmente	penalmente	B	B	--	19	mod
21	.	.	F	FS	--	2	punc

Tabella 1 - Un esempio di annotazione linguistica.

La tabella si interpreta come segue: il risultato della fase di tokenizzazione è visibile nella seconda colonna ('forma'), suddivisa in tante righe quante sono le occorrenze di forma di parola (tokens) riconosciute nel testo; a ciascuna è attribuito un numero progressivo riportato nella prima colonna ('id'). Nella fase di annotazione morfo-sintattica, l'informazione relativa a ciascun token viene

arricchita con l'attribuzione della categoria morfo-sintattica che la parola ha nel contesto specifico (colonne CPos e FPos)[8], insieme ad eventuali specificazioni morfologiche associate a determinate categorie (es. tratti flessivi come persona, genere, numero, ecc.) , (colonna 'tratti morfologici'), e il lemma corrispondente (colonna 'lemma'). Ad esempio, la forma 'dichiarazioni' (Id=7) viene ricondotta al relativo lemma 'dichiarazione', che è annotato con la categoria sostantivo (S) e viene inoltre specificato che si tratta di una forma plurale (num=p) e femminile (gen=f).

Le ultime due colonne riportano il risultato del livello di annotazione sintattica, che fornisce una descrizione della frase in termini di relazioni binarie di dipendenza tra parole (tipicamente, relazioni binarie asimmetriche tra una testa e un dipendente, come "soggetto", "oggetto diretto", "modificatore", etc.). In base a questa annotazione, la colonna 7 ('testa sintattica') contiene per ogni parola l'identificatore univoco della forma che costituisce la testa da cui dipende (0 per il verbo della proposizione principale, assunto come radice dell'albero sintattico), mentre l'ultima colonna ('tipo di relazione') specifica il tipo di dipendenza. Questa formalizzazione rende possibile ricavare, ad esempio, che il sostantivo 'dichiarazioni' (id=7) è il soggetto del verbo 'sanzionate' della subordinata passiva (id=19), il quale costituisce la testa della relazione.

Il risultato dell'annotazione può inoltre essere graficamente visualizzato, come mostra la Figura 1: in essa, la struttura sintattica della frase annotata è rappresentata come una serie di nodi lessicali (i singoli tokens), messi in collegamento da archi di dipendenza marcati con il nome del tipo di relazione di dipendenza che lega la testa al dipendente.

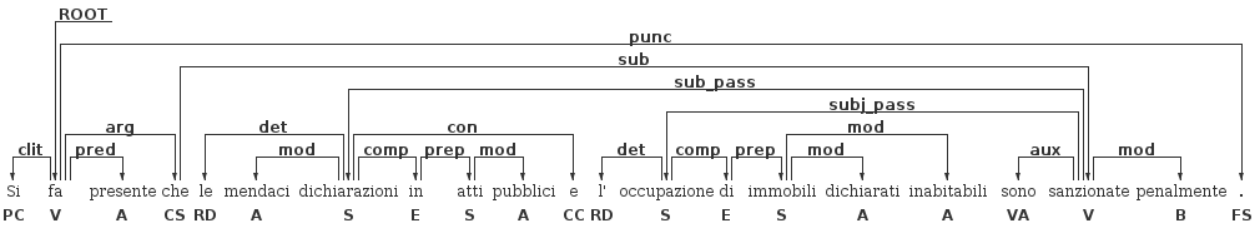

Figura 1 - Un esempio di rappresentazione grafica dell'annotazione sintattica a dipendenze.

Nell'ambito del presente studio, l'annotazione linguistica è stata condotta con gli strumenti software integrati nella piattoforma *LinguA* (*Linguistic Annotation pipeline*), una catena di strumenti statistici di Trattamento Automatico del Linguaggio sviluppati congiuntamente dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa[9]. Tali strumenti rappresentano lo "stato dell'arte" per la lingua italiana, in quanto risultati come i più

[8] Per ogni token viene riconosciuta la categoria morfo-sintattica generale (CPos) e eventuali sottocategorie (FPos). Ad esempio, alla forma (token) 'le' viene associata la categoria articolo (R) e viene ulteriormente specificato che si tratta di un articolo determinativo (RD). Allo stesso modo, il token '.' viene annotato come un segno di punteggiatura (F) di fine periodo (FS).

[9] Una demo di *LinguA* è disponibile alla pagina: < http://linguistic-annotation-tool.italianlp.it/>

precisi e affidabili nell'ambito della campagna di valutazione di strumenti per l'analisi automatica dell'italiano, EVALITA-2009. In particolare, il componente di analisi morfo-sintattica, descritto in Dell'Orletta (2009), ha ottenuto un'accuratezza del 96,34%[10] nell'identificazione simultanea della categoria grammaticale e dei tratti morfologici associati. Per quanto riguarda l'analisi sintattica a dipendenze, il modulo deputato è il parser *DeSR* (Attardi et al., 2009) che ha raggiunto livelli di LAS[11] e UAS[12] pari rispettivamente a 83,38% e 87,71%, in linea con lo state dell'arte rispetto a questo compito.

3. Un caso di studio: monitoraggio di un 'corpus parallelo monolingue' di testi amministrativi

Il corpus oggetto di monitoraggio linguistico è costituito da 89 testi amministrativi "allineati", ovvero disponibili nella loro versione originale e nella relativa versione semplificata, prodotta da linguisti esperti nelle tematiche della semplificazione del linguaggio amministrativo[13]. Si tratta di documenti amministrativi che, pur appartenendo a tipologie distinte per grado di formalità (es. lettere al cittadino – nella forma di concessioni, nulla osta, autorizzazioni ecc… – modulistica, bandi di concorso) sono accomunati da quella che Viale (2008: 106-109) identifica come prospettiva orientata al destinatario esterno all'amministrazione; sono infatti questi i testi su cui

[10] L'accuratezza è calcolata come il rapporto tra il numero di tokens classificati correttamente e il numero totale di tokens analizzati.

[11] 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.

[12] 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.

[13] La maggior parte di questi testi deriva da esperienze accademiche e corsi di aggiornamento professionale indirizzati al personale amministrativo, organizzati dal Dipartimento di Linguistica dell'Università degli studi di Padova, dove da tempo è attiva una linea di ricerca sul tema della semplificazione del linguaggio amministrativo (il lettore interessato può visitare la pagina di riferimento, citata in nota 2: <http://www.maldura.unipd.it/buro/index.html>, che contiene un'ampia bibliografia dedicata e una panoramica delle attività svolte dal Dipartimento).
Nel dettaglio, i testi raccolti (e le relative versioni rielaborate) provengono dalle seguenti fonti: 22 testi fanno parte del «Corpus TACS (testi amministrativi chiari e semplici)», la cui versione elettronica è disponibile alla pagina: <http://www.maldura.unipd.it/buro/tacs.html>; inoltre è stata documentata nella pubblicazione cartacea interna all'amministrazione comunale che ha collaborato con il Dipartimento: *Semplificazione del linguaggio amministrativo. Esempi di scrittura per le comunicazioni ai cittadini*, a cura di Michele A. Cortelazzo, (1999); 22 testi sono stati tratti dal volume *Il Comune scrive chiaro. Come semplificare le comunicazioni al cittadino*, a cura di Michele A. Cortelazzo, (2005); 35 testi derivano dal progetto di ricerca «Comuniversità», promosso dal Consorzio Interuniversitario sulla Formazione (CO.IN.FO) che ha coinvolto il personale di alcune università italiane nella definizione di un «repertorio sistematico di testi standard universitari, redatti con tecniche di scrittura chiara, semplice ed efficace», cfr. il documento programmatico disponibile all'indirizzo Internet: <http://www.coinfo.net/documenti/RicercheIntervento/Allegati/Ricerca_Intervento_ComUniversit%C3%A0.pdf>; 9 testi provengono dalla versione elettronica di alcuni esercizi di riscrittura contenuti nei più noti manuali sulla semplificazione di scrittura amministrativa, reperita in rete; infine, è stato considerato come unico testo la proposta di riformulazione del documento ufficiale delle *Istruzioni per le operazioni degli uffici elettorali di sezione*, pubblicate dal Ministero dell'Interno nel 2009, disponibile online all'indirizzo <http://www.maldura.unipd.it/buro/index.html>, seguendo i puntatori "istruzioni elettorali" / "traduzione in italiano corrente"; anch'essa è stata poi pubblicata nel volume *Le istruzioni per le operazioni degli uffici elettorali di sezione tradotte in italiano*, Cortelazzo / Di Benedetto / Viale (2008).

maggiormente si sono concentrate le attività dei linguisti impegnati nella semplificazione del linguaggio amministrativo, con l'obiettivo di creare dei modelli di "buona" scrittura da poter essere riutilizzati internamente all'amministrazione, improntati ad uno stile comunicativo chiaro ed efficace.

La scelta di raccogliere un corpus così costituito è funzionale agli scopi che hanno orientato la presente indagine di monitoraggio linguistico. Se si assume che il corpus dei testi originali e quello delle relative riscritture rappresentino due possibili "modelli" di scrittura burocratica – rispettivamente "complesso" e "semplice" – la domanda che ci si è posti è in che modo sia possibile rintracciare, esplorando l'output dell'annotazione automatica multi-livello, non solo le differenze ma anche le eventuali similarità tra le due varietà indagate. Mentre le differenze permetterebbero di svelare, traducendoli in una metrica computazionale, i tratti tipici del "burocratese" - dunque quelle strutture linguistiche "inutilmente" complesse su cui il revisore ha agito per rendere il testo originale più comprensibile - le similarità assumerebbero lo status di marcatori di questo genere testuale, soprattutto laddove evidenzino tendenze diverse da quelle specifiche di altri generi e varietà linguistiche. A questo scopo, il corpus amministrativo è stato comparato non solo internamente, ma anche rispetto ad altri corpora, assunti come rappresentativi di altrettanti generi testuali e utilizzati in funzione di «monitor corpus» secondo i presupposti della metodologia di monitoraggio comparativo descritta in Montemagni (2013). I generi qui considerati sono cinque: prosa letteraria, linguaggio giornalistico, materiali didattici, linguaggio scientifico, linguaggio legislativo, ciascuno dei quali - a sua volta - internamente suddiviso in due sotto-corpora, considerati esemplificativi di una varietà di lingua, rispettivamente, "semplice" e "complessa". La distinzione nei due registri è stata definita sulla base del lettore di riferimento, per quanto riguarda i primi quattro corpora[14] - e dei risultati di uno studio di monitoraggio linguistico comparativo, per il corpus dei testi di ambito legislativo[15]. Di particolare interesse nel valutare le similarità e le differenze tra le due varietà di testi amministrativi oggetto di indagine sarà il confronto con i due corpora di ambito giornalistico, Due Parole[16] e La Repubblica: sono questi, infatti, i corpora, esemplificativi della lingua comune, sui quali è stato addestrato l'indice READ-IT per definire i due possibili poli di leggibilità, rispettivamente "semplice" e "complesso", verso cui un testo scritto può

[14] Maggiori dettagli su questi corpora, e la rispettiva distinzione in varietà semplice vs complessa, sono contenuti in Dell'Orletta / Montemagni / Venturi (2013).

[15] Il corpus scelto come rappresentativo del linguaggio legislativo è tratto da una collezione più ampia di testi di ambito giuridico, raccolta e descritta in Venturi (in c.s), in cui l'autrice ha monitorato il profilo linguistico di tipologie diverse di testi giuridici italiani da una prospettiva linguistico-computazionale. Tra i risultati del suo studio è emerso che il testo della Costituzione italiana riporta valori più simili ai testi giornalistici classificati come testi di "facile lettura" rispetto ad una serie di caratteristiche estratte dall'output dell'annotazione linguistica automatica ed esemplificative del livello di leggibilità.

[16] *Due Parole* (Piemontese, 1996) è un periodico di "facile lettura" redatto con criteri di scrittura controllata e indirizzato ad un pubblico di lettori con lievi deficit cognitivi o un basso livello di scolarizzazione.

avvicinarsi. Pertanto, se si può ragionevolmente ipotizzare che un documento amministrativo redatto nel più tradizionale stile "burocratese" esibisca un profilo di leggibilità ancora più complesso di quello dei testi giornalistici di "difficile lettura", meno scontato è prevedere in che modo la semplificazione di questi testi, che pur rimangono di ambito settoriale, tenda ad influenzarne il profilo di leggibilità.

La tabella II illustra i corpora utilizzati in questo studio; per ogni genere, all'interno della colonna **corpus**, sono riportati i due sotto-corpora, rappresentativi, il primo, della varietà "semplice", il secondo, di quella "complessa".

Genere	**Corpus**	**Etichetta di riferimento**	**Numero documenti**	**Numero tokens**
Letteratura	Narrativa per bambini (Marconi et al., 1994)	Narr_child	101	19,370
	Narrativa per adulti (Marinelli et al., 2003)	Narr_adult	327	471,421
			Tot: 428	**Tot: 490,791**
Giornalismo	Due Parole (Piemontese, 1996)	Due Par	322	73,314
	La Repubblica (Marinelli et al., 2003)	La Rep	321	232,908
			Tot: 643	**Tot: 306,222**
Materiali didattici	Materiali didattici per la scuola primaria (Dell'Orletta et al. , 2011b)	Edu_child	127	48,036
	Materiali didattici per la scuola secondaria superiore (Dell'Orletta et al., 2011 b)	Edu_adult	70	48,103
			Tot: 197	**Tot: 96,139**
Prosa scientifica	Articoli italiani di *Wikipedia*, sezione "Ecologia e Ambiente"	Wiki	293	205,071
	Articoli scientifici di vari settori scientifici (es. cambiamenti climatici, linguistica)	Scient_art	84	471,969
			Tot: 377	**Tot: 677,040**
Linguaggio legislativo	Atti legislativi in materia ambientale	Norm_acts	553	1,309,866
	Costituzione Italiana (1947)	It_Const	1	10,487
			Tot: 554	**Tot: 1,320,353**
Linguaggio burocratico-amministrativo	Testi burocratici originali	Bur_orig	89	61,208
	Testi burocratici semplificati	Bur_simpl	89	43,780
			Tot: 178	**Tot: 104,988**

Tabella II - I corpora analizzati

4. Il monitoraggio linguistico: caratteristiche selezionate

Come descritto nella sezione 2, l'annotazione multi-livello del testo consente un'esplorazione ad ampio raggio della struttura linguistica sottostante al testo stesso; ovviamente, la scelta delle caratteristiche rintracciabili nel testo annotato, che possono contribuire ad un'indagine di monitoraggio, varia in relazione agli scopi del fenomeno che si intende indagare. Nel caso dello studio qui discusso, l'interesse era duplice: i) identificare la presenza di "marcatori" del linguaggio burocratico; ii) individuare gli interventi di semplificazione apportati dal revisore al testo burocratico originario.

Alla luce di queste domande di ricerca, la selezione dei tratti monitorati è stata motivata da due ordini di fattori:

- il loro "potere" predittivo nel formalizzare parametri di complessità del testo a diversi livelli di descrizione linguistica, così come già emerso nell'ambito di studi condotti con metodi e strumenti di Trattamento Automatico del Linguaggio volti alla valutazione automatica della leggibilità di testi nella lingua comune [cfr., il già citato READ-IT, Dell'Orletta et al. (2011a)], all'identificazione di generi e varietà testuali, cfr. Dell'Orletta et al. (2013), e al monitoraggio delle competenze scolastiche, cfr. Dell'Orletta et al. (2011b).
- la capacità di esplicitare le indicazioni qualitative emerse in letteratura tanto sulle peculiarità del linguaggio burocratico italiano, quanto sulle linee guida sulla semplificazione.

Partendo da questi presupposti, la tabella III sintetizza le principali caratteristiche linguistiche che hanno ricevuto conferma dall'analisi di monitoraggio della loro implicazione nelle domande di ricerca iniziali; come si nota, esse sono state organizzate in base alle fasi di annotazione linguistica progressivamente più raffinate da cui derivano, dunque: divisione in frasi e tokenizzazione, lemmatizzazione e annotazione morfo-sintattica, annotazione sintattica a dipendenze. Anche nella loro discussione si manterrà tale articolazione, fermo restando che si tratta di un criterio descrittivo che, pur permettendo di discriminare il ruolo, l'affidabilità e il livello di profondità esplicativa derivante da ciascun livello, non intende oscurare la consapevolezza che i fenomeni linguistici, soprattutto quando coinvolgono proprietà sintattiche, possono essere compresi adeguatamente solo se si considera l'interazione tra i diversi livelli, come nel caso della subordinazione.

Tipo di caratteristica	**Livello di annotazione linguistica**	**Tratto monitorato**
Superficiale	Divisione del testo in frasi/ Tokenizzazione	Lunghezza media della frase
Lessicale	Lemmatizzazione/ Analisi morfo-	• Densità Lessicale

	sintattica	• Percentuale di parole appartenenti al Vocabolario di Base del *Grande Dizionario italiano dell'uso* (De Mauro, 2000) e loro distribuzione nei repertori d'uso (fondamentale, alto uso, alta disponibilità) • Rapporto Type/Token
Morfo-sintattica	Analisi morfo-sintattica	• Distribuzione delle categorie morfo-sintattiche (lessicali e funzionali) • Rapporto Nomi/Verbi
Sintattica	Analisi sintattica a dipendenze	Caratteristiche basate sulla struttura dell'albero sintattico: • Profondità media dell'intero albero sintattico; • Lunghezza media delle relazioni di dipendenza sintattica (link sintattici); • Lunghezza media del link sintattico più lungo;
		Caratteristiche basate sul tipo di dipendenza: • Distribuzione dei tipi di dipendenza sintattica
		Caratteristiche proprie della subordinazione (derivanti dalla struttura dell'albero sintattico e dal tipo di dipendenza): • Rapporto principali/subordinate; • Profondità media delle clausole subordinate.

Tabella III – Le caratteristiche monitorate

5. Analisi delle distribuzioni quantitative: similarità e differenze all'interno del corpus dei testi amministrativi

Partiamo innanzitutto dall'analisi delle similarità tra le due varietà di testi amministrativi, ossia di quelle caratteristiche che distinguono il corpus burocratico nella sua globalità, e che dunque si possono interpretare come non implicate nella semplificazione.

Una prima indicazione in questa direzione emerge comparando, nei vari corpora di confronto considerati[17], le distribuzioni riportate da un sotto-insieme di categorie morfo-sintattiche primarie (cfr. Tabella IV); è un dato – questo – in linea con la letteratura di *register variation* secondo cui

[17] Nella maggioranza dei casi si considera il dato medio riportato da ciascun corpus di confronto, dunque collassando la distinzione interna nelle due varietà "semplice" e "complessa": di questa si darà conto solo se significativa ai fini della discussione.

«systematic differences in the relative use of core linguistic features provide the primary distinguishing characteristics among registers», Biber (1995: 36).

	Giornalismo		*Materiali didattici*		*Letteratura*		*Prosa scientifica*		*Linguaggio legislativo*		***Linguaggio Burocratico***	
	Due Par	Rep	Edu_ Child	Edu_ Adult	Narr_ Child	Narr_ Adult	Wiki	Scient Art	Norm acts	It_ Cost	Bur_ simp	Bur_ orig
Aggettivi	5,92	6,40	6,61	8,81	5,93	6,38	8,71	8,99	8,16	8,40	5,72	5,98
	6,16		7,76		6,15		8,85		8,28		5,85	
Avverbi	3,52	4,83	5,72	5,86	6,43	5,38	4,15	3,81	1,43	2,29	1,91	2,14
	4,18		5,79		5,90		3,98		1,86		2,03	
Congiunzioni	3,69	3,61	4,37	5,01	4,83	4,36	3,75	3,44	4,13	5,29	3,09	2,75
	3,65		4,69		4,60		3,60		4,71		2,92	
Preposizioni	15,28	16,41	13,89	15,25	12,09	12,34	16,46	17,63	20,64	18,63	18,07	20,42
	15,85		14,57		12,21		17,05		19,64		19,25	
Pronomi	2,32	3,76	5,11	5,61	6,88	6,13	3,18	2,88	2,09	2,40	2,74	3,18
	3,04		5,36		6,51		3,03		2,25		2,96	
Nomi	29,30	27,19	23,17	22,99	21,96	24,08	28,46	28,41	30,27	30,16	30,52	29,82
	28,24		23,08		23,02		28,44		30,21		30,17	
Verbi	13,66	12,89	15,05	12,67	15,83	14,96	10,65	10,60	8,59	11,50	11,05	11,12
	13,28		13,86		15,39		10,62		10,04		11,09	

Tabella IV – Distribuzione delle categorie morfo-sintattiche primarie nei corpora considerati

In particolare, focalizzando l'attenzione sulle parti del discorso di classe aperta, si osserva la percentuale più bassa sia di aggettivi sia di avverbi, tanto nel corpus burocratico considerato globalmente (aggettivi: 5,85; avverbi: 2,03), tanto nelle due varietà interne (Bur_simp: aggettivi: 5,72; avverbi: 1,91 / Bur_orig: aggettivi: 5,98; avverbi: 2,14). Diametralmente opposta è la distribuzione dei nomi - una delle categorie complessivamente più rappresentate in tutti i corpora (infra) - che, sia nei testi originali sia nei semplificati, si approssima al dato medio (cfr. Bur_simp: 30,52; Bur_orig: 29,82; media: 30,17). Rispetto alla categoria dei verbi, si rileva invece una frequenza media molto bassa, e simile nelle due varietà, (Bur_simp: 11,05; Bur_orig: 11,12; media: 11,09), più vicina ai testi legislativi e scientifici rispetto a quelli giornalistici, didattici e di narrativa, che registrano i valori più elevati. Spostando l'attenzione alle categorie funzionali, un ulteriore tratto distintivo dei testi burocratici è dato dalla frequenza, molto alta, delle preposizioni. Anche questo è un dato atteso alla luce della percentuale elevata di nomi, e conferma quanto già osservato da Biber (1988) sull'esistenza di associazioni sistematiche nella distribuzione di nomi, preposizioni e aggettivi attributivi che connotano i testi dalla spiccata funzione informativa. A tal proposito, è interessante notare che se il dato empirico emerso in questo studio avvalora la correlazione tra nomi e preposizioni, lo stesso non si rileva per quanto riguarda gli aggettivi che, come in precedenza osservato, sono la categoria meno rappresentata; pertanto, in questo caso, una peculiarità del genere burocratico sembra essere instanziata da un pattern di associazione negativo.

Sebbene l'informazione sulla distribuzione delle categorie morfo-sintattiche in isolamento fornisca alcuni dati preliminari allo studio della variazione linguistica, è tuttavia più interessante considerare come alcune parti del discorso si rapportino l'una all'altra; in proposito, si è scelto di presentare il dato relativo alla proporzione tra nomi e verbi (tabella V). Nella letteratura sulla varietà linguistica, tale misura è associata generalmente alla distinzione sull'asse diamesico, con il parlato che registra valori più bassi nell'andamento di questo parametro rispetto allo scritto, dove invece i nomi predominano sui verbi, cfr. Biber (1995); Voghera (2005). Questa distinzione si è dimostrata significativa per cogliere differenze e similarità tra generi e varietà testuali anche all'interno della stessa dimensione diamesica; focalizzandosi su corpora di produzione scritta annotati automaticamente, Montemagni (2013) ha osservato che la proporzione tra nomi e verbi varia in maniera tale da discriminare testi informativi, da un lato, e testi di scrittura creativa, dall'altro, che riportano valori quasi comparabili a quelli del parlato. Ancora, nell'ambito dei soli testi con marcata funzione informativa, il rapporto nomi/verbi è più basso nei testi giornalistici rispetto a quelli accademici, cfr. Biber (1993).

La tendenza verso una correlazione tra la predominanza dei nomi e il carattere spiccatamente informativo del testo risulta confermata anche dai dati qui ottenuti: il corpus burocratico riporta infatti un rapporto nomi/verbi molto alto (pari a 2,72), simile a quello della prosa scientifica e inferiore di soli 0,35 punti alla media dei testi di ambito legislativo, tipicamente caratterizzati da valori elevati[18]. È interessante osservare come non sia apprezzabile alcuna variazione rilevante internamente al corpus burocratico; al contrario, la proporzione dei nomi sui verbi in Bur_simp eccede addirittura, sebbene di poco più di 0,08 punti, quella attestata in Bur_orig.

Genere	**Corpus**	Rapporto Nomi/Verbi	
Giornalismo	Due Par	2,14:1	2,13:1
	La Rep	2,11:1	
Materiali didattici	Edu_child	1,55:1	1,67:1
	Edu_adult	1,81:1	
Letteratura	Narr_child	1,39:1	1,50:1
	Narr_adult	1,61:1	
Prosa scientifica	Wiki	2,67:1	2,68:1
	Scient_art	2,69:1	
Linguaggio legislativo	It_const	2,62:1	3,07:1
	Norm_acts	3,52:1	
Linguaggio burocratico	Bur_simp	2,76:1	**2,72:1**
	Bur_orig	2,68:1	

Tabella V – Proporzione tra nomi e verbi nei corpora analizzati

[18] Cfr. i dati riportati da Venturi, in c.s.

L'uso così massiccio di nomi oltre ad essere una "spia" della funzione altamente informativa del testo burocratico, riflette tanto il fatto che i documenti amministrativi contengono riferimenti a persone, uffici, istituzioni ecc..., spesso necessari per la completezza delle informazioni trasmesse, quanto il grado di astrazione che caratterizza questi testi, il cui focus sono prevalentemente soggetti inanimati (leggi, regolamenti, divieti, permessi, pagamenti e così via), piuttosto che soggetti animati e concreti, Raso (2005: 112-113).

Procedendo con l'analisi dell'output estratto congiuntamente dal testo lemmatizzato e morfo-sintatticamente annotato, è stato possibile investigare il contributo alla definizione di marcatori del genere burocratico apportato da alcune caratteristiche che catturano, in modo diverso, aspetti di complessità lessicale. È il caso della "type/token ratio", la misura che mette a rapporto il numero delle occorrenze di unità del vocabolario di un testo (al denominatore) con il numero di parole tipo (al numeratore): il suo valore oscilla nell'intervallo tra 0 e 1. Si tratta di un parametro che formalizza la varietà lessicale di un testo e che tende ad aumentare all'aumentare della complessità di un testo. Come dimostra la tabella VI, la type/token ratio[19] nel corpus burocratico medio (pari a 0,69) è più bassa di quella attestata all'interno di generi più descrittivi, quali la letteratura e i materiali didattici; al contrario, si avvicina a quella riportata dalla prosa scientifica (0,78) e al corpus di testi legislativi (0,46), il cui valore si approssima al limite negativo che questo parametro può assumere. Ancora una volta, la differenza interna alle due varietà di testi amministrativi è statisticamente irrilevante (Bur_simp: 0,68 / Bur_orig: 0,70). Questo dato fornisce una conferma empirica della stretta parentela tra la lingua della burocrazia e i linguaggi tecnico-scientifici; come osserva Gotti (2005: 33) le lingue speciali (o linguaggi specialistici) sono caratterizzate dalla proprietà della 'monoreferenzialità', ovvero da un elevato formalismo nella designazione semantica, che limita l'uso di sinonimi o perifrasi per indicare lo stesso referente, pena il rischio di distorcere il messaggio originale o sollevare possibili ambiguità. Pertanto, il valore empirico qui rilevato amplia le possibili interpretazioni di questo parametro, suggerendo che una bassa type/token ratio non solo contraddistingue testi volutamente scritti in un linguaggio più semplice[20] ma traduce anche, quantitativamente, la precisione richiesta dai linguaggi tecnico-scientifici.

Inoltre, l'equiparazione del valore assunto dalla type/token ratio all'interno delle due varietà del corpus burocratico sembra suggerire come, rispetto a questo parametro, i testi originali fossero già

[19] Il valore è calcolato sulle prime 100 parole del testo.

[20] Si consideri in proposito la differenza interna al corpus giornalistico, con *Due Parole* che riporta un valore di type/token ratio pari a 0,66 e *La Repubblica* a 0,86.

mediamente ben formati, e dunque conformi al suggerimento, frequente nei manuali sulla semplificazione della scrittura amministrativa, ad evitare possibili ambiguità nel testo[21].

Genere	Corpus	Type/Token Ratio	
Giornalismo	Due Par	0,66	0,76
	La Rep	0,86	
Materiali didattici	Edu_child	0,80	0,80
	Edu_adult	0,81	
Letteratura	Narr_child	0,81	0,81
	Narr_adult	0,80	
Prosa scientifica	Wiki	0,77	0,78
	Scient_art	0,80	
Linguaggio legislativo	It_const	0,49	0,46
	Norm_acts	0,44	
Linguaggio burocratico	Bur_simp	0,68	**0,69**
	Bur_orig	0,70	

Tabella VI – Valore del rapporto tra parole-unità e totale delle occorrenze nei corpora analizzati

Sempre rimanendo sul piano del lessico, se la type/token ratio è associata alla varietà lessicale di un testo, un parametro che invece ne qualifica la sua ricchezza è la densità lessicale, ottenuta dal rapporto tra parole "contenuto" (o categorie semanticamente "piene") sul totale delle occorrenze di parole. Anche in questo caso, la letteratura sulla leggibilità dimostra che valori superiori di densità lessicale sono associati ad un maggior carico informativo e, dunque, ad una maggiore complessità del testo. Tuttavia, pur limitandoci all'analisi delle produzioni scritte[22], il "potere" esplicativo di questo parametro nel definire una demarcazione tra generi e varietà di lingua più o meni complessi non è così immediato. I testi informativi, infatti, quali quelli scientifici e accademici, tendono ad essere lessicalmente più "densi", come viene confermato anche dai dati di questo studio (cfr. tabella VII), con il corpus di articoli scientifici che ottiene i valori più elevati di densità lessicale (0,577). Altrettanto informativi sono i testi legislativi e quelli amministrativi, che tuttavia riportano valori più bassi, con la distinzione interna al corpus burocratico statisticamente non significativa (Bur_orig: 0,538; Bur_simp: 0,544, t=0,158, p>.05). In questo caso, si può pensare che a far variare il parametro in senso negativo non sia tanto la frequenza, pur elevata, di parole contenuto, quanto l'ampia attestazione di categorie funzionali, quali le preposizioni (cfr. tabella IV).

[21] A pagina 29 della *Guida alla redazione dei testi amministrativi. Regole e suggerimenti*, cit., si legge a questo proposito: «In un atto amministrativo è opportuno evitare l'ambiguità e raggiungere il massimo di esplicitezza: è consigliabile pertanto, anche a costo di numerose ripetizioni, usare sempre lo stesso termine per designare la stessa azione, lo stesso concetto o la stessa persona».

[22] Sul versante della dimensione diamesica, un dato ben attestato è la maggior "leggerezza" lessicale del parlato rispetto allo scritto, cfr. Halliday (1995). Questa tendenza è legata all'influenza dei fattori extra-linguistici nelle produzioni orali e, in particolare, all'immediatezza del flusso conversazionale che impedisce al parlante di pianificare gli enunciati ('utterances') con la stessa accuratezza di quanto avviene nello scritto; ciò si traduce in false partenze, disfluenze, frasi incomplete, un uso superiore di interiezioni e di elementi funzionali del linguaggio.

Genere	Corpus	Densità Lessicale	
Giornalismo	Due Par	0,564	0,564
	La Rep	0,564	
Materiali didattici	Edu_child	0,558	0,557
	Edu_adult	0,556	
Letteratura	Narr_child	0,568	0,573
	Narr_adult	0,578	
Prosa scientifica	Wiki	0,584	0,577
	Scient_art	0,571	
Linguaggio legislativo	It_const	0,555	0,543
	Norm_acts	0,533	
Linguaggio burocratico	Bur_simp	0,544	**0,544**
	Bur_orig	0,538	

Tabella VII – Valore del rapporto tra parole contenuto e totale delle occorrenze di parole (densità lessicale) nei corpora analizzati

Veniamo infine alla trattazione di un ultimo parametro che è possibile rintracciare automaticamente dal testo lemmatizzato e morfo-sintatticamente annotato e che fornisce un'indicazione qualitativamente più raffinata del tipo di vocabolario usato nei testi: la distribuzione del lessico rispetto al **Vocabolario di Base** (VdB), De Mauro (2000). Come si nota dalla tabella VIII, se nel corpus legislativo il VdB è abbondantemente il meno rappresentato (35,60%), la lingua della pubblica amministrazione continua ad apparire come una lingua per "esperti"; la rappresentatività del VdB, tanto nel corpus degli originali (Bur_orig: 58,33%) quanto nelle riscritture (59,29%), è infatti molto più vicina a quella riportata dai testi scientifici e superiore di ben 10 punti percentuali alla frequenza attestata nel corpus giornalistico, anche nella sola varietà più complessa (La Rep: 67,09).

Genere	Corpus	Percentuale di lemmi appartenenti al Vocabolario di Base	
Giornalismo	Due Par	74,58	70,84
	La Rep	67,09	
Materiali didattici	Edu_child	74,57	73,57
	Edu_adult	72,56	
Letteratura	Narr_child	73,95	71,76
	Narr_adult	69,57	
Prosa scientifica	Wiki	60,77	55,44
	Scient_art	50,11	
Linguaggio legislativo	It_const	54,87	35,60
	Norm_acts	16,34	
Linguaggio burocratico	Bur_simp	59,29	**58,81**
	Bur_orig	58,33	

Tabella VIII – Frequenza percentuale di lemmi appartenenti al Vocabolario di Base nei corpora analizzati

Tuttavia, se si considera la ripartizione interna al VdB nei tre repertori d'uso – fondamentale, alto uso, alta disponibilità – a livello del Lessico Fondamentale emerge una distinzione che, sebbene contenuta, è statisticamente significativa (cfr. tabella IX). In Bur_simp, infatti, la frequenza percentuale del Lessico Fondamentale è superiore (67,12%) a Bur_orig (64,94%): tale differenza sembra riflettere il tentativo del redattore del testo semplificato di attingere, quando possibile, a parole più familiari al lettore comune così da favorire la sua comprensione[23].

Genere	**Corpus**	**Percentuale di lemmi appartenenti al Lessico Fondamentale**	
Giornalismo	Due Par	75,06	73,53
	La Rep	72,00	
Materiali didattici	Edu_child	73,02	73,15
	Edu_adult	73,29	
Letteratura	Narr_child	76,84	76,31
	Narr_adult	75,78	
Prosa scientifica	Wiki	68,18	67,04
	Scient_art	65,89	
Linguaggio legislativo	It_const	70,03	66,69
	Norm_acts	63,36	
Linguaggio burocratico	Bur_simp	67,12	**65,63**
	Bur_orig	64,94	

Tabella IX – Frequenza percentuale di lemmi appartenenti al repertorio del Lessico Fondamentale nei corpora analizzati

Concludendo la discussione sui dati del monitoraggio che esplicitano le similarità tra i due profili di testi amministrativi indagati, prendiamo in considerazione alcune caratteristiche di complessità strutturale del testo selezionate dal livello più profondo di analisi automatica disponibile: l'annotazione sintattica a dipendenze. Anche a questo livello, certamente il più ricco e raffinato rispetto al contributo fornito alla ricostruzione del profilo linguistico di un testo, sono emerse alcune similarità significative tra le due versioni. Si tratta probabilmente del dato meno atteso: la letteratura sottolinea infatti come sia soprattutto l'uso di strutture sintattiche infrequenti e poco rappresentate nella lingua comune a caratterizzare negativamente il "burocratese". Sulla base di queste indicazioni si è cercato di verificare, in primo luogo, come fossero distribuite quelle caratteristiche, rintracciabili automaticamente nel testo sintatticamente annotato, che rendono conto dell'uso della subordinazione all'interno del corpus indagato.

[23] Ad esempio, la regola 26 delle *Trenta Regole*, cfr. nota 2, esorta a usare parole comuni in quanto: «Chi legge un testo deve poter capire tutte le parole per riuscire a ricostruirne il senso completo [...] Quanto meno numerose sono in un testo le parole del vocabolario di base, tanto meno numerose sono le persone in grado di comprenderlo». Anche la *Guida alla redazione degli atti amministrativi*, cit., ricalca questo concetto, affermando che «quando è possibile, occorre scegliere le parole del vocabolario di base, preferendole a quelle più rare», p. 25.

La subordinazione, soprattutto di grado elevato, è infatti un marcatore per definizione di maggior complessità strutturale, oltre ad essere riconosciuto come carattere tipico della prosa burocratica. I risultati empirici sembrano però smentire, almeno in parte, l'idea che la semplificazione di un testo di ambito settoriale, quale quello amministrativo, possa essere raggiunta semplicemente diminuendo il numero e l'incassamento di frasi subordinate. Consideriamo, ad esempio, la proporzione tra frasi principali e frasi subordinate, riportata nella quinta colonna della tabella X: come si può osservare, tanto Bur_orig quanto Bur_simp fanno un ampio uso della costruzione ipotattica, superiore a quello attestato in tutti gli altri corpora di confronto (Bur_simp: 0,56; Bur_orig: 0,60). Cosa ancora più interessante, il grado di incassamento gerarchico delle subordinate[24], tipicamente superiore nei testi più complessi, è praticamente invariato nelle due versioni (cfr. 'profondità media delle clausole subordinate', ultima colonna della tabella X).

Genere	**Corpus**	**Clausole Principali**	**Clausole subordinate**	**Rapporto Principali vs Subordinate**		**Profondità media delle clausole subordinate**	
Giornalismo	Due Par	73,55	26,14	0,36	0,42	1,01	1,09
	La Rep	67,33	32,36	0,48		1,17	
Materiali didattici	Edu_child	69,94	28,42	0,41	0,48	0,98	1,08
	Edu_adult	63,55	35,04	0,55		1,17	
Letteratura	Narr_child	68,32	30,69	0,45	0,48	1,20	1,16
	Narr_adult	65,77	33,92	0,52		1,11	
Prosa scientifica	Wiki	72,92	26,74	0,37	0,40	0,90	1,03
	Scient_art	69,12	29,70	0,43		1,17	
Linguaggio legislativo	It_const	86,07	13,93	0,16	0,26	1,03	1,11
	Norm_acts	73,39	26,61	0,36		1,18	
Linguaggio burocratico	Bur_simp	63,63	35,24	0,56	**0,58**	**0,95**	**0,95**
	Bur_orig	61,58	37,29	0,60		**0,96**	

Tabella X – Proporzione tra clausole principali e subordinate e profondità media delle clausole subordinate nei corpora analizzati.

Se da un lato questi dati sono utili per isolare dei "pattern" di complessità sintattica diversamente declinati a seconda della tipologia testuale, come già osservato da Voghera (2001: 69):

> non tutta la subordinazione è uguale: ciò che costituisce un forte elemento di complessità non è la semplice presenza di una subordinata, ma la combinazione tra subordinazione e vari fattori: ordine relativo tra principale e subordinata; grado di incassatura della subordinata; rapporto di corrispondenza tra concatenazione degli eventi e sequenza delle clausole.

[24] Questo dato può essere ricavato dal testo sottoposto ad annotazione sintattica a dipendenze andando a monitorare i sotto-alberi di clausole subordinate ricorsivamente incassate, cfr. Dell'Orletta e Montemagni (2012).

Proprio rispetto ad alcuni di questi parametri, e più precisamente nella loro "manifestazione" inferibile dall'analisi automatica multi-livello del testo, è stato possibile rintracciare tendenze differenti tra il profilo dei testi amministrativi originali e quello delle riscritture.

Una prima marcata distinzione è fornita già dai livelli di annotazione più superficiali[25], che hanno consentito il calcolo della lunghezza media dei periodi in ciascun corpus (tabella XI). Si tratta ovviamente di un parametro molto grezzo rispetto alla caratterizzazione della complessità di un testo, nonostante sia proprio quello più utilizzato per formalizzare il grado di complessità sintattica dagli indici di leggibilità tradizionali. La correlazione positiva tra la lunghezza della frase e la maggior complessità sintattica è ribadita, del resto, anche dalle analisi tradizionali sui caratteri peculiari del "burocratese"[26] e confermata qui dal dato empirico. A fronte di una lunghezza media delle frasi pari a 26,72 parole nei testi originali, le riscritture non superano mediamente le 20 parole, una differenza risultata statisticamente significativa ($p<0.01$ al t-test).

Genere	Corpus	Lunghezza media del periodo	
Giornalismo	Due Par	19,20	22,87
	La Rep	26,54	
Materiali didattici	Edu_child	23,64	27,64
	Edu_adult	31,63	
Letteratura	Narr_child	16,96	17,61
	Narr_adult	18,25	
Prosa scientifica	Wiki	25,80	28,73
	Scient_art	31,65	
Linguaggio legislativo	It_const	16,59	20,79
	Norm_acts	24,99	
Linguaggio burocratico	Bur_simp	**20,00**	23,36
	Bur_orig	**26,72**	

Tabella XI – Lunghezza media del periodo (in parole) nei corpora analizzati

Tuttavia, l'accorciamento della frase non si è necessariamente tradotto in una riduzione della presenza di subordinate, come hanno mostrato i dati della tabella X. Se, come discusso poco sopra, l'equivalenza tra una nozione generale di subordinazione e l'idea di complessità sintattica è forse troppo sommaria, ci siamo chiesti in che modo altre caratteristiche estratte dal testo annotato forniscano degli "indizi" ulteriori per raffinare questa equivalenza. A questo proposito, torniamo all'output dell'annotazione morfo-sintattica e alla distribuzione delle parti del discorso. Al livello di categorie morfo-sintattiche primarie (cfr. tabella IV), Bur_orig e Bur_simp mostravano distribuzioni comparabili. Quando però si restringe il focus dell'analisi alla distinzione di alcune di queste

[25] La distinzione del testo in frasi e la divisione delle frasi in *tokens*, cfr. § 2.

[26] Cfr. Fortis, (2005: 65): «La lunghezza dei periodi costituisce forse la caratteristica dello stile amministrativo che balza più immediatamente agli occhi, e quanto più una frase è lunga, maggiore è la probabilità che sia sintatticamente complessa»; Cortelazzo / Viale (2006: 2118): «l'unione tra complessità lessicale e complessità morfologica genera 'ipertrofia': la lingua burocratica utilizza più parole di quella comune per dire le stesse cose».

categorie nelle loro sotto-classi e ai tratti di specificazione morfologica[27] emergono differenze significative non ricavabili dal dato aggregato. Tali differenze rivelano aspetti interessanti rispetto al tipo di subordinazione usata in questi testi: è il caso delle congiunzioni, della specificazione dei tratti di modo e persona nei verbi e della distribuzione dei pronomi.

Per quanto riguarda la distinzione interna alla categoria delle congiunzioni (tabella XII), si osserva una differenza contenuta, ma comunque statisticamente rilevante, nel rapporto tra congiunzioni coordinanti e subordinanti, con una maggior occorrenza di congiunzioni subordinanti in Bur_simp (0,99) rispetto a Bur_orig (0,77).

Genere	**Corpus**	**Congiunzioni subordinanti**	**Congiunzioni coordinanti**	**Rapporto subordinanti/coordinanti**	
Giornalismo	Due Par	0,67	3,02	0,22	0,79
	La Rep	0,91	2,70	0,34	
Materiali didattici	Edu_child	0,99	3,39	0,29	1,02
	Edu_adult	1,06	3,96	0,27	
Letteratura	Narr_child	1,41	3,43	0,41	1,36
	Narr_adult	1,31	3,05	0,43	
Prosa scientifica	Wiki	0,57	3,18	0,18	0,63
	Scient_art	0,69	2,75	0,25	
Linguaggio legislativo	It_Const	0,72	4,57	0,16	0,59
	Adm_acts	0,45	3,68	0,12	
Linguaggio burocratico	Bur_simp	**0,99**	2,10	0,47	0,88
	Bur_orig	**0,77**	1,98	0,40	

Tabella XII – Proporzione tra congiunzioni coordinanti e subordinanti nei corpora analizzati

Per quanto riguarda invece la frequenza dei modi verbali, consideriamo ad esempio l'occorrenza dei participi. A questo proposito, il primo dato ottenuto dall'annotazione morfo-sintattica rivelava che in Bur_orig i participi sfioravano quasi il 30% (29,79), mentre in Bur_simp la loro occorrenza era inferiore di circa 3 punti % (26,09). Tuttavia, poiché il participio passato in italiano è usato anche per la formazione dei tempi verbali composti, è stato necessario ripulire il dato iniziale, escludendo dall'analisi i participi all'interno dei tempi composti[28]. Il risultato è fornito in tabella XIII ed evidenzia una prevalenza ancora più marcata di participi in funzione verbale nei testi amministrativi originali (31,69) rispetto alle riscritture (26,02).

Genere	**Corpus**	**% Verbi participiali**	
Giornalismo	Due Par	5,74	10,16
	La Rep	14,57	
Materiali didattici	Edu_child	10,25	10,87
	Edu_adult	11,50	

[27] Si tratta del dato riportato nelle colonne 'FPos' e 'tratti morfologici' della tabella I, cfr. § 2.

[28] Anche questo dato è ricavabile dall'annotazione automatica, incrociando l'output dell'annotazione morfo-sintattica e quello dell'analisi sintattica a dipendenze ed escludendo i sotto-alberi di verbi etichettati come participi e dipendenti da una testa verbale di tipo ausiliare o modale.

Letteratura	Narr_child	8,62	9,07
	Narr_adult	9,51	
Prosa scientifica	Wiki	17,94	19,26
	Scient_art	20,59	
Linguaggio legislativo	It_Const	19,71	29,28
	Adm_acts	38,85	
Linguaggio burocratico	Bur_simp	**26,02**	28,81
	Bur_orig	**31,69**	

Tabella XIII – Frequenza percentuale di participi verbali nei corpora considerati

Tanto i dati sulla distribuzione delle congiunzioni, quanto quelli sul participio possono contribuire a rivelare due strategie, diverse per grado di complessità, di organizzazione del discorso ipotattico nei testi amministrativi originali e nelle riscritture, esemplificati dai casi riportati in (I) e (II), tratti dai rispettivi corpora. In particolare, la diversa proporzione di congiunzioni subordinanti sembra suggerire come l'autore dei testi semplificati abbia scelto di ricorrere più frequentemente a marcatori subordinanti espliciti per introdurre una clausola subordinata, strategia ritenuta cognitivamente più semplice perché esplicita i legami logici tra le proposizioni [cfr. la discussione e i riferimenti riportati in Fiorentino, (2007: 11-37)], laddove invece il testo originale si affidava ad altri espedienti, tra cui locuzioni proposizionali complesse (es. ai fini di, allo scopo di ecc..) o subordinate implicite, quali appunto le participiali. L'abuso di frasi implicite, con l'effetto di oscurità e densità informativa che ne deriva, è considerato proprio una cifra stilistica del "burocratese", mutuata dal linguaggio giuridico (e infatti abbondantemente rappresentate nel corpus legislativo qui considerato), e tipicamente sconsigliata dai manuali di semplificazione.
Il confronto ottenuto empiricamente va dunque nella stessa direzione delle osservazioni qualitative.

Esempi:

(I) riscritture:

(a) ***Se**, invece, preferite mantenere il regime del diritto di superficie*, vi ricordiamo che la convenzione preliminare che avete sottoscritto vi obbliga a chiedere al Comune di Schio l'autorizzazione preventiva per qualunque passaggio di proprietà, affitto, cambio societario, ecc.

(b) I proprietari di autoveicoli e i titolari di patente non sono obbligati a cambiare l'indirizzo su libretto di circolazione e patente, ***perché** l'obbligo è previsto solo nel caso di effettivo cambio di abitazione*.

(c) 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.

(II) testi originali:

(a) La medesima circolare ministeriale suggerisce altresì che il Comune, ***allo scopo di** evitare contestazioni* che potrebbero comportare il ritiro dei documenti [...]

(b) La variazione anagrafica in esame non comporta per i proprietari di autoveicoli e per i titolari di patente di guida l'obbligo di fare aggiornare la carta di circolazione e la patente di guida, ***in quanto** tale obbligo è previsto dal Codice della Strada soltanto per i casi di trasferimento effettivo di abitazione.*

(c) Si ricorda che, ***mantenendo*** *il regime del diritto di superficie*, qualunque passaggio di proprietà, affitto, cambio societario, ecc. dovrà essere autorizzato dal Comune di Schio [...]

Sempre rimanendo sui dati relativi al livello più "granulare" di annotazione morfo-sintattica, due ulteriori caratteristiche contribuiscono alla qualificazione delle operazioni di semplificazione manuale apportate al testo amministrativo di partenza: la distribuzione dei tratti flessivi di persona nei verbi e la distribuzione dei pronomi nelle rispettive sotto-classi (personali, dimostrativi, clitici, ecc.).

Il primo aspetto è trattato dalla tabella XIV, dalla quale un dato spicca con particolare evidenza: la differenza nella frequenza percentuale di verbi alla prima persona plurale (sesta colonna della tabella XIV), praticamente assenti in Bur_orig e ampiamente rappresentati in Bur_simp, anche rispetto a tutti gli altri corpora considerati. È invece la terza persona singolare (quinta colonna della tabella XIV) a connotare distintamente Bur_orig. Anche in questo caso il dato quantitativo non è casuale e conferma, piuttosto, la tendenza alla "spersonalizzazione" che caratterizza lo stile burocratico, «presumibilmente dettata dall'intento di conferire al documento un tono ufficiale, autorevole e solenne», Fortis (2005: 69); la terza persona, infatti, accorda tanto con i soggetti tipicamente inanimati che il burocrate sceglie per identificarsi e rivolgersi al cittadino (es. questo ufficio, questa amministrazione ecc.), quanto con l'uso di frasi impersonali. Al contrario, la prima persona plurale è più consona ad uno stile comunicativo focalizzato sul lettore[29].

Genere	Corpus	1 Sing	2 Sing	3 Sing	1 Pl	2 Pl	3 Pl
Giornalismo	Due Par	3,60	0,27	20,31	3,10	0,01	21,31
	La Rep	1,62	0,64	29,60	1,01	0,12	9,51
Materiali didattici	Edu_child	1,83	0,64	35,14	0,63	0,29	18,36
	Edu_adult	1,22	0,73	38,04	2,14	0,07	9,58
Letteratura	Narr_child	3,77	2,40	38,54	2,16	1,09	12,52
	Narr_adult	3,62	1,55	39,09	1,77	0,36	7,84
Prosa scientifica	Wiki	0,42	0,88	32,34	0,18	0,02	16,64
	Scient_art	0,43	0,47	26,73	1,49	0,03	12,09
Linguaggio legislativo	It_Const	0,19	0,39	49,71	0,00	0,00	23,30
	Adm_acts	2,08	0,63	40,84	0,01	0,14	26,89
Linguaggio burocratico	Bur_simp	2,32	0,44	16,78	**13,57**	0,50	3,83
	Bur_orig	2,77	0,84	24,64	**1,48**	0,01	5,20

Tabella XIV – Distribuzione dei tratti flessivi di persona nei verbi all'interno dei corpora analizzati

[29] Cfr., a questo proposito, quanto riporta la regola 19 delle *Trenta regole* citate in nota 2 sull'uso delle forme impersonali nella scrittura burocratica: «Con che cosa si può sostituire l'impersonale? Con una forma personale che abbia per soggetto il nome dell'ufficio, o quello dell'amministrazione oppure con una forma verbale di prima persona plurale, senza indicazione del soggetto (quindi: «vi trasmettiamo», «vi informiamo», «abbiamo respinto la richiesta»). Questa seconda soluzione coniuga la necessità di non mettere in evidenza la persona dello scrivente, perché scrive non a nome proprio ma a nome dell'amministrazione, con l'opportunità di usare comunque una forma comune e diretta, come può essere la prima persona plurale».

Questa tendenza è stata ulteriormente qualificata misurando l'andamento delle distribuzioni dei pronomi, in particolare confrontando i pronomi personali e i pronomi clitici: a questo proposito, la tabella XV rivela un pattern distribuzionale opposto, con i primi che predominano nei testi semplificati (Bur_simp: 0,55 vs Bur_orig: 0,19), mentre i clitici nei testi originali (Bur_simp: 1,21 vs Bur_orig: 1,76).

Genere	**Corpus**	**% Pronomi personali**		**% Pronomi Clitici**	
Giornalismo	Due Par	0,08	0,14	0,96	1,28
	La Rep	0,21		1,59	
Materiali didattici	Edu_child	0,72	0,61	2,23	2,17
	Edu_adult	0,51		2,10	
Letteratura	Narr_child	0,84	0,75	3,62	3,38
	Narr_adult	0,66		3,15	
Prosa scientifica	Wiki	0,15	0,14	1,12	2,14
	Scient_art	0,12		1,05	
Linguaggio legislativo	It_Const	0,13	0,11	0,96	0,69
	Adm_acts	0,08		0,42	
Linguaggio burocratico	Bur_simp	**0,55**	0,37	**1,21**	1,48
	Bur_orig	**0,19**		**1,76**	

Tabella XV – Distribuzione percentuale dei pronomi personali e clitici nei corpora analizzati.

Mentre la scelta della prima persona plurale e l'uso di pronomi personali riflettono una modalità di scrittura più diretta e orientata al destinatario, il dato isolato sulla frequenza del clitico non è immediatamente interpretabile come "spia" della presenza di costruzioni impersonali, dal momento che il pronome clitico può svolgere ruoli sintattici diversi nelle varietà dell'italiano standard (accusativo, dativo, partitivo, riflessivo), con il solo clitico "si" usato in funzione impersonale; pertanto, anche in questo caso è stato necessario confrontare il dato dell'annotazione morfo-sintattica con l'output dell'analisi sintattica a dipendenze. In linea con le previsioni, questo controllo "incrociato" ha rivelato la preponderanza di dipendenze clitiche[30] in Bur_orig rispetto a Bur_simp (1,53 vs 0,45), come riportano i dati della tabella XVI.

Genere	**Corpus**	**% Dipendenze clitiche**	
Giornalismo	Due Par	0,48	0,67
	La Rep	0,86	
Materiali didattici	Edu_child	1,19	1,19
	Edu_adult	1,19	
Letteratura	Narr_child	1,52	1,36
	Narr_adult	1,20	

[30] In base al tagset sintattico di riferimento, infatti, gli archi di dipendenza marcati come "clit" identificano la relazione tra un pronome clitico e una testa verbale usata in funzione pronominale, mentre la relazione tra una testa verbale e un clitico in funzione accusativa o dativa è marcata da due etichette distinte, rispettivamente, "obj" e "comp_ind".

Prosa scientifica	Wiki	0,81	0,77
	Scient_art	0,73	
Linguaggio legislativo	It_Const	0,47	0,37
	Adm_acts	0,27	
Linguaggio burocratico	Bur_simp	**0,45**	0,99
	Bur_orig	**1,53**	

Tabella XVI – Distribuzione percentuale di dipendenze sintattiche di tipo clitico nei corpora analizzati.

Gli esempi, tratti rispettivamente dal corpus dei testi originali (III) e semplificati (IV), esemplificano come una frase impersonale sia stata tipicamente rielaborata nel processo di riscrittura e in che modo l'esito finale sia diversamente marcato nell'output dell'annotazione sintattica a dipendenze (figure II e III).

Esempi:

(III) testo originale:

> **Si comunica** alla S.V. che presso il Comando di Polizia Municipale sono stati consegnati i seguenti documenti [...]

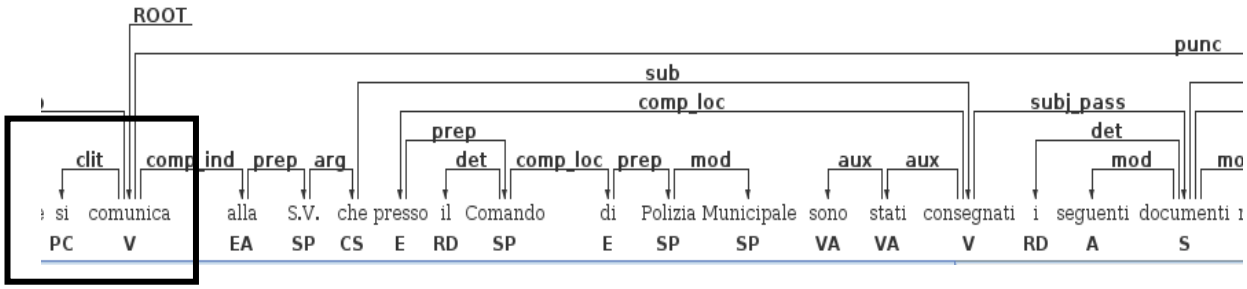

Figura II – rappresentazione grafica dell'annotazione linguistica della frase nell'esempio (III)

(IV) riscrittura:

> **Le comunichiamo** che la Sua carta d'identità è stata ritrovata [...]

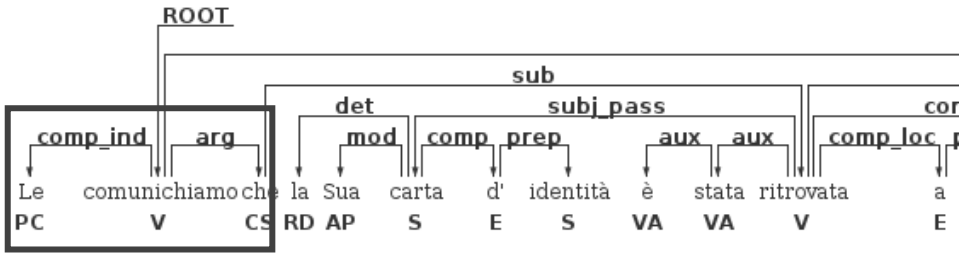

Figura III – rappresentazione grafica dell'annotazione linguistica della frase nell'esempio (IV)

Va sottolineato che queste ultime caratteristiche discusse (tratto di persona nei verbi, pronomi clitici e dipendenze sintattiche clitiche) non sono tanto implicate in una nozione "generale" del concetto di complessità, valida "across genres", ma rappresentano piuttosto delle scelte di tipo stilistico (e più precisamente, il loro correlato in una prospettiva linguistico-computazionale) la cui variazione,

rispetto al genere indagato, permette di riconoscere un esempio di "burocratese", distinguendolo invece da un testo amministrativo che, pur complesso, aderisce ai requisiti di un linguaggio chiaro, semplice e maggiormente comprensibile.

Tuttavia, è interessante notare che i due profili di testi amministrativi variano anche in relazione a parametri di complessità sintattica già risultati predittivi del livello di leggibilità di testi scritti nella lingua comune (cfr. Dell'Orletta, 2011a); si tratta, in particolare, di quelle caratteristiche che cercano di trattare a livello computazionale alcuni principi formali legati alla località delle dipendenze sintattiche, la cui implicazione nei processi di comprensione ha ricevuto ampie conferme dagli studi psicolinguistici[31] e dalla ricerca linguistica di ispirazione cognitiva, cfr. Rizzi (2003, 2006). È il caso della profondità dell'albero sintattico e della lunghezza delle relazioni di dipendenza sintattica.

La tabella XVII prende in considerazione i dati derivanti dal primo fattore, esplicitato nel parametro "media delle altezze massime" [32].

Genere	**Corpus**	**Media delle altezze massime**	
Giornalismo	Due Par	5,29	5,90
	La Rep	6,51	
Materiali didattici	Edu_child	5,54	6,45
	Edu_adult	7,36	
Letteratura	Narr_child	4,51	4,54
	Narr_adult	4,57	
Prosa scientifica	Wiki	6,47	7,04
	Scient_art	7,62	
Linguaggio legislativo	It_Const	4,73	5,44
	Adm_acts	6,15	
Linguaggio burocratico	Bur_simp	**5,96**	6,64
	Bur_orig	**7,32**	

Tabella XVII – Media delle altezze massime dell'albero sintattico nei corpora considerati.

Rispetto a questo parametro, i dati rivelano una variazione interna ai due profili statisticamente significativa (Bur_orig: 7,32 vs Bur_simp: 5,96), che rispecchia la tendenza generale verso una diminuzione del valore di questo parametro dalla varietà "complessa" a quella "semplice" all'interno di ciascun genere considerato. Addirittura, prendendo come riferimento il genere giornalistico (sul quale, come già ricordato, è stato addestrato l'indice di leggibilità READ-IT), si

[31] Alcuni riferimenti sono: Yngve (1960); Frazier (1985); Gibson (1998); De Vincenzi (1996).

[32] Tale parametro è ottenuto dalla media del numero di archi (ossia relazioni di dipendenza) che intercorrono tra una foglia (rappresentata da parole del testo senza dipendenti) e la radice dell'albero sintattico.

nota come il testo amministrativo derivante dalla semplificazione si avvicini al valore medio dei testi di *Due Parole* (5,29).

Simile è l'andamento dei quelle caratteristiche che esplicitano aspetti di complessità della frase legati alla presenza di dipendenze sintattiche più lunghe[33], associate ad una maggior difficoltà di elaborazione da parte del lettore (cfr. nota 31). Tali aspetti sono esplicitati dai parametri "lunghezza media dei link" (terza e quarta colonna della tabella XVIII) e "lunghezza dei link massimi" (quinta e sesta colonna della tabella XVIII).

Genere	**Corpus**	**Lunghezza media dei link**		**Lunghezza dei link massimi**	
Giornalismo	Due Par	2,16	2,27	7,91	9,10
	La Rep	2,39		10,28	
Materiali didattici	Edu_child	2,24	2,39	8,89	10,69
	Edu_adult	2,54		12,50	
Letteratura	Narr_child	2,25	2,33	6,63	7,03
	Narr_adult	2,40		7,43	
Prosa scientifica	Wiki	2,46	2,47	9,88	10,92
	Scient_art	2,48		11,96	
Linguaggio legislativo	It_Const	2,34	2,68	6,75	9,51
	Adm_acts	3,03		12,28	
Linguaggio burocratico	Bur_simp	**2,26**	2,36	**8,69**	10,17
	Bur_orig	**2,45**		**11,65**	

Tabella XVIII – Media delle lunghezze delle relazioni di dipendenza e delle lunghezza massime nei corpora considerati.

Come per le altezze massime, si osserva nuovamente che un valore più basso è riportato dal corpus che, all'interno di ciascun genere, ne costituisce la varietà "semplice"; il corpus burocratico non fa eccezione.

Anche in questo caso, dunque, un parametro rintracciabile nel testo automaticamente annotato si dimostra efficace nel cogliere distinzioni tra modalità di scrittura burocratica diversa, che impattano in maniera differente sulla chiarezza del testo finale. Il valore più alto delle lunghezze sintattiche medie può riflettere, ad esempio, l'abuso di incisi e frasi parentetiche, tipico della sintassi del "burocratese", che compromettono la leggibilità del testo (come nell'esempio in V), soprattutto quando si frappongono tra il soggetto e il verbo, tra il verbo e i suoi componenti o tra altre unità logiche del periodo, Fortis (2005: 67). All'opposto, la sua diminuzione permette di inferire che la semplificazione ha agito in maniera tale da ricostruire l'ordine non marcato dei costituenti,

[33] La lunghezza è qui calcolata in termini di numero di parole che separano l'elemento dipendente della relazione dalla sua testa.

operazione che può tradursi anche in una divisione della frase in due proposizioni autonome, come accaduto in (VI).

(V) testo originale:

Risulta **che** *per essere in regola con l'iscrizione fino all'a.a. 2002/2003 e poter quindi adire all'esame generale di laurea della sessione autunnale deve essere* **regolarizzata** la seconda rata dell'a.a. 1998/1999 e deve essere pagata la tassa di ricognizione dall'a.a. 1999/2000 all'a.a 2002/2003.

(VI) riscrittura:

Inoltre Lei non risulta regolarmente iscritto all'Università dall'a.a. 1998-1999 all'a.a. 2001-2002.
Nel caso intenda laurearsi nell'a.a. 2002 – 2003 (sessioni autunnale e straordinaria) deve regolarizzare l'iscrizione prima di presentare la domanda di laurea, versando l'importo totale di 2882 euro.

6. Alcune riflessioni finali

Questo studio si è proposto di offrire un contributo alla descrizione delle peculiarità della lingua della pubblica amministrazione da una prospettiva metodologica innovativa, basata sull'uso delle tecnologie linguistico-computazionali.

I dati emersi su base empirica sono senza dubbio parziali e richiedono una conferma con un corpus di dimensioni più ampie e dotato di una maggior differenziazione interna che renda conto delle molteplici varietà di tipologie testuali in cui si manifesta il linguaggio amministrativo. Tuttavia, l'analisi di questi risultati ha dimostrato non solo come essi avvalorino le analisi linguistiche tradizionali sui caratteri generali di questo genere, ma anche come possano contribuire ad un loro ulteriore arricchimento; una riprova – questa – della possibilità di conciliare fruttuosamente approccio qualitativo e quantitativo tanto nello studio della variazione linguistica tra generi e varietà testuali, quanto per l'approfondimento di concetti centrali dell'indagine linguistica, quale la nozione di complessità in rapporto a diverse tipologie di testo.

In questo scenario, il caso di studio qui presentato, oltre ad offrire interessanti spunti di riflessione teorica, può avere importanti ripercussioni sul piano applicativo; si pensi, in particolare, alla specializzazione di un indice di leggibilità "avanzato" come READ-IT sui parametri di complessità caratterizzanti questa tipologia di testi[34]. Tale indice si offrirebbe così come un ausilio concreto verso la promozione di una comunicazione amministrativa più chiara e semplice, prerequisito, a sua volta, per una pubblica amministrazione realmente semplificata e accessibile al cittadino.

[34] Un esempio di applicazione di READ-IT ai testi giuridici, tra cui alcuni testi amministrativi estratti dal corpus presentato in questo articolo, è stato recentemente proposto da Brunato/ Venturi (in c.s), dove le problematiche poste dalla lingua del diritto nelle sue diverse declinazioni sono state inquadrate nell'ottica del più generale tema dell'accessibilità e del contributo che possono offrire strumenti di valutazione della leggibilità del testo fondati su tecnologie linguistico-computazionali.

Bibliografia

Attardi, Giuseppe – Dell'Orletta, Felice – Simi, Maria – Turian, Joseph
2009 «Accurate Dependency Parsing with a Stacked Multilayer Perceptron», in *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia, Evalita.

Beccaria, Gian Luigi
1992 [1998] *Italiano antico e nuovo*. Milano, Garzanti.

Berruto, Gaetano
1987 *Sociolinguistica dell'italiano contemporaneo*, Roma, La Nuova Italia Scientifica

Biber, Douglas
1988 *Variation across speech and writing*. Cambridge & New York, Cambridge University Press.

Biber, Douglas
1993 «Using Register-diversified Corpora for General Language studies», in *Computational Linguistics Journal*, 19 (2), pp. 219-241.

Biber, Douglas
1995 *Dimensions of register variation: A cross-linguistic comparison*, Cambridge & New York, Cambridge University Press.

Brunato, Dominique – Venturi, Giulia
2014 «Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici», in Daniela Tiscornia, Francesco Romano e Maria Teresa Sagri (a cura di), fascicolo monografico di *Informatica e Diritto*, numero 2014/1, in c.s.

Calvino, Italo
1965 «Per ora sommersi dall'antilingua», Il Giorno, 3 febbraio (ora in *Una pietra sopra*, Torino, Einaudi, 1980, pp, 122-126).

Cortelazzo, Michele A.
1999 *Semplificazione del linguaggio amministrativo. Esempi di scrittura per le comunicazioni ai cittadini*, Padova, Comune di Padova.

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

Cortelazzo, Michele A.
2005 *Il Comune scrive chiaro. Come semplificare la comunicazione al cittadino. Con 24 esempi di testi rielaborati e le istruzioni per scrivere con stile*, Santarcangelo di Romagna, Maggioli.

Cortelazzo, Michele A. – Viale, Matteo
2006 «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 (Hg.) Gerhard Ernst - Martin-Dietrich Gleßgen - Christian Schmitt und Wolfgang Schweickard, *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, pp. 2112-2123.

Cortelazzo, Michele A. – Di Benedetto, Chiara – Matteo, Viale (a cura di)
2008 *Le "Istruzioni per le operazioni degli uffici elettorali di sezione" tradotte in italiano. Omaggio al Ministro dell'Interno*, Padova, Cleup.

De Mauro, Tullio
2000 *Grande dizionario italiano dell'uso (GRADIT)*, Torino, UTET.

De Vincenzi, Marica
1991 «Syntactic Parsing Strategies in Italian: The Minimal Chain Principle», Dordrecht; Boston, Kluwer Academic Publishers

Dell'Orletta, Felice
2009 «Ensemble system for Part-of-Speech tagging», in *Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)*, Reggio Emilia, Evalita.

Dell'Orletta, Felice – Montemagni, Simonetta – Venturi, Giulia
2011a «READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification», in *Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)*, Edinburgh, 30 luglio 2011, pp. 73-83.

Dell'Orletta, Felice – Montemagni, Simonetta – Vecchi, Eva Maria - Venturi, Giulia
2011b «Tecnologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria», in G.C. Bruno - I. Caruso – M. Sanna - Vellecco, I. (a cura di), *Percorsi migranti: uomini, diritto, lavoro, linguaggi*, Milano, McGraw-Hill, pp. 319-336.

Dell'Orletta, Felice – Montemagni, Simonetta
2012 «Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scolastico», in *SLI 2010 – XLIV Congresso Internazionale di Studi della Società di Linguistica Italiana* (Viterbo, Università degli Studi della Tuscia, 27-29 settembre 2010). Proceedings, vol. Linguistica Educativa pp. 343–359. Silvana Ferreri (ed.). Bulzoni Editore

Dell'Orletta, Felice – Montemagni, Simonetta – Venturi, Giulia
2013 «Linguistic Profiling of Texts Across Textual Genre and Readability Level. An Exploratory Study on Italian Fictional Prose», in *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)*, 7-11 September, Hissar, Bulgaria, pp. 189-197.

Fiorentino, Giuliana
2007 «*Web usability* e semplificazione linguistica nella scrittura amministrativa», in Federica Venier (a cura di), *Rete Pubblica*, pp. 11-37.

Fioritto, Alfredo (a cura di)
1997 *Manuale di stile*, Bologna, il Mulino.

Franceschini, Fabrizio – Gigli, Sara (a cura di)
2003 *Manuale di scrittura amministrativa*, Roma: Agenzia delle Entrate.

Frazier, Lyn
1985 «Syntactic complexity», in D.R. Dowty - L. Karttunen - A.M. Zwicky (a cura di), *Natural Language*

Parsing , Cambridge University Press, Cambridge, UK.

Fortis, Daniele
2005 «Il linguaggio amministrativo italiano», in *Revista de Liengua i dret*, n.43, pp. 47-116.

Gibson, Edward
1998 «Linguistic complexity: Locality of syntactic dependencies», in *Cognition*, 68(1), pp. 1-76.

Gotti, Maurizio
2005 *Investigating Specialized Discourse*. Bern: Peter Lang.

Halliday, Michael Alexander Kirkwood
1985 *Spoken and Written Language.* Geelong, Vic.: Deakin University Press.

Istituto di Teorie e Tecniche dell'informazione giuridica (ITTIG) e Accademia della Crusca
2011 *Guida alla redazione degli atti amministrativi. Regole e suggerimenti*.

Lucisano, Piero – Piemontese, Maria Emanuela
1988 «Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana», in *Scuola e Città*, vol. 3, pp. 57-68.

Marinelli, Rita – Biagini, Lisa – Bindi, Remo – Goggi, Sara – Monachini, Monica – Orsolini, Paola – Picchi, Eugenio – Rossi, Sergio – Calzolari, Nicoletta – Zampolli, Antonio
2003 «The Italian PAROLE corpus: an overview», in Antonio Zampolli et al. (a cura di), *Computational Linguistics in Pisa*, XVI-XVII(1), Pisa-Roma, IEPI, pp. 401-421.

Marconi, Lucia – Ott, Michela – Pesenti, Elia – Ratti, Daniela – Tavella, Mauro
1994 *Lessico Elementare*, Zanichelli, Bologna

Montemagni, Simonetta
2013 «Tecnologie linguistico-computazionali e monitoraggio della lingua italiana», in *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, Anno XLII, Numero 1, pp. 145-172.

Nelli, Roberto Paolo
2009 «Forme innovative di relazione e di comunicazione», in Katia Giusepponi (a cura di), *Gestione e controllo delle amministrazioni pubbliche. Strumenti operativi e percorsi d'innovazione*, Giuffrè Editore, Milano, pp. 899-950.

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

Piemontese, Maria Emanuela
1999 «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, pp. 269-292.

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

Rizzi, Luigi
2006 «Sintassi: Le strutture», in Alessandro Laudanna e Miriam Voghera (a cura di), *Il linguaggio: Strutture linguistiche e processi cognitivi*, Laterza, Bari, 205-229.

Rizzi, Luigi
2003 «Some Elements of the Study of Language as a Cognitive Capacity», in Dimitri, ed., N., M. Basili, I. Gilboa, (eds.), *Cognitive Processes and Economic Behaviour*, Routledge, London and New York, 104-136,

Serianni, Luca
2003 *Italiani Scritti*. Bologna, Il Mulino.

Sobrero, Alberto A.
1993 «Lingue speciali», in Alberto A. Sobrero (a cura di), *Introduzione all'italiano contemporaneo. La variazione e gli usi*, Roma - Bari, pp. 237-277.

Venturi, Giulia
2013 «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)*, in c.s.

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

Voghera, Miriam
2001 «Riflessioni su semplificazione, complessità e modalità di trasmissione: sintassi e semantica», in Maurizio Dardano et al. (a cura di), *Scritto e parlato. Metodi, testi e contesti*, Atti del Colloquio Internazionale di Studi, Aracne, Roma, pp.65-78.

Voghera, Miriam
2005 «La misura delle categorie sintattiche», in Isabella Chiari – Tullio De Mauro (a cura di) *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne, Roma, pp.125-138.

Yngve, Victor H.A.
1960 «A model and an hypothesis for language structure», in *Proceedings of the America Philosophical Society*, pp. 444-466

Sitografia

Guida alla redazione degli atti amministrativi. Regole e suggerimenti.
<http://www.pacto.it/content/view/416/48/>; [data ultima consultazione: 01/09/2014]

Linguaggio amministrativo chiaro e semplice (Maldura, Università degli studi di Padova) <http://www.maldura.unipd.it/buro/index.html>; [data ultima consultazione: 01/09/2014]

LinguA - Linguistic Annotation Pipeline (ItaliaNLP Lab) <http://linguistic-annotation-tool.italianlp.it/>; [data ultima consultazione: 01/09/2014]

Progetto Comuniversità – Comunicazione Istituzionale nelle Università. Raccolta sistematica di modelli testuali <http://www.coinfo.net/documenti/RicercheIntervento/Allegati/Ricerca_Intervento_ComUniversit%C3%A0.pdf>; [data ultima consultazione: 01/09/2014]

READ-IT <http://www.italianlp.it/demo/read-it/> [data ultima consultazione: 01/09/2014]