CORPUS «ITAIST»: NOTE PER LO SVILUPPO DI UNA RISORSA LINGUISTICA PER LO STUDIO DELL'ITALIANO ISTITUZIONALE PER IL DIRITTO DI ACCESSO CIVICO

Daniela Vellutino, *Nicola Cirillo*[1]

1. CORPUS "ITAIST": UNA RISORSA LINGUISTICA PER L'ITALIANO ISTITUZIONALE

Il corpus "ItaIst" è una risorsa linguistica che si sta sviluppando nell'ambito del progetto PRIN 2020 "VerbACxSS: su verbi analitici, complessità, verbi sintetici, e semplificazione. Per l'accessibilità"[2], costituita da testi istituzionali scritti in lingua italiana, con la finalità di sviluppare un modello linguistico di fondazione per l'AI generativa per l'italiano istituzionale.

Nel contributo si presentano le prime note per lo sviluppo di una parte del corpus "ItaIst", la risorsa linguistica "ItaIst-DdAC_GRU" (*Italiano istituzionale per il Diritto di Accesso Civico del dominio terminologico Gestione Rifiuti Urbani*). La risorsa è costituita dai testi istituzionali dei documenti scritti in lingua italiana per lo specifico dominio terminologico, in relazione alla comunicazione specialistica normativa e amministrativa e alle attività d'informazione e comunicazione delle amministrazioni pubbliche per il diritto di accesso civico[3].

Il diritto di accesso civico (semplice e generalizzato) è un istituto giuridico introdotto da poco più di dieci anni dell'ordinamento italiano per consentire a chiunque di accedere ai dati, ai documenti e alle informazioni delle pubbliche amministrazioni senza necessità di dimostrare un interesse qualificato.

I documenti originali sono stati raccolti secondo il modello di classificazione dei testi istituzionali per la comunicazione pubblica e istituzionale "Modello CPI" (Vellutino,

[1] Università degli Studi di Salerno.
A fini di assegnazione formale, vanno attribuiti a Daniela Vellutino i §§ 1, 3, 4, 4.1, 5, 6 e a Nicola Cirillo i §§ 2, 5.1, 5.2.

[2] Progetto di Rilevante Interesse Nazionale – PRIN 2020 "VerbACxSS: su verbi analitici, complessità, verbi sintetici e semplificazione. Per l'accessibilità" (2020, Prot. 2020BJKB9M), finanziato dal Ministero dell'Università e della Ricerca. Coinvolge l'Università Roma Tre (PI: Anna Pompei), l'Università del Molise (coordinamento dell'unità di ricerca: Giuliana Fiorentino) e l'Università di Salerno (coordinamento dell'unità di ricerca: Daniela Vellutino).

[3] In Italia l'istituto giuridico del diritto accesso civico semplice è disciplinato dal D.Lgs 33/2013 "Riordino della disciplina riguardante il diritto di accesso civico e gli obblighi di pubblicità, trasparenza e diffusione di informazioni da parte delle pubbliche amministrazioni"; mentre il diritto di accesso civico generalizzato dal D.Lgs 97/2016 "Revisione e semplificazione delle disposizioni in materia di prevenzione della corruzione, pubblicità e trasparenza, correttivo della legge 6 novembre 2012, n. 190 e del decreto legislativo 14 marzo 2013, n. 33, ai sensi dell'articolo 7 della legge 7 agosto 2015, n. 124, in materia di riorganizzazione delle amministrazioni pubbliche". Il diritto di accesso civico semplice obbliga le PA a pubblicare sui siti web istituzionali alcuni contenuti informativi minimi per la trasparenza amministrativa, tra i quali, le informazioni ambientali; mentre il diritto di accesso civico generalizzato consente a chiunque di richiede alle amministrazioni pubbliche dati, informazioni e documenti tenendo conto delle limitazioni legate alla tutela degli interessi pubblici e privati contenuti nell'art. 5-bis del decreto. Per approfondimenti https://foia.gov.it/normativa/cose-il-foia.

2012, 2014, 2018) che distingue i testi istituzionali in relazione alle esigenze pragmatiche collegate ai diversi contesti comunicativi. I contesti comunicativi sono normati dalle leggi italiane in materia di trasparenza amministrativa, informazione, comunicazione e diritto di accesso civico.

In base a questo modello di classificazione, i testi istituzionali sono distinti in due categorie: testi dei linguaggi istituzionali speciali del diritto e dell'amministrazione e testi dei linguaggi istituzionali mediali delle attività d'informazione e comunicazione delle pubbliche amministrazioni. I testi istituzionali speciali sono comunicati attraverso gli strumenti della pubblicità legale, mentre i testi istituzionali mediali sono comunicati attraverso media istituzionali quali siti web e canali social, materiale pubblicitario e a mezzo stampa.

Questa prima fase di studio è consistita nel monitoraggio linguistico condotto mediante la raccolta dei testi istituzionali che sono stati repertoriati seguendo uno schema di metadatazione basato sul modello CPI e progettato per registrare sia i dati amministrativi che i dati linguistici.

Lo schema di metadatazione dei testi istituzionali rappresenta un primo livello di annotazione linguistica utile a repertoriare i testi istituzionali e tracciare il profilo delle differenti testualità dei linguaggi istituzionali speciali e mediali.

La risorsa linguistica creata, il corpus "ItaIst-DdAC_GRU", sarà utilizzata per osservare i fatti linguistici caratteristici dei testi "fonte" normativi e amministrativi dei linguaggi istituzionali speciali e quelli dei testi mediali che usano le informazioni contenute nei testi fonte e le riformulano per adattarle linguisticamente alle differenti testualità mediali. Inoltre, il corpus sarà utilizzato per annotare informazioni linguistiche per le seguenti finalità di ricerca e di didattica:

1. creazione di un corpus parallelo costituito da coppie di frasi complesse-semplici per sviluppare un modello di fondazione per l'AI generativa per i testi istituzionali mediali;
2. creazione di risorse linguistiche terminologiche quali lessici istituzionali, glossari e schede terminologiche attraverso l'estrazione della terminologia di dominio;
3. tracciamento della frequenza, distribuzione e uso dei termini nei diversi tipi di testo istituzionali speciali e mediali;
4. rilevamento dei fatti linguistici che connotano i testi dei linguaggi istituzionali speciali e mediali;
5. sviluppo di modelli linguistici descrittivi e di metodi glottodidattici per fornire agli studenti e alle studentesse le competenze linguistiche e comunicative necessarie per diventare professionisti della comunicazione pubblica e istituzionale[4].

2. STATO DELL'ARTE

I corpora sono uno strumento fondamentale per lo studio dei fenomeni linguistici sia dell'italiano standard che dei linguaggi istituzionali speciali e mediali. In questi anni, però, l'attenzione degli studi è stata rivolta al linguaggio normativo e, prevalentemente, amministrativo.

Uno dei corpora principali sviluppati per lo studio della lingua italiana della pubblica amministrazione è PAWaC! (Public Administration Web as Corpus) (Passaro, Lenci,

[4] Le competenze linguistiche necessarie per la formazione nell'ambito della comunicazione pubblica e istituzionale sono state descritte nella norma tecnica dell'Ente Italiano di Normazione UNI 111483:2021 "Figure professionali operanti nell'ambito della comunicazione" (rilasciata il 9 settembre 2021).

2015). Questo corpus, realizzato nell'ambito del progetto SEMPLICE[5] (SEMantic Instruments for PubLIc Administrators and CitizEns) della Regione Toscana, contiene 4.172 documenti, 3.043.842 frasi e 25.218.385 token. Le tipologie di documenti che costituiscono il corpus sono delibere, determine, bandi e altri atti amministrativi, raccolti tramite il *crawling*[6] dell'Albo Pretorio online di 277 Comuni toscani.

La caratteristica più spesso sottolineata dell'italiano amministrativo è sicuramente la scarsa leggibilità dei tipi di testo. A ragione di ciò sono stati sviluppati corpora paralleli con coppie di frasi: il testo originale e la relativa riscrittura semplificata. Ne è un esempio il corpus parallelo SIMPITIKI (Tonelli *e al.*, 2016), formato da 1.166 coppie di frasi.

Nello specifico SIMPITIKI è composto da due sotto-corpora. Il primo, denominato *wiki*, contiene 575 coppie di frasi estratte da Wikipedia, mentre il secondo, denominato *PA*, è composto da 591 coppie di frasi tratte dai documenti amministrativi del comune di Trento e semplificate a mano seguendo le operazioni proposte da Brunato *et al.* (2015).

L'annotazione manuale individua le operazioni di semplificazione delle frasi mostrate in Tabella 1.

Tabella 1. *Operazioni di semplificazione individuate in SIMPITIKI (Tonelli e al., 2016)*

Operazione		**Esempio**
Separazione		[…] è quello dei genitori che dovrà essere autocertificato […] […] è quello dei genitori. Esso dovrà essere autocertificato […]
Unione		I Salmi sono per il giudaismo il testo della fede pura per eccellenza. Sono il fondamento e la ragione […] I Salmi sono per il giudaismo il testo della fede pura per eccellenza, il fondamento e la ragione [...]
Permutazione		[…] la cancellazione della domanda di ammissione al nido dalla graduatoria. […] la cancellazione dalla graduatoria della domanda di ammissione al nido.
Inserimento	Verbo	In funzione della tipologia dell'opera, oggetto di richiesta di autorizzazione […] In funzione della tipologia dell'opera che è oggetto di richiesta di autorizzazione […]
	Soggetto	Il modulo contiene dichiarazioni […] e pertanto, ai sensi della normativa vigente, deve essere: Il modulo contiene dichiarazioni […] e pertanto, ai sensi della normativa vigente, il modulo deve essere:
	Altro	REQUISITI PER L'ACCESSO REQUISITI PER L'ACCESSO AL SERVIZIO

[5] Sito web del progetto SEMPLICE: http://www.progettosemplice.it/.

[6] Il termine *crawling* denota l'accesso a siti web per ricercare e raccogliere dati attraverso un apposito software denominato *crawler*.

Cancellazione	Verbo	[…] devono essere posseduti all'atto della domanda e <u>trovare conferma</u> al momento della chiusura dei termini. […] devono essere posseduti all'atto della domanda e al momento della chiusura dei termini.
	Soggetto	[…] può accettare fin da subito e senza alcuna altra formalizzazione il posto nel nido di prima scelta qualora <u>lo stesso</u> si renda disponibile. […] può accettare fin da subito e senza alcuna altra formalizzazione il posto nel nido di prima scelta qualora si renda disponibile.
	Altro	<u>Limitatamente</u> ai richiedenti residenti nella zona di S. Lazzaro […] Ai richiedenti residenti nella zona di S. Lazzaro […]
Trasformazione	Sostituzione (parola)	L'utente dovrà <u>individuare</u> una delle due L'utente dovrà <u>indicare</u> una delle due
	Sostituzione (sintagma)	Pertanto, ogni variazione intervenuta <u>successivamente</u> <u>al</u>la domanda di ammissione […] Pertanto, ogni variazione intervenuta <u>dopo</u> la domanda di ammissione […]
	Ripresa anaforica	[…] il mancato inserimento della <u>stessa</u> nella graduatoria. […] il mancato inserimento della <u>domanda</u> nella graduatoria.
	Nome – verbo	La richiesta dei servizi di <u>anticipo</u> e <u>posticipo</u> di orario […] La richiesta di <u>anticipare</u> e <u>posticipare</u> l'orario […]
	Verbo – nome	[…] le varie informazioni <u>che occorrono</u> al master […] le varie informazioni <u>necessarie</u> al master
	Diatesi	[…] tutta la documentazione <u>prevista</u> dalla tipologia di riferimento […] […] tutta la documentazione <u>che prevede</u> la tipologia di riferimento […]
	Tempo verbale	Gli elementi emersi dagli approfondimenti di cui sopra <u>andranno</u> descritti […] Gli elementi emersi dagli approfondimenti di cui sopra <u>vanno</u> descritti […]

Analizzando le caratteristiche del corpus, gli autori hanno notato come la sostituzione lessicale di termini ad alto specialismo e la sostituzione Nome-Verbo siano più frequenti nel sotto-corpus *PA* che in *wiki*.

Collegato ai corpora SIMPITIKI e PAWaC di più recente sviluppo, è il corpus Admin-It (Miliani *e al.*, 2022), un corpus parallelo di frasi semplificate tratte da testi amministrativi. Esso è composto da 736 coppie di frasi ed è diviso in tre sotto-corpora: Admin-It$_{OP}$, Admin-It$_{RS}$, Admin-It$_{RD}$. Nel dettaglio, Admin-It$_{OP}$ contiene 588 coppie di frasi, estratte direttamente da SIMPITIKI, che sono state semplificate applicando delle precise operazioni (cfr. Tabella 1). Admin-It$_{RS}$ contiene 100 coppie di frasi, estratte da siti web di alcuni comuni italiani e dal corpus PAWaC. Le frasi semplificate sono state riscritte dagli autori seguendo le *Trenta regole per una buona scrittura amministrativa* proposte dal linguista Michele Cortellazzo (2021). Infine, Admin-It$_{RD}$ è composto da 48 coppie di frasi, selezionate da alcuni testi amministrativi semplificati manualmente da Cortelazzo (1998; 1999).

I corpora PAWaC, SIMPITIKI e Admin-It sono accessibili e liberamente scaricabili sul web.[7]

3. Italiano Istituzionale per il diritto di Accesso Civico

Oggetto del nostro studio sono i testi istituzionali per il diritto di accesso civico, vale a dire quei testi che rispondono agli obblighi informativi per garantire questo diritto. Pertanto, riguardano gran parte degli usi della varietà dell'italiano istituzionale[8].

> la varietà della lingua nazionale attestata dagli usi linguistici delle comunicazioni ufficiali delle organizzazioni del settore pubblico e privato. Si colloca nello spazio socio-pragmalinguistico dell'italiano contemporaneo nell'area della lingua standard. I suoi usi sono molteplici e articolati in un *continuum* che attraversa le dimensioni della variazione. In particolare, nel polo alto formale si collocano gli usi dei linguaggi istituzionali speciali – legali, amministrativi, tecnico-scientifici, tecnico-operativi –; lungo la direttrice della variazione diafasica, i differenti usi dei linguaggi mediali delle attività d'informazione e comunicazione, le diverse forme di scrittura oralizzante dei social media e l'oralità conversazionale delle interazioni faccia a faccia presso gli sportelli informativi o via chatbot. Dal punto di vista tipologico-strutturale, è caratterizzato da un ricco repertorio testuale e da incessante dinamismo neologico con interferenze inter-intralinguistiche per il costante afflusso delle terminologie specialistiche spesso in forma di unità lessicali superiori, che possono anche ridursi in acronimi, dando origine a varianti con diversi gradi di specialismo (Vellutino: in stampa).

In particolare, i testi istituzionali repertoriati nel corpus "ItaIst-DdAC_GRU" sono quelli individuati nello schema di classificazione dei tipi di testo istituzionali per le attività d'informazione e comunicazione pubblica e istituzionale – Modello CPI – illustrato in Tabella 2.

[7] PAWaC! è distribuito con licenza CC BY 4.0 su: https://data.europa.eu/data/datasets/elrc_1282.
SIMPITIKI è distribuito con licenza CC BY 4.0 su: https://github.com/dhfbk/simpitiki.
Admin-It: è distribuito su: https://github.com/Unipisa/admin-It.
[8] Vellutino (in stampa).

Tabella 2. *Classificazione dei testi istituzionali per le attività d'informazione e comunicazione pubblica e istituzionale- Modello CPI (D.Vellutino)*

Obblighi informativi e finalità Legge 150/2000 Art. 1. Comma 5	**Funzione pragmatica**	**Contesto comunicativo**	**Tipo di testo istituzionale per struttura operativa Legge 150/2000**	**Vincolo interpretativo**
Linguaggi istituzionali speciali del diritto e dell'ammnistrazione				
Illustrare le disposizioni normative	Prescrittiva	• Pubblicità legale (Legge 69/2009) Gazzetta ufficiale Bollettini ufficiali Albo Pretorio online	• Atti legislativi - URP • Atti amministrativi - URP	Molto vincolanti
Favorire la conoscenza dei procedimenti amministrativi	Strumentale-regolativa	• Comunicazione per la trasparenza amministrativa e il diritto di accesso civico (Legge 241/1990; D.lgs. 33/2013; D.lgs 97/2016; Delibera ARERA 444/2019/R/RIF)	• Testi tecnico-operativi – URP	Mediamente vincolanti
Linguaggi istituzionali mediali per attività d'informazione e comunicazione pubblica e istituzionale				
Illustrare le attività delle istituzioni e il loro funzionamento	Informativa	• Comunicazione pubblica per l'accountability (D.lgs. 33/2013; D.lgs 97/2016)	• Testi per l'accountability - Portavoce	Mediamente vincolanti
Favorire l'accesso ai servizi pubblici		• Informazione per la pubblica utilità (Legge 150/2000)	• Testi informativi – Ufficio stampa	
Promuovere conoscenze su temi di interesse pubblico	Persuasiva	• Pubblicità istituzionale • Marketing territoriale • Brand image/identity (Legge 150/2000)	• Testi delle campagne di comunicazione istituzionale	Poco vincolanti
Promuovere l'immagine delle amministrazioni				

Il modello di classificazione dei testi istituzionali CPI individua per ogni obbligo informativo e ogni finalità delle attività d'informazione e comunicazione delle pubbliche amministrazioni i diversi tipi di testo istituzionali, in base alla distinzione tra linguaggi istituzionali speciali del diritto e dell'amministrazione, e linguaggi istituzionali mediali.

I tipi di testo istituzionali sono distinti in relazione alla loro funzione pragmatica, al contesto comunicativo disciplinato dalle norme in materia e al tipo di vincolo interpretativo[9] che condiziona la struttura del tipo di testo.

4. METODOLOGIA PER LA COSTRUZIONE DEL CORPUS ITAIST-DDAC_GRU

In questa prima fase di sviluppo del corpus ItaIst-DdAC_GRU, si è deciso di seguire la classificazione dei testi istituzionali del modello CPI perché consente di individuare i testi dei documenti da repertoriare in modo da tracciare il passaggio dalla "rigidità" dei tipi di testo dei linguaggi istituzionali speciali alla "elasticità" dei tipi di testo dei linguaggi istituzionali mediali.

Il monitoraggio linguistico per costruire il corpus è il risultato di un procedimento di tipo esplorativo: i testi dei documenti di gestione del servizio di raccolta dei rifiuti urbani sono stati richiesti agli studenti e alle studentesse del corso "Comunicazione pubblica e linguaggi istituzionali" esercitando il diritto di accesso civico[10].

Pertanto, la quantità di testi istituzionali raccolta per creare il corpus non è stata definita a priori né è stato previsto un loro bilanciamento.

4.1. *Richieste di Accesso civico*, *raccolta e metadatazione dei testi istituzionali*

Gli studenti e le studentesse hanno esercitato il diritto di accesso civico semplice (Art. 5, Art. 40 D.Lgs. 33/2013) richiedendo i seguenti documenti quando non presenti sui siti web istituzionali dei propri comuni di residenza:

- Piano Economico Finanziario (PEF) del servizio di raccolta differenziata.
- Documentazione dell'appalto della gestione dei rifiuti urbani.
- Modello unico Dichiarazione Ambientale (MUD).
- Atti amministrativi sulla gestione dei rifiuti.
- Piano di comunicazione della raccolta differenziata.

Questi documenti sono testi tecnico-amministrativi che contengono dati e informazioni che riguardano la trasparenza amministrativa e l'accountability e, dunque, devono essere riformulati in forme testuali utili per la comunicazione pubblica e istituzionale.

I testi raccolti per essere repertoriati al fine di creare il corpus ItaIst-DdAC_GRU sono stati soggetti ad una prima operazione di metadatazione, vale a dire che ogni documento è stato descritto attraverso set di metadati amministrativi e linguistici.

Questa operazione di metadatazione rappresenta il primo livello di annotazione utile per creare il profilo di questa risorsa linguistica secondo il paradigma FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*)[11], seguendo i principi di interoperabilità e di conoscenza approfondita del contesto documentale e linguistico utile per la condivisione dei risultati delle ricerche sia nell'ambito scientifico che in quello delle amministrazioni pubbliche.

[9] Per le nozioni di "rigidità" e "elasticità" dei testi e di vincolo interpretativo si rimanda a Sabatini (1998).
[10] Il monitoraggio linguistico è stato realizzato nel corso del primo semestre dell'anno accademico 2022-2023 (ottobre 2022 - marzo 2023) del corso di laurea magistrale *Corporate Communication e Media* dell'Università degli Studi di Salerno.
[11] Per il dibattito sulla necessità di creazione di dati FAIR si veda Wilkinson *et al.*. (2016).

Pertanto, agli studenti e alle studentesse è stato chiesto di metadatare i documenti raccolti attraverso le richieste di accesso civico e il monitoraggio dei siti web istituzionali dei comuni.

In tal modo, gli studenti e le studentesse hanno appreso le tecniche di classificazione e di annotazione dei testi istituzionali per realizzare una prima annotazione del corpus ItaIst-DdAC_GRU. Successivamente la metadatazione dei documenti è stata standardizzata, riveduta e corretta dagli autori di questo articolo.

I metadati amministrativi associati ai documenti sono i seguenti:

- Nome originale del file.
- Link alla pagina web su cui è presente il documento.
- Link per il download del documento.
- Ente che ha prodotto il documento.
- Tipo di documento.
- Titolo.
- Data di produzione del documento.
- Data di download del documento.
- Formato.

I metadati linguistici dello schema di classificazione dei testi istituzionali proposto da Vellutino (2018) sono i seguenti:

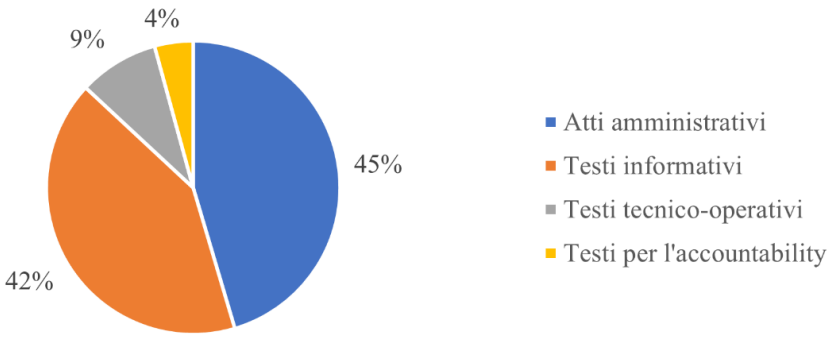
- Obblighi informativi e finalità.
- Funzione pragmatica.
- Contesto comunicativo.
- Tipo di testo istituzionale.

5. CORPUS "ITAIST-DDAC_GRU"

Il corpus ItaIst-DdAC_GRU è attualmente composto da 306 testi provenienti da 27 comuni campani. Tutti i testi riguardano il dominio terminologico della gestione dei rifiuti urbani.

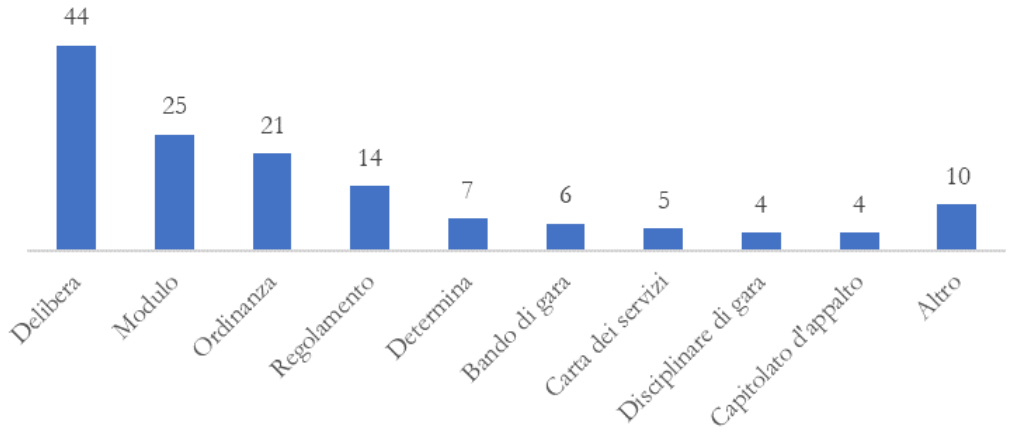
La maggior parte di essi appartiene alle tipologie testuali degli atti amministrativi (140) e dei testi informativi (127). Altre tipologie testuali repertoriate sono i testi tecnico-operativi (26) e, meno rappresentati, i testi per l'*accountability* (13).

Figura 1. *Numero di documenti per tipo di testo*



Le principali tipologie di atti amministrativi presenti nel corpus sono rappresentate dai testi dei linguaggi istituzionali speciali, in particolare, gli atti amministrativi: delibere, moduli, determine, ordinanze, regolamenti, bandi, disciplinari di gara, capitolati d'appalto, carte dei servizi. Appartengono ai tipi di testo dei linguaggi istituzionali speciali anche i testi tecnico-operativi relativi alla gestione dei rifiuti quali il Modello Unico di Dichiarazione Ambientale (MUD) e il Piano Economico Finanziario (PEF) del servizio di gestione dei rifiuti urbani.

Figura 2. *Numero di documenti per tipologia di atti amministrativi*

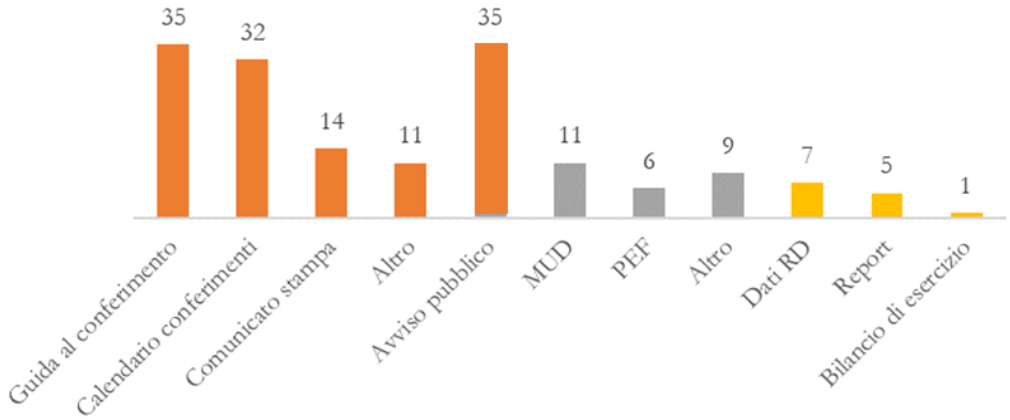


È da rilevare che alcuni testi per l'accountability sono in forma di tabelle di dati relativi alla raccolta differenziata (RD), report e testi contabili quali i bilanci d'esercizio.

Tra i testi istituzionali informativi troviamo gli avvisi pubblici che, dal punto di vista pragmalinguistico, rappresentano una forma testuale per definizione amministrativa e per funzione informativa (Vellutino, 2018).

I testi informativi presenti nel corpus sono per la maggior parte calendari e guide per il conferimento dei rifiuti e i testi istituzionali mediali quali comunicati stampa e news per il sito web istituzionale.

Figura 3. *Numero di documenti per tipologia di testi informativi (arancione), testi tecnico-operativi (grigio) e testi per l'accountability (giallo)*



Per quanto riguarda le dimensioni, il corpus contiene 1.379.843 frasi e 1.379.843 token.

È da rilevare che gli atti amministrativi sono, in media, molto più lunghi delle altre tipologie testuali (7.293 token). Seguono i testi tecnico-operativi (7.066 token), i testi per l'accountability (1.702 token) e, infine, i testi informativi (1.212 token) che sono i più brevi.

5.1. *Estrazione del testo e preprocessing*

Per estrarre il testo dai documenti raccolti abbiamo sviluppato uno script Python in grado di trattare quattro tipologie differenti di file. Lo script tratta i PDF testuali attraverso la libreria *pdfplumber* e i documenti in formato DOCX attraverso la libreria *textract*. Più complesso è il trattamento di immagini e scansioni PDF. Per estrarre il testo da tali documenti è necessario un software OCR (*Optical Character Recognition*). A questo scopo abbiamo optato per *Tesseract*, software open source, sponsorizzato da Google. Infine, lo *scraping*[12] dei comunicati stampa e delle news dei siti web istituzionali è stato effettuato tramite la libreria *Beautiful Soup*. È molto comune che le tabelle vengano rimosse nel processo di costruzione di un corpus. In questo caso, abbiamo scelto invece di conservarle in quanto esse rappresentano una fonte preziosa per estrarre la terminologia specialistica.

Una volta estratto il testo dai documenti, si è poi reso necessario modificarne la formattazione per ricostruire le frasi. A questo scopo la scelta è ricaduta sull'utilizzo del modulo *sentence-splitter* sviluppato da Koehn (2005) per il trattamento del corpus Europarl. Infine, abbiamo ottenuto l'annotazione linguistica delle frasi (*part-of-speech tagging*, *lemmatizzazione* e *depency parsing*) grazie al parser *stanza* (Manning *et al.*, 2014).

5.2. *Alcuni esempi di filiera documentale del continuum dai linguaggi istituzionali speciali ai mediali*

In questa fase, il corpus ItaIst-DdAC_GRU non contiene il collegamento tra il testo fonte e il relativo testo informativo. Osserviamo comunque alcuni casi in cui questo collegamento è stato rilevato. Infatti, le disposizioni delle ordinanze si traducono nel contenuto informativo di opuscoli e calendari per la raccolta differenziata. Di seguito è riportata una breve analisi di alcuni estratti.

Figura 4. *Estratto da un'ordinanza*

MODALITA DI CONFERIMENTO:

UTENZE DOMESTICHE (nuclei familiari): il sacchetto riposto all'interno del mastello o bidone carrellato, contenente la frazione di rifiuto raccolta separatamente, deve essere tenuto all'interno della proprietà privata e deve essere esposto all'esterno sulla pubblica via, esclusivamente in corrispondenza al numero civico della propria abitazione, dopo le **ore 21.00 del giorno precedente e fino alle 5.00 del giorno di ritiro (raccolta)**

Il testo in Figura 4 è composto da un singolo periodo molto lungo e complesso: il sintagma nominale soggetto comprende almeno tre relative implicite e i due verbi principali sono al passivo, preceduti dal verbo *dovere*. Il lessico contiene tecnicismi in forma

[12] Il termine *scraping* denota l'estrazione dei dati dalle pagine web, individuandone la posizione ed eliminando i tag HTML superflui.

di poliematiche quali *bidone carrellato*, *frazione di rifiuto raccolta separatamente* e *pubblica via*. Per semplificarlo, sarebbe sufficiente rimuovere l'aggettivo *carrellato*, sostituire *frazione di rifiuto raccolta separatamente* con *rifiuto differenziato* e il termine *pubblica via* con *strada*.

Figura 5. *Estratto da un opuscolo*

ORARIO DI CONFERIMENTO

Entro le ore 5:00 del giorno indicato, a partire dalle ore 21:00 del giorno precedente.

Esporre le attrezzature (bidoni, mastelli, sacchi) all'esterno della propria abitazione o condominio, in corrispondenza del numero civico.

L'estratto dall'opuscolo in Figura 5, invece, è di più immediata fruizione. Presenta i due nuclei informativi in due diverse frasi, di cui una nominale. I sintagmi sono brevi e privi di relative, la modalità deontica è espressa attraverso l'infinito con valore d'imperativo e il lessico è decisamente più semplice.

Figura 6. *Estratto da un'ordinanza*

R – RIFIUTI IMGOMBRANTI E RAEE DI DIMENSIONI SUPERIORI AI 25 CM

I rifiuti **Ingombranti** ed i **Raee** possono essere conferiti dal **Lunedì** al **Sabato** presso i centri comunali di raccolta denominati "Arechi" e "Fratte" dalle ore **8,00** alle ore **17,00** nel periodo invernale (1° novembre – 30 aprile) e dalle ore **8,00** alle ore **19,00** nel periodo **estivo** (1° maggio – 31 ottobre) o su prenotazione per ritiro a domicilio al numero **0892882036** o al numero verde **800809303**.

È stato osservato che non sempre i testi informativi sono più accessibili di quelli amministrativi. Nell'esempio seguente, infatti, è il testo dell'ordinanza in Figura 6 ad essere più semplice di quello dell'opuscolo in Figura 7. Esso è composto da un singolo periodo contenente sintagmi nominali semplici e la coordinazione è prevalente rispetto alla subordinazione. L'unico elemento che appesantisce il testo è il verbo principale al passivo, preceduto dal verbo *potere*.

Figura 7. *Estratto da un opuscolo*

Utenze domestiche – Come conferire

I rifiuti ingombranti e i RAEE possono essere conferiti su sede stradale, previo appuntamento per il ritiro, chiamando al numero 089 288 2036 (attivo dal LUNEDI al SABATO dalle 9 alle 19). La prenotazione può essere fatta pure con l'APP Junker.

Il conferimento può avvenire anche presso i Centri di Raccolta Comunale Fratte e Arechi, dal LUNEDI al SABATO dalle ore 8 alle 17 nel periodo invernale e dalle 8 alle 19 nel periodo estivo.

Nel testo dell'opuscolo troviamo tre periodi tenuti insieme da riprese anaforiche e tre verbi principali, tutti in forma passiva, preceduti dal modale *potere*. Due di essi, *fare una prenotazione* e *il conferimento avviene*, sono costruzioni a verbo supporto[13], che notoriamente pongono problemi di leggibilità (Cortelazzo, 2021). L'uso dei verbi sintetici *prenotare* e

[13] Per informazioni sugli aspetti linguistici delle costruzioni a verbo supporto, cfr. Gross (1981) e Mel'čuk (2004).

conferire avrebbe reso il testo più leggibile. Inoltre, il tecnicismo specifico presente nell'ordinanza, *prenotazione per ritiro a domicilio*, è qui reso con un'inutilmente complessa perifrasi: […] *conferiti su sede stradale*, *previo appuntamento per il ritiro*. Essa contiene due tecnicismi collaterali (*sede stradale* e *previo*) e viene comunque ripresa anaforicamente dal nome *prenotazione* nella frase seguente.

6. CONCLUSIONI

Il corpus ItaIst-DdAC_GRU rappresenta un primo studio che rende evidente quanto lo sviluppo di risorse linguistiche sia utile per la ricerca sul profilo linguistico dell'italiano istituzionale e per la didattica della comunicazione pubblica.

Le amministrazioni pubbliche hanno già intrapreso la strada dell'uso dell'AI nei processi decisionali e amministrativi; non c'è traccia, però, di studi su come progettare modelli linguistici adeguati a garantire il diritto di accesso civico.

Per questa ragione le risorse linguistiche sono sempre più necessarie per progettare modelli linguistici di fondazione per l'AI generativa per l'italiano istituzionale. È necessario, però, approfondire lo studio sullo sviluppo degli schemi di metadatazione e di annotazione seguendo i principi FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) in modo che sia possibile condividere i prodotti e i risultati della ricerca all'interno della comunità scientifica collaborando anche con le amministrazioni pubbliche.

Nell'ambito del progetto PRIN 2020 "VerbACxSS: su verbi analitici, complessità, verbi sintetici e semplificazione. Per l'accessibilità", in collaborazione con le altre unità di ricerca, abbiamo avviato il ciclo di seminari "Risorse linguistiche per l'accessibilità. Accessibilità alle risorse linguistiche" con lo scopo di promuovere la discussione scientifica e il confronto con coloro che sviluppano i servizi pubblici digitali.

RIFERIMENTI BIBLIOGRAFICI

Brunato D. (2015), *A study on linguistic complexity from a computational linguistics perspective. a corpus-based investigation of italian bureaucratic texts*. Tesi di dottorato non pubblicata, Università di Siena.

Cortelazzo M. A. (1998), "Semplificazione del linguaggio amministrativo", in *Quaderni del Comune di Trento*. Progetti, 3.

Cortelazzo M. A. (2021), *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*, Carocci, Roma.

Cortelazzo M. A., Pellegrino F., Viale M. (1999), *Semplificazione del linguaggio amministrativo. Esempi di scrittura per le comunicazioni ai cittadini*, Comune di Padova.

Gross M. (1981), "Les bases empiriques de la notion de prédicat sémantique", in *Langages*, 63, numero monografico: *Formes syntaxiques et prédicats sémantiques*, a cura di Guillet A., Leclère C., pp. 7-52.

Manning C. D., Surdeanu M., Bauer J., Finkel J., Bethard S. J., McClosky D. (2014), "The Stanford CoreNLP Natural Language Processing Toolkit", in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Mel'čuk I. (2004), "Verbes supports sans peine", in *Lingvisticae Investigationes*, 27, 2, pp. 203-217.

Miliani M., Auriemma S., Alva-Manchego F., Lenci A. (2022), "Neural Readability Pairwise Ranking for Sentences in Italian Administrative Language", in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* ,Volume 1: Long Papers, pp. 849-866.

Passaro L. C., Lenci A. (2015), "Extracting terms with extra", in *Proceedings of EUROPHRAS 2015*, Tradulex, pp. 188-196.

Sabatini F. (1998), "«Rigidità-esplicitezza» vs «elasticità-implicitezza»: possibili parametri massimi per una tipologia dei testi", in Skytte G., Sabatini F. (a cura di), *Linguistica testuale comparativa. In memoriam Maria Elisabeth Conte*. Atti del Convegno interannuale della Società di Linguistica Italiana (Copenaghen, 5-7 febbraio 1998), Museum Tusculanum Press, Copenaghen, pp. 141-172.

Tonelli S., Aprosio A. P., Saltori F. (2016), "SIMPITIKI: A simplification corpus for Italian", in *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*, Accademia University Press, Torino.

Vellutino D., Marano F., Elia A. (2012), "L'italiano istituzionale e le sue varietà d'uso pubblico. Aspetti lessicali nei tipi di testo d'informazione e comunicazione delle pubbliche amministrazioni", in Bianchi P., De Blasi N., De Caprio C., Montuori F. (a cura di), *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali*, Franco Cesati Editore, Firenze, pp. 539-550

Vellutino D. (2014), "Esercizi di stile per il diritto di accesso civico", in Ruffino G., Macaluso F. P. (a cura di), *La lingua variabile nei testi letterari, artistici e funzionali contemporanei. Analisi, interpretazione, traduzione*. Atti del XIII Congresso della Società Internazionale di Linguistica e Filologia Italiana, Centro studi filologici e linguistici siciliani, Palermo, pp. 1-16.

Vellutino D. (2015), "Risorse linguistiche e Open Data per la comunicazione pubblica della gestione dei rifiuti urbani", in Vellutino D., Zanola M.T., *Comunicare in Europa. Lessici istituzionali e terminologie specialistiche*, EDUCatt - Ente per il Diritto allo studio universitario dell'Università Cattolica, Milano,pp. 217-245

Vellutino D., Maslias R., Rossi F. (2016), "Verso l'interoperabilità semantica di IATE. Studio preliminare per il dominio *Gestione dei rifiuti urbani*", in Zanola M. T., Diglio C., Grimaldi C., *Terminologie specialistiche e diffusione dei saperi*, EDUCatt - Ente per il Diritto allo studio universitario dell'Università Cattolica, Milano, pp. 221-240.

Vellutino D. (2018), *L'italiano istituzionale per la comunicazione pubblica*, il Mulino, Bologna.

Vellutino D. (2021), Insegnare gli usi dell'italiano istituzionale per la comunicazione pubblica", in *Lingue e Linguaggi*, 41, pp. 279-296.

Vellutino (in stampa), "Italiano istituzionale", in Vedovelli M., Serena E. (a cura di), *Dizionario dell'italiano L2: insegnamento, apprendimento, ricerca*, Pacini Editore, Pisa.

Wilkinson M. D. *et al.* (2016), The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18.