Cov-I-Cor: un corpus di italiano istituzionale relativo alla gestione dell'emergenza sanitaria

Laura Occhipinti

1. Introduzione

La situazione di emergenza legata alla pandemia da Covid-19 ha messo in evidenza la necessità di un'indagine sul linguaggio istituzionale, varietà della lingua nazionale relativa alle comunicazioni ufficiali delle istituzioni.

Un'analisi del linguaggio istituzionale basata su *corpus* può essere interessante per evidenziare strategie e fenomeni linguistici problematici e/o complessi che potrebbero aver influito su una corretta comunicazione e dunque comprensione dei testi da parte dei cittadini. A questo proposito è stato costruito il *corpus* Cov-I-Cor, *corpus* di italiano istituzionale relativo alla gestione dell'emergenza sanitaria.

La comunicazione è sempre centrale in ogni tipologia di interazione che preveda uno scambio di informazioni efficace e lo è, ancor di più, in contesti di emergenza caratterizzati dall'incertezza. In queste situazioni, infatti, alla chiarezza espositiva, e dunque al messaggio, corrisponde una vera e propria risposta attiva da parte dei destinatari che può influire fortemente sull'andamento dell'emergenza stessa. Raffaella Bombi in *Comunicazione istituzionale e Covid-19 tra ricerca e formazione* sottolinea che «la pandemia, come numerosi altri eventi politici, economici, sociali e le trasformazioni scientifico-tecnologiche, […] ha riportato alla luce il tema della *crisis* e *risk communication*, con particolare attenzione per la comunicazione tra lo Stato e i cittadini»[1]. Il tema non

1 Bombi 2021, p. 5.

DOI 10.35948/DILEF/978-88-6032-750-5.14

Laura Occhipinti, Alma Mater Studiorum Università di Bologna - Dipartimento di Filologia Classica e Italianistica (FICLIT), laura.occhipinti3@unibo.it, 0009-0007-8799-4333

è nuovo dato che l'Organizzazione Mondiale della Sanità (Oms), già a partire dal 2017, ben prima quindi della diffusione del Covid-19, aveva evidenziato la necessità di una comunicazione del rischio chiara e non ambigua, nonostante la consapevolezza dei contesti incerti in cui avviene, attraverso la pubblicazione delle linee guida relative alla comunicazione in contesti di emergenza. Il documento, *Communicating risk in public health emergencies A WHO (World Health Organization) guideline for emergency risk communication (ERC) policy and practice*, facilmente reperibile sul sito dell'OMS[2], è stato elaborato per i responsabili politici e decisionali nelle situazioni di emergenza e pone al centro numerosi elementi e strategie che sono stati talvolta sottovalutati da parte delle istituzioni o non sempre messi in pratica nel giusto modo, in questa, come in altre occasioni di emergenza.

La pianificazione di un testo e la comprensione possono infatti influire fortemente sul successo (o insuccesso) delle indicazioni e informazioni fornite alla popolazione per rallentare e/o fermare la diffusione dei contagi. Le istituzioni, dunque, sono centrali in processi comunicativi che dovrebbero mirare al coinvolgimento e alla rassicurazione della popolazione colpita, che ha bisogno di sentirsi coinvolta e di fidarsi degli "addetti ai lavori", nazionali, ma anche locali, esistenti per la risposta all'emergenza.

Si è ritenuto dunque interessante costruire una risorsa che miri a raccogliere i testi istituzionali nazionali, scritti, con cui è stata gestita l'emergenza sanitaria. Alle scelte relative alla tipologia di italiano istituzionale preso in esame (§ 2), punto non poco problematico considerando l'ampia varietà del campo di analisi, seguirà una descrizione dettagliata del *corpus* (§ 3), con le diverse fasi di costruzione e raccolta dei testi (§ 4). Si concluderà con una panoramica delle possibilità di analisi da poter condurre su questo *corpus* e sugli obiettivi ultimi del mio lavoro di ricerca (conclusioni).

2 Si rimanda al sito dell'Oms per la lettura integrale del documento https://apps.who.int/iris/bitstream/handle/10665/259807/9789241550208-eng.pdf?sequence=2&isAllowed=y.

2. Il linguaggio istituzionale

Per avere un campione rappresentativo e bilanciato della varietà di italiano istituzionale che si vuole analizzare, sono necessarie una serie di scelte, qualitative *in primis* e quantitative[3] poi, abbastanza problematiche. In effetti, quello del linguaggio istituzionale è un campo di indagine molto ampio: più che di italiano istituzionale è possibile, se non necessario, parlare di linguaggi istituzionali, a seconda «degli scopi e delle sfere di attività delle specifiche organizzazioni del settore pubblico e privato nelle loro comunicazioni ufficiali»[4]. L'attenzione alla selezione del dato linguistico si rivela essenziale dal momento che quest'ultimo rappresenta «l'evidenza empirica su cui fondare lo sviluppo di modelli e teorie linguistiche»[5].

Ogni linguaggio istituzionale è caratterizzato da una serie di scelte linguistiche e stilistiche legate al contesto extralinguistico, ossia dipendenti dai fattori reali in cui viene utilizzato, che ne determinano forma linguistica e finalità. In primis, è importante ricordare che i destinatari a cui la comunicazione pubblica si rivolge sono molteplici e vanno dal singolo cittadino alla collettività. Con il termine *collettività* si intende «una pluralità di persone considerate nel loro insieme»[6], come un tutt'uno e dunque si fa riferimento sia all'insieme dei cittadini che alle altre istituzioni, enti o gruppi di persone autonomamente individuabili. Sebbene dunque «tutti i linguaggi istituzionali [siano] mezzi espressivi funzionali a far conoscere i contenuti informativi degli atti giuridici e amministrativi al fine di far attuare le loro disposizioni ai vari livelli di governo»[7], non è possibile pensarli come un agglomerato unico.

3 Lenci 2005, scheda informativa 1.3 (*Costruire un corpus*).

4 Vellutino 2018, p. 84.

5 Lenci 2005, p. 23.

6 Si veda https://www.treccani.it/vocabolario/collettivita/.

7 Vellutino 2018, p. 85.

La prima macro-distinzione di cui bisogna tener conto nel quadro teorico necessario alla definizione dei criteri di selezione che precede la raccolta vera e propria dei testi è quella tra linguaggi istituzionali speciali e mediali[8]. Con i primi ci si riferisce ai linguaggi normativo e amministrativo che, per definizione, hanno delle caratteristiche più rigide legate al codice del diritto che li regola. I testi di queste varietà sono classificati come «vincolanti»[9] da Sabatini, in relazione al rapporto tra autore e destinatario: «l'autore determina un vincolo più o meno forte nell'interpretazione del testo da parte del destinatario»[10].

Con l'espressione linguaggi istituzionali mediali, invece, si intendono i linguaggi più propriamente legati alla comunicazione e informazione sociale: i linguaggi che si propongono di dare informazioni di pubblica utilità, promuovendo la cittadinanza attiva; il linguaggio giornalistico, finalizzato a informare gli organi di stampa; il linguaggio pubblicitario, finalizzato a comunicare l'identità istituzionale e utilizzato dalle campagne di comunicazione pubblica per far conoscere diritti, servizi, opportunità e sensibilizzare ai temi di interesse generale[11].

Si è deciso di soffermarsi sui linguaggi istituzionali vincolanti, pensando alla necessità di un intervento che miri ad affrontare il problema "alla base", a partire dai testi sorgente che regolano la collettività. Ad avvalorare questa scelta risuonano le parole presenti nella Direttiva sulla semplificazione dei testi amministrativi, redatta dal Dipartimento della funzione pubblica nel 2005:

> I numerosi atti prodotti dalle pubbliche amministrazioni, sia interni (circolari, ordini di servizio, bilanci) sia esterni, devono prevedere l'utilizzo di un linguaggio comprensibile, evitando espressioni burocratiche e termini tecnici. Anche gli atti amministrativi in senso stretto, che producono effetti giuridici diretti e immediati per i destinatari, devono essere progettati e scritti

8 *Ibidem*.

9 Sabatini 1990, p. 97.

10 Raso 2005, p. 57.

11 Per una disamina completa si rimanda a Vellutino 2018, pp. 95-111.

pensando a chi li legge. […] Devono, perciò, essere sia legittimi ed efficaci dal punto di vista giuridico, sia comprensibili, cioè di fatto efficaci, dal punto di vista comunicativo[12].

Per il concetto stesso di vincolante «il destinatario tendenzialmente non può dare contributi interpretativi al testo e quindi esso deve fornire al lettore tutte le istruzioni per una decodifica precisa»[13]: il linguaggio deve essere chiaro, lineare e comprensibile.

Per bilanciare la raccolta dei testi ai fini della rappresentatività del campione, si è provato a distinguere l'ambito amministrativo, inteso come linguaggio della Pubblica Amministrazione, e quello giuridico, più genericamente inteso come linguaggio che attiene alla sfera giuridico-normativa, cercando di individuare le tipologie di testi connessi a questi due campi. Ad oggi, però, manca una classificazione generale delle tipologie testuali che rientrano nei linguaggi istituzionali vincolanti[14] e spesso c'è confusione o sovrapposizione tra gli ambiti: «i testi del diritto e dell'amministrazione costituiscono un oggetto "dai confini più permeabili e sfrangiati", non sempre facilmente circoscrivibili nelle griglie ristrette di precise tipologie testuali»[15]. In effetti, il confine tra linguaggio giuridico e amministrativo è molto difficile da delimitare, dal momento che molti testi si collocano a metà, basti pensare alle ordinanze, che presentano caratteristiche normative e applicative. Dal punto di vista linguistico, inoltre, il confine sembra ancora meno netto: il linguaggio amministrativo è molto influenzato dal linguaggio

12 Dipartimento della funzione pubblica, *Direttiva sulla semplificazione dei testi amministrativi*, p. 2.

13 Raso 2005, p. 57.

14 Per la classificazione più esaustiva presente in letteratura si rimanda a Viale 2008. La classificazione proposta risulta abbastanza soddisfacente per quanto riguarda i testi amministrativi, ma alcune tipologie testuali possono rientrare in entrambi gli ambiti. Il criterio utilizzato per distinguerli, che si basa sulla tipologia di destinatario a cui si rivolgono, non è sufficiente: ad entrambi gli ambiti corrispondono sia testi "esterni" che testi "interni".

15 Lubello 2017, p. 11.

del diritto e questo implica che molte forme linguistiche ormai risiedano in entrambe le varietà.

Per i fini più ampi della mia ricerca, che mira ad automatizzare le procedure di semplificazione dei linguaggi istituzionali vincolanti, non è dunque necessaria una separazione tra le due varietà[16]. D'altronde, come sostenuto da Raso, i testi amministrativi, anche definiti burocratici, «risentono delle caratteristiche dei testi che ne costituiscono la fonte, cioè i testi legislativi e l'ampia produzione di norme che a essi si collega»[17].

Diversi sono i libri e gli articoli scientifici che trattano il linguaggio giuridico e quello amministrativo come un unico oggetto, basti citare a titolo d'esempio un saggio di Piero Fiorelli del 1994[18] o il più recente lavoro di Lubello del 2017:

> si tratta di settori strettamente contigui, [il linguaggio amministrativo], peraltro, è storicamente una variante particolarmente estesa del linguaggio giuridico, con cui intrattiene un legame strutturale, dal momento che quest'ultimo rappresenta la fonte primaria della normativa burocratica [...]. A differenza del linguaggio giuridico, quello amministrativo [...] non è tecnico-specialistico stricto sensu e conosce un variegato spettro di realizzazioni testuali, di contesti d'uso e di destinazioni, applicandosi a un ambito molto ampio di comunicazione: ad accomunare testi molto eterogenei non sono né l'emittente né il destinatario, ma un insieme di scelte linguistiche che delineano un codice scritto formale, tendenzialmente conservativo nelle sue strutture.

16 Dal momento che l'obiettivo più ampio della ricerca è quello di costruire una risorsa parallela semplificata che possa contribuire all'avanzamento dell'automatizzazione dei processi di semplificazione, a interessarci sono principalmente i fenomeni linguistici presenti, come ribadito nel corpo del testo, in entrambi gli ambiti linguistici. Inoltre, è bene ricordare che i dati necessari per le procedure di semplificazione automatica sono, ad oggi molto pochi, in particolare per la lingua italiana. Questo rappresenta sicuramente uno dei problemi centrali nell'avanzamento e nello sviluppo di questo *task* linguistico.

17 Raso 2005, p. 29.

18 Fiorelli 1994.

Si è deciso dunque di bypassare questa divisione e di procedere alla raccolta dei testi su base tematica, in modo da rendere il *corpus* rappresentativo e bilanciato per un'analisi linguistica della gestione pandemica da parte delle istituzioni nazionali. L'interesse di base, infatti, è quello di evidenziare i punti di oscurità e complessità linguistica presenti nelle decisioni che sono state prese e messe in atto, sia in ambito normativo che amministrativo, con lo scopo di proporne una semplificazione, necessaria per rendere i testi più accessibili e fruibili per la popolazione. Questo obiettivo, già teorizzato a partire dagli anni '90[19], si rivela cruciale in questa fase storica, dal momento che la comprensione delle misure adottate è alla base della democraticità di un Paese ed è sinonimo di sostenibilità sociale.

3. *Corpus*

Il linguaggio istituzionale che si è scelto di analizzare è dunque quello relativo alle istituzioni nazionali. Questa decisione si è rivelata essenziale per ridurre il campo di indagine e condurre un'analisi più focalizzata. Si è deciso di escludere i testi prodotti da istituzioni regionali, provinciali e comunali, per non incorrere in fenomeni ipotetici legati alla variazione diatopica del testo, nonostante ci sia la consapevolezza che questa varietà ha una possibilità minore, seppur non nulla, di incorrere in problematiche di questo tipo[20]. Inoltre, in un'ottica di sem-

19 Il tema della semplificazione normativa e amministrativa è al centro di un dibattito molto ampio, non solo contemporaneo e relativo a un'unica area: interessa tutte le amministrazioni occidentali. In Italia, il tema viene portato avanti e dibattuto principalmente da professori universitari che, a partire dagli anni '90, con la figura di Sabino Cassese, hanno proposto una serie di iniziative e linee guida per rendere i testi più fruibili per i cittadini. Le varie iniziative «non si sono mai trasformate, però, in una vera e propria campagna, caratterizzata da sistematicità e continuità» (Cortelazzo 2021, p. 64). Per una rassegna sulle iniziative passate si veda Cortelazzo 2021 pp. 63-74 e Lubello 2017, pp. 98-110.

20 Diversi sono gli studi diacronici e sincronici che sono stati condotti sui testi istituzionali di una regione o città specifica. Si cita a titolo di esempio il saggio di Nobili 2021.

plificazione che si rivolga potenzialmente[21] all'intera popolazione, si è deciso di concentrarsi sui testi a cui tutti i cittadini italiani, o residenti in Italia, sono stati esposti. Sono state inoltre escluse le audizioni o le conferenze-stampa, sebbene siano state frequenti durante questo "biennio pandemico", dal momento che fanno riferimento a un canale di trasmissione più vicino al campo dell'oralità, evitando così variazioni problematiche legate all'asse diamesico.

I testi raccolti, dunque, sono tutti i testi nazionali ufficiali scritti, prodotti per fronteggiare l'emergenza sanitaria, in un arco di tempo che va dal 21 gennaio 2020, data delle prime informazioni relative al Covid provenienti dalla Cina, ancor prima che arrivasse in Italia[22], al 31 marzo 2022, che segna la fine dello stato di emergenza nel nostro Paese.

Nell'arco di questi due anni, o poco più, sono stati prodotti – esclusivamente a livello nazionale – e raccolti 904 testi, atti prodotti da diverse istituzioni, tra cui Governo, presidente del Consiglio, presidente della Repubblica, Protezione Civile, Aifa (Agenzia italiana del farmaco) e i diversi ministeri.

Si è deciso di selezionare questi testi perché riconosciuti come fonti ufficiali delle comunicazioni a tema Covid. La gran parte di questi atti, in effetti, è stata pubblicata sulla Gazzetta Ufficiale della Repubblica Italiana. Quest'ultima può essere considerata la fonte ufficiale di conoscenza delle norme in vigore in Italia: è uno «strumento di diffusione, informazione e ufficializzazione di testi legislativi, atti pubblici e privati che devono giungere con certezza a conoscenza dell'intera comunità»[23], come si legge nella schermata ufficiale della Gazzetta. Affinché

21 È necessaria l'aggiunta dell'avverbio *potenzialmente* dal momento che si è consapevoli che parlare di semplificazione in senso assoluto potrebbe essere limitante. Consapevoli che la semplificazione dovrebbe essere fatta *ad personam*, o almeno prestando attenzione alle necessità di gruppi più o meno ampi di individui, sulla base del loro *background* culturale, si è optato per una semplificazione generale guidata dalle linee guida prodotte durante questi anni dalle istituzioni e dai progetti universitari portati avanti.

22 Il primo caso ufficialmente certificato sul suolo italiano risale al 29 gennaio 2020, con relativo annuncio di Giuseppe Conte.

23 Si rimanda al sito ufficiale https://www.gazzettaufficiale.it/.

una legge ufficiale o un provvedimento entrino in vigore è necessaria la pubblicazione in Gazzetta, motivo per cui i testi che vengono pubblicati dovrebbero essere accessibili all'intera popolazione.

Secondo l'art. 54 della Costituzione italiana, in effetti, «tutti i cittadini hanno il dovere di essere fedeli alla Repubblica»[24], ma, per esserne fedeli, è necessario che comprendano tutti i principi e le leggi che regolamentano la vita della e nella Repubblica. È utile ancora ricordare l'art. 5 del nostro Codice penale: «Nessuno può invocare a propria scusa l'ignoranza della legge»[25]. Questa norma riprende il principio *ignorantia iuris (legis) non excusat* e sembra non ammettere eccezioni. A questo proposito è intervenuta la Corte costituzionale che ha ammesso la scusabilità dell'ignoranza della legge[26] nei casi in cui questa ignoranza sia inevitabile, citando a esempio casi di assoluta oscurità legislativa, nonché linguistica. Il confine tra ciò che è oscuro o chiaro non è così nettamente delineabile e non può prescindere dal contesto globale in cui si inserisce: una disposizione che potrebbe essere chiara per qualcuno, con un certo grado di istruzione e un dato contesto socioeconomico, potrebbe essere del tutto oscura per qualcun altro con caratteristiche totalmente opposte.

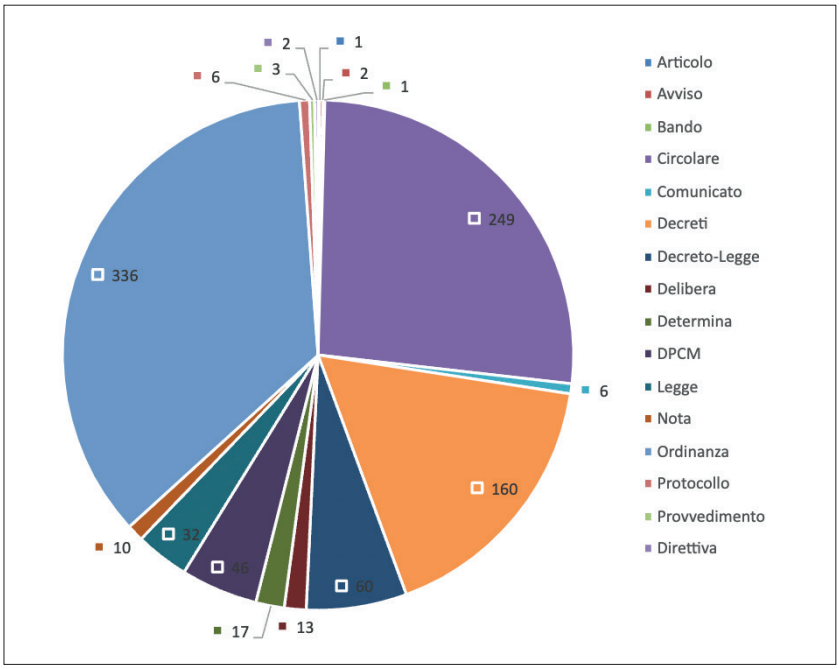
Cov-I-Cor, dunque, può essere descritto come un *corpus* monolingue (italiano), scritto, specialistico[27], sincronico. Contiene al suo interno 904 testi, per un numero complessivo di parole pari a circa 4.000 *token*.

24 Costituzione italiana, art. 54. Si rimanda a https://www.senato.it/istituzione/la-costituzione.

25 Codice penale, art. 5. Si rimanda a https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.versione=2&art.idGruppo=1&art.flagTipoArticolo=1&art.codiceRedazionale=030U1398&art.idArticolo=5&art.idSottoArticolo=1&art.idSottoArticolo1=10&art.dataPubblicazioneGazzetta=1930-10-26&art.progressivo=0.

26 Si veda la sentenza n. 364 del 1988 prodotta dalla Corte costituzionale.

27 Si parla di *corpus* specialistico perché è rappresentativo di una data varietà dell'italiano, ma si vuole chiarire che il linguaggio di cui ci stiamo occupando è sicuramente un linguaggio settoriale ma non del tutto speciale, dal momento che contiene al suo interno diversi linguaggi speciali, afferenti ai diversi campi di cui si occupa.



4. Raccolta testi

Dopo aver fissato i criteri di progettazione del *corpus*, è stato necessario procedere all'acquisizione dei testi: la raccolta è avvenuta con diverse modalità. È importante sottolineare che tutti i testi presi in considerazione sono di natura pubblica, motivo per cui non è stato necessario richiedere permessi particolari legati alla questione, problematica in questo ambito, del *copyright* e, inoltre, a partire dalla legge n. 69 del 18 giugno 2009[28] e dal decreto legislativo del 7 marzo 2005 n.

28 La legge n. 69/2009 e, in particolare, l'art. 32, ha stabilito che dal 1° gennaio 2010 «gli obblighi di pubblicazione di atti e provvedimenti amministrativi aventi effetto di pubblicità legale si intendono assolti con la pubblicazione, da parte delle am-

82[29], tutti i documenti di interesse per questa ricerca sono pubblici e consultabili online. Tutto ciò ha sicuramente facilitato la raccolta dei testi che è avvenuta, *in primis*, attraverso l'utilizzo di BootCat (Bootstrapping Corpora and Terms)[30], la cui funzione principale è quella di creare in maniera semi-automatica *corpora* specialistici di dimensioni medio-piccole estraendo i testi dal web sulla base di *query* fornite dal linguista. Questo strumento è stato utilizzato per provare ad automatizzare il processo di ricerca dei testi dal web: sono state inserite delle *tuple*[31] di parole chiave da utilizzare come *query* che successivamente sono state trovate e aperte nel browser. Attraverso questa ricerca sono stati evidenziati rapidamente i siti attinenti alle parole fornite in input ed è iniziata l'analisi puntuale di ogni singolo sito web; sono stati così progressivamente esaminati tutti i siti web delle Istituzioni a livello nazionale. Il primo sito emerso dalla ricerca con BootCat è stato sicuramente quello del Governo, o meglio della Presidenza del Consiglio dei ministri, a cui sono seguiti quello del Parlamento, della Presidenza della Repubblica, e i siti dei singoli ministeri, a partire dal Ministero della Salute, punto di riferimento centrale in questa emergenza pandemica. Hanno affiancato il lavoro del Ministero della Salute sicuramente la Protezione Civile, l'Istituto superiore di sanità e l'Aifa, motivo per cui si è ricorso alla consultazione anche dei siti di queste istituzioni. In effetti, numerosi sono stati anche i testi promulgati dai non pochi commissari nominati durante questa fase di emergenza.

Tutti i testi rilevati erano già in formato elettronico; si è proceduto a scaricarli e a categorizzarli per data, tipologia e istituzione di riferi-

---

ministrazioni e degli enti pubblici obbligati, nei propri siti informatici, o nei siti informatici di altre amministrazioni ed enti pubblici obbligati, ovvero di loro associazioni».

29 Decreto legislativo n. 82 del 7 marzo 2005; si rimanda alla fonte ufficiale per un approfondimento, https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-03-07;82.

30 Baroni-Bernardini, 2004.

31 Si rimanda alla pagina ufficiale di BootCat relativa a "*Corpus creation mode*": https://docs.sslmit.unibo.it/doku.php?id=bootcat:help:corpus_creation_mode#custom_tuples_advanced.

mento in un documento Excel; questo ha permesso di avere una panoramica chiara del materiale su cui lavorare.

Dopo aver raccolto i materiali digitali, è stato necessario trasformarli nel formato più adeguato al trattamento computazionale, provvedendo alla loro codifica e annotazione. In effetti i testi scaricati e/o salvati non erano in formato txt, ma tutti in formato pdf. A volte i testi disponibili erano già in formato pdf, altre volte è stato necessario salvare le pagine html. Al formato pdf, però, corrispondevano sia documenti in formato testuale, sicuramente meno problematici per la conversione e normalizzazione testuale, sia in formato OCR: ci si è trovati dinanzi a immagini scannerizzate di documenti, la cui conversione è più problematica e presenta documenti che successivamente necessitano di una "pulizia" maggiore.

È stato necessario trasformare tutti i testi in formato txt, un documento di testo standard contenente testo non formattato[32]. Questa operazione è essenziale sia ai fini dell'etichettatura del testo sia per rendere consultabile i testi senza incorrere in problemi di interscambiabilità e riusabilità: «presenta l'innegabile vantaggio di poter essere gestito da programmi diversi indipendentemente dal sistema operativo»[33].

La conversione dei testi è avvenuta in diversi modi. La maggior parte dei documenti è stata trasformata mediante l'utilizzo di una libreria Python chiamata PyPDF2[34] che ha permesso la conversione automatica dei file dal formato pdf a txt. Per alcuni pdf, che erano stati generati da Mac, non è stato possibile convertire i file in questo modo, a causa di problemi di incompatibilità. Pertanto, si è ricorsi all'utilizzo di BootCat che permette di trasformare facilmente i testi in questo

32 Si ricorda che «i formati come doc (il formato dei file di Microsoft Word) o pdf (il formato dei file di Acrobat Adobe) [...] presentano la peculiarità di strutturare un testo digitale in maniera estremamente fruibile per il lettore umano, con tutte le informazioni di formattazione utili alla sua composizione editoriale e alla sua visualizzazione» (Lenci 2005, p. 67).

33 Lenci 2005, p. 67.

34 https://pypi.org/project/PyPDF2/.

formato. Per i file, invece, contenenti immagini di testo scannerizzate, è stato necessario ricorrere a sistemi di riconoscimento OCR. Essendo pochi i documenti in questo formato, è stato utilizzato Adobe Pro[35] per riconoscere i caratteri e per salvarli in un nuovo file con la stessa estensione.

È stato necessario poi ripulire i testi, intervenendo su tutti quegli elementi che potrebbero interferire con la consultabilità e linearità del testo:

- sono state eliminate tutte le tabelle presenti, i grafici, le figure, con le relative *caption*, gli indici, le appendici e i numeri di pagina: tutti quegli elementi di difficile leggibilità per un computer e che rappresentano dunque rumore, soprattutto nel cambio codifica;
- tutte le note sono state posizionate alla fine del documento, in modo che non interrompessero il corpo centrale del testo;
- le parole che sono state separate erroneamente nella trasformazione dei documenti, o che si interrompevano nell'andata a capo tra una riga e l'altra, sono state ripristinate;
- sono stati cancellati tutti i simboli relativi a codifiche particolari e dunque non appartenenti al formato Unicode UTF-8;
- sono state sostituite tutte le vocali apostrofate[36] con vocali accentate e tutti i *bullet* particolari nei punti elenco con il carattere "-";
- sono stati cancellati tutti i tab e gli spazi "superflui" presenti nel corpo del testo.

Questa fase di normalizzazione del testo è avvenuta in gran parte manualmente, ma alcune delle operazioni sono state effettuate attraverso il riconoscimento di pattern ricorrenti nei documenti mediante la libreria re[37] di Python, che consente, attraverso le espressioni regolari di eliminare o trasformare i pattern individuati nel *corpus*.

35 https://www.adobe.com/it/acrobat/how-to/ocr-software-convert-pdf-to-text.html.

36 Formattazione utilizzata dalla Gazzetta Ufficiale.

37 https://docs.python.org/3/library/re.html.

Una volta ripuliti, i testi sono stati annotati in formato CONLL-U: schema di annotazione afferente alle Universal Dependencies[38]. I testi sono stati automaticamente annotati mediante l'utilizzo della catena di analisi Stanza[39] e della libreria os[40], necessaria per iterare sui file presenti nelle diverse cartelle. Ogni file, dunque, è stato letto e poi annotato su un nuovo *file script* e presenta il testo segmentato in frasi e per ogni frase c'è un'annotazione in colonne con dati relativi a ID, FORM, LEMMA, UPOS, XPOS, FEATS, HEAD, DEPREL DEPS, MISC.

Si è optato per questo schema di annotazione per due motivi principali: il primo fa riferimento al *task* che ci proponiamo, relativo alla semplificazione linguistica automatica; il secondo fa riferimento alla rappresentatività interlinguistica delle UD.

Diversi studi in letteratura, infatti, mostrano che la semplificazione automatica basata su dipendenze (*dependency-based*) ottiene risultati migliori rispetto alle semplificazioni *constituency-based*[41]. Inoltre, la costruzione del *corpus* parallelo che ne seguirà potrebbe essere di aiuto per lo sviluppo di sistemi di semplificazione automatica anche per lingue diverse dall'italiano, per l'interscambiabilità dei dati.

5. Conclusioni

In questo lavoro è stato presentato e descritto il *corpus* Cov-I-Cor. Dopo aver motivato la necessità di un *corpus* specialistico di questo tipo, sono stati illustrati nel dettaglio i criteri teorici che hanno portato alla deli-

38 Le *Universal Dependencies* costituiscono un tentativo di unificazione e conformità rappresentativa interlinguistica: si mira a un'annotazione coerente che riesca a fornire un inventario di categorie e linee guida che facilitino lo sviluppo di un parser multilingue. Per un approfondimento si consulti https://universaldependencies.org/.

39 https://stanfordnlp.github.io/stanza/index.html.

40 https://docs.python.org/3/library/os.html#os-file-dir.

41 Per un approfondimento si veda Siddharthan 2010 e Siddharthan 2011.

neazione del campione in esame e la metodologia con cui questi testi sono stati raccolti, ripuliti e annotati.

La costruzione di questa risorsa rappresenta solo il primo step di un lavoro più ampio che mira alla costruzione di un *corpus* parallelo, a partire da una sottosezione di Cov-I-Cor, che presenterà testi originali e semplificati allineati. Il proposito è quello di automatizzare alcuni dei meccanismi di semplificazione proposti e di fornire uno strumento che possa aiutare il cittadino – nella comprensione dei testi ufficiali – e chi scrive ad avere un modello concreto di testo semplificato. Questo strumento può essere molto utile sia per i propositi di semplificazione in generale, sia in questa fase transitoria in cui le numerose linee guida fornite ai "comunicatori ufficiali" non sono state realmente messe in atto nel processo di ideazione e scrittura dei testi.

Bibliografia

Alebachew *et al.* 2017 = World Health Organization, *Communicating risk in public health emergencies. A WHO guideline for emergency risk communication (ERC) policy and practice*, Geneva, Schweiz, 2017.

Atkins-Clear-Ostler 1992 = Sue Atkins, Jeremy Clear, Nicholas Ostler, *Corpus Design Criteria*, in «Literary and Linguistic Computing», volume 7, 1992, pp. 1-16.

Baroni-Bernardini 2004 = Marco Baroni, Silvia Bernardini, *BootCaT: Bootstrapping Corpora and Terms from the Web*, in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.

Baroni-Ueyama 2006 = Marco Baroni, *Building general- and special-purpose corpora by Web crawling*, in Proc. *13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, Tokyo, 2006, pp. 31-40.

Biber 1993 = Douglas Biber, *Representativeness in corpus design*, in «Journal of Literary and Linguistic Computing», 8 (4), 1993.

Bombi 2021 = *La comunicazione istituzionale ai tempi della pandemia. Da sfida a opportunità*, a cura di Raffaella Bombi, *Lingue, culture e testi*, Roma, Il Calamo, 2021.

Brunato-Dell'Orletta-Venturi 2022 = Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, *Linguistically-Based Comparison of Different*

Approaches to Building Corpora for Text Simplification: A Case Study on Italian, «Frontiers in Psychology», vol. 13, Switzerland, Frontiers Research Foundation, 2022.

Brunato *et al.* 2019 = Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, *Trattamento Automatico della Lingua per la comunicazione della Pubblica Amministrazione*, in Proceedings of Ital-IA 2019, 18-19 March, Rome, Italy, 2019.

Cortelazzo 2021 = Michele A. Cortelazzo, *Il linguaggio amministrativo*, Roma, Carocci editore, 2021.

De Marneffe *et al.* 2021= Marie-Catherin de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman, *Universal Dependencies*, in «Computational Linguistics», vol. 47, no. 2, 2021, pp. 255-308.

Dipartimento della funzione pubblica 2005 = Dipartimento della funzione pubblica, *Direttiva sulla semplificazione del linguaggio delle Pubbliche amministrazioni*, Roma, 2005.

Fioritto 1997 = Alfredo Fioritto, *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*, Bologna, il Mulino, 1997.

Lenci-Montemagni-Pirelli 2016 = Alessandro Lenci, Simonetta Montemagni, Vito Pirelli, *Testo e computer. Elementi di linguistica computazionale* (IV edizione), Roma, Carocci editore, 2016.

Lubello 2015 = Sergio Lubello, *Il linguaggio burocratico*, Roma, Carocci editore, 2015.

Nobili 2021 = Claudio Nobili, *Per lo studio dell'italiano burocratico in area campana: ancora sul progetto CUR e presentazione di CorTIBuS*, in Bombi 2021.

Raso 2005 = Tommaso Raso, *La scrittura burocratica*, Roma, Carocci, 2005.

Sabatini 1999 = Francesco Sabatini, *Rigidità-esplicitezza vs elasticità-implicitezza: possibili parametri massimi per una tipologia dei testi*, in *Linguistica testuale comparativa*, a cura di Francesco Sabatini, Gunver Skytte, Copenhagen, Museum Tusculanum Press, 1999.

Siddharthan 2010 = Advaith Siddharthan, *Complex lexico-syntactic reformulation of sentences using typed dependency representations*, in Proceedings of the 6th International Natural Language Generation Conference (INLG 2010), Dublin, Ireland, 2010.

Siddharthan 2011 = Advaith Siddharthan, *Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different Generation Strategie*, in Proceedings of the 13th European Workshop on Natural Language Generation, Nancy, France, 2011.

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

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

Riassunto In questo contributo vengono presentate e discusse le decisioni teoriche e metodologiche che sono state prese per la costruzione del *corpus* Cov-I-Cor, una risorsa linguistica che raccoglie i testi istituzionali utilizzati per gestire l'emergenza sanitaria da Covid-19.

Abstract This paper presents and discusses the theoretical and methodological decisions that were made in the construction of the Cov-I-Cor *corpus*. Cov-I-Cor is a linguistic resource that collects institutional texts used to manage Covid-19 health emergency.