Annotazione semi-automatica di due corpora rappresentativi di una varietà linguistica attraverso lo schema delle *Universal Dependencies*

Roberto Cannarella,
matr. [redacted]
Informatica umanistica,
Linguistica computazionale II

Sommario

1. Introduzione: scopi e struttura della relazione

L'annotazione linguistica del testo è un compito cruciale nel panorama della linguistica computazionale. Poiché consiste nella codifica, al livello del testo, di informazioni linguistiche appartenenti a gradi diversi di analisi[1], essa rende di fatto esplicita una serie di informazioni 'latenti', che sorreggono il testo in quanto unità discorsiva, logica e sintattica; informazioni che, di per sé, sono tutte più o meno evidenti ai fruitori "umani" del testo, ma che non sono accessibili invece ai fruitori "non umani" dello stesso, cioè le macchine. Queste ultime, proprio grazie all'*encoding* linguistico appena descritto, possono, attraverso algoritmi di *machine learning* che compiono classificazione probabilistica, 'imparare'[2] a classificare, ad esempio, una parola in quanto soggetto (analisi sintattica) o come sostantivo (analisi morfosintattica[3]). Il risultato di questo processo di apprendimento è la costruzione di un *language model*, che è poi utilizzabile per analizzare e annotare nuovi testi. Il modello così addestrato è, dunque, in grado di accedere a un campionario di informazioni linguistiche variegate, cosa che ha risvolti applicativi numerosi e differenti, che vanno, ad esempio, dal monitoraggio dell'interlingua[4] a nuovi modi di definire la leggibilità di un testo[5]; la disponibilità di queste informazioni linguistiche è anche utile "di per sé", perché permette l'esplorazione l'analisi quantitativa dei tratti linguistici rilevanti dei testi (*linguistic profiling*), specie se appartenenti a domini e registri non-standard.

Proprio testi appartenenti a un dominio non-standard sono quelli facenti parte di due corpora, ParlaMint-IT_2015-03-11-LEG17-Sed-407-5 e ParlaMint-IT_2020-03-11-LEG18-Sed-200-5 (da qui in poi, rispettivamente, solo **P.2015** e **P.2020**), la cui annotazione semi-automatica, condotta insieme a Luca Baù, è oggetto di questa relazione. I due corpora presentano la trascrizione parziale di due sedute diverse del Senato della Repubblica, fra di loro temporalmente distanti di cinque anni. Sono distanti anche tematicamente (in P.2015 si dibatte sulle adozioni, in P.2020 sulle rivolte in carcere che hanno avuto luogo all'inizio dell'emergenza epidemiologica

---

[1] A. LENCI, S. MONTEMAGNI, V. PIRRELLI, *Testo e computer. Elementi di linguistica computazionale*, Carocci editore, Roma, 2016, p. 211 ss.

[2] L'"apprendimento" consiste, almeno nella sua accezione più basilare e canonica, nell'imparare, da parte del modello, a "dosare" i *parametri* (anche detti *pesi*) associati alle *feature* (solitamente definite a monte) in maniera tale da minimizzare (dopo un certo numero di *epoche*) l'errore empirico riportato dalla funzione di classificazione (*ipotesi*) del modello.

[3] L'annotazione linguistica può avvenire infatti a ciascun livello dell'analisi linguistica e testuale (ivi compresa l'analisi pragmatica), anche se va detto che i livelli ai quali solitamente si realizza l'annotazione sono la morfologia, la sintassi e la semantica. A uno stesso livello di annotazione, poi, le teorie linguistiche seguite per definire *cosa* debba essere annotato e *come* lo si debba annotare possono essere anche molto diverse: si pensi alla differenza fra annotazione sintattica a costituenti e annotazione sintattica a dipendenze, oppure alla differenza fra l'annotazione semantica dei "sensi" di una parola (cfr. i *synset* di WordFrame) e l'annotazione dei *frame* semantici (cfr. FrameNet).

[4] Cfr. S. MONTEMAGNI, F. DELL'ORLETTA, G. VENTURI, *Esplorazioni computazionali nello spazio dell'interlingua: verso una nuova metodologia di indagine*, in R. BOMBI, V. ORIOLES, (a cura di), "Atti del XLVIII Congresso Internazionale di Studi della Società di Linguistica Italiana" (SLI 2014), 2016.

[5] Cfr. K. COLLINS-THOMPSON, *Computational Assessment of Text Readability: A Survey of Current and Future Research*, 2014.

da COVID-19), oltre che linguisticamente (come si vedrà). Ciò che accomuna entrambi i corpora è il loro appartenere a una varietà dell'italiano caratterizzata da tratti discordanti: da una parte sono testi trascritti, quindi appartenenti al parlato; d'altro canto, i tratti tipici del parlato (ad es. la tendenza all'informalità e gli errori dovuti alla produzione *online* dei testi) sono sostituiti da un certo controllo, anche retorico, della forma.

Prima ancora dell'annotazione morfosintattica e sintattica sono state prese delle decisioni comuni in riferimento alla segmentazione (*sentence splitting*), alla tokenizzazione e al riconoscimento delle parole sintattiche dei due corpora. Fatto ciò, è stato possibile passare alla loro annotazione morfosintattica e sintattica. Trattandosi di un processo di annotazione semi-automatica, i corpora sono stati, dopo l'annotazione automatica realizzata attraverso il modello udpide-ud2.5-191206 addestrato sulla *treebank* ISDT (**italian-isdt-ud-2.5-191206.udpipe**[6]), rivisti in maniera manuale e incrementale, considerando tutti i livelli di annotazione previsti dallo schema di annotazione di riferimento, le *Universal Dependencies*[7]: l'analisi morfosintattica (lemmatizzazione, *part-of-speech tagging* e individuazione delle *feature*) e l'analisi sintattica a dipendenze (*dependency/syntactic parsing*). La revisione di questi livelli è stata condotta in primo luogo in maniera individuale da parte dei due annotatori; una volta calcolato l'*inter-annotator agreement* (cfr. *infra*, §1.1), si è proceduto alla creazione delle versioni *gold standard* dei due corpora, coordinando le scelte e le intuizioni linguistiche riguardanti la gestione dei token annotati in modo erroneo o anomalo dal modello[8]. Durante questa fase di passaggio ai *gold*, si è fatto più volte riferimento alla *treebank* ISDT[9] per adattare la gestione di alcuni fenomeni linguistici opachi al modo in cui essi sono stati precedentemente trattati, da altre/i annotatrici/tori, in essa. Le scelte compiute ai vari livelli linguistici per la creazione del *gold* sono descritte nel corso di questa relazione, in particolare nelle varie sotto-sezioni della §2: in §2.1, sono descritte le modifiche apportate in merito a *sentence splitting* e tokenizzazione; in §2.2, le modifiche in merito all'annotazione morfosintattica; in §2.3, sono illustrate invece le modifiche riguardanti l'annotazione a dipendenze. Si segnala che, per le visualizzazioni dei legami di dipendenza sintattica riportate in queste sezioni (soprattutto nella §2.3), è stato utilizzato lo strumento conllu.js[10].

La sezione §3, infine, presenta i risultati dell'uso dei corpora manualmente rivisti come *test set*. Essi, infatti, sono stati utilizzati per la verifica dell'accuratezza di uno stesso modello, **ud-**

[6] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131.
[7] https://universaldependencies.org/.
[8] Come si ridirà in §1.1, va segnalato che la revisione circa il *sentence splitting* e la tokenizzazione è stata sì manuale, ma fin da subito in collaborazione fra i due annotatori, laddove per il resto la revisione è stata, in una prima fase, manuale e totalmente individuale. Questo perché, per potere fare annotare al modello i due corpora e potere poi confrontare gli esiti delle revisioni manuali, condizione fondamentale era che, almeno a questi livelli generali (token e segmentazione), i file dei due annotatori fossero equivalenti.
[9] Consultabile all'indirizzo: http://match.grew.fr/?corpus=UD_Italian-ISDT@2.8.
[10] http://spyysalo.github.io/conllu.js/.

2.5-191206.udpipe, addestrato su due corpora differenti: 1) nel primo caso, il modello addestrato sulla già citata *treebank* ISDT, di carattere generale (**isdt-ud-2.5-191206.udpipe**), il quale corrisponde di fatto al modello utilizzato per l'annotazione automatica; 2) nel secondo caso, il modello addestrato sulla *treebank* POSTWITA, contenente *tweet* in lingua italiana (**italian-postwita-ud-2.5-191206.udpipe**[11]). I risultati di questa analisi sono interpretati qualitativamente nella sotto-sezione §3.1.

1.1. Verso i *gold standard* corpora: il calcolo dell'*inter-annotator agreement*

La creazione dei corpora *gold standard* è avvenuta attraverso la collaborazione dei due annotatori dopo una prima revisione manuale individuale; la revisione manuale è stata svolta in maniera individuale per quanto riguarda, in particolare, l'annotazione morfosintattica e l'annotazione sintattica. Trattandosi di due revisioni manuali, è stato possibile calcolare il cosiddetto grado di *agreement* fra di esse, il quale è interpretabile come una stima del grado di affidabilità dello schema di annotazione utilizzato. Di seguito i valori di accordo calcolati:

	2015		2020	
	POS	DEP	POS	DEP
Average observed agreement	98.56 %	89.06 %	98.43 %	88.24 %
Kappa (κ)	98.38 %	89.01 %	98.23 %	88.18 %

Com'è possibile notare, due sono gli indici che sono stati utilizzati per questa stima: l'*average observed agreement* (che corrisponde al rapporto tra il numero di elementi trattati allo stesso modo nelle due annotazioni e il numero totale di elementi annotati) e il *k* di Cohen, calcolato come segue:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)},$$

dove *P(A)* sta per il numero di volte (rispetto al totale) in cui i due annotatori hanno fatto la stessa scelta (*relative observed **a**greement*) e *P(E)* sta per la probabilità ***e**xpected* che i due annotatori

[11] Disponibile allo stesso indirizzo della nota [**6**].

possano avere fatto una stessa scelta in maniera casuale. Nonostante il κ di Cohen sia una stima più raffinata, si nota che, in questo caso, i valori stimati sono molto simili a quelli stimati attraverso l'*average observed agreement*. In generale si nota anche che, prevedibilmente, nell'annotazione morfosintattica è stato raggiunto un grado di accordo molto alto, vicino al 100% (~98%), e che per l'annotazione a dipendenze il valore cala (~88%). Tale scarto fra i due livelli di annotazione linguistica è simile nei due corpora (considerando solo i κ, è di 9.37 nel caso di P.2015 e di 10,05 nel caso di P.2020), quindi è considerabile come dovuto, più che alle caratteristiche di uno specifico corpus, a quelle del linguaggio preso in esame, che spesso presenta costruzioni complesse o comunque interpretabili in più modi, specie per quanto riguarda le attribuzioni di dipendenza. Il valore registrato per l'*agreement* sintattico è, in ogni caso, decisamente accettabile, considerato che 0.8 (quindi 80%) è il valore solitamente riconosciuto come il 'valore soglia' a partire dal quale l'affidabilità dello schema utilizzato è garantita.

2. Creazione dei corpora *gold standard*

In questa sezione si descrivono le modifiche che si è congiuntamente deciso di applicare nei confronti dell'annotazione automatica, ovvero le varie modifiche che hanno condotto, infine, alla definizione dei due corpora *gold standard*. La descrizione del processo di revisione avviene in maniera incrementale, considerando progressivamente tutti i livelli di elaborazione testuale e di annotazione previsti dal progetto delle *Universal Dependencies*.

2.1. *Sentence splitting* e tokenizzazione

Il *sentence splitting*, o segmentazione in frasi, consiste nella suddivisione (*splitting*) del testo in frasi (*sentences*). Per l'individuazione di frasi e periodi, il marcatore principale è rappresentato dai segni di punteggiatura, i quali possono comunque avere un 'peso' e un significato diversi: se la presenza di un punto fermo è agilmente ricollegabile alla fine di una frase[12], più opaca risulta ad esempio la trattazione dei doppi punti e del punto e virgola. Nel progetto *Universal Dependencies*, questi ultimi sono solitamente trattati come segni di punteggiatura forte (cioè indicanti la fine di frase) e, per quanto questa tendenza sia stata rispettata anche nel lavoro qui descritto, è stato necessario prendere, in un paio di casi, delle decisioni diverse, di cui si riportano le motivazioni.

Il primo segno di punteggiatura preso in esame sono i due punti. Il modello li ha considerati, tendenzialmente, come un segno di punteggiatura forte, dividendo in due frasi distinte anche costruzioni molto brevi come, ad esempio, la frase P.2020§33[13]:

P.2020§33: Prendiamo in considerazione alcune soluzioni:

Il valore di segno forte è stato mantenuto anche in un altro caso come la frase P.2020§42[14]. Vi è un caso, invece, in cui si è deciso di riunire due frasi che erano state separate dal modello: si tratta delle frasi P.2015§23-24, di seguito riportate:

[12] Con le dovute eccezioni, chiaramente: il punto delle abbreviazioni (es. "Sig."), ad esempio, non indica di per sé la fine di una frase.
[13] Per riferirsi agli oggetti dell'annotazione, nel corso di questa relazione si usa la convenzione "Nome-File§Numero-Frase#Numero-Token".
[14] P.2020§42: "Dico di più, signor Ministro:".

P.2015§23: Nella formulazione del capoverso 5- bis vi è un problema, che, ad una prima lettura, nella globalità dell'articolato, potrebbe essere visto in modo superficiale, mentre questo testo di legge verrà utilizzato per quello che dice parola per parola (perché questo è quello che ho imparato in Commissione giustizia:

P.2015§24: un articolato viene valutato e soppesato parola per parola e proposizione per proposizione).

Ciò che fin da subito salta all'occhio è che la scissione separa due segni di punteggiatura c.d. bilanciata[15], cioè le parentesi, il che è già di per sé problematico. Le due frasi scisse, inoltre, sebbene possano comunque funzionare in isolamento sia da un punto di vista sintattico che semantico, sono comunque, sotto entrambi i punti di vista, profondamente interdipendenti. A livello funzionale, infatti, la seconda frase è una vera e propria apposizione del "quello" che chi parla ha "imparato in Commissione giustizia". Riunire le due frasi, oltre a evitare di separare le due parentesi, permette quindi di esplicitare, al livello più avanzato delle *dependencies*, questo rapporto di `appos`: ciò risulta utile anche in un'ottica di estrazione dell'informazione o della conoscenza, ottica applicativa che, com'è noto, ha guidato, insieme a quella tipologica e cross-linguistica, diverse delle scelte compiute nel progetto delle *Universal Dependencies*[16].

Per quanto riguarda il punto e virgola, la tendenza, sia da parte del modello che al momento della revisione, è stata quella di considerare questo segno come forte e di dividere dunque le frasi. Esempi di divisioni compiute dal modello sono le frasi P.2020§2-3[17] e P.2015§14-15[18]; un caso in cui si è invece intervenuti, per uniformità col resto dell'annotazione, sono le seguenti due frasi, trattate dal modello come un'unica *sentence* (i.e., ex-P.2015§40):

[15] Cfr. https://universaldependencies.org/u/dep/punct.html, in cui: "Paired punctuation marks (e.g. quotes and brackets, sometimes also dashes, commas and other) should be attached to the same word unless that would create non-projectivity. This word is usually the head of the phrase enclosed in the paired punctuation."

[16] Ad esempio, la scelta di focalizzarsi sulle *content words* come nodi principali, e il loro uso come teste al posto delle parole funzionali, ha risposto a tutti questi bisogni: da un punto di vista tipologico, permette un più facile confronto fra lingue diverse, perché lingue diverse trasmettono il contenuto informativo in maniera meno variabile rispetto al contenuto funzionale (ad es. costruzioni preposizionali); da un punto di vista applicativo, rende più facile l'estrazione di informazione, perché lo strumento deve esplorare un numero tendenzialmente più basso di rami sintattici (con una crescita tendenzialmente lineare) per raggiungerle. Cfr. J. NIVRE, *Towards a Universal Grammar for Natural Language Processing*, in A. GELBUKH (ed.), "Proceedings of CICLing 2015, Part I, LNCS 9041", pp. 3–16, Springer International Publishing Switzerland, 2015, e M.C. DE MARNEFFE, J. NIVRE, *Dependency Grammar*, in "Annual review of linguistics", 5:197, pp. 197-218, 2019.

[17] P.2020§2: "Se andremo in questa direzione, Fratelli d'Italia è pronta a fare il proprio dovere;". P.2020§3: "in caso contrario, troverà in noi sempre dei forti oppositori."

[18] P.2015§14: "Nell'emendamento 1.104 di modifica del capoverso 5- bis prevedo pertanto uno sperimentato, idoneo e comprovato progetto di reinserimento del minore nella propria famiglia d'origine, quantomeno allo scopo di chiedere che vi sia questa accortezza relativamente ad un piano di recupero verificato;". P.2015§15: "altrimenti, si rischia di passare da un affido prolungato, ma non idoneo, ad un'adozione, senza la possibilità di intervento da parte dei familiari.".

P.2020§40: È evidente che gli atti compiuti in questi giorni sono gravi ed ingiustificabili;

P.2020§41: vanno puniti con grande fermezza gli autori delle devastazioni.

La separazione è avvenuta anche perché si tratta di unità semanticamente indipendenti, per quanto brevi, e legate al massimo da un rapporto paratattico. Con lo stesso criterio è avvenuta la separazione delle frasi P.2020§28-29[19], inizialmente considerate dal modello come una sola unità. Vi è un caso, invece, in cui il modello non ha effettuato lo *splitting* e ciò è stato mantenuto. È il seguente:

P.2020§39: Il Gruppo Partito Democratico esprime la più sentita solidarietà, il sostegno e la vicinanza agli agenti della polizia penitenziaria, impegnati in questi giorni e in queste ore in un difficile ruolo; agli operatori, che pure in carcere stanno cercando di lavorare per attenuare le tensioni; ai direttori, che sono di fronte a un compito molto difficile, spesso con strumenti limitati.

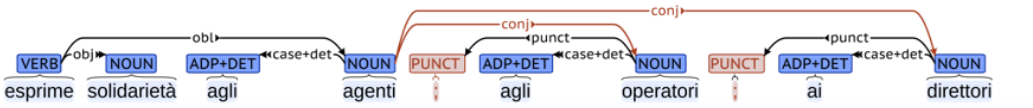

Figura 1: visualizzazione di una parte della frase **P.2020§39**

In questo caso la frase è stata preservata come unica perché, trattandosi di un vero e proprio elenco di complementi fra di loro coordinati, le frasi post-punto e virgola non presentano, al loro livello più alto, teste verbali. Sia da un punto di vista funzionale che semantico sono quindi porzioni totalmente dipendenti dal verbo della principale ("esprime").

Il successivo controllo ha riguardato la tokenizzazione, cioè l'individuazione delle unità lessicali e testuali minime. Al riguardo non è stato necessario intervenire: il modello ha correttamente individuato e scisso tutti i token, e anche unità lessicali complesse come i nomi propri composti da più parole grafiche sono state dal modello divise, in accordo con le linee guida del progetto UD (ad esempio, nella P.2020§39 riportata subito sopra, i token che formano il nome "Gruppo Partito Democratico" sono appunto tre token diversi: #2-3-4[20]). L'unico caso in cui è stato necessario intervenire è il trattamento del "dei" partitivo plurale, nei casi in cui è utilizzato come una sorta di articolo indeterminativo plurale. Il modello lo ha interpretato sistematicamente come una preposizione articolata, dunque come un c.d. *multiword token*[21] (di fatto, quindi, non si

[19] **P.2020§28**: "Lo dico con l'esperienza di chi non ha esitato, quando necessario, ad infliggere decine di ergastoli e migliaia di anni di reclusione;", **P.2020§29**: "non sono temi su cui fare populismo."
[20] Il loro legame a livello lessicale non è comunque perso: viene poi espresso, al livello delle dipendenze, usando la relazione `flat:name`, che specializza `flat` (https://universaldependencies.org/u/dep/-flat.html).
[21] "[…] it is important to note that the basic units of annotation are *syntactic* words (not phonological or orthographic words), which means that we systematically want to split off clitics, as in Spanish *dámelo* = *da me lo*, and undo contractions, as in French *au* = *à le*. We refer to such cases as *multiword tokens* because

tratta di un errore di tokenizzazione in senso stretto, visto che è comunque considerato come un solo token, bensì di un errore di riconoscimento delle parole sintattiche). Il seguente esempio (P.2020§3#9-10, poi solo P.2020§3#9) illustra la differenza dopo la revisione, avvenuta in accordo col modo in cui lo stesso token è stato già trattato nella *treebank* ISDT:

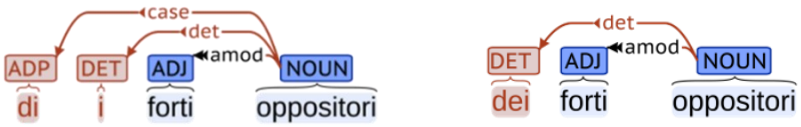

Figure 2-3: esempio di trattamento del "dei" partitivo plurale (**P.2020§39**)

Alla luce di queste modifiche nella segmentazione del testo in frasi e token, si riportano i risultati della verifica della correttezza del modello isdt-ud-2.5-191206.udpipe (lo stesso dell'annotazione automatica iniziale):

Metric	**Precision**	**Recall**	**F1 Score**
Tokens	100.00	100.00	100.00
Sentences	93.75	96.77	95.24
Words	99.73	99.86	99.79

Metric	**Precision**	**Recall**	**F1 Score**
Tokens	100.00	100.00	100.00
Sentences	96.23	92.73	94.44
Words	99.87	99.93	99.90

I valori si riferiscono rispettivamente al corpus P.2015 e al corpus P.2020. "Sentences" si riferisce alle modifiche al livello del *sentence splitting* descritte, "Words" alle modifiche alle parole sintattiche (dunque al "dei" descritto nel paragrafo precedente). Non si registrano, in riferimento a questi valori, differenze significative fra i due corpora.

2.2. *Part-of-speech tagging*: analisi morfosintattica

Una volta definiti i confini di frase e i token, è stato possibile fare annotare in maniera automatica i due corpora, rivedere l'annotazione in maniera individuale e infine definire delle scelte condivise da entrambi gli annotatori. In questa sezione si definiscono le tipologie principali di modifiche che sono state apportate per quanto riguarda l'annotazione morfosintattica, per

a single orthographic *token* corresponds to multiple (syntactic) *words*" (https://universaldependencies.org/u/overview/tokenization.html).

quanto riguarda quindi l'assegnazione della parte del discorso e i correlati task di lemmatizzazione e di definizione dei tratti morfologici. Per comodità, è stato deciso di seguire, per la trattazione, l'ordine con cui sono proiettate le informazioni nel formato CoNLL-U[22]: lemmatizzazione, riconoscimento delle parti del discorso, specificazione dei tratti morfologici.

La *lemmatizzazione* automatica, cioè il riconducimento da parte del modello dei token alle loro forme non marcate (lemmi), ha presentato in diversi casi criticità. Una prima categoria di problemi è rappresentata da tipologie di nomi molto presenti nel linguaggio preso in considerazione: i nomi propri di persona (soprattutto nomi di parlamentari e/o ministri) e i titoli civili. Nel caso dei nomi di persona, si tratta semplicemente di problemi di copertura del vocabolario del modello. Si consideri il seguente esempio:

> **P.2015§18#21-28**: quando vi era ancora la senatrice Alberti Casellati

Qui, "Casellati" viene ricondotto erroneamente alla forma *casellati, nonostante si tratti di un cognome. È interessante notare come il resto dell'informazione linguistica sia stato ricostruito correttamente: il token è indicato giustamente come un nome proprio governato da "Alberti" attraverso una dipendenza `flat:name`. Nonostante fosse impossibile al modello riconoscere il lemma in sé, quindi, l'informazione contestuale ha comunque permesso di ricostruirne il contenuto morfologico e sintattico del token. Il problema coi titoli civili è simile, ma reso più ambiguo dal fatto che essi *sono presenti* nel vocabolario del modello, *ma* non lo sono nella forma in cui li si trova nei testi, cioè con la lettera iniziale maiuscola. Si consideri l'esempio:

> **P.2020§4#1-6+9-10+23-26**: Signor Presidente, signor Ministro, […] quanto avvenuto […] è molto preoccupante.

Questi due vocativi consecutivi sono interessanti perché permettono di ricostruire, in parte, il sistema inferenziale del modello. Il token #1, "Signor", è stato ricondotto al lemma "*Signor", laddove il token #3, "signor", è stato ricondotto al corretto "signore": questo mette già in luce l'importanza che la maiuscola ricopre per l'individuazione del lemma. Il secondo vocativo non è comunque privo di problemi: il token #4, "Ministro", è stato erroneamente ricondotto al lemma "*Ministro", piuttosto che a "ministro". Anche in questo caso la maiuscola ha avuto sicuramente un'influenza, ma c'è probabilmente dell'altro. Si noti infatti che il token #2 della frase, "Presidente", si trova esattamente nella stessa costruzione, ha anch'esso la maiuscola, eppure è stato lemmatizzato correttamente come "presidente". Si può ipotizzare che quel "signor" posto in

[22] https://universaldependencies.org/format.html.

precedenza a "Ministro" abbia, poiché lemmatizzato correttamente come "signore", spinto il modello a interpretare la costruzione alla stregua di un "signor Rossi", e quindi a ricondurre "Ministro" a una forma base con la lettera maiuscola iniziale.

La ricostruzione del lemma fatica pure quando ha a che fare con verbi coniugati, anche in questo caso per problemi di copertura. Ad esempio, i token P.2015§1#73, "chiuderebbero", P.2015§6#44, "eviteremo", P.2020§37#19, "andrebbero", non sono noti al modello, che quindi tenta di ricostruirli, rispettivamente, come "*chiudeere", "*eviteere", "*andre". Si noti che, nel caso di "andrebbero", il tentativo di costruzione dell'infinito è avvenuto semplicemente rimuovendo il suffisso *-bbero*, processo che, se si fosse realizzato in questo modo anche nel caso di "chiuderebbero" (dove invece il modello ha anche aggiunto una *-e-* interfissale fra radice e desinenza), avrebbe effettivamente permesso di derivare l'infinito corretto (chiudere|bbero)[23]. Problematiche sono anche le parole in *-ino*: in un caso, token come "bambino" (P.2015§25#14) e "ragazzino" (P.2015§25#16) sono interpretati (solo al livello della lemmatizzazione) come congiuntivi 3ª p.pl., e ricondotti dunque ai lemmi "*bambare" e "*ragazzare"[24]. Il token "ragazzino", #64 della stessa frase, viene poi trattato come "ragazzino", piuttosto che come "ragazzo": nonostante nelle UD il tratto del diminutivo non sia esplicitabile per i sostantivi e nonostante non ci siano esempi di questo caso nella *treebank* ISDT, si è comunque scelto, in accordo con ciò che avviene in altre *treebank* dell'italiano, di ricondurre il token alla forma base "ragazzo".

La lemmatizzazione crea problemi anche con le parole in *-io* al plurale: "colloqui" (P.2020§9#9) diventa "*colloquo", "contagi" (P.2020§34#36) diventa "*contago". Vi sono poi sostantivi e aggettivi interpretati come participi passati: è il caso di "sindacati" sostantivo (P.2020§12#18), ricondotto a "sindacare"; anche in questo caso gioca un ruolo importante l'informazione contestuale[25]. Un ultimo elemento interessante per quanto riguarda la lemmatizzazione è rappresentato dal token "II" (P.2015§24#41), ricondotto dal modello alla forma sciolta "secondo", tra l'altro nella stessa frase in cui un altro token equivalente (#37) è stato reso come "II". La lemmatizzazione nella forma 'sciolta' non è di per sé scorretta, ma è molto

[23] Fra i numerosi altri esempi, si possono citare anche "*rimarre" per "rimarrà" (P.2015§2#59), "*redere" per "resti" (P.2015§27#20), "*aggiungare" per "aggiungano" (P.2020§22#4).

[24] Va comunque segnalata la particolarità della costruzione in cui ciò accade: "La dichiarazione riguardante un minore, in base alla quale egli, *bambino o ragazzino che sia*, non tornerà mai più nella sua famiglia naturale […]". In effetti, nella "frase incriminata" *vi è* un congiuntivo, come giustamente inferito dal modello alla luce dell'informazione contestuale. È interessante che però (specie nel caso di "bambino") non abbia comunque prevalso l'informazione "già nota" (il lemma "bambino") rispetto all'informazione contestuale/ricostruita.

[25] Si consideri ciò che precede: "Perché non sono state date chiare e precise direttive su una comunicazione anticipata che *rassicurasse detenuti e sindacati* […]". Tanto "detenuti" quando "sindacati" sono stati interpretati come complementi predicativi del verbo "rassicurare", a cui sono coerentemente legati, secondo il modello, attraverso una relazione `xcomp`. È un collegamento sintatticamente sensato, specie considerato che gli infiniti a cui sono ricondotti ("detenere" e "sindacare") sono noti al modello. A riprova dell'importanza del contesto, si segnala che la stessa forma "sindacati" è, in un altro punto (P.2020§19#45), preceduta da un articolo e correttamente lemmatizzata come "sindacato".

rara in ISDT, dove è utilizzata solo in 3 delle 29 attestazioni del token "II"[26]: per uniformità con la *treebank* si è dunque deciso di utilizzare "II" come lemma.

Seguendo, come detto, l'"ordine CoNLL-U", ci si concentrerà adesso sul riconoscimento delle parti del discorso, cioè il *part of speech tagging*[27]. L'esito dell'annotazione automatica risulta particolarmente problematico nel caso di parole come "che", la cui parte del discorso dipende dall'uso in-contesto: alla luce dell'approccio funzionalista delle UD, in base all'uso "che" può essere PRON ("la persona che gioca"), ADV ("che bello!"), SCONJ ("credo che tu abbia ragione") e ADP ("mi piace più X che Y")[28]. Questa disambiguazione è complessa per il modello, che solitamente ha associato a "che" parti del discorso errate. Si riporta un esempio, in cui il token #7 è presentato rispettivamente nella sua descrizione post-annotazione automatica e post-annotazione rivista:

> **P.2015§27#1-9+20-34**: È quanto mai inopportuno allora, che l'inciso […] resti in relazione alla disposizione che «il minore sia dichiarato adottabile».

```
7 che   che   PRON  PR   PronType=Rel 4    nsubj _   _
7 che   che   SCONJ CS   _     20    mark  _   _
```

Altri esempi di "che" erroneamente annotati come pronomi sono P.2015§1#29[29], P.2015§10#4[30] e P.2020§35#38[31] (complessivamente, accade 8 volte in P.2015 e 1 volta in P.2020). Si segnala inoltre che in P.2015§2#58[32] il modello ha fatto il contrario, cioè ha annotato un pronome relativo come SCONJ. Criticità simili si incontrano col token "perché" che, oltre alla funzione subordinante (SCONJ), può anche trovarsi in una principale, dove va annotato come ADV (ad es. in P.2020§12#1, "*Perché* non sono state date chiare e precise direttive […]?"), o col token "dopo",

[26] Tutti e tre i casi appartengono al sotto-corpus 7_WIKIShake.
[27] Pur volendo tenere i diversi livelli di analisi distinti, risulterà inevitabile, in alcuni casi, specificare anche per quale relazione di dipendenza sintattica si è deciso, dopo essere intervenuti sulla morfosintassi, di optare.
[28] È un grado di variabilità segnalato, in riferimento al caso del *that* inglese ma anche, più in generale, in riferimento a parole ad alta frequenza (le quali spesso acquisiscono più funzioni morfosintattiche, sfuggendo così a classificazioni rigide), anche in C. D. MANNING, *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?*, in Alexander Gelbukh (ed.), "Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608", pp. 171-189, Springer, 2011.
[29] P.2015§1#19-42: "[…] se accade - ci sono sentenze che lo testimoniano - *che* dopo sette anni i figli sballottati possono rientrare nel nucleo originario, […]".
[30] P.2015§10: "È per questo che ritengo assolutamente indispensabile modificare il capoverso 5- bis.".
[31] P.2020§35#1-9+38-58: "Voglio infine porre all'attenzione del Governo […] che in queste ore complesse l'emergenza coronavirus, oltre che nelle carceri, rischia di produrre effetti critici anche […]".
[32] P.2015§2#22: "[…] perché chi può negare che si creino legami affettivi tra gli affidatari e i ragazzi, che riconoscono tanti come secondi genitori o verso cui sentono un senso di gratitudine per tutto quello che hanno fatto che rimarrà per tutta la vita?".

che può ricoprire tanto il ruolo di ADV quanto di ADP (come ad es. in P.2015§1#50-52, "dopo due anni").

Altri token che hanno richiesto una certa attenzione al momento del *part of speech tagging* sono "bis" e i termini stranieri. Il token "bis" occorre 5 volte in P.2015, nell'espressione "capoverso 5-bis" (un esempio è la frase §6). Il modello lo ha annotato 4 volte come NOUN e 1 come ADV; la questione si è complicata consultando dizionari online, dove a volte è definito come aggettivo[33] e a volte come avverbio[34]. Alla fine, si è deciso di utilizzare il tag ADV, così come nell'unico esempio disponibile in ISDT (scelta che ha, chiaramente, un'influenza anche per quanto riguarda il livello delle dipendenze sintattiche, per cui si è usato `advmod`). Per quanto riguarda le parole straniere (presenti solo in P.2020§55), essendo sempre opaco il confine fra prestito occasionale e prestito integrato nel sistema linguistico, per decidere se usare il tag X o il tag NOUN il criterio guida è stato principalmente grammaticale, come anche indicato nelle linee guida UD[35]. Si riporta la frase P.2020§55:

> **P.2020§55**: Bene il triage, così come l'introduzione delle norme sanitarie di cui ci ha parlato, i provvedimenti adottati e la costituzione della task force che ha preannunciato per affrontare la questione.

Il token "triage" è utilizzato in una costruzione perfettamente aderente all'italiano, ed è stato quindi considerato (come già fatto dal modello) come NOUN. Si è deciso di operare diversamente per "task force". Il modello considerava "task" NOUN e "force" ADJ (dunque in una relazione di `amod` con "task"); tentava di proiettare, insomma, una struttura tipica dell'italiano, quella [Nome+Aggettivo], in una struttura che però in inglese è [Nome/Modificatore+Nome] (tipica dei composti, con testa a destra). Chiaramente, "forzare" una struttura italiana in un'espressione inglese sarebbe sintatticamente sbagliato; allo stesso modo, è parso anche poco congeniale tentare di ricostruire "in loco" la sintassi inglese (con "task" modificatore di "force"), perché vorrebbe dire utilizzare criteri appartenenti a una lingua diversa da quella del testo in analisi. Per questo motivo, si è infine deciso di annotare sia "task" che "force" come X (quindi come unità "estranee", non del tutto trattabili), rendendo "force" governato, in un rapporto `flat:foreign`[36], da "task".

[33] Ad esempio, nella versione de "il Sabatini Coletti" consultabile su corriere.it: https://dizionari.corriere.it/-dizionario_italiano/B/bis.shtml.
[34] È avverbio, ad esempio, secondo Treccani: https://www.treccani.it/vocabolario/bis/.
[35] "A special usage of X is for cases of code-switching where it is not possible (or meaningful) to analyze the intervening language grammatically (and where the dependency relation flat:foreign is typically used in the syntactic analysis). This usage does not extend to ordinary loan words which should be assigned a normal part-of-speech" (https://universaldependencies.org/u/pos/X.html).
[36] Un'altra relazione adatta avrebbe potuto essere `compound`.

Un'ultima classe di difficoltà relative al *part of speech tagging* è rappresentata dalla distinzione fra aggettivi e verbi al participio passato, che non è sempre chiara[37]. Si consideri l'esempio seguente:

> **P.2020§16#1-16**: Si tratta di una situazione risaputa, già al limite e più volte denunciata,

In questo come in altri casi, il criterio per distinguere le due classi è stata la presenza di strutture argomentali o modificatori, che solitamente indica la presenza di un verbo. In questo caso si tratterebbe di "più volte", che è stato utilizzato anche come 'segnale' per decidere di annotare come verbo il token 6 della frase P.2020§15 ("Questo è stato più volte *segnalato*."). Un caso opposto, in cui invece si è passati da VERB ad ADJ, è token #15 della frase seguente:

> **P.2020§43#1-24**: penso che vada anche verificato fino in fondo se ci sia stato un disegno destabilizzante su tutto il territorio nazionale per diffondere le rivolte

In questo caso, per considerare "destabilizzante" come verbo, risulterebbe necessario considerare "su tutto il territorio" come un suo argomento. È quanto fa il modello, che infatti lega "tutto" a "destabilizzante" attraverso una relazione `obl`. A livello di interpretazione testuale, tuttavia, risulta abbastanza chiaro che si tratta di una struttura che modifica piuttosto "ci sia stato".

L'ultimo livello di analisi morfologica preso in esame è rappresentato dalla definizione delle *features*, cioè dei tratti morfologici che esplicitano informazioni come il "genere" e il "grado" degli aggettivi. Una categoria che necessita particolare attenzione sono le parole ambigeneri, cioè le parole (sostantivi e aggettivi) che sono usate sia al femminile che al maschile senza modificare il suffisso flessivo. Nel caso delle UD, la descrizione di parole di questo tipo *non* prevede che si codifichi il genere, anche quando esso è ricavabile dal contesto in cui sono usate. Alla luce di ciò, è stato rimosso il tratto `Gender=Masc` nella descrizione del sostantivo "minore" (P.2015§5#47) e in quella dell'aggettivo "immuni" (P.2020§11#31), informazione che è possibile teorizzare il modello abbia inferito dal resto del contesto. Vi è un caso meno semplice, cioè quello di "Presidente": prima di tutto, si riporta che il modello associa solitamente l'informazione del genere maschile, e che il maschile è sistematicamente codificato anche in ISDT. Detto ciò, in un caso, P.2015§16#2, il modello non codifica il genere, il che ha stimolato un po' di riflessione su come trattare più in generale questo sostantivo. Il problema sta nel fatto

[37] Sempre in C. D. MANNING, *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?*, p. 177, si riporta questa stessa distinzione, fra aggettivo e verbo, come esempio di casi in cui il tag è "sottospecificato, ambiguo, o poco chiaro rispetto al contesto".

che, se è vero che "presidente" *è* ad oggi di fatto un termine ambigenere, o "di genere comune" (ed è quindi utilizzabile indistintamente come "il/la presidente"), è altrettanto vero che ad essere più precisi *sta diventando* ambigenere; si tratta cioè di un cambiamento linguistico in corso, che è anche stato anche molto discusso[38]. Inoltre, un femminile marcato ("presidentessa"), seppur con una carica connotativa a volte negativa, *è* attestato, e andrebbe anch'esso ricondotto al lemma "presidente" (con la specifica, nei tratti, del genere femminile), di fatto sancendo una potenziale distinzione fra forma maschile (non marcata) e forma femminile (marcata). Si tratta insomma di una zona un po' opaca: da una parte vi è un cambiamento ancora in corso, dall'altra un femminile tendente al disuso ma storicamente attestato. Alla fine, si è deciso (anche per uniformità con ISDT) di codificare il genere maschile, ma sostantivi come questo meriterebbero probabilmente una rivisitazione d'insieme (considerato anche che, sempre in ISDT, per token anch'essi ambigeneri come "psicanalista" e "docente", invece, non viene codificato il maschile)[39].

Si segnala infine che, sempre per quanto riguarda i tratti morfologici:

- è stato rimosso il tratto `Gender=Fem` nel caso del "lei" usato come forma di cortesia (attestato 6 volte in P.2020);
- il token P.2015§13#42, "fine", è stato interpretato dal modello come "*la* fine", con quindi associato tratto `Gender=Fem`, laddove si tratta de "*il* fine";
- similmente a quanto accaduto con sostantivi come "sindacati" (di cui si è discusso in merito alla lemmatizzazione), il modello interpreta male la forma "rivolte": occorre 4 volte in P.2020, e in tutti i casi, nonostante la lemmatizzazione sia corretta, il token è trattato come un participio passato piuttosto che come un sostantivo, con relativa associazione di *features* errate, che veicolano informazione verbale;
- ad alcune forme verbali il modello ha associato tratti errati: è il caso di "eviteremmo" (P.2015§6#44) e "aggiunga" (P.2020§22#4), token ai quali è stato associato il tratto `Mood=Ind` al posto di `Mood=Sub`.

Per i risultati sul grado di correttezza nel *part of speech tagging* dei modelli testati sulle versioni modificate dei corpora, si rimanda a §3.

2.3. *Syntactic parsing*: analisi delle dipendenze sintattiche

[38] Cfr. al riguardo V. GHENO, *Femminili singolari. Il femminismo è nelle parole*, effequ, Firenze, 2019.
[39] Nel caso del corpus P.2020, la codifica del maschile è resa ancora più critica dal fatto che, nei vocativi in cui compare la parola, il referente è la presidente Casellati.

Il livello di analisi che si prenderà adesso in considerazione è quello dell'analisi sintattica. Il progetto *Universal Dependencies* prevede, come il nome fa intendere, un'analisi a dipendenze, un'analisi basata, cioè, su relazioni binarie asimmetriche instaurate fra un elemento testa (detto anche *governor*) e un elemento dipendente. Queste relazioni definiscono un albero che ha come elemento fondamentale una radice, *root*, caratterizzato dal fatto di non dipendere a sua volta da nessun elemento. Si tratta tipicamente di una testa verbale, ma in alcuni casi può trattarsi di un sostantivo o di un aggettivo; ad esempio, quando il verbo "essere" è usato con funzione copulativa, è il nome del predicato a ricoprire il ruolo di radice. Proprio casi come questo, in cui il riconoscimento della radice sintattica risulta poco banale, hanno messo in difficoltà il modello. Si consideri l'esempio seguente, la frase P.2020§5:

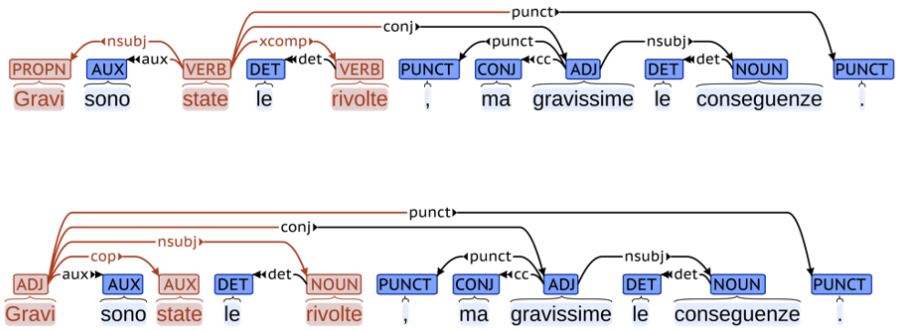

Figura 4 – Visualizzazione della frase **P.2020§5***, nell'analisi automatica e nell'analisi rivista*

In questo esempio, si nota un verbo essere usato come copula di un aggettivo ("[g]ravi"), il quale è in più dislocato alla sinistra della frase. Proprio la dislocazione, con conseguente interpretazione di "Gravi" come un nome proprio, stimola un'analisi automatica totalmente errata, in base alla quale un ipotetico individuo Gravi è soggetto di "sono state", e "rivolte", piuttosto che essere riconosciuto come soggetto, è interpretato come un participio passato legato a "sono state" attraverso una relazione `xcomp`. Nella figura 4 è visualizzabile la versione rivista e corretta della stessa frase[40].

Riconoscere la radice risulta particolarmente ostico anche in quei casi in cui il periodo risulta sintatticamente complesso, ad esempio perché gli alberi sintattici sono molto alti e le relazioni di dipendenza lunghe. Ciò è, in particolare, importante per il corpus P.2015, che ha,

[40] Vale la pena segnalare un caso complementare, sempre legato al riconoscimento della radice nel caso della copula. La frase P.2020§54 recita: "Il tema oggi, come lei ha detto, è attrezzare meglio le carceri per garantire questi colloqui."; il modello aveva annotato "attrezzare" come `root`, in maniera apparentemente coerente con come sono trattati i predicati nominali. In realtà, il sito delle UD riporta anche che: "[…] the `cop` relation is not used when the nonverbal predicate has the form of a clause", quindi esattamente in un caso come quello della frase P.2020§54, in cui la radice è, in definitiva, il verbo essere.

come si vedrà anche più in dettaglio, frasi tendenzialmente più complesse. Si consideri il seguente esempio (per permettere di osservare pienamente la complessità del periodo, lo si riporta sia per esteso che, abbreviato, in una sua visualizzazione):

> **P.2015§29**: Lo spirito di questo disegno di legge, come riconoscerà la stessa prima firmataria, la senatrice Puglisi, è quello di fare in modo che laddove, con percorso autonomo, ci sia la dichiarazione di adottabilità del bambino (che sia adottato dalla famiglia affidataria, che sia adottato da un'altra famiglia) sia garantita comunque la possibilità per la famiglia affidataria di mantenere i legami affettivi e, qualora la famiglia affidataria abbia i requisiti per chiedere l'adozione, la valutazione venga fatta anche in relazione al prolungato periodo di affidamento.

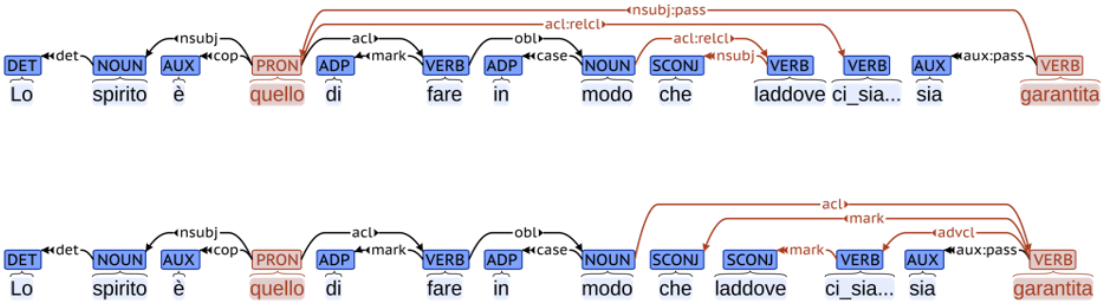

Figura 5 – Visualizzazione della struttura della frase **P.2015§29***, nell'analisi automatica e nell'analisi rivista*

In questo caso è la ricchezza ipotattica del periodo a 'confondere' il modello: a causa di una serie di errori a cascata nel riconoscimento delle dipendenze, scatenati dall'anomala interpretazione come verbo, e quindi come elemento più alto di una subordinata relativa, di "laddove", "garantita" viene annotato come *governor* di "quello", e quindi come radice del periodo. Altri esempi, fra i molti, di mancato riconoscimento della radice sono la frase P.2015§34[41] e la frase P.2020§12[42]; in entrambi casi, sembra essere la dislocazione, dovuta all'uso di "perché" a inizio frase, ad avere creato difficoltà al *parser*.

Un altro problema, di rilevanza comparabile, è il riconoscimento delle dipendenze `csubj`, e quindi delle subordinate soggettive. In maniera sistematica, il modello le interpreta, alla luce del "che" che le introduce, come subordinate completive (`ccomp` o `xcomp`); questo accade sia per periodi piuttosto complessi (come quello visualizzato di seguito), in cui i nodi nella relazione csubj sono effettivamente distanti, che in casi in cui i nodi sono invece prossimi.

[41] P.2015§34: "ad esempio, perché non concordare con la magistratura di sorveglianza il permesso temporaneo di restare a casa per i detenuti in semilibertà che in carcere tornano solo a dormire e potrebbero essere vettori di contagi?". Il token "esempio" era stato annotato come `root` al posto di "concordare".
[42] P.2020§12: "Perché non sono state date chiare e precise direttive su una comunicazione anticipata che rassicurasse detenuti e sindacati nel garantire valide alternative come filtri sanitari, telefonate, colloqui via Skype e altre soluzioni?". Il token "colloqui" era stato annotato come `root` al posto di "date".

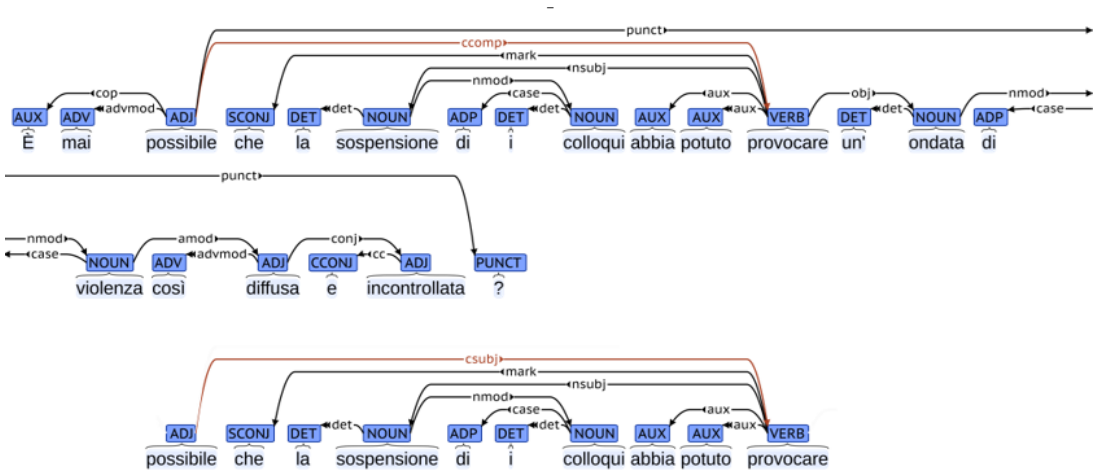

Figura 6 – Visualizzazione della frase **P.2015§29***, nell'analisi automatica e nell'analisi rivista*

Esempi di frasi più semplici in cui si presenta lo stesso errore sono la frase P.2020§40, "È evidente che gli atti compiuti in questi giorni sono gravi ed ingiustificabili;", dove "gravi" è stato annotato come `ccomp` di "evidente" piuttosto che come `csubj`, e la frase P.2020§31(#1-5), "Mi ha stupito sapere che […]", dove "sapere" è stato annotato come `xcomp` di "stupito".

Un'altra serie di casi in cui è risultato necessario intervenire manualmente sono quei casi di dipendenza sintattica difficili da ricostruire perché necessitano di adottare un criterio semantico-logico, oltre che sintattico. Si considerino, ad esempio, i primi 6 token della frase P.2015§1, "Ho portato solo due esempi,", e la visualizzazione della sua annotazione automatica:

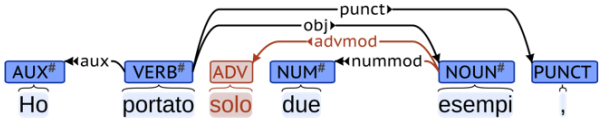

Figura 7 - Visualizzazione dei token **P.2015§1#1-6** *in seguito all'annotazione automatica*

Il modello interpreta "solo" come un modificatore avverbiale (`advmod`) del token "esempi". Nonostante la differenza sia, allo scritto, sottile, l'interpretazione della frase ci ha spinti a considerare l'avverbio come un modificatore del numerale "due": nel resto dell'enunciato, chi parla giustifica il suo "portare solo due esempi" dicendo che "ha una visione parziale delle cose"

(#10-16); questa visione parziale diventa quindi la spiegazione di un numero limitato (*solo due*, appunto) di esempi riportati[43]. Si riporta la visualizzazione dell'annotazione rivista:

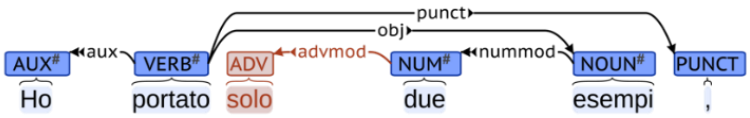

Figura 8 – Visualizzazione dei token **P.2015§1#1-6** *in seguito all'annotazione automatica*

Diverse modifiche sono state apportate anche in riferimento al pronome "si" espletivo. Il riconoscimento della relazione base `expl` è in linea di massimo avvenuto correttamente (anche se a volte con indicazione della testa sbagliata); più numerose sono le difficoltà riscontrate per rintracciare le sotto-specificazioni `expl:pass` ed `expl:impers`, il cui confine risulta piuttosto sfumato anche per annotatori umani. Si consideri l'esempio seguente, che presenta due "si":

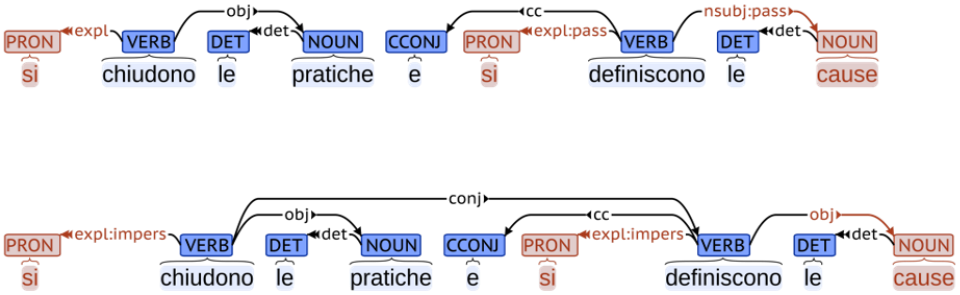

Figura 9 – Visualizzazione della frase **P.2015§1***, nell'analisi automatica e nell'analisi rivista*

Il valore dei due "si" non è scontato: a una prima occhiata, sembrerebbe possibile interpretarli tanto come passivi (con, quindi, "pratiche" e "cause" in funzione di soggetti passivi) quanto come impersonali. Alla luce dell'analisi della frase completa[44], si è deciso di optare per la relazione `expl:impers` e di interpretare i sostantivi post-verbali come oggetti: questo perché il "si" a cui

[43] Si segnala che un uso molto simile della dipendenza `advmod` è attestato anche fra gli esempi delle linee guida UD: si tratta, in particolare, dell'ultimo esempio in https://universaldependencies.org/u/dep/-advmod.html, "*About* 200 people came to the party".

[44] "Ho portato solo due esempi, perché il sottoscritto ha una visione parziale delle cose, ma se accade - ci sono sentenze che lo testimoniano - che dopo sette anni i figli sballottati possono rientrare nel nucleo originario, non facciamo l'errore di dire che dopo due anni questi figli vengono adottati dalla famiglia affidataria perché tanti magistrati, anche in buona fede, con questo provvedimento chiuderebbero tanti fascicoli e farebbero statisticamente anche «bella figura» perché si chiudono le pratiche e si definiscono le cause."

si fa riferimento è interpretabile come un insieme generico, non definito, di "magistrati", trattandosi (questa azione di chiusura delle pratiche e di definizione delle cause) del motivo per cui proprio i magistrati "farebbero bella figura". Un esempio in cui invece, di fronte alla mancata sotto-specificazione da parte del modello, si è deciso di impiegare `expl:pass`, è il "si" (#27) presente nella frase P.2015§2, che si riporta di seguito. In questo caso, "legami affettivi" è stato interpretato come il soggetto passivo che "viene creato".

> **P.2015§2#1-36**: Per tutti questi argomenti, che avrò magari anche elencato in modo confuso, credo sia opportuno mantenere questa previsione, perché chi può negare che si creino legami affettivi tra gli affidatari e i ragazzi […]

Un'altra classe di errori che vale la pena segnalare è l'assegnazione della dipendenza `compound`, che è utilizzata per indicare espressioni polirematiche e composti endocentrici (laddove `flat` è utilizzata per le costruzioni esocentriche e `fixed` per le costruzioni grammaticali fossilizzate). Il problema del riconoscimento di questa relazione è spesso collegato alle difficoltà di analisi descritte precedentemente, ad esempio per quanto riguarda la lemmatizzazione. Si considerino questi due token:

Figura 10 – Visualizzazione della dipendenza fra token 59 e 60 della frase **P.2015§23**, *nell'analisi automatica e nell'analisi rivista*

Il token "giustizia" è stato lemmatizzato dal modello come un aggettivo e ricondotto al lemma "*giustizio": non sorprende allora che la relazione compound non sia stata riconosciuta e che al suo posto il modello abbia tentato una relazione `amod`. Sempre dovuto a problemi di lemmatizzazione è l'errore nell'esempio seguente:

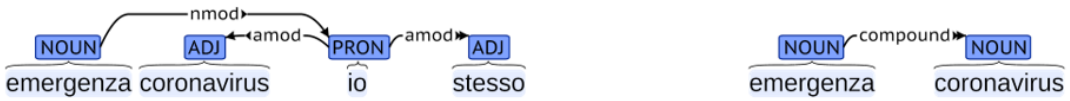

Figura 11 – Visualizzazione della dipendenza fra token 27-28(-29-30) della frase **P.2020§16**, *nell'analisi automatica e nell'analisi rivista*

Il vocabolario del modello, non conoscendo per ovvi motivi storici il lemma "coronavirus", attribuisce al token corrispondente la funzione più probabile, cioè quella di aggettivo (si noti che

infatti è in posizione post-nominale). Questo stimola un errore nell'interpretazione sintattica, in base alla quale, tra l'altro, il token "coronavirus" ha il pronome "io" come *governor* e non il più scontato (e corretto) "emergenza"[45]. Per concludere questa sezione sulla composizione, si riporta l'interpretazione, errata e rivista, di un composto che appartiene al linguaggio settoriale preso in esame, e dimostra come anche i composti possano essere distintivi di un particolare dominio, e dunque non banali per modelli che, su quel dominio, non sono addestrati:

Figura 12 – Visualizzazione della dipendenza fra token 11, 12 e 13 della frase **P.2020§53***, nell'analisi automatica e nell'analisi rivista*

L'ultima costruzione sintattica che si prende adesso in esame è il vocativo (`vocative`). È un esempio di costruzione piuttosto semplice per un annotatore umano, ma che sembra creare problemi all'annotatore automatico. Nel linguaggio preso in esame, appartenente a un'oralità "controllata", è utilizzato più volte perché permette di attirare l'attenzione altrui. Si consideri l'esempio seguente[46]:

Figura 13 – Visualizzazione della dipendenza fra token 1, 2 e 3 della frase **P.2020§8***, nell'analisi automatica e nell'analisi rivista*

La frase continua con "voglio sottolineare", ed è proprio "sottolineare" l'elemento di cui, secondo il modello, "Signor" sarebbe il soggetto. Vi sono altri casi equivalenti o simili (in uno, cioè in P.2020§8#1-3, "Ministro" viene collegato con la corretta dipendenza `flat:name`, ma "Signor" è comunque interpretato come soggetto della radice verbale della frase), fra i quali vale la pena segnalare che nella frase P.2020§42 ("Dico di più, signor Ministro"), il vocativo è stato annotato

[45] La struttura della frase ha sicuramente avuto un'influenza in questo senso. Si noti la lunghezza della relazione (di 14 token) che intercorre fra l'"io" preso in esame e la sua testa verbale, "firmato": "[…] al punto che proprio alcuni giorni prima dell'emergenza coronavirus *io* stesso, insieme ai rappresentanti di tutti i Gruppi di maggioranza, ho *firmato* un disegno di legge per aumentare la possibilità di colloqui telefonici in carcere."

[46] Nell'esempio si può notare anche una modifica apportata per quanto riguarda i `compound`, relazione appena affrontata. "Signor Ministro" è infatti un'espressione che ricorre più volte nei testi in esame, e che il modello ha annotato in maniere differenti in base ai casi (variabilità che, al momento della revisione manuale, è stata neutralizzata optando per la forma visualizzabile nell'esempio).

come complemento oggetto del verbo "dire" (per motivi legati, probabilmente, al suo essergli post-posto). Si riporta la visualizzazione dell'esempio appena descritto:

Figura 14 – Visualizzazione della dipendenza fra token 1, 4, 5 e 6 della frase **P.2020§42***, nell'analisi automatica e nell'analisi rivista*

Una questione complessa è rappresentata dall'annotazione della punteggiatura, rispetto cui è stato necessario un intervento sistematico (circa la metà delle relazioni `punct` sono state infatti modificate). Ad esempio, la regola, prevista dalle *Universal Dependencies*, in base alla quale un segno di punteggiatura posto fra unità coordinate dovrebbe essere collegato all'elemento congiunto subito successivo, spesso non viene rispettata (solo per la gestione delle *frasi* coordinate, quindi non includendo casi di sostantivi coordinati fra di loro, sono state necessarie 11 modifiche). Particolarmente problematica è risultata la punteggiatura bilanciata: secondo le linee guida, "i segni di punteggiatura bilanciata (virgolette, parentesi, a volte anche trattini, virgole ed altri) dovrebbero essere attaccati alla stessa parola a meno che non si crei non-proiettività" e "questa parola è solitamente la testa del sintagma incluso nella punteggiatura". Ciò accade con difficoltà nell'analisi automatica. Si riporta un esempio di intervento[47]:

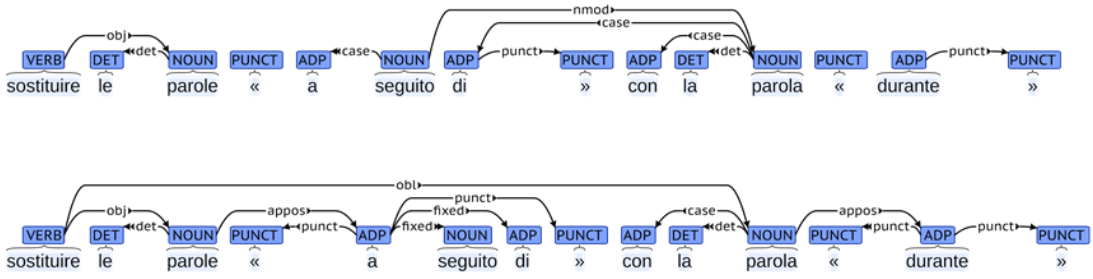

Figura 15 – Visualizzazione di **P.2020§42#54-67***, nell'analisi automatica e nell'analisi rivista*

Può essere interessante, infine, confrontare i processi di revisione dei due corpora. Il corpus P.2015 ha richiesto la modifica di 90 dipendenze totali, il corpus P.2020 di 76. Si noti che,

[47] Nella visualizzazione dell'annotazione automatica, token come "«" (#57) non hanno alcun legame semplicemente perché dipendono da token non visualizzati (sia #57 che #65 dipendono infatti da #50).

nonostante una lunghezza in token molto simile (1457 sono i token di P.2015, 1522 quelli di P.2020), il secondo corpus contiene un numero di periodi molto più alto: 55, contro i 31 di P.2015. Ciò fa sì che, nel caso di P.2020, la lunghezza media dei periodi sia sensibilmente più bassa rispetto a quella di P.2015. Alla luce di queste considerazioni, risulta possibile ipotizzare che vi sia una correlazione fra il numero di errori commessi dall'annotatore automatico e la complessità, o più semplicemente la lunghezza, della frase[48].

Ciò non deve comunque far credere che, per quanto riguarda le performance dell'annotatore automatico, ogni frase breve sia banale e ogni frase lunga sia molto difficile. La difficoltà delle frasi più lunghe è, in un certo senso, matematica o quanto meno probabilistica: più aumentano i token, più è probabile che ci siano fra i vari elementi rapporti di interdipendenza difficili da gestire e che aumentino i tratti contestuali che il modello cerca di considerare per compiere la sua classificazione (ad esempio, aumenta il numero di n-grammi considerati). Appurata questa possibilità, va comunque detto che molto dipende dal tipo di periodo e dal tipo di costruzioni linguistiche utilizzate. Ad esempio, vi sono frasi come la P.2015§24[49], di lunghezza medio-alta (58 token), che sono state analizzate in maniera quasi perfetta; d'altro canto, ci sono anche frasi, come la P.2020§15 (che si allega a conclusione di questa sezione), estremamente brevi (7 token), in cui l'analisi errata di anche solo un elemento genera, a cascata, tutta una serie di problemi, a livelli diversi di analisi:

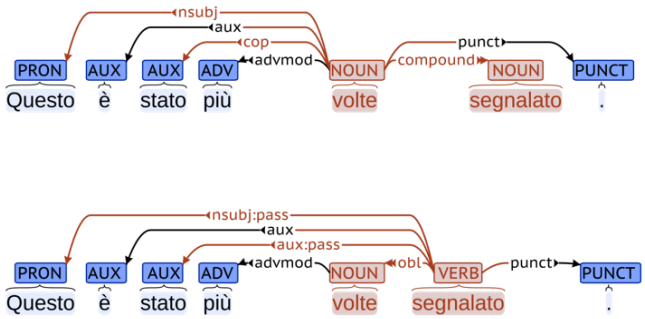

Figura 16 – Visualizzazione di **P.2020§15**, *nell'analisi automatica e nell'analisi rivista*

In questa breve frase si ritrovano, insieme, alcuni dei problemi che sono stati affrontati nella sezione §3: "segnalato" è analizzato come nome, dunque 'salta' il suo riconoscimento come radice; la sotto-specificazione di una categoria (in questo caso `nsubj`) non viene compiuta dal

[48] Riguardo le caratteristiche dei due corpora si ritornerà in §3.1, dove saranno presentate in maniera più organica per discutere i risultati della verifica della correttezza dei modelli.
[49] P.2015§24: "In questo testo, così come licenziato dalla Commissione, abbiamo una pericolosa relazione con il concetto per cui «il minore sia dichiarato adottabile, ai sensi delle disposizioni del capo II del titolo II», che è un percorso che dev'essere considerato assolutamente autonomo da qualunque altro elemento."

modello; “volte” di “più volte”, essendo stato riconosciuto come `root` al posto di “segnalato”, diventa il *governor* della copula precedente, di “segnalato”, e anche del segno di punteggiatura; la relazione che dovrebbe legare “più volte” alla radice, `obl`, non può, nel caso dell’annotazione automatica, essere rintracciata. Tutto questo in soli sette token.

3. Verifica della correttezza dei modelli

In questa sezione ci si concentrerà sull'ultima fase del lavoro, cioè l'utilizzo dei *gold standard* corpora a questo punto costruiti come *test set* di due modelli: il modello su cui è stata effettuata l'annotazione semi-automatica, isdt-ud-2.5-191206.udpipe**,** e il modello italian-postwita-ud-2.5-191206.udpipe, addestrato sulla *treebank* di *tweet* POSTWITA (da qui in poi, rispettivamente solo U.ISDT e U.POSTWITA). I valori di seguito riportati[50] sono stati ricavati utilizzando lo script distribuito in occasione dello *CoNLL 2018 Shared task: Multilingual Parsing from Raw Text to Universal Dependencies*[51]. Per la correttezza del modello iniziale nel *sentence splitting* e la tokenizzazione, si rimanda a §2.1.

Si riportano prima di tutto i dati legati al *part of speech tagging*. Si segnala che "UFeats" si riferisce alla correttezza, in termini di F1 score, delle features morfologiche, e che "AllTags" invece tiene conto dell'informazione morfologica in generale (UPOS+XPOS+feats):

	2015		2020	
F1 Score	**U.ISDT**	**U.POSTWITA**	**U.ISDT**	**U.POSTWITA**
UPOS	97.60	93.75	96.71	94.55
XPOS[52]	97.32	92.52	96.06	93.43
UFeats	97.74	92.31	97.11	91.33
AllTags	96.57	89.09	95.40	89.22
Lemmas	98.42	96.84	97.57	94.88

Risulta subito evidente l'alto grado di correttezza nel caso di U.ISDT, nel caso sia di P.2015 che di P.2020. Il valore a cui il modello tende nelle varie metriche è il 97% ($\mu = 97.53$, $\sigma = 0.67$), valore solitamente riconosciuto come la "soglia" oltre cui l'annotazione automatica non sembra,

[50] Nel corso della relazione si farà riferimento solo all'F1 score, in quanto media armonica di *precision* e *recall*. Nell'appendice è possibile trovare, per quanto riguarda l'analisi sintattica, le tabelle con tutte le misure di accuratezza disponibili: Precision, Recall, F1 Score, AligndAcc. Per l'analisi morfosintattica, i valori sono gli stessi a prescindere dal tipo di metrica, quindi non sono stati riportati.

[51] https://universaldependencies.org/conll18/evaluation.html.

[52] Nel corso della relazione non si è fatto esplicito riferimento alle modifiche apportate alle XPOS (http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf) semplicemente perché sono motivate da quelle altre modifiche che, invece, sono state trattate. Ad esempio, una modifica che è stata più volte effettuata in riferimento alle XPOS è l'utilizzo del tag FB, che è un tag utilizzato per indicare segni di punteggiatura bilanciata; quest'ultima, come si è visto, non è sempre riconosciuta dal modello, e chiaramente la sua revisione manuale ha implicato anche la sostituzione del tag precedentemente definito dal modello (FF).

al momento, potere andare. Questo valore peggiora nel caso di P.2020 ($\mu = 96.57$, $\sigma = 0.85$), anche se non di molto. È interessante osservare ciò che succede quando si usa il modello addestrato su POSTWITA: i valori assoluti sono in tutti i casi peggiori dei corrispettivi di U.ISDT (il che è prevedibile), ma il loro peggioramento, nel confronto fra P.2015 e P.2020, è decisamente meno importante. Infatti, il valore medio registrato per P.2015 ($\mu = 92.902$, $\sigma = 2.79$) si ripresenta quasi intatto anche per P.2020 ($\mu = 92.682$, $\sigma = 2.38$), con un peggioramento medio molto basso (.22, contro un .96 di peggioramento di U.ISDT).

Si considerino adesso i valori registrati per l'accuratezza dell'analisi sintattica[53]:

	2015		**2020**	
F1 Score	**U.ISDT**	**U.POSTWITA**	**U.ISDT**	**U.POSTWITA**
UAS	86.82	77.28	85.28	78.06
LAS	84.08	72.00	82.13	72.86
CLAS	79.21	65.34	75.10	63.65
MLAS	76.91	58.36	72.20	56.35
BLEX	77.45	62.60	71.94	57.54

Prima di tutto, si nota che i valori sono, in generale, decisamente più bassi di quelli rilevati rispetto all'analisi morfosintattica. Questo non è sorprendente: l'analisi a dipendenze è un compito generalmente più complesso, con un più alto grado di variabilità e interpretabilità (anche l'*inter-annotator agreement* sintattico calcolato, infatti, era sensibilmente più basso di quello morfosintattico). Si nota anche che, pure in questo caso, i valori assoluti di U.ISDT sono generalmente più alti di quelli di U.POSTWITA. Anche in questo caso, inoltre, si registra, nel passaggio da P.2015 a P.2020, un peggioramento evidente del modello ISDT: da un F1 score medio di circa 80% ($\mu = 80.894$, $\sigma = 4.35$) si passa, in P.2020, a un valore medio di circa 77% ($\mu = 77.33$, $\sigma = 6.05$). E, anche in questo caso, il peggioramento di U.POSTWITA, per quanto rilevante, risulta più basso: se l'F1 score, rispetto a P.2015, è di circa il 67% di accuratezza ($\mu = 67.11$, $\sigma = 7.53$), esso scende a un valore di 65% ($\mu = 65.69$, $\sigma = 9.51$) nel caso di P.2020. Il

[53] Per quanto riguarda le metriche utilizzate: UAS ("Unlabeled Attachment Score") tiene conto delle relazioni di dipendenza individuate correttamente fra i nodi, a prescindere dalla correttezza delle categorie ad esse associate; LAS ("Labeled Attachment Score") tiene conto anche della label associata alla relazione; CLAS ("Content-Word Labeled Attachment Score") tiene conto della correttezza delle sole relazioni di dipendenza fra parole contenuto; BLEX ("BiLeXical Dependency Score") è come CLAS ma tiene conto anche del lemma del dipendente.

peggioramento nell'analisi del secondo corpora è quindi di 1.42, contro un peggioramento di 3.564 nel caso di U.ISDT.

3.1. Una possibile interpretazione dei risultati: *linguistic profiling* dei due corpora

Sistematicamente, quindi, i modelli si comportano peggio sul corpus più recente, e sistematicamente U.POSTWITA si comporta 'meno peggio' di U.ISDT. Il tentativo di interpretare questi risultati deve passare inevitabilmente per un'analisi delle caratteristiche linguistiche dei due testi, che permettano di metterne in luce le differenze. In realtà, una prima descrizione, concernente semplicemente la lunghezza media delle frasi, è stata presentata già nella sezione §3.1. Adesso si presenteranno questi stessi dati in maniera più sistematica[54]:

	P.2015	**P.2020**
N di periodi	31	55
N di parole	1457	1522
L media dei periodi	47	28
L media delle parole	4.6	4.6
Lemmi in VdB (Fond.)	84.8%	83.8%
Lemmi in VdB (Alto Uso)	6.2%	8.2%
Lemmi in VdB (Alta Disp.)	0.4%	1.1%
Lemmi fuori VdB	8.6%	6.9%
Altezza media albero sint.	4.2	3.5
L media relazioni	3.3	2.8
L massima relazione	145	93

L'ipotesi, precedentemente avanzata, che il corpus P.2020 sia tendenzialmente "più semplice" di P.2015 sembra trovare conferma in questi dati. Per quanto riguarda il livello lessicale, il 93.1% dei lemmi di P.2020 appartengono al Vocabolario di Base, contro il 91.4% dei lemmi di P.2015. Anche a livello sintattico, i tratti solitamente associati a un alto grado di complessità testuale e sintattica (altezza dell'albero sintattico e la lunghezza delle relazioni di dipendenza) sono più caratteristici di P.2015 che di P.2020. Questa maggiore 'semplicità' formale potrebbe rendere lo stile di P.2020 almeno in parte più vicino a quello dei *tweet* su cui è addestrato U.POSTWITA, il che spiegherebbe perché il modello si comporta abbastanza bene, almeno rispetto alla performance dello stesso su P.2015. Si noti, comunque, che la semplicità formale di P.2020, come gli *score* appena riportati mettono in evidenza, *non* implica necessariamente un'analisi realizzata

[54] Soprattutto per quanto riguarda i valori da "lunghezza media delle parole" in poi, si specifica che i valori sono stati calcolati, manualmente, da Luca Baù.

più facilmente: è quanto è già stato detto alla fine della §2.3, e quanto è esemplificato dal fatto che, *in ogni caso*, entrambi i modelli si comportano peggio sul corpus formalmente più semplice (per quanto un modello 'peggiori meno' rispetto all'altro).

Viene infine da chiedersi a cosa possano essere dovute queste differenze nella forma, in ogni caso presenti. Va detto prima di tutto che i campioni presi in esame sono di dimensioni contenute, quindi non è semplice tentare generalizzazioni scientificamente valide; inoltre, trattandosi di trascrizioni di testi orali, un certo grado di variabilità (ad esempio nell'uso della punteggiatura) è inevitabile e non necessariamente significativo. Ciononostante, le differenze nello stile possono essere spiegate, almeno in parte, considerando i temi di cui si discute, e in particolare notando che, in P.2020, il tema affrontato sono le rivolte nelle carceri all'inizio dell'emergenza epidemiologica: sicuramente un argomento scottante, che può avere stimolato, vista l'urgenza, una comunicazione più diretta e incisiva rispetto a quella di P.2015.

4. Conclusioni

Nonostante gli interventi risultati necessari, è possibile considerare l'analisi automatica svolta dal modello U.ISDT come generalmente soddisfacente. Per quanto riguarda la varietà linguistica che caratterizza i testi, alcune caratteristiche (ad es. lemmi particolari come "bis", composti come "decreto-legge" e strutture sintattiche come le dislocazioni e i vocativi) sono state effettivamente gestite con difficoltà dal modello; in generale, però, quello in esame è un linguaggio controllato, seppur appartenente all'orale, e piuttosto vicino allo standard, quindi i suoi tratti più 'critici' rappresentano principalmente casi particolari e isolati. Più grave risulta la gestione spesso errata di strutture che sono tipiche anche dello standard, come ad esempio le subordinate soggettive: la sistematicità con cui costruzioni come questa sono state gestite in maniera errata può fare pensare che si tratti di mancanze interne del modello, dovute all'assenza di *training data* annotati rilevanti.

Inoltre, le differenze nelle capacità dei modelli di analizzare i due corpora hanno permesso di evidenziare come le caratteristiche linguistiche del testo (in particolare, le caratteristiche inerenti la complessità sintattica) possano effettivamente influenzare la bontà dell'annotazione automatica stessa, che è, almeno in parte, condizionata dalle caratteristiche dei dati linguistici su cui il modello che la effettua è stato addestrato. In questo senso, si potrebbero forse aggiungere alla *treebank* ISDT nuovi dati appartenenti a domini come la lingua di Twitter, così da aumentare il grado di copertura dei modelli su di essa addestrati.

5. Appendice

Tabella 1 - ParlaMint2015: analisi sintattica con U.ISDT

Metric	Precision	Recall	F1 Score	AligndAcc
UAS	86.82	86.82	86.82	86.82
LAS	84.08	84.08	84.08	84.08
CLAS	78.73	79.70	79.21	79.70
MLAS	76.45	77.38	76.91	77.38
BLEX	76.99	77.93	77.45	77.93

Tabella 2 – ParlaMint 2020: analisi sintattica con U.ISDT

Metric	Precision	Recall	F1 Score	AligndAcc
UAS	85.28	85.28	85.28	85.28
LAS	82.13	82.13	82.13	82.13
CLAS	75.10	75.10	75.10	75.10
MLAS	72.20	72.20	72.20	72.20
BLEX	71.94	71.94	71.94	71.94

Tabella 3 – ParlaMint 2015: analisi sintattica con U.POSTWITA

Metric	Precision	Recall	F1 Score	AligndAcc
UAS	77.28	77.28	77.28	77.28
LAS	72.00	72.00	72.00	72.00
CLAS	65.70	64.99	65.34	64.99
MLAS	58.68	58.04	58.36	58.04
BLEX	62.95	62.26	62.60	62.26

Tabella 4 – ParlaMint 2020: analisi sintattica con U.POSTWITA

Metric	Precision	Recall	F1 Score	AligndAcc
UAS	78.06	78.06	78.06	78.06
LAS	72.86	72.86	72.86	72.86
CLAS	64.21	63.11	63.65	63.11
MLAS	56.84	55.86	56.35	55.86
BLEX	58.04	57.05	57.54	57.05

6. Bibliografia, sitografia e strumenti utilizzati

- K. COLLINS-THOMPSON, *Computational Assessment of Text Readability: A Survey of Current and Future Research*, 2014.
- V. GHENO, *Femminili singolari. Il femminismo è nelle parole*, effequ, Firenze, 2019.
- A. LENCI, S. MONTEMAGNI, V. PIRRELLI, *Testo e computer. Elementi di linguistica computazionale*, Carocci editore, Roma, 2016.
- C. D. MANNING, *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?*, in Alexander Gelbukh (ed.), "Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608", pp. 171-189, Springer, 2011.
- S. MONTEMAGNI, F. DELL'ORLETTA, G. VENTURI, *Esplorazioni computazionali nello spazio dell'interlingua: verso una nuova metodologia di indagine*, in R. BOMBI, V. ORIOLES, (a cura di), "Atti del XLVIII Congresso Internazionale di Studi della Società di Linguistica Italiana" (SLI 2014), 2016.
- J. NIVRE, *Towards a Universal Grammar for Natural Language Processing*, in A. GELBUKH (ed.), "Proceedings of CICLing 2015, Part I, LNCS 9041", pp. 3–16, Springer International Publishing Switzerland, 2015, e M.C. DE MARNEFFE, J. NIVRE, *Dependency Grammar*, in "Annual review of linguistics", 5:197, pp. 197-218, 2019.
- Universal Dependencies 2.5 Models for UDPipe (2019-12-06): https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131.
- ISST-TANL Tagsets: http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf. Sito delle Universal Dependencies: https://universaldependencies.org/.
- Sito per la consultazione della *treebank* ISDT: http://match.grew.fr/-?corpus=UD_Italian-ISDT@2.8.
- conllu.js: http://spyysalo.github.io/conllu.js/.
- CoNLL 2018 Shared Task Evalutation script: https://universaldependencies.org/-conll18/evaluation.html.