UNIVERSITÀ DI PISA

INFORMATICA UMANISTICA

DIPARTIMENTO DI FILOLOGIA, LETTERATURA E LINGUISTICA

Revisione e analisi dell'annotazione linguistica di un corpus di tweets

CORSO DI LINGUISTICA COMPUTAZIONALE II

Prof.ssa Simonetta Montemagni

Dott.ssa Giulia Venturi

Matricola: 587958

Studentessa: Silvia Cuozzo

Anno accademico: 2019/2020

1. Progetto e strumenti utilizzati

La seguente relazione riguarda il progetto di annotazione linguistica di testi rappresentativi di una specifica varietà d'uso della lingua italiana: la lingua di Twitter. Il progetto è stato assegnato durante il corso di Linguistica computazionale II tenuto dalla prof.ssa Montemagni e dalla Dott.ssa Venturi, ed è stato realizzato in collaborazione con la mia collega, Camilla Zucchi.

Ci è stato assegnato un corpus di tweets di 2.500 token riguardanti tre argomenti differenti ma molto attuali: l'ultima edizione 2020 di Sanremo, il Coronavirus e i Fridays for future ovvero gli scioperi organizzati per la sensibilizzazione sul cambiamento climatico. Seguendo le linee guida assegnateci durante il corso, abbiamo scaricato e installato la catena di annotazione linguistica (UDPipe) e il modello addestrato sulla *treebank PoSTWITa*. Il primo passaggio dell'annotazione automatica ha riguardato la tokenizzazione dei file con successiva revisione manuale, fatta individualmente, per poi procedere al confronto. Il secondo passaggio, invece, ha riguardato l'annotazione automatica del *Parts-of-speech tagging e syntactic parsing* con conseguente revisione manuale di: lemmatizzazione, annotazione morfo-sintattica e sintattica. Per entrambe le fasi abbiamo focalizzato l'attenzione sulla tipologia degli errori riscontrati.

Per la revisione manuale abbiamo utilizzato diversi strumenti come supporto per il nostro lavoro:

- Le linee guida presenti sul sito delle Universal Dependencies, facendo attenzione agli esempi e alle spiegazioni, delle relazioni di dipendenza e utilizzo del PoStagset[1];
- Un editor di testi presente su Ubuntu e visual code studio per correggere gli errori;
- UD annotatrix per visualizzare gli alberi di dipendenza.

Successivamente, abbiamo verificato l'accordo delle annotazioni fatte da me e la mia collega mediante l' *Inter-Annotator agreement* calcolato con uno script in python.

Prima di effettuare l'ultima fase del progetto, ci siamo confrontate sulle nostre rispettive revisioni e abbiamo discusso di alcuni casi di disaccordo, dopodiché abbiamo unificato il nostro corpus di *tweets*. A questo punto, abbiamo verificato il

[1] https://universaldependencies.org/u/dep/
https://universaldependencies.org/u/pos/
http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

livello di accuratezza di UDPipe nell'analisi dei *tweets* impiegando un modello addestrato sulla stessa varietà di lingua (*postwita*) e uno di varietà diversa (*isdt*), effettuato quindi due diversi confronti rispetto allo stesso *file gold.* Per portare a termine questo tipo di lavoro, è stato utilizzato lo script di valutazione *CONLL 2018.*

Per l'analisi statistica abbiamo utilizzato sia R sia Exel, infine il nostro lavoro è stato rivisto e migliorato anche grazie ai seminari, durante i quali abbiamo avuto la possibilità di confrontarci con i nostri colleghi e risolvere dubbi comuni riguardanti i punti salienti del progetto, permettendoci così di proseguire con maggiore facilità.

2. Analisi del corpus

Il nostro corpus contiene *tweets*, letteralmente "cinguettio", messaggi di breve lunghezza pubblicati dagli utenti su *Twitter,* servizio di notizie e microblogging. Il testo del messaggio può contenere sia *hashtag* ovvero parole precedute dal simbolo del cancelletto *#*, che permettono di creare un collegamento ipertestuale con tutti i messaggi che condividono lo stesso hashtag e di conseguenza trattano dello stesso argomento, sia *menzioni* ovvero nomi utenti preceduti dalla *@*. *Twitter* ha ormai raggiunto una grande popolarità, utilizzato per discutere di argomenti d'attualità da persone di tutte le età. Analizzando i *tweets* del nostro corpus, possiamo affermare che da un punto di vista qualitativo, la tendenza della lingua è molto vicina al parlato per la presenza di costruzioni impersonali, mancanza di soggetti espliciti, costruzioni paratattiche e quindi di poche subordinate, l'uso di interiezioni e di vocativi. In particolare, l'uso del maiuscolo principalmente nel file di "Sanremo2020" sottolinea una volontà dell'utente di avvicinarsi al parlato, quasi come se stesse urlando, scelta dettata probabilmente dalla volontà di trasmettere maggiore enfasi. Abbiamo anche riscontrato una sottile differenza: il file "Sanremo 2020" ha un linguaggio molto colorito e ironico attribuibile ad un target di utenti più giovani. Esso è diverso dal linguaggio degli altri due file, in cui è presente un linguaggio leggermente più formale, ciò è imputabile alle diverse tematiche trattate. Infatti, "Sanremo2020" tratta argomenti leggeri e divertenti, la maggior parte dei *tweets* si concentrano sul personaggio che ha creato più scalpore durante quest'edizione di Sanremo, ovvero Achille Lauro, quindi sono commenti riguardanti il suo modo di vestire e la sua performance. Altri riguardano concorrenti in gara come Piero Pelù e commenti sul programma in generale. Diverso, invece, il contenuto degli altri due file: in "Coronavirus", i *tweets* variano da i problemi di gestione della pandemia che stiamo vivendo, al numero di vittime negli ospedali e nelle case di riposo; invece in "Fridays

for Future" trattano degli incendi in Australia, degli scioperi che hanno interessato tutta l'Italia, l'importanza di investire in risorse rinnovabili.

Da un punto di vista quantitativo, i dati originari si presentavano in questo modo:

Sanremo2020

- 936 tokens
- 440 parole tipo
- 31 frasi totali
- frase più lunga: 95 tokens
- frase più breve: 1 token

Coronavirus

- 909 tokens
- 464 parole tipo
- 38 frasi totali
- frase più lunga: 54 tokens
- frase più corta: 1 token

Fridays for Future

- 934 tokens
- 498 parole tipo
- 42 frasi totali
- frase più lunga: 63 tokens
- frase più breve: 1 token

A seguito della nostra revisione, il corpus ha subito delle modificazioni:

Sanremo2020

- 969 tokens
- 364 parole tipo
- 29 frasi
- frase più lunga: 62 tokens
- frase più breve: 22 tokens

- type/token ratio: 0.38
- densità lessicale: 0.40

Coronavirus

- 921 tokens
- 467 parole tipo
- 32 frasi totali
- frase più lunga: 59 tokens
- frase più breve: 13 tokens
- type/token ratio: 0.52
- densità lessicale: 0.38

Fridays for Future

- 940 tokens
- 500 parole tipo
- 24 frasi
- frase più lunga: 67 tokens
- frase più breve: 26 tokens
- type/token ratio: 0.53
- densità lessicale : 0.37

Come è possibile notare, per tutti e tre i file c'è stato un incremento anche se minimo dei token totali dovuto ad errori di tokenizzazione dell'algoritmo, riconducibili principalmente ad una mancata separazione del token dalla punteggiatura precedente o consecutiva. Il numero di frasi è diminuito poiché, anche in questo caso, l'algoritmo ha erroneamente splittato *tweets* che noi abbiamo unito, confuso dalla presenza di hashtag o menzioni che spesso ha considerato come *tweet* a sé. Completamente diversi sono i numeri di token della frase più lunga e più breve, ancora una volta conseguenza di un'errata separazione dei *tweets*.

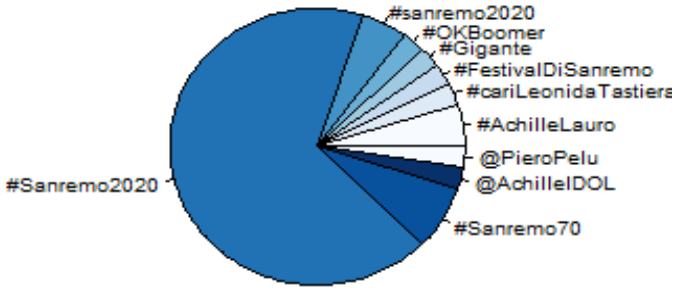
Per la versione finale del corpus, ho calcolato per ciascun file: la *Type/Token Ratio* ovvero l' indice di varietà lessicale del testo, ottenuto dal rapporto tra il numero di parole tipo al numeratore e il totale di occorrenze di unità del vocabolario al denominatore, restituendo un valore che oscilla tra 0 e 1. Sia "Coronavirus" sia "Fridays for future" hanno un valore che supera lo 0.50 per tanto possiamo

affermare che è un testo lessicalmente vario, invece il file “Sanremo2020” ha un valore dello 0.38 che testimonia un vocabolario poco variegato.

Un altro indice calcolato è la Densità lessicale che mette in luce invece aspetti della ricchezza del vocabolario, ottenuto dal rapporto tra il numero delle parole semanticamente piene e il totale dei tokens del testo. La densità lessicale risulta invece alquanto omogenea per tutti e tre i file.

Infine, grazie all’ausilio di R ho calcolato e rappresentato la frequenza degli elementi metalinguistici, hashtag e menzioni, in tutti e tre i file.

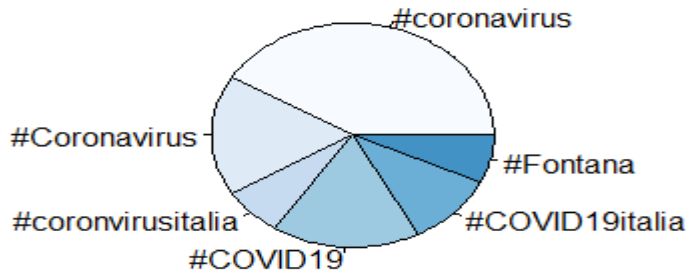
Sanremo2020



#AchilleLauro	#cariLeonidaTastiera	#FestivalDiSanremo	#Gigante	#OKBoomer
2	1	1	1	1
#sanremo2020	#Sanremo2020	#Sanremo70	@AchilleIDOL	@PieroPelu
2	27	3	1	1

In “Sanremo2020” gli elementi metalinguistici sono in totale 40 e costituiscono il 4,12% del testo. L’hashtag usato maggiormente è #Sanremo2020 con una frequenza pari a 27. Presenti solo due menzioni con una frequenza pari a 1, riferite ad Achille Lauro e Piero Pelù.

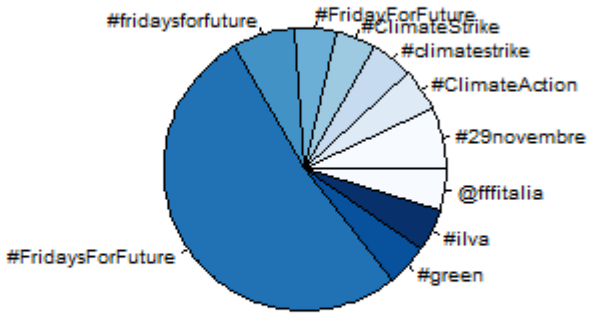
Coronavirus




```
#coronavirus      #Coronavirus #coronvirusitalia          #COVID19     #COVID19italia          #Fontana
           12                 5                  2                 5                  3                  2
```

Nel file “Coronavirus” gli elementi metalinguistici totali sono 77 e costituiscono l’8,36% del testo. In questo grafico, ho inserito gli hashtag più frequenti ovvero con una frequenza superiore a 1. Tutti gli hashtag fanno riferimento al coronavirus, quello con maggiore frequenza pari a 12 è #coronavirus, tranne l’hashtag riguardante il presidente della regione Lombardia, la regione più colpita durante la pandemia.

Fridays for future




```
      #29novembre     #ClimateAction     #climatestrike     #ClimateStrike  #FridayForFuture #fridaysforfuture
                3                  2                  2                  2                 2                  3
#FridaysForFuture            #green              #ilva         @fffitalia
               22                  2                  2                  2
```

Infine, anche nel grafico di “Fridays for future” ho inserito gli hashtag e menzioni maggiormente frequenti su un totale di 125 elementi che costituiscono il 13.29% del testo. L’hashtag con maggiore frequenza pari a 22 è #FridaysForFuture.

Ciò che possiamo osservare è che gli hashtag sono molto più frequenti delle menzioni, infatti esse hanno quasi sempre frequenza 1 tranne nel caso di @fffitalia. Inoltre, il file con un maggior numero di elementi metalinguistici risulta essere “Fridays for future”.

2.1 Sentence Splitting

E’ possibile suddividere il lavoro di revisione effettuato sul corpus in ben cinque fasi, che di seguito verranno analizzate nel dettaglio:

- Sentence Splitting
- Tokenizzazione
- Lemmatizzazione
- PoS Tagging
- Relazione di dipendenza

La *sentence splitting*, la separazione del testo in frasi, è la prima operazione svolta dall’algoritmo in un processo di annotazione linguistica automatica. Questa fase ha richiesto particolare attenzione poiché abbiamo riscontrato diversi errori di separazione dei *tweets*, alcuni dovuti alla presenza di hashtag e menzioni che vengono analizzati dall’algoritmo come *tweets* indipendenti (Fig.1) . Essa risulta problematica anche in presenza di token inizianti per maiuscola all’interno della frase(Fig.2) oppure a causa della punteggiatura. Io e la mia collega, abbiamo deciso di unire i *tweets* che ci sembravano logicamente connessi; inoltre l’algoritmo considerava l’elemento *pic.twitter.com* ovvero le immagini come un elemento di unione tra i *tweets*, noi invece ci siamo accordate sul considerarlo un elemento di chiusura.

```
# newpar
# sent_id = 5
# text = #Coronavirus #Cengio
1       #Coronavirus    _       _       _       _       _       _       _       _
2       #Cengio _       _       _       _       _       _       _
```

Fig. 1, Coronavirus (non revisionato)

```
# newpar
# sent_id = 3
# text = Io sono stata zitta tutta la sera ma adesso prendo la macchina e vado corcare di botte tutti quelli che hanno fischiato
1       Io      _       _       _       _       _       _       _       _
2       sono    _       _       _       _       _       _       _       _
3       stata   _       _       _       _       _       _       _       _
4       zitta   _       _       _       _       _       _       _       _
5       tutta   _       _       _       _       _       _       _       _
6       la      _       _       _       _       _       _       _       _
7       sera    _       _       _       _       _       _       _       _
8       ma      _       _       _       _       _       _       _       _
9       adesso  _       _       _       _       _       _       _       _
10      prendo  _       _       _       _       _       _       _       _
11      la      _       _       _       _       _       _       _       _
12      macchina        _       _       _       _       _       _       _       _
13      e       _       _       _       _       _       _       _       _
14      vado    _       _       _       _       _       _       _       _
15      corcare _       _       _       _       _       _       _       _
16      di      _       _       _       _       _       _       _       _
17      botte   _       _       _       _       _       _       _       _
18      tutti   _       _       _       _       _       _       _       _
19      quelli  _       _       _       _       _       _       _       _
20      che     _       _       _       _       _       _       _       _
21      hanno   _       _       _       _       _       _       _       _
22      fischiato       _       _       _       _       _       _       _       _

# sent_id = 4
# text = Achille Lauro #Sanremo2020
1       Achille _       _       _       _       _       _       _       _
2       Lauro   _       _       _       _       _       _       _       _
3       #Sanremo2020    _       _       _       _       _       _       _       SpacesAfter=\n\n
```

Fig. 2, Sanremo2020 (non revisionato)

```
# sent_id = 22
# text = @RespSocialeRai @fffitalia @F4F_Turin #FridaysForFuture #22Dicembre pic.twitter.com/dQKWIjhOXg Anche l' #arte sciopera per i cambiamenti climatici. Succede
oggi a @Recanati_
1       @RespSocialeRai _       _       _       _       _       _       _       _
2       @fffitalia      _       _       _       _       _       _       _       _
3       @F4F_Turin      _       _       _       _       _       _       _       _
4       #FridaysForFuture       _       _       _       _       _       _       _       _
5       #22Dicembre     _       _       _       _       _       _       _       _
6       pic.twitter.com/dQKWIjhOXg      _       _       _       _       _       _       _       SpacesAfter=\n\n
7       Anche   _       _       _       _       _       _       _       _
8       l'      _       _       _       _       _       _       _       _
9       #arte   _       _       _       _       _       _       _       _
10      sciopera        _       _       _       _       _       _       _       _
11      per     _       _       _       _       _       _       _       _
12      i       _       _       _       _       _       _       _       _
13      cambiamenti     _       _       _       _       _       _       _       _
14      climatici       _       _       _       _       _       _       _       SpaceAfter=No
15      .       _       _       _       _       _       _       _       _
16      Succede _       _       _       _       _       _       _       _
17      oggi    _       _       _       _       _       _       _       _
18      a       _       _       _       _       _       _       _       _
19      @Recanati_      _       _       _       _       _       _       _       _

# sent_id = 23
# text = , nelle sale dedicate a #Leopardi di Villa Colloredo Mels. L'opera scelta è “Leopardi sul letto di morte” (G. Ciaranfi, 1838-1902). Non sarà visitabile.
#29novembre #FridaysForFuture #museumsforfuture pic.twitter.com/6CZiHZWlSh raga ma la finite di dare per scontato che tutta la gente scesa a manifestare oggi l’abbia
fatto solo per saltare scuola? #FridayForFuture #FridaysForFuture
```

Fig.3, Fridays for future (non revisionato)

Nella Fig.3 appartenente al file “Fridays For Future”, abbiamo un chiaro esempio di errata sentence splitting causata dalla presenza di hashtag e menzioni. Abbiamo deciso, dopo un’attenta riflessione, di ricondurre le menzioni, gli hashtag e l’immagine iniziale al *tweet* precedente, anche per coerenza con gli argomenti trattati; di far iniziare quindi il 22 *tweet* con “Anche l’#arte” ed unirlo al tweet successivo (23), facendolo terminare con il *pic.twitter.com*, da noi ritenuto

elemento conclusivo. Il resto, invece, costituisce un altro *tweet*. Di seguito la nostra correzione.

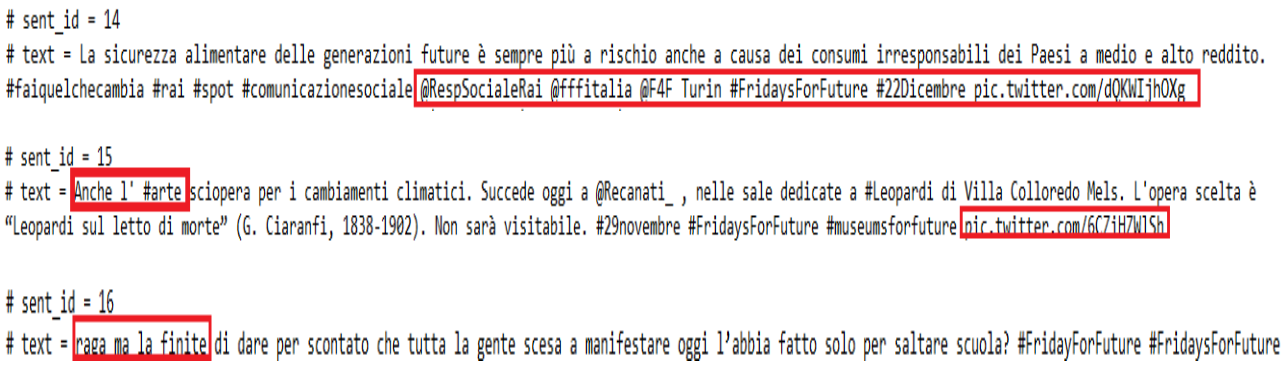
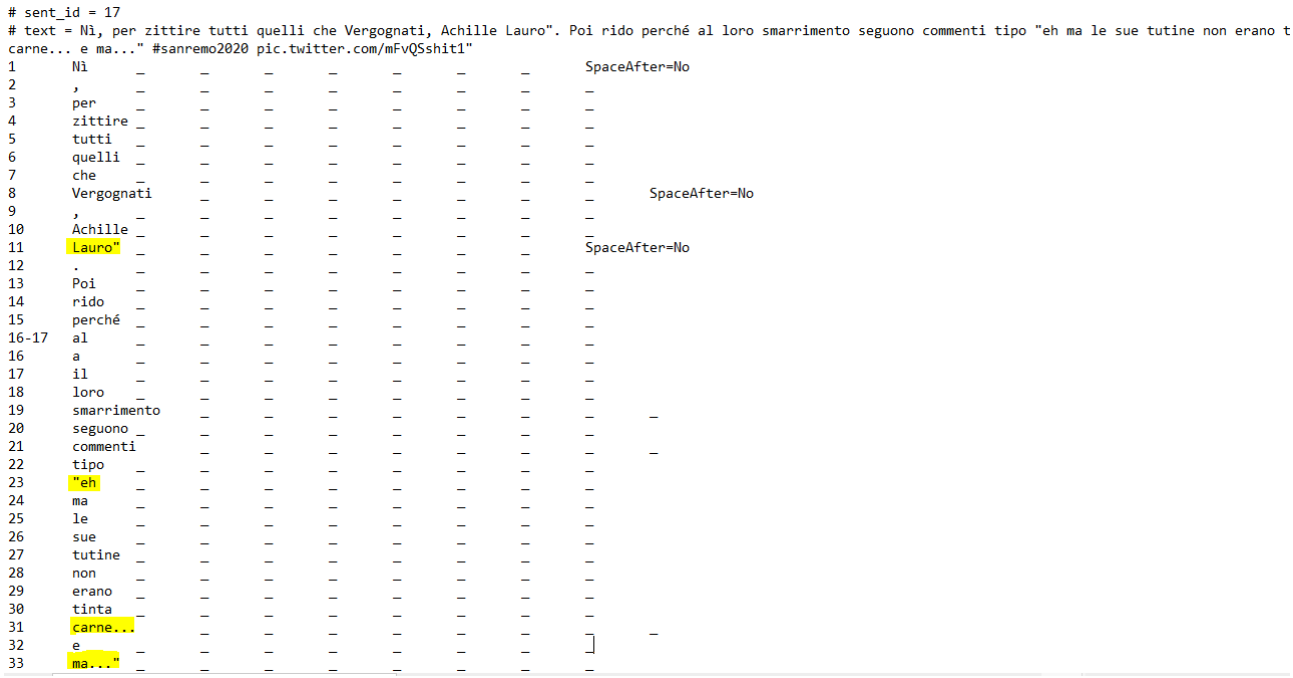


Fig.4, Fridays for future (revisionato)

In generale, la sentence splitting risulta maggiormente problematica per i file "Sanremo2020" e "Fridays for future" a causa della maggiore presenza di punteggiatura, hashtag, menzioni e maiuscole.

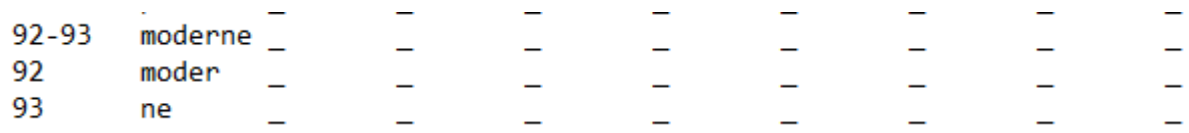
2.2 Errori di tokenizzazione

La tokenizzazione riguarda la divisione di sequenze di caratteri in unità minime definite appunto token, sono le unità di base per i successivi livelli di analisi. Nel corpus abbiamo notato molti errori, tutti riconducibili ad un'errata segmentazione dei token in presenza di punteggiatura, quest'ultima infatti sia che lo preceda o segua viene legata ad esso. Al momento della revisione, separando la punteggiatura, abbiamo ottenuto un aumento dei token per frase e anche a livello complessivo. Di seguito riporto alcuni esempi significativi.



```
# sent_id = 17
# text = Nì, per zittire tutti quelli che Vergognati, Achille Lauro". Poi rido perché al loro smarrimento seguono commenti tipo "eh ma le sue tutine non erano t
carne... e ma..." #sanremo2020 pic.twitter.com/mFvQSshit1"
1       Nì      _       _       _       _       _       _       _       SpaceAfter=No
2       ,       _       _       _       _       _       _       _       _
3       per     _       _       _       _       _       _       _       _
4       zittire _       _       _       _       _       _       _       _
5       tutti   _       _       _       _       _       _       _       _
6       quelli  _       _       _       _       _       _       _       _
7       che     _       _       _       _       _       _       _       _
8       Vergognati      _       _       _       _       _       _       _       SpaceAfter=No
9       ,       _       _       _       _       _       _       _       _
10      Achille _       _       _       _       _       _       _       _
11      Lauro"  _       _       _       _       _       _       _       SpaceAfter=No
12      .       _       _       _       _       _       _       _       _
13      Poi     _       _       _       _       _       _       _       _
14      rido    _       _       _       _       _       _       _       _
15      perché  _       _       _       _       _       _       _       _
16-17   al      _       _       _       _       _       _       _       _
16      a       _       _       _       _       _       _       _       _
17      il      _       _       _       _       _       _       _       _
18      loro    _       _       _       _       _       _       _       _
19      smarrimento     _       _       _       _       _       _       _       _
20      seguono _       _       _       _       _       _       _       _
21      commenti        _       _       _       _       _       _       _       _
22      tipo    _       _       _       _       _       _       _       _
23      "eh     _       _       _       _       _       _       _       _
24      ma      _       _       _       _       _       _       _       _
25      le      _       _       _       _       _       _       _       _
26      sue     _       _       _       _       _       _       _       _
27      tutine  _       _       _       _       _       _       _       _
28      non     _       _       _       _       _       _       _       _
29      erano   _       _       _       _       _       _       _       _
30      tinta   _       _       _       _       _       _       _       _
31      carne...        _       _       _       _       _       _       _       _
32      e       _       _       _       _       _       _       _       _
33      ma..."  _       _       _       _       _       _       _       _
```

Fig.5, Sanremo2020 (non revisionato)



```
92-93   moderne _       _       _       _       _       _       _       _
92      moder   _       _       _       _       _       _       _       _
93      ne      _       _       _       _       _       _       _       _
```

Fig. 6, Sanremo2020 (non revisionato)

In questa figura, possiamo notare come l'algoritmo tokenizzi erroneamente la parola "moderne" considerando evidentemente "ne" come un clitico.

Un altro caso particolare che ci ha incuriosite è il seguente, in cui l'algoritmo unisce diverse parole, segni di punteggiatura e numeri in un unico token. Effettuando delle ricerche, abbiamo scoperto che " " è il codice del *thin space*, chiaramente non decodificato; abbiamo quindi separato i cinque token differenti.

```
# newpar
# sent_id = 7
# text = Coronavirus , nuova chiusura almeno fino al 18 aprile e poi riapertura a tappe: tutte le ipotesi in campo per la fine del lockdown
https://www.ilsole24ore.com/art/nuova-chiusura-e-poi-riapertura-tappe-ipotesi-campo-la-fine-lockdown-AD6a1sG …
1       Coronavirus     _       _       _       _       _       _       _       _
2       ,       _       _       _       _       _       _       _       _
3       nuova   _       _       _       _       _       _       _       _
4       chiusura        _       _       _       _       _       _       _       _
5       almeno  _       _       _       _       _       _       _       _
6       fino    _       _       _       _       _       _       _       _
7-8     al      _       _       _       _       _       _       _       _
7       a       _       _       _       _       _       _       _       _
8       il      _       _       _       _       _       _       _       _
9       18      _       _       _       _       _       _       _       _
10      aprile  _       _       _       _       _       _       _       _
11      e       _       _       _       _       _       _       _       _
12      poi     _       _       _       _       _       _       _       _
13      riapertura      _       _       _       _       _       _       _       _
14      a       _       _       _       _       _       _       _       _
15      tappe: tutte      _       _       _       _       _       _       _       _
```

Fig.7, Coronavirus (non revisionato)

2.3 Errori di lemmatizzazione

Il compito successivo dell'algoritmo è di assegnare ad ogni token, il lemma corrispondente. Analizzando il corpus abbiamo notato che, spesso, l'algoritmo sbaglia, assegnando a volte anche lemmi "fantasiosi".

```
# newpar
# sent_id = 29
# text = I fischi non erano fischi da teatro ma fischi da concerto dai su non si può fischiare ad Achille su da no #sanremo2020 #FestivalDiSanremo
1       I       il      DET     RD      Definite=Def|Gender=Masc|Number=Plur|PronType=Art       2       det     _       _
2       fischi  fischio NOUN    S       Gender=Masc|Number=Plur 5       nsubj   _       _
3       non     non     ADV     BN      PronType=Neg    5       advmod  _       _
4       erano   essere  AUX     V       Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin    5       cop     _       _
5       fischi  fiscare NOUN    S       Gender=Masc|Number=Plur 0       root    _       _
6       da      da      ADP     E       _       7       case    _       _
7       teatro  teatro  NOUN    S       Gender=Masc|Number=Sing 5       nmod    _       _
8       ma      ma      CCONJ   CC      _       9       cc      _       _
9       fischi  fiscare NOUN    S       Gender=Masc|Number=Plur 7       conj    _       _
10      da      da      ADP     E       _       11      case    _       _
11      concerto        concerto        NOUN    S       Gender=Masc|Number=Sing 9       nmod    _       _
12-13   dai     _       _       _       _       _       _       _
12      da      da      ADP     E       _       18      mark    _       _
13      i       il      DET     RD      Definite=Def|Gender=Masc|Number=Plur|PronType=Art       18      det     _       _
14      su      su      ADV     B       _       18      advmod  _       _
15      non     non     ADV     BN      PronType=Neg    18      advmod  _       _
16      si      si      PRON    PC      Clitic=Yes|Person=3|PronType=Prs        18      expl    _       _
17      può     potere  AUX     VM      Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   18      aux     _       _
18      fischiare       fischiare       VERB    V       VerbForm=Inf    9       acl     _       _
19      ad      ad      ADP     E       _       20      case    _       _
20      Achille Achilla PROPN   SP      _       18      obl     _       _
21      su      su      ADP     E       _       23      case    _       _
22      da      da      ADP     E       _       23      case    _       _
23      no      no      INTJ    I       _       20      nmod    _       _
24      #sanremo2020    #sanremo2020    SYM     SYM     _       5       dep     _       _
25      #FestivalDiSanremo      #FestivalDiSanremo      VERB    V       Mood=Ind|Number=Plur|Person=1|Tense=Fut|VerbForm=Fin    5       parataxis       _
SpacesAfter=\n\n
```

Fig. 8, Sanremo2020 (non revisionato)

Abbiamo riscontrato diverse difficoltà da parte dell'algoritmo per i nomi comuni e nomi propri finenti in *-e* che riconduce in maniera errata ad un lemma terminante in *-a/ -o*, come nel caso di "Achille", oppure nel caso di "vacanze" lemmatizzato come "vacanzo".

Problemi anche con nomi e verbi contenenti la "h", in questo caso riconosce che si tratta di un nome ma lo riconduce al verbo inesistente "fiscare".

```
# newpar
# sent_id = 17
# text = Codognè si ferma per un minuto come tutta l'Italia per onorare coloro che hanno perso la vita al Coronavirus http://wveneto.blog/2020/03/31/codogne-si-ferma-
per-un-minuto-come-tutta-litalia-per-onorare-coloro-che-hanno-perso-la-vita-al-coronavirus/ … pic.twitter.com/kLYAwL2ciZ #coronavirus contro #Amazon . il colosso
licenzia il capo delle proteste dei lavoratori che avevano chiesto più protezione :(
1       Codognè Codognè SCONJ   CS      _       3       mark    _       _
2       si      si      PRON    PC      Clitic=Yes|Person=3|PronType=Prs        3       expl:impers     _       _
3       ferma   fermare VERB    V       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    _       _
4       per     per     ADP     E       _       6       case    _       _
5       un      uno     DET     RI      Definite=Ind|Gender=Masc|Number=Sing|PronType=Art       6       det     _       _
6       minuto  minuto  NOUN    S       Gender=Masc|Number=Sing 3       obl     _       _
7       come    come    ADP     E       _       9       case    _       _
8       tutta   tutto   DET     DI      PronType=Ind    9       det     _       _
9       l'Italia        l'Italia        PROPN   SP      _       6       nmod    _       _
10      per     per     ADP     E       _       11      mark    _       _
11      onorare onorarire       VERB    V       VerbForm=Inf    3       advcl   _       _
```

Fig.9, Coronavirus (non revisionato)

```
# sent_id = 5
# text =  saluta i partecipanti alla marcia per il clima #FridaysForFuture . #FFFSiena https://youtu.be/HqhBJXpkJZ8di @YouTube
1                       PROPN   SP      _       2       nsubj   _       _
2       saluta  salutare        VERB    V       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    _       _
3       i       il      DET     RD      Definite=Def|Gender=Masc|Number=Plur|PronType=Art       4       det     _       _
4       partecipanti    partecipante    NOUN    S       Gender=Masc|Number=Plur 2       obj     _       _
5-6     alla    _       _       _       _       _       _       _
5       a       a       ADP     E       _       7       case    _       _
6       la      il      DET     RD      Definite=Def|Gender=Fem|Number=Sing|PronType=Art        7       det     _       _
7       marcia  marcio  NOUN    S       Gender=Fem|Number=Sing  2       obl     _       _
8       per     per     ADP     E       _       10      case    _       _
9       il      il      DET     RD      Definite=Def|Gender=Masc|Number=Sing|PronType=Art       10      det     _       _
10      clima   clima   NOUN    S       Gender=Masc|Number=Sing 7       nmod    _       _
11      #FridaysForFuture       #FridaysForFuturo       PROPN   SP      _       10      nmod    _       _
12      .       .       PUNCT   FS      _       2       punct   _       _
13      #FFFSiena       #FFFSiena       SYM     SYM     _       2       dep     _       _
14      https://youtu.be/HqhBJXpkJZ8di  https://youtu.be/HqhBJXpkJZ8di  PROPN   SP      _       2       parataxis       _       _
15      @YouTube        @YouTube        PROPN   SP      _       14      flat:name       _       SpacesAfter=\n\n

# newpar
# sent_id = 6
# text = Le idee geniali per dare una mano all'ambiente a volte artivano dai rifiuti... (una metafora?)
1       Le      il      DET     RD      Definite=Def|Gender=Fem|Number=Plur|PronType=Art        2       det     _       _
2       idee    idea    NOUN    S       Gender=Fem|Number=Plur  0       root    _       _
3       geniali geniale ADJ     A       Number=Plur     2       amod    _       _
4       per     per     ADP     E       _       5       mark    _       _
5       dare    dare    VERB    V       VerbForm=Inf    2       acl     _       _
6       una     uno     DET     RI      Definite=Ind|Gender=Fem|Number=Sing|PronType=Art        7       det     _       _
7       mano    mano    NOUN    S       Gender=Fem|Number=Sing  5       obj     _       _
8       all'ambiente    all'ambiente    ADV     B       _       5       advmod  _       _
9       a       a       ADP     E       _       10      case    _       _
10      volte   volta   NOUN    S       Gender=Fem|Number=Plur  5       obl     _       _
11      artivano        artire  VERB    V       Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin    5       advcl   _       _
```

Fig.10, Fridays for future (non revisionato)

In due degli esempi riportati, riscontriamo problemi riguardanti la lemmatizzazione del verbo. Nel primo caso aggiunge erroneamente la desinenza *-ire*, nel secondo caso invece notiamo come un errore ortografico dell'utente, "artivano" presumibilmente "arrivano", abbia ripercussioni anche nella fase di lemmatizzazione. Infine, nel caso di "marcia", l'algoritmo identifica correttamente che si tratta di un NOUN ma lo lemmatizza come se fosse un aggettivo.

Per quanto concerne gli hashtag, un'osservazione preliminare ha portato me e la mia collega a ritenere la lemmatizzazione degli hashtag invariata rispetto al token, a volte infatti nomi e verbi dopo il # venivano lemmatizzati.

2.4 Errori di Part-of- Speech tagging e syntactic parsing

Una cattiva lemmatizzazione inevitabilmente condiziona e produce degli errori a cascata nei livelli di analisi successivi.

Una particolare attenzione e revisione è stata necessaria per i link, ai quali di volta in volta veniva assegnato un tag morfosintattico differente (aggettivo, nome proprio, verbo), invece per quanto concerne le relazioni di dipendenza la maggior parte delle volte erano identificati in maniera corretta come *dep,* dal momento che non appartengono alla struttura sintattica della frase.

Problematica si è rivelata anche la trattazione delle relazioni di dipendenza di menzioni e hashtag nel caso in cui essi non ricoprano una funzione sintattica. Gli hashtag, in particolare, vengono utilizzati nei social per sottolineare il tema trattato mediante richiami a parole, sigle, luoghi e per tanto si riferirebbero almeno idealmente all'intero *tweet*. Alla luce di questo ragionamento, io e la mia collega, abbiamo deciso di comune accordo di riferirli alla *root* e di utilizzare rispettivamente *parataxis:hashtag* e *vocative:mention*.

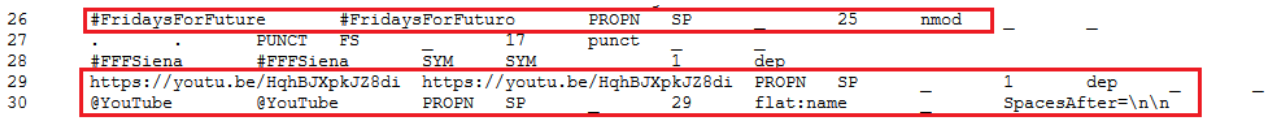


Fig.11, Fridays for future (non revisionato)

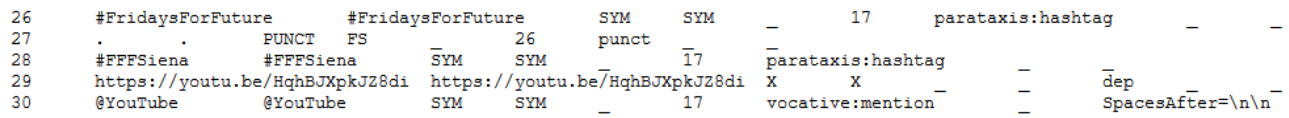
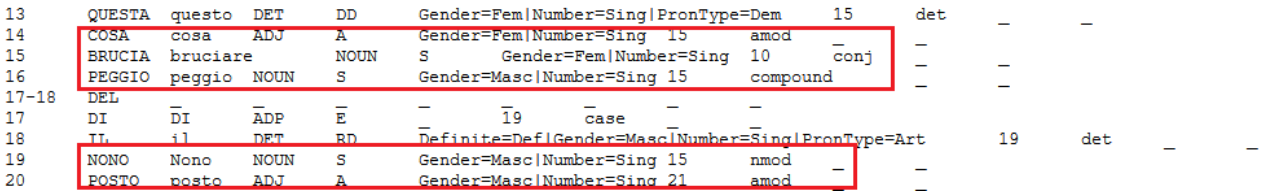


Fig.12, Fridays for Future (revisionato)

Il prossimo esempio, mette in evidenza il problema degli errori a cascata. Possiamo notare che, nonostante la lemmatizzazione dei token presi in esame sia giusta, il parser commette degli errori sia per i tag morfosintattici sia per quelli sintattici.

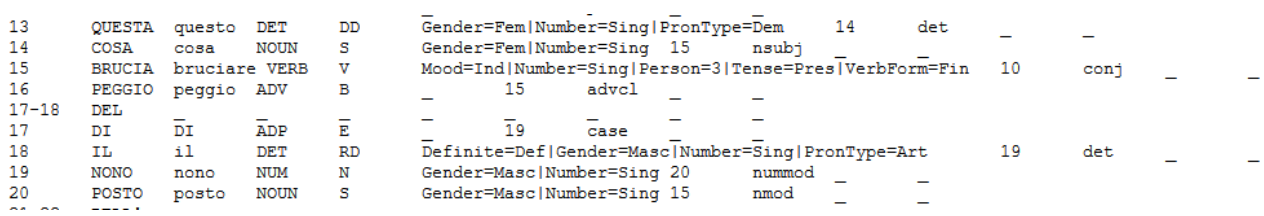


13	QUESTA	questo	DET	DD	Gender=Fem|Number=Sing|PronType=Dem	15	det	_	_	
14	COSA	cosa	ADJ	A	Gender=Fem|Number=Sing	15	amod	_	_	
15	BRUCIA	bruciare	NOUN	S	Gender=Fem|Number=Sing	10	conj	_	_	
16	PEGGIO	peggio	NOUN	S	Gender=Masc|Number=Sing	15	compound	_	_	
17-18	DEL	_	_	_	_	_	_	_		
17	DI	DI	ADP	E	_	19	case	_	_	
18	IL	il	DET	RD	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	19	det	_	_	
19	NONO	Nono	NOUN	S	Gender=Masc|Number=Sing	15	nmod	_	_	
20	POSTO	posto	ADJ	A	Gender=Masc|Number=Sing	21	amod	_	_	

Fig. 13, Sanremo2020 (non revisionato)

Abbiamo quindi corretto il tag morfosintattito ADJ A di “COSA” come NOUN S e il tag sintattico *amod* in *nsubj.* Per quanto riguarda “BRUCIA” abbiamo assegnato il tag corretto VERB V, modificato le features di un nome in quelle di un verbo; “PEGGIO” ha subito il passaggio dal tag morfosintattico NOUN S in ADV B e anche il tag sintattico è stato modificato in *advcl.*

Infine, “NONO” viene classificato erroneamente come NOUN S ma invece è NUM N e abbiamo corretto non solo il tag sintattico *nmod* in *nummod* ma anche la dipendenza, riferendolo al token successivo “POSTO”. Anche quest’ultimo è erroneamente classificato come ADJ A, lo abbiamo per tanto modificato in NOUN S e trasformato il tag sintattico da *amod* in *nmod.* Di seguito, la nostra correzione.



13	QUESTA	questo	DET	DD	Gender=Fem|Number=Sing|PronType=Dem	14	det	_	_	
14	COSA	cosa	NOUN	S	Gender=Fem|Number=Sing	15	nsubj	_	_	
15	BRUCIA	bruciare	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	10	conj	_	_	
16	PEGGIO	peggio	ADV	B	_	15	advcl	_	_	
17-18	DEL	_	_	_	_	_	_	_		
17	DI	DI	ADP	E	_	19	case	_	_	
18	IL	il	DET	RD	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	19	det	_	_	
19	NONO	nono	NUM	N	Gender=Masc|Number=Sing	20	nummod	_	_	
20	POSTO	posto	NOUN	S	Gender=Masc|Number=Sing	15	nmod	_	_	

Fig.14, Sanremo2020 (revisionato)

Un altro errore che l’algoritmo commette di frequente è quello di classificare token inizianti per maiuscola come PROPN, come evidenziato negli esempi che seguono.

```
14      il      il      DET     RD      Definite=Def|Gender=Masc|Number=Sing|PronType=Art       15      det     _       _
15      campionato      campionato      NOUN    S       Gender=Masc|Number=Sing 11      obj     _       _
16      di      di      ADP     E       _       17      case    _       _
17      Serie   Serie   PROPN   SP      _       15      nmod    _       _
18      A       A       ADP     E       _       19      case    _       _
```

Fig.15, Coronavirus (non rivisto)

```
16      Madre   madre   PROPN   SP      _       14      obj     _       _
17      Terra   Terra   PROPN   SP              16      flat:name
```

Fig. 16, Fridays for future (non revisionato)

```
31      Demand  Demand  PROPN   SP      _       30      flat:name       _       _
32      Action  Action  PROPN   SP      _       30      flat:name       _       _
```

Fig. 17, Fridays for future (non revisionato)

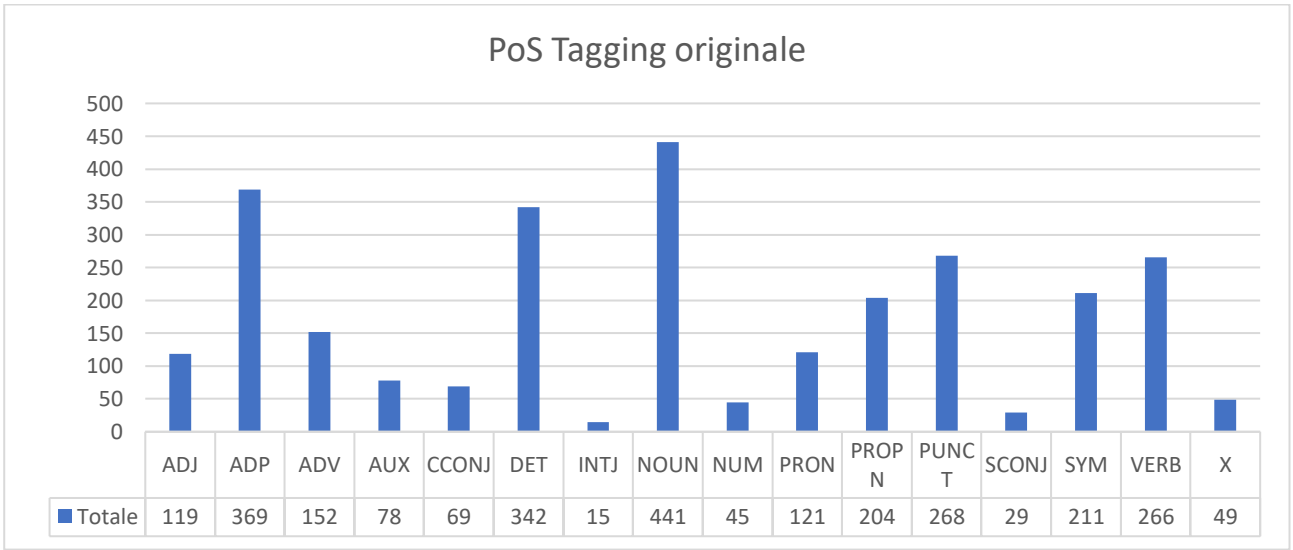


Grafico 1, Pos Tagging corpus originale

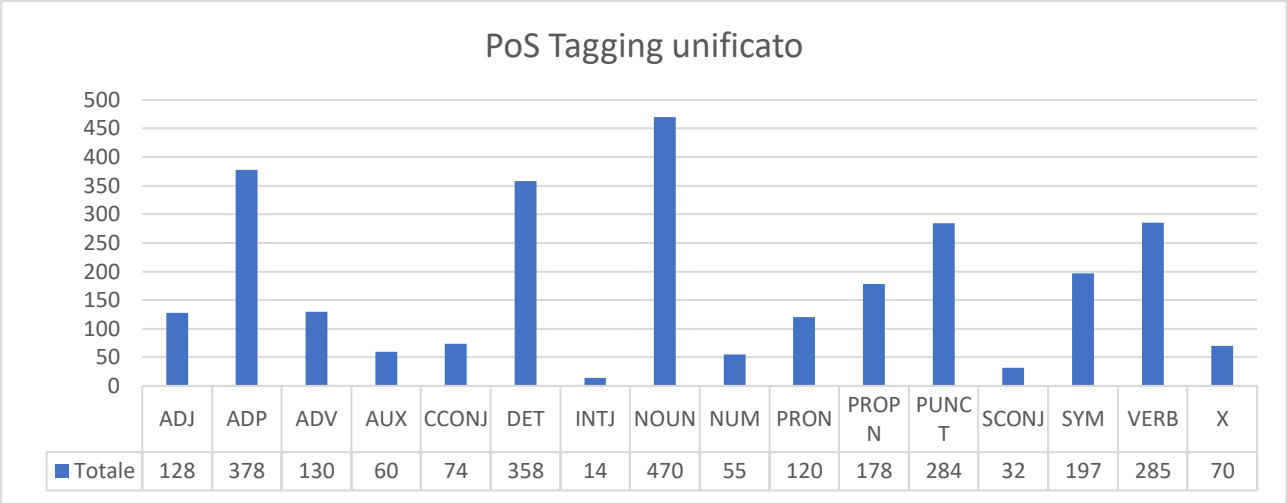


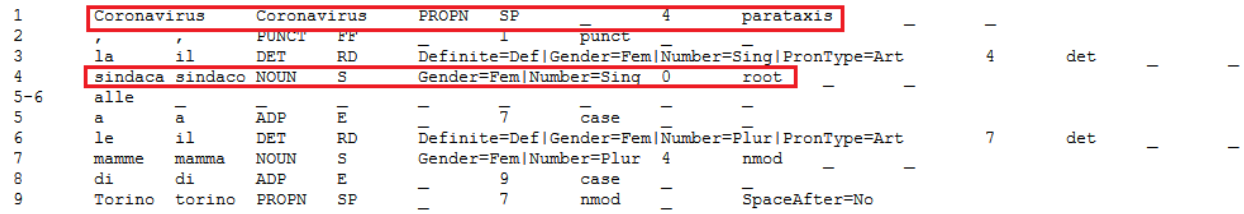
Grafico 2, Pos tagging corpus unificato

In questi due grafici, si mostrano le frequenze assolute dei tag morfosintattici assegnati dall'algoritmo (Grafico 1) e quelle successive alla nostra revisione (Grafico 2). Nella maggior parte dei casi, gli incrementi o le riduzioni non sono particolarmente evidenti. Notiamo un incremento di VERB (+19), un decremento di PROPN (- 26) da attribuire, almeno in parte, alle modifiche precedentemente discusse apportate ai link e non solo, di conseguenza abbiamo un aumento (+21) del tag X assegnato ai link e al *pic.twitter.com*. L'incremento dei NOUN (+29) può essere attribuito alla modifica da PROPN a NOUN.

Un'altra problematica da noi riscontrata è stata la presenza dell'ellissi verbale, fenomeno tipico della lingua scritta ma anche di quella parlata, che il parser non riesce sempre ad identificare, assegnando di conseguenza tag sintattici sbagliati.

```
# text = Coronavirus , la sindaca alle mamme di Torino: L'ora d'aria per i bambini? Mi spiace restate in cas…
https://torino.repubblica.it/cronaca/2020/03/31/news/coronavirus_l_appello_di_mammeatorino_alla_sindaca_un_ora_d_aria_per_i
_nostri_bambini-252750480/ … di @repubblica Brava signora sindaca. Giusto così. @c_appendino #Torino"
1       Coronavirus     Coronavirus     PROPN   SP      _       0       root    _       _
2       ,       ,       PUNCT   FF      _       1       punct   _       _
3       la      il      DET     RD      Definite=Def|Gender=Fem|Number=Sing|PronType=Art        4       det     _       _
4       sindaca sindaco NOUN    S       Gender=Fem|Number=Sing  1       appos   _       _
5-6     alle    _       _       _       _       _       _       _
5       a       a       ADP     E       _       7       case    _       _
6       le      il      DET     RD      Definite=Def|Gender=Fem|Number=Plur|PronType=Art        7       det     _       _
7       mamme   mamma   NOUN    S       Gender=Fem|Number=Plur  4       nmod    _       _
8       di      di      ADP     E       _       9       case    _       _
9       Torino  torino  PROPN   SP      _       7       nmod    _       SpaceAfter=No
```

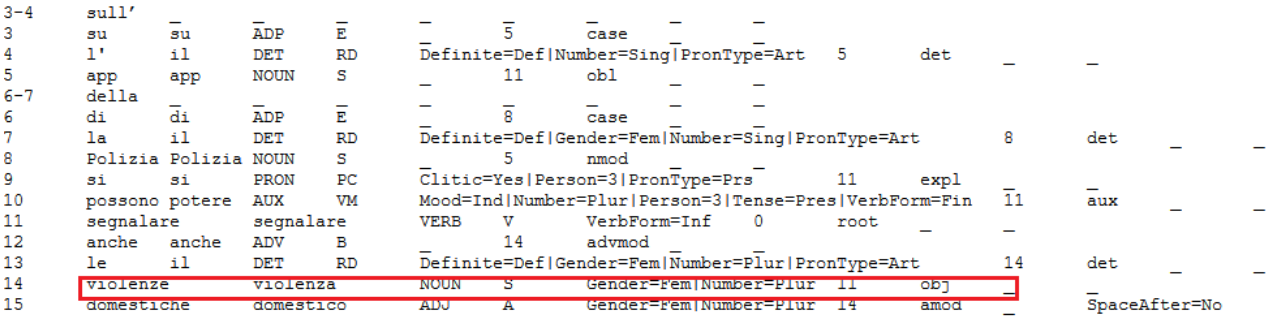
Fig.18, Coronavirus (non revisionato)



1	Coronavirus	Coronavirus	PROPN	SP	_	4	parataxis	_	_			
2	,	,	PUNCT	FF	_	1	punct	_	_			
3	la	il	DET	RD	Definite=Def|Gender=Fem|Number=Sing|PronType=Art					4	det	_	_
4	sindaca	sindaco	NOUN	S	Gender=Fem|Number=Sing	0	root	_	_			
5-6	alle	_	_	_	_	_	_					
5	a	a	ADP	E	_	7	case	_	_			
6	le	il	DET	RD	Definite=Def|Gender=Fem|Number=Plur|PronType=Art					7	det	_	_
7	mamme	mamma	NOUN	S	Gender=Fem|Number=Plur	4	nmod	_	_			
8	di	di	ADP	E	_	9	case	_	_			
9	Torino	torino	PROPN	SP	_	7	nmod	_	SpaceAfter=No			

Fig.19, Coronavirus (revisionato)

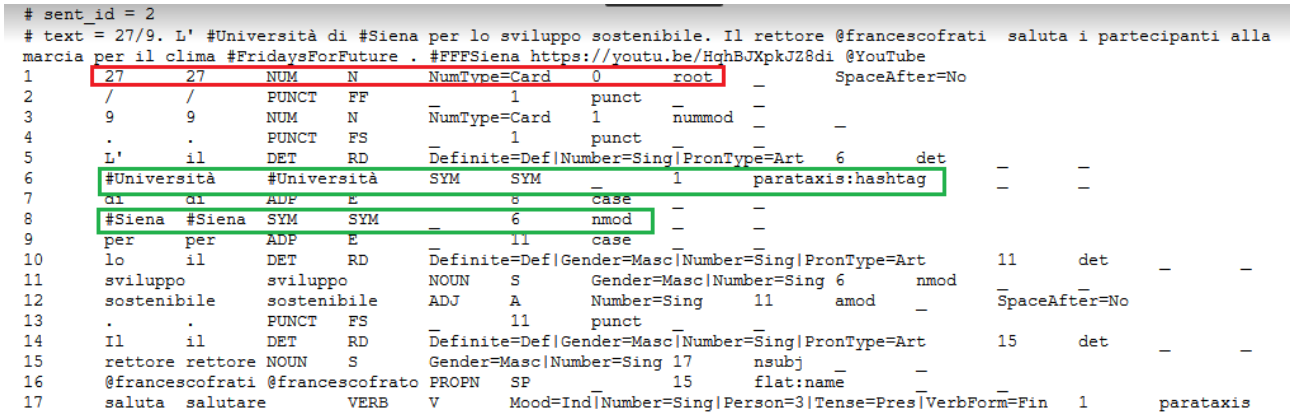
In questi due esempi, si può notare come il parser identifichi erroneamente come *root* "Coronavirus" anziché "sindaca", in mancanza di verbo la *root* è nominale. Inoltre, spesso abbiamo avuto *tweets* inizianti con parole, come nel caso di "Coronavirus" in questo esempio, non aventi legami sintattici con il resto della frase. Io e la mia collega abbiamo discusso a lungo, le nostre idee erano discordanti ma grazie al confronto con i nostri colleghi durante i seminari, siamo giunte alla conclusione di considerarli *parataxis* e farli dipendere dalla *root*.



3-4	sull'	_	_	_	_	_	_	_	_			
3	su	su	ADP	E	_	5	case	_	_			
4	l'	il	DET	RD	Definite=Def|Number=Sing|PronType=Art					5	det	_	_
5	app	app	NOUN	S	_	11	obl	_	_			
6-7	della	_	_	_	_	_	_	_	_			
6	di	di	ADP	E	_	8	case	_	_			
7	la	il	DET	RD	Definite=Def|Gender=Fem|Number=Sing|PronType=Art					8	det	_	_
8	Polizia	Polizia	NOUN	S	_	5	nmod	_	_			
9	si	si	PRON	PC	Clitic=Yes|Person=3|PronType=Prs					11	expl	_	_
10	possono	potere	AUX	VM	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin					11	aux	_	_
11	segnalare		segnalare		VERB	V	VerbForm=Inf	0	root	_	_		
12	anche	anche	ADV	B	_	14	advmod	_	_			
13	le	il	DET	RD	Definite=Def|Gender=Fem|Number=Plur|PronType=Art					14	det	_	_
14	violenze		violenza		NOUN	S	Gender=Fem|Number=Plur	11	obj	_	_		
15	domestiche		domestico		ADJ	A	Gender=Fem|Number=Plur	14	amod	_	SpaceAfter=No		

Fig. 20, Coronavirus (non revisionato)

In questo caso, il parser ha difficoltà nel riconoscere un sostantivo in funzione di soggetto postposto al verbo, riconosce infatti "violenze" come un complemento oggetto, dal momento che è quella la sua canonica posizione all'interno delle frasi.



```
# sent_id = 2
# text = 27/9. L' #Università di #Siena per lo sviluppo sostenibile. Il rettore @francescofrati  saluta i partecipanti alla
marcia per il clima #FridaysForFuture . #FFFSiena https://youtu.be/HqhBJXpkJZ8di @YouTube
1       27      27      NUM     N       NumType=Card    0       root    _       SpaceAfter=No
2       /       /       PUNCT   FF      _       1       punct   _       _
3       9       9       NUM     N       NumType=Card    1       nummod  _       _
4       .       .       PUNCT   FS      _       1       punct   _       _
5       L'      il      DET     RD      Definite=Def|Number=Sing|PronType=Art   6       det     _       _
6       #Università     #Università     SYM     SYM     _       1       parataxis:hashtag       _       _
7       di      di      ADP     E       _       8       case    _       _
8       #Siena  #Siena  SYM     SYM     _       6       nmod    _       _
9       per     per     ADP     E       _       11      case    _       _
10      lo      il      DET     RD      Definite=Def|Gender=Masc|Number=Sing|PronType=Art       11      det     _       _
11      sviluppo        sviluppo        NOUN    S       Gender=Masc|Number=Sing 6       nmod    _       _
12      sostenibile     sostenibile     ADJ     A       Number=Sing     11      amod    _       SpaceAfter=No
13      .       .       PUNCT   FS      _       11      punct   _       _
14      Il      il      DET     RD      Definite=Def|Gender=Masc|Number=Sing|PronType=Art       15      det     _       _
15      rettore rettore NOUN    S       Gender=Masc|Number=Sing 17      nsubj   _       _
16      @francescofrati @francescofrato PROPN   SP      _       15      flat:name       _       _
17      saluta  salutare        VERB    V       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   1       parataxis
```

Fig.21, Fridays for future (non revisionato)

Nell'esempio sopra riportato, possiamo notare come l'algoritmo riscontri delle difficoltà nell'assegnare la *root*, dal momento che la frase non segue la consueta struttura SVO, ma inizi con un periodo composto da una data e il periodo successivo con un ellissi verbale. Interessante notare come il parser non identifichi correttamente la funzione sintattica del primo hashtag identificandolo come *parataxis: hashtag,* invece identifica bene *nmod* del secondo hashtag.

3. Inter- Annotator Agreement

Un'altra fase del progetto ha riguardato il calcolo dell'*Inter- Annotator Agreement* ovvero il calcolo del grado di accordo tra me e Camilla sia per l'assegnazione delle PoS sia per l'assegnazione delle dipendenze, mediante lo script in python.

PoSTag	*Dipendenze*
Sanremo2020	
Average observed agreement: 0.97898989899 Kappa: 0.978892358504	Average observed agreement: 0.929221435794 Kappa: 0.928908833255

Coronavirus	
Average observed agreement: 0.965029821074 Kappa: 0.964473887746	Average observed agreement: 0.909452736318 Kappa: 0.90887959413
Fridays for future	
Average observed agreement: 0.981681034483 Kappa: 0.979794279742	Average observed agreement: 0.934358974359 Kappa: 0.934012245804

Come si può notare dai risultati riportati in tabella, l'accordo supera in tutti i casi il valore soglia, ma i valori risultano più alti per quanto riguarda il livello delle *PoS* rispetto a quello delle dipendenze.

Ciò potrebbe dipendere dal fatto che il livello di analisi delle dipendenze è stato quello che ha richiesto maggiore attenzione e maggiori dubbi, casi di disaccordo risolti dopo attente riflessioni, ricerche e confronto con gli esempi riportati sui siti messi a disposizione dalle docenti.

In particolare, i valori più bassi in termini di accordo delle dipendenze sono quelli di "Coronavirus", ciò probabilmente perché è il file in cui sono presenti più *tweets* contenenti elissi verbali e periodi brevi consecutivi che hanno causato maggiore difficoltà decisionale.

Un altro argomento di discussione è stata la presenza di token iniziali, non legati sintatticamente alla frase, inizialmente eravamo indecise tra il tag *obl* e *dislocated*, ma grazie al confronto avuto durante i seminari abbiamo optato per il tag *parataxis.* Abbiamo poi concordato per l'uso di *nmod* per i complementi di specificazione, tempo e spazio. Disaccordo si è avuto anche nell'assegnazione delle dipendenze della punteggiatura, riferita poi generalmente al token precedente.

Inoltre, inizialmente gli hashtag erano stati fatti dipendere dal token precedente o successivo, poi siamo giunte alla conclusione di farli dipendere dalla *root.*

Per quanto riguarda le PoS, attenzione particolare è stata data ad alcune parole straniere presenti nel testo, la maggior parte di esse erano hashtag ma altre erano semplici token. Abbiamo discusso sul modo di trattarle perché, sebbene siano parole straniere, ormai sono entrata a far parte del lessico comune, parliamo di prestiti linguistici provenienti principalmente dall'inglese. Per queste parole, abbiamo deciso di non assegnare il tag morfosintattico SW ma di trattarle come parole ormai integrate nel lessico dell'italiano.

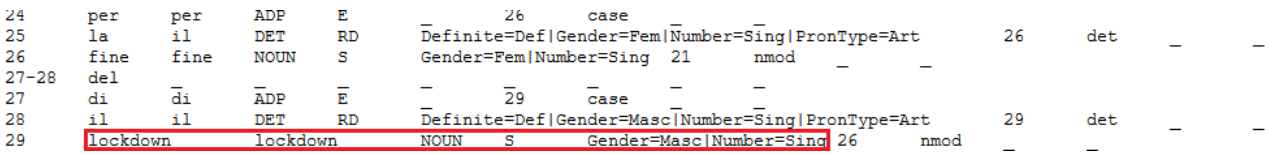


Fig. 22, Coronavirus (revisionato)

Fig. 23, Sanremo (revisionato)

4. Script di valutazione

L'ultima fase del progetto ha riguardato la valutazione dell'analisi dei *tweets* impiegando un modello addestrato sulla stessa varietà di lingua (modello Postwita) e uno su una varietà diversa (isdt), dunque sono stati fatti due confronti rispetto allo stesso *file-gold*. Per effettuare questi confronti abbiamo utilizzato lo script di valutazione *Conll 2018.* Di seguito, riporto gli output dello script di valutazione.

Sanremo2020

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	97.75	95.49	96.61	
Sentences	38.71	41.38	40.00	
Words	97.20	94.95	96.06	
UPOS	86.64	84.63	85.62	89.14
XPOS	88.79	86.74	87.75	91.35
UFeats	91.81	89.68	90.73	94.46
AllTags	84.05	82.11	83.07	86.47
Lemmas	90.52	88.42	89.46	93.13
UAS	60.78	59.37	60.06	62.53
LAS	49.89	48.74	49.31	51.33
CLAS	37.57	39.77	38.64	41.34
MLAS	33.09	35.04	34.04	36.42
BLEX	35.24	37.31	36.25	38.78

Fig.24, Poswita

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	98.13	97.91	98.02	
Sentences	14.00	24.14	17.72	
Words	97.57	97.37	97.47	
UPOS	75.84	75.68	75.76	77.73
XPOS	75.42	75.26	75.34	77.30
UFeats	80.38	80.21	80.30	82.38
AllTags	70.04	69.89	69.97	71.78
Lemmas	84.39	84.21	84.30	86.49
UAS	51.79	51.68	51.74	53.08
LAS	41.03	40.95	40.99	42.05
CLAS	27.18	29.55	28.31	30.41
MLAS	21.95	23.86	22.87	24.56
BLEX	25.44	27.65	26.50	28.46

Fig. 25, Isdt

Coronavirus

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	92.70	91.57	92.14	
Sentences	39.47	53.57	45.45	
Words	93.05	91.19	92.11	
UPOS	85.88	84.17	85.02	92.30
XPOS	86.63	84.91	85.76	93.10
UFeats	88.66	86.90	87.77	95.29
AllTags	83.64	81.97	82.80	89.89
Lemmas	89.52	87.74	88.62	96.21
UAS	60.00	58.81	59.40	64.48
LAS	52.73	51.68	52.20	56.67
CLAS	37.71	37.64	37.68	41.10
MLAS	32.65	32.58	32.61	35.58
BLEX	35.65	35.58	35.61	38.85

Fig.26, Poswita

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	92.36	95.12	93.72	
Sentences	8.33	17.86	11.36	
Words	91.75	94.44	93.08	
UPOS	76.78	79.04	77.89	83.68
XPOS	75.66	77.88	76.76	82.46
UFeats	79.12	81.45	80.27	86.24
AllTags	70.98	73.06	72.00	77.36
Lemmas	84.83	87.32	86.05	92.45
UAS	55.91	57.55	56.71	60.93
LAS	49.90	51.36	50.62	54.38
CLAS	37.01	37.08	37.04	39.76
MLAS	28.79	28.84	28.81	30.92
BLEX	33.64	33.71	33.68	36.14

Fig.27, Isdt

Fridays for future

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	95.54	93.72	94.62	
Sentences	2.38	4.17	3.03	
Words	95.19	93.15	94.16	
UPOS	82.49	80.73	81.60	86.67
XPOS	82.28	80.51	81.39	86.44
UFeats	87.96	86.08	87.01	92.41
AllTags	78.88	77.19	78.03	82.87
Lemmas	88.62	86.72	87.66	93.10
UAS	57.66	56.42	57.03	60.57
LAS	50.11	49.04	49.57	52.64
CLAS	37.17	36.70	36.93	39.37
MLAS	30.30	29.91	30.10	32.09
BLEX	34.94	34.50	34.72	37.01

Fig.28, Poswita

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	95.56	96.52	96.04	
Sentences	0.00	0.00	0.00	
Words	95.11	95.72	95.41	
UPOS	78.72	79.23	78.98	82.77
XPOS	77.45	77.94	77.69	81.43
UFeats	77.77	78.27	78.01	81.77
AllTags	69.57	70.02	69.80	73.15
Lemmas	85.21	85.76	85.49	89.60
UAS	54.04	54.39	54.22	56.82
LAS	46.06	46.36	46.21	48.43
CLAS	33.52	33.03	33.27	34.95
MLAS	23.65	23.30	23.48	24.66
BLEX	31.28	30.83	31.05	32.62

Fig. 29, Isdt

Diverse metriche di valutazione sono state computate per le diverse parti dell'annotazione (Tokens, Sentences, Words) : la *Precision* ovvero la misura della correttezza delle risposte del sistema, la *Recall* ovvero la misura della copertura del sistema, la *F1 Score* ovvero la media armonica tra *Precision* e *Recall*; infine *AligndAcc* indica la percentuale dei valori del modello riconosciuti rispetto al *file-gold*.

I valori sono in gran parte accettabili, notiamo come il modello Isdt sebbene sia un modello non addestrato sulla lingua dei *tweets* presenti alcuni valori più bassi rispetto al modello Poswita, in altri casi valori simili. Inoltre, il valore delle *sentence*

in alcuni file risulta essere molto basso, ciò potrebbe derivare dal fatto che sia il modello Poswita sia il modello Isdt creano tantissime frasi rispetto alla nostra *sentence splitting*, ad esempio il file gold "Fridays for future" contiene 24 frasi, invece il modello Poswita ne individua 42 e Isdt 59.

CONCLUSIONI

In conclusione, l'algoritmo ha difficoltà per quanto riguarda la *sentence splitting* a causa della presenza di hashtag e menzioni di cui non comprende la collocazione giusta, in particolare inizio e fine frase; problematica anche la corretta identificazione della funzione sintattica di questi elementi metalinguistici e delle loro dipendenze.

Le proposte di miglioramento potrebbero riguardare la fase di lemmatizzazione, in cui abbiamo riscontrato diversi errori, la fase di tokenizzazione dove i problemi riguardano principalmente la punteggiatura che viene legata al token precedente o successivo, nonché l'assegnazione delle PoS ai link e all'elemento *pic.twitter.com* che risulta quasi sempre sbagliata.