Relazione del progetto di Linguistica computazionale II

Eleonora Rossi

2020/21

1 Introduzione

La seguente relazione ha lo scopo di descrivere le varie fasi che hanno portato alla realizzazione dell'annotazione linguistica del corpus fornito e successivamente di esaminare in modo critico i risultati ottenuti, riportando osservazioni e analisi su aspetti particolari emersi.

In particolare la sezione 2 descrive il coprus su cui è stata effettuata l'annotazione linguistica; nella sezione 3 verranno descritti gli strumenti utilizzati in fase di annotazione e successive valutazioni; la sezione 4 fornisce una descrizione sommaria delle pipeline seguita in fase di sviluppo del progetto. La sezione 5 discute aspetti e considerazioni emerse durante la fase di revisione del corpus annotato automaticamente; la sezione 6 mostra l'accordo tra i due annotatori con relativa analisi di differenze tra le due annotazioni; la sezione 7 infine discute dei diversi livelli di accuratezza dell'annotazione automatica con i modelli ISDT e Postwita confrontandola con il gold costruito.

2 Corpus utilizzato

Il corpus utilizzato è una trascrizione di due sedute del Senato della Repubblica italiana avvenute in due periodi storici differenti. Tali dibattiti sono contenuti nei seguenti file:

- ParlaMint-IT_2015-03-11-LEG17-Sed-407-4: dibattito dell'11 marzo 2015
- ParlaMint-IT_2020-03-11-LEG18-Sed-200-4: dibattito dell'11 marzo 2020

In particolar modo il primo dibattito tratta delle problematiche relative all'affidamento di minori; mentre il secondo dibattito tratta delle problematiche emerse nelle carceri a seguito delle misure restrittive attuate al fine di contenere il diffondersi del coronavirus nelle strutture stesse.

Per quanto riguarda invece il registro presente all'interno del corpus, trattandosi appunto di trascrizioni di dibattici, è possibile incontrare molti aspetti tipici del parlato, come frasi mal formate, alcuni elementi colloquiali come incisi costituiti da "come dicevo", vocativi e così via. Dall'altro lato però lo stile si avvicina molto anche allo scritto poiché, trattandosi appunto di dibattici parlamentari, si è mantenuto sempre un tono piuttosto controllato e formale, discostandosi quindi dal parlato colloquiale e quotidiano. Dunque è possibile affermare che si tratta di un corpus particolare con un registro caratterizzato da duplici varietà linguistiche.

Andando un po' più nello specifico, di seguito sono riportate una serie di valori descrittivi relativi al corpus già segmentato e tokenizzato:

	Seduta 2015	Seduta 2020	**Totale** 2015 + 2020
Numero di **tokens**	1393	1438	2831
Numero di **words**	1480	1505	2985
Numero di **frasi**	42	51	93

Tabella 1: Tabella descrittiva del corpus

Dall'analisi dei dati nella tabella, si osserva che in generale il secondo testo relativo alla seduta del 2020 è costituito da un numero di tokens e frasi superiore rispetto al primo.

3 Strumenti utilizzati

Nelle varie fasi del progetto sono stati utilizzati diversi strumenti al fine di agevolare il più possibile la revisione e di conseguenza l'annotazione del coprus:

- La revisione manuale del corpus annotato automaticamente è stata fatta, facendo uso di: Notepad ++ e UD Annotatrix per la rappresentazione grafica dell'albero delle dipendenze
- La Treebank di riferimento è stata consultata attraverso il seguente portale online: http://match.grew.fr/

- L'inter-Annotator Agreement è stato calcolato impiegando uno script in python: script.py
- Il confronto tra il gold corpus e le analisi effettuate automaticamente con i modelli ISDT e Postwita è stato effettuato utilizzando lo script di valutazione distribuito in occasione dello CoNLL 2018 Shared task: *Multilingual Parsing from Raw Text to Universal Dependencies*

Il corpus è stato annotato seguendo le linee guida di Universal Dependencies (UD), un'iniziativa nata per sviluppare Treebank annotate in maniera coerente tra lingue diverse, secondo il modello ISDT.

4 Fasi del progetto

Le fasi del progetto possono essere riassunte nel modo seguente:

Prima fase: Annotazione automatica del corpus testuale attraverso la catena di annotazione UDPipe, utilizzando il modello ISDT

Seconda fase: Revisione manuale dell'annotazione automatica

Terza fase: Verifica dell'accordo tra i due revisori manuali

Quarta fase: Costruzione di un gold corpus

Quinta fase: Uso del gold corpus per la verifica della correttezza dell'analisi automatica dello stesso corpus utilizzando 2 modelli di UDPipe (ISDT e Postwita) addestrati su due diverse varietà della lingua italiana

La prima e la seconda fase del progetto sono procedute alternativamente: il corpus è stato in primis segmentato e tokenizzato attraverso la catena UDPipe, successivamente abbiamo revisionato il corpus correggendo eventuali errori di tokenizzazione o di sentence splitting; successivamente al corpus verticalizzato sono stati applicati il POS tagging e il syntactic parsing automatici e l'output è stato a sua volta revisionato manualmente.

La terza fase invece ha comportato, come già indicato, l'utilizzo di uno script per il calcolo dell'Inter-Annotator Agreement, ovvero il grado di accordo fra gli annotatori; questo viene calcolato a partire dal medesimo corpus revisionato autonomamente dalla mia collega ed io. Successivamente è stato costruito un gold corpus sulla base delle revisione effettuate e con quest'ultimo è stato calcolata la correttezza dell'annotazione automatica utilizzando i modelli ISDT e Postwita.

5 Revisione dell'annotazione

Partendo dalla segmentazione e tokenizzazione del corpus fino alla revisione dell'annotazione sintattica, è possibile rintracciare sia quantitativamente che qualitativamente gli errori fatti dal sistema.

5.1 Sentence splitting e tokenizzazione

Il primo step, come già anticipato, consiste in una prima revisione con particolare attenzione alla divisione del corpus in frasi. Da tale revisione è emerso un solo errore effettivo di segmentazione, il quale è presente nel testo relativo al 2015:

> L'affidamento dovrebbe essere totalmente diverso dall'adozione, e non perché sia una misura di serie **B.** È un gesto nobilissimo quello delle famiglie che accettano dei bambini in affidamento, [...]

Il sistema non ha segmentato tale frase probabilmente perché, a causa della presenza di una sola lettera scritta in maiuscolo, ha considerato il punto di fine frase come un punto di abbreviazione.

Mentre per quanto riguarda le modifiche, che abbiamo deciso di attuare rispetto a quanto è stato fatto dal sistema, emergono due casi: innanzitutto abbiamo deciso di seguire le linee guida proposte da UDPipe, ovvero considerare in generale i due punti e il punto e virgola come segni di punteggiatura forti, che portano quindi alla separazione delle frasi. Modifiche di questo genere sono state fatte solo per il testo relativo alle Sedute del 2020, poiché nel file del 2015 era presente un unica frase costituita da due punti che il sistema ha separato correttamente. Di seguito alcuni esempi di frasi che abbiamo separato:

- occorre mettere gli agenti di polizia in condizione di difendersi**;** occorre un piano nazionale sulle strutture penitenziarie.
- Non c'è più la lotta armata in questo Paese e la mafia ha abbassato il livello della sua violenza**;** quindi, aumenta la presenza di detenuti nelle nostre carceri in un contesto in cui, invece, il nostro Paese ha dimostrato di saper mettere in campo condizioni di sicurezza adeguate.

Il secondo caso riguarda alcune eccezioni ritrovate in entrambi i testi:

- abbiamo deciso di **unire** frasi separate da **punto e virgola** (e di considerare di conseguenza questo come un segno di punteggiatura debole) in caso di liste; tale punteggiatura infatti può essere sostituta da una virgola mantenendo il senso della frase

  > Da questo punto di vista, abbiamo la necessità di stare sicuramente uniti**;** di prendere misure straordinarie e - ripeto - anche dure nei confronti di chi si è reso responsabile di atti di violenza all'interno delle strutture.

- abbiamo deciso di **unire** frasi separate da **due punti** nel momento in cui la subordinata presente nella frase successiva ai due punti è retta dalla principale presente prima dei due punti:

  > Da lei avremmo preteso e pretendiamo qualcosa di più**:** una presa di posizione politica, perché lei rappresenta la responsabilità di questo

> Dicastero, una posizione che andasse al di là degli scontati ma dovuti apprezzamenti alla polizia penitenziaria e alle forze di polizia in genere.

Di seguito una tabella riassuntiva degli errori individuati in fase di revisione e delle modifiche da noi effettuate:

	Errori	**Modifiche**	
		Unione	Separazione
Seduta 2015	1	0	0
Seduta 2020	0	2	2
Totale	1	4	

Tabella 2: Tabella della distribuzione di errori e modifiche di segmentazione delle frasi

Una volta completato questo primo step, abbiamo proseguito con la revisione del corpus dal punto di vista della tokenizzazione.
In questo caso abbiamo riscontrato una sola tipologia di errore che tuttavia si è ripresentato più volte, ovvero il sistema non ha mai riconosciuto la costruzione dell'articolo partitivo. Secondo le linee guida delle Universal Depency, gli articoli partitivi dovrebbero essere mantenuti uniti e conseguentemente etichettati come articoli; il sistema tuttavia li ha sempre considerati come preposizioni complesse, separandoli di conseguenza in preposizione semplice e articolo.
Segue un esempio tratto dal corpus:

> Ci sono **delle** supplenze, **dei** trasferimenti, **degli** spostamenti e **degli** avanzamenti di carriera [...]

Tokenizzazione del sistema:

```
1 Ci _ _ _ _ _ _ _ _ _
2 sono _ _ _ _ _ _ _ _ _
3-4 delle _ _ _ _ _ _ _ _ _
3 di _ _ _ _ _ _ _ _ _
4 le _ _ _ _ _ _ _ _ _
5 supplenze _ _ _ _ _ _ _ _
6 ,_ _ _ _ _ _ _ _ _
7-8 dei _ _ _ _ _ _ _ _ _
7 di _ _ _ _ _ _ _ _ _
8 i _ _ _ _ _ _ _ _ _
9 trasferimenti _ _ _ _ _ _ _ _
10 ,_ _ _ _ _ _ _ _ _
11-12 degli _ _ _ _ _ _ _ _ _
11 di _ _ _ _ _ _ _ _ _
12 gli _ _ _ _ _ _ _ _ _
13 spostamenti _ _ _ _ _ _ _ _ _
14 e _ _ _ _ _ _ _ _ _
15-16 degli _ _ _ _ _ _ _ _ _
15 di _ _ _ _ _ _ _ _ _
16 gli _ _ _ _ _ _ _ _ _
17 avanzamenti _ _ _ _ _ _ _ _ _
18 di _ _ _ _ _ _ _ _ _
19 carriera _ _ _ _ _ _ _
```

Modifica:

```
1 Ci _ _ _ _ _ _ _ _ _
2 sono _ _ _ _ _ _ _ _ _
3 delle _ _ _ _ _ _ _ _ _
4 supplenze _ _ _ _ _ _ _ _
5 ,_ _ _ _ _ _ _ _ _
6 dei _ _ _ _ _ _ _ _ _
7 trasferimenti _ _ _ _ _ _ _ _
8 ,_ _ _ _ _ _ _ _ _
9 degli _ _ _ _ _ _ _ _ _
10 spostamenti _ _ _ _ _ _ _ _ _
11 e _ _ _ _ _ _ _ _ _
12 degli _ _ _ _ _ _ _ _ _
13 avanzamenti _ _ _ _ _ _ _ _ _
14 di _ _ _ _ _ _ _ _ _
15 carriera _ _ _ _ _ _ _
```

Durante questa fasi di revisione abbiamo anche deciso di apportare alcune modifiche per quanto riguarda evidenti errori di battitura nella stesura della trascrizione; le modifiche effettuate sono le seguenti:

- Correzione di E in È → **E** chiaro che dopo un trauma di questo genere il problema emerge.
- Correzione di INTERVIENE in INTERVENIRE → Ma non dovremmo piuttosto intervenire su ciò che sta alla radice, cioè sull'articolo 403 del codice civile, prima di **interviene** sulle conseguenze?
- Correzione di DELLE in DELLA → [...] che si intende inserire all'articolo 4 **delle** legge n. 184 del 1983.

Segue una tabella contenente il numero di errori presenti in questo livello:

	Errori	Percentuale di errori nel corpus	Errori di battitura
Seduta 2015	9	82 %	3
Seduta 2020	2	18 %	0
Totale	11	100 %	3

Tabella 3: Tabella della distribuzione di errori e modifiche di tokenizzazione

5.2 Annotazione POS e dipendenze

Ho deciso di presentare nello stesso paragrafo sia la revisione dell'annotazione POS che dell'annotazione a dipendenze poiché l'assegnamento scorretto di etichette POS a determinati token ha portato spesso, in fase di annotazione delle dipendenze, ad errori nei token stessi e in quelli ad essi attigui; dunque è interessante analizzare appunto questo fenomeno di errori a cascata, procedendo attraverso un'analisi in parallelo delle revisioni effettuate nei confronti dei due livelli di annotazione.

Come sappiamo la lemmatizzazione avviene dopo che è stata effettuata la disambiguazione attraverso il riconoscimento da parte del tagger del Part-Of-Speech del token, infatti è possibile notare come lo scorretto riconoscimento dal parte del tagger dell'etichetta e delle features porta ad uno scorretto riconoscimento del lemma:

- il token crea è stato etichettato come ADJ (invece di VERB) e dunque è stato lemmatizzato come crea invece di creare
- al token affezioni sono state assegnate features scorrette e ciò può aver portato il tagger a lemmatizzare come affezione invece di affezionare
- il token intese è stato etichettato come ADJ (invece di VERB) e dunque è stato lemmatizzato come intesa, invece di intendere

Tuttavia possiamo individuare alcune eccezioni, in cui nonostante sia stato identificato correttamente il POS del token, è stato assegnato il lemma scorretto. Di seguito alcuni esempi:

- il token Istituto è stato etichettato correttamente come NOUN, ma lemmatizzato come Istituto

- il token Signor è stato etichettato correttamente come NOUN, ma lemmatizzato come Signor
- il token andassimo è stato etichettato correttamente come VERB e sono state individuate le corrette features Mood=Sub|Number=Plur|Person=1|Tense=Imp|VerbForm=Fin, tuttavia è stato lemmatizzato come andassimo

Possiamo poi osservare anche casi in cui sono stati individuato come lemmi parole mal formate come:

- il token risorse è stato lemmatizzato come risorso
- il token profeti è stato lemmatizzato come profete
- il token famiglia è stato lemmatizzato come familiere
- il token sacrosante è stato lemmatizzato come sacrosante
- il token idonee è stato lemmatizzato come idonee

Da notare che negli utili due casi nell'individuazione del lemma scorretto potrebbe aver contribuito anche una assegnamento scorretto delle features.

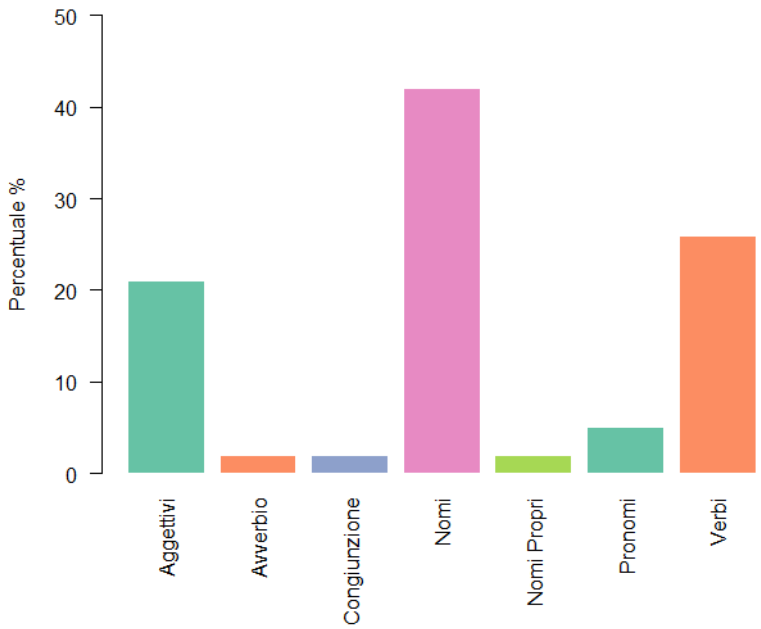

Figura 1: Rappresentazione grafica della distribuzione degli errori nei lemmi

In generale osservando la tabella che segue è possibili notare che il maggior numero di lemmi scorretti è individuabile nel secondo testo (70% di errori rispetto al 30% del primo testo), mentre dal grafico (Fig. 1) è possibile vedere come la maggior parte degli errori nella lemmatizzazione riguardano nomi (42%), verbi (26%) e aggettivi (21%).
Gli errori di lemmatizzazione nei nomi tuttavia sono dovuti a correzioni fatte a seguito del cambio di etichetta da PROPN a NOUN (15 casi su 24), poiché la mia collega ed io abbiamo deciso, in fase di confronto, di etichettare come nomi comuni tutti quei nomi con iniziale maiuscola che possono essere flessi, ad esempio è stato etichettato come NOUN:

- il token Ministro con lemma ministro
- il token Presidente con lemma presidente
- il token Aula con lemma aula

Inoltre tra i verbi lemmatizzati scorrettamente troviamo un caso particolare ovvero verbi al participio che possono essere utilizzati all'interno di una frase anche con funzione aggettivale, tra i lemmi scorretti troviamo:

- il token intese etichettato erroneamente come ADJ e quindi lemmatizzato come intesa
- il token armata anch'esso etichettato erroneamente come ADJ e quindi lemmatizzato come armato

Lo stesso problema vale per gli aggettivi sotto forma di participio, spesso confusi come verbi:

- il token dovuto etichettato erroneamente come VERB e quindi lemmatizzato come dovere
- il token premeditati etichettato erroneamente come VERB e quindi lemmatizzato come premeditare

Ci sono poi casi in cui, nonostante sia stato effettivamente riconosciuta la giusta etichetta di aggettivo o verbo, è stato comunque sbagliato il lemma, un esempio è il seguente:

> il token generalizzato è stato correttamente etichettato come ADJ
> ma lemmatizzato come generalizzare

Di seguito una tabella riassuntiva degli errori dei lemmi individuati in fase di revisione, facendo anche una distinzione tra le due Sedute:

	Freq assoluta		Percentuale $\frac{f}{tot_{Seduta}}$	
	Seduta 2015	Seduta 2020	Seduta 2015	Seduta 2020
Verbi	7	8	41 %	20 %
Aggettivi	3	9	18 %	22.5 %
Nomi	6	18	35 %	45 %
Nomi Propri	0	1	0 %	2.5 %
Pronomi le	1	2	6 %	2.5 %
Congiunzione si	0	1	0 %	2.5 %
Avverbio completamente	0	1	0 %	2.5 %
Totale errori nelle sedute	17	40	100 %	100 %

Di seguito invece la stessa tabella riassuntiva senza fare distinzioni tra le due Sedute:

	Freq assoluta	Percentuale $\frac{f}{tot_{corpus}}$
Verbi	15	26 %
Aggettivi	12	21 %
Nomi	24	42 %
Nomi Propri	1	2 %
Pronomi le	3	5 %
Congiunzione si	1	2 %
Avverbio completamente	1	2 %
Totale errori nel corpus	57	100 %

Per quanto riguarda l'analisi morfosintattica, dalle tabelle che seguono è possibile vedere che in generale il tagger ha commesso più errori in fase di POS tagging nel testo relativo alla Seduta 2020 (circa il 66% degli errori commessi sono nel testo del 2020, rispetto al 33% del testo del 2015); ma è possibile anche notare che in tutto il corpus di circa 2831 tokens solo il 3% dei tokens sono stati etichettati scorrettamente.

Di seguito una tabella riassuntiva della distribuzione di errori individuati nei POS dei token rispetto alle due sedute:

	Freq assoluta		Percentuale $\frac{f}{tot_{Seduta}}$	
	Seduta 2015	Seduta 2020	Seduta 2015	Seduta 2020
VERB	5	6	17 %	10 %
ADJ	8	10	28 %	17 %
NOUN	7	29	24 %	50 %
PRON	2	4	7 %	7 %
SCONJ	4	4	14 %	7 %
CCONJ	0	2	0 %	3 %
ADV	1	0	3 %	0 %
ADP	1	1	3 %	2 %
AUX	1	2	3 %	3 %
Totale errori	29	58	100 %	100 %

Di seguito invece la stessa tabella riassuntiva rispetto però a tutto il corpus:

	Freq assoluta	Percentuale $\frac{f}{tot_{corpus}}$
VERB	11	13 %
ADJ	18	21 %
NOUN	36	41 %
PRON	6	7 %
SCONJ	8	9 %
CCONJ	2	2 %
ADV	1	1 %
ADP	2	2 %
AUX	3	3 %
Totale errori nel corpus	87	100 %

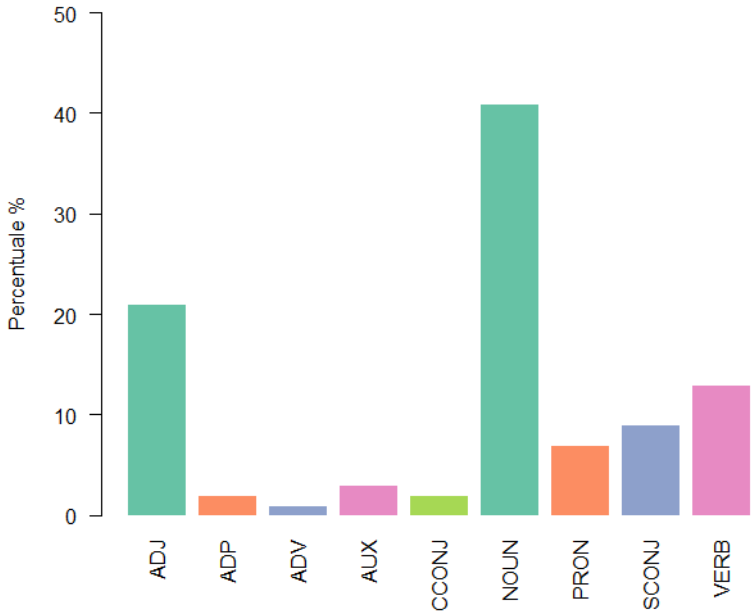

Figura 2: Rappresentazione grafica della distribuzione degli errori di POS tagging rispetto a tutto il corpus

Il grafico invece mostra come in tutto il corpus il maggior numero di errori è stato fatto nell'etichettare i nomi (41%), di questi abbiamo un maggior numero nella Seduta del 2020 (29 casi rispetto ai 7 del testo del 2015, quindi l'81% degli errori nei NOUN si trova nel testo del 2020).
Seguono poi gli aggettivi (21%), spesso confusi dal tagger come nomi (10 casi su 18, quindi il 55% delle volte), verbi (5 casi su 18, quindi il 28% delle volte) o altre categorie varie (3 casi su 18, quindi 17% delle volte).
Anche i verbi sono stati spesso etichettati in maniera scorretta (13%), i quali venivano spesso scambiati per aggettivi (5 casi su 11, ovvero il 45% delle volte), ausiliari (4 casi su 11, ovvero il 36% delle volte) o nomi (2 casi su 11, ovvero il 18% delle volte).

Andando più nello specifico dell'analisi morfosintattica, possiamo prendere di nuovo in considerazione come oggetto di studio il caso riguardante gli aggettivi sotto forma di participio, spesso confusi con verbi, e i verbi al participio, spesso confusi come aggettivi, su cui è necessario fare ulteriori osservazioni.
Questa distinzione è stata motivo di dibattito e confronto tra la mia collega ed io; confrontando infatti il gold corpus da noi creato e ciò che ha etichettato il tagger è possibile notare alcune di discrepanze:

Sentence	Token	Gold corpus	POS originale
37	intese	VERB	ADJ
2	armata	ADJ	VERB
10	prese	VERB	ADJ
39	premeditati	ADJ	VERB

Questo genere di errore ha portato ad un effetto a cascata nell'annotazione a dipendenze:

- Non c'è più la lotta armata in questo Paese e la mafia ha abbassato il livello della sua violenza;
  in questa frase armata è stata considerata dal parser come acl, mentre in realtà si tratta di una relazione amod
- Al contrario, se gli stessi fossero stati responsabilizzati e coinvolti, se si fosse detto loro che le misure prese erano finalizzate anzitutto a tutelare le loro condizioni di salute [...]
  in questo caso prese è stato considerato come amod, quando invece è acl
- Non possiamo arrenderci di fronte a questi atti premeditati di destabilizzazione dello Stato..
  anche in questo caso premeditati è da considerare come dipendente della relazione amod e non acl

Nelle figure che seguono invece è presente la frase contenente l'errore relativo al token intese, il quale ha portato a ulteriori errori all'interno della struttura sintattica (intendeva dipendente della relazione ccomp e non obl, con testa intese, non esistono)

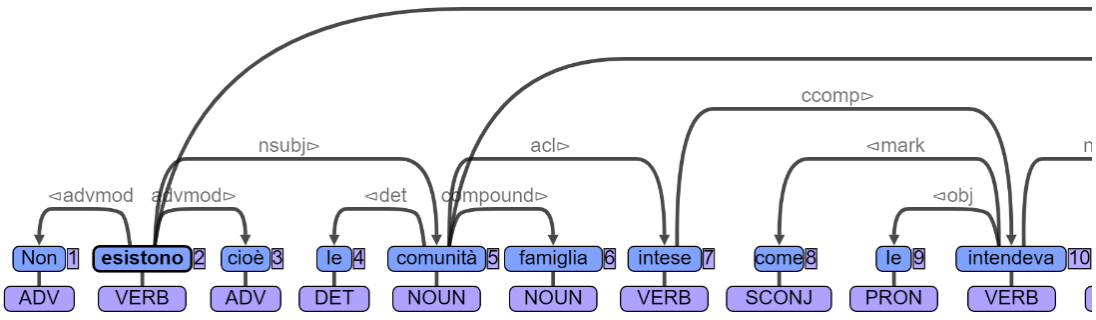

Figura 3: Annotazione delle dipendenze del gold corpus

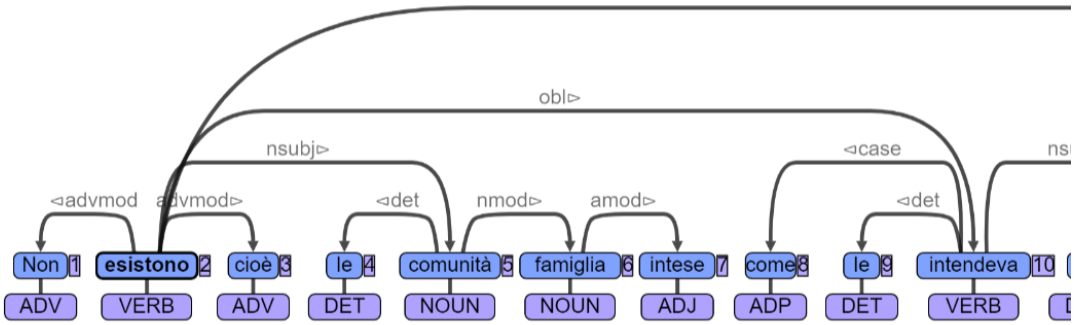

Figura 4: Annotazione delle dipendenze automatica

Un altro fenomeno piuttosto interessante che si è presentato più volte è la scorretta interpretazione da parte del POS tagger del token che; in quattro casi è stata assegnata una etichetta POS sbagliata:

1. SCONJ → PRON: Dal momento che le situazioni di affidamento dovrebbero servire temporaneamente per consentire poi il rientro nella famiglia originaria, il problema va spostato sugli aiuti e sulle attenzioni, **che** la comunità dovrebbe mettere al centro della propria azione:

2. PRON → SCONJ: Mettiamo il caso che i tribunali dei minori, volendo smaltire (questo è anche il leitmotiv), velocizzare e rendere più efficiente il sistema della giustizia, **decidessero**, se noi ponessimo alcuni termini temporali, **che** decorsi due anni alla famiglia affidataria vengano dati in adozione i figli.

3. PRON → SCONJ: Naturalmente, condivido quanto detto dal presidente Casini, ovvero che non si può imputare a lei, signor Ministro, la condizione delle carceri in questo momento, però sicuramente mi chiedo come si faccia a non **comprendere** - da parte di chi dirige le carceri in Italia - **che** in questa situazione così drammatica basta pochissimo per scatenare il caos.

4. PRON → SCONJ: È da anni **che** se ne parla.

Da queste frasi possiamo infierire che c'è stato un margine di dubbio da parte del sistema per quanto riguarda la possibilità che il token che sia una congiunzione subordinante o un pronome. In genere viene assegnata l'etichetta SCONJ nel momento in cui il token che è preceduto da un verbo, e su questa base è possibile dedurre che nelle frasi 2 e 3, l'errore da parte del sistema sia da imputare al fatto che è presente un inciso tra il verbo e il token che. Il problema si ripresenta anche con l'ultima frase, poiché in questo caso il token in questione non è preceduto da un verbo, ma da un nome e questo può aver portato il sistema a etichettare tale token come un pronome.
Come per il caso precedente anche qui l'errore in fase di POS tagging porta

a una scorretta analisi da parte del parser, in questo caso riporterò solo un esempio:

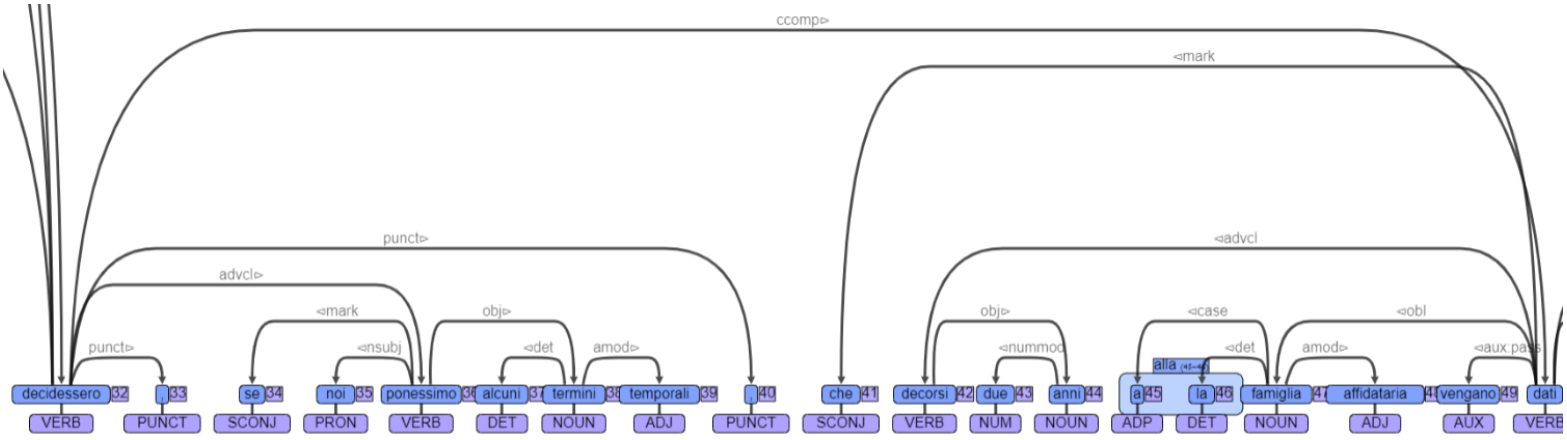

Figura 5: Annotazione a dipendenze del gold corpus

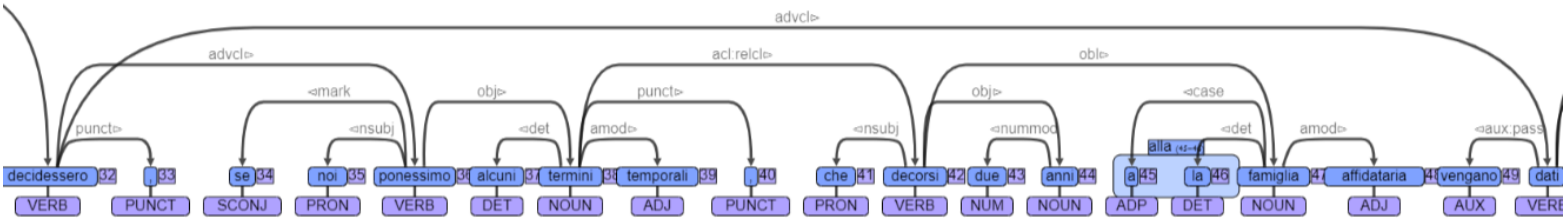

Figura 6: Annotazione a dipendenze del raw corpus

Un ultimo errore piuttosto comune che vorrei presentare è il POS tagging del token come. In fase di confronto la mia collega ed io abbiamo riscontrato molte diversità nel modo in cui noi e il tagger abbiamo etichettato il token come.
Dal confronto tra gold corpus e annotazione automatica possiamo vedere le seguenti discordanze:

1. ADV → SCONJ: Ma **come** dice la legge, giustamente, nelle parti che non vengono modificate, l'affidamento dovrebbe fare esattamente l'opposto dell'adozione e prevedere un progetto che consenta a questi bambini di ritornare nelle loro famiglie originarie.

2. ADP → SCONJ: Non esistono cioè le comunità famiglia intese **come** le intendeva la norma originaria, ovvero degli ambienti che replicano l'ambiente genitoriale.

3. ADP → CCONJ: Signor Ministro, abbiamo ritenuto un atteggiamento debole i suoi appelli alla calma rivolti verso i detenuti, **così come** non sono più sufficienti, per noi, solo i complimenti alle forze di polizia.

4. ADP → SCONJ **Come** le dicevo, vogliamo risposte forti.

Come per il token che, anche in questo caso la linea guida generale prevede che, se il token come introduce una frase, allora è da considerare come congiunzione; il problema tuttavia si crea nel momento in cui dobbiamo decidere se classificare tale token come una congiunzione coordinante o subordinante. All'interno della Treebank non abbiamo trovato molti esempi di come come congiunzione coordinante, per questo abbiamo deciso di utilizzare l'etichetta CCONJ solo nella terza frase dove il come è preceduto da così, e di prediligere l'etichetta SCONJ. Inoltre è possibile vedere dal grafico (Fig. 7) che segue che il tagger ha etichettato come per la maggior parte delle volte come ADP, anche in casi in cui era evidente che fosse una congiunzione, come ad esempio nella frase 4.

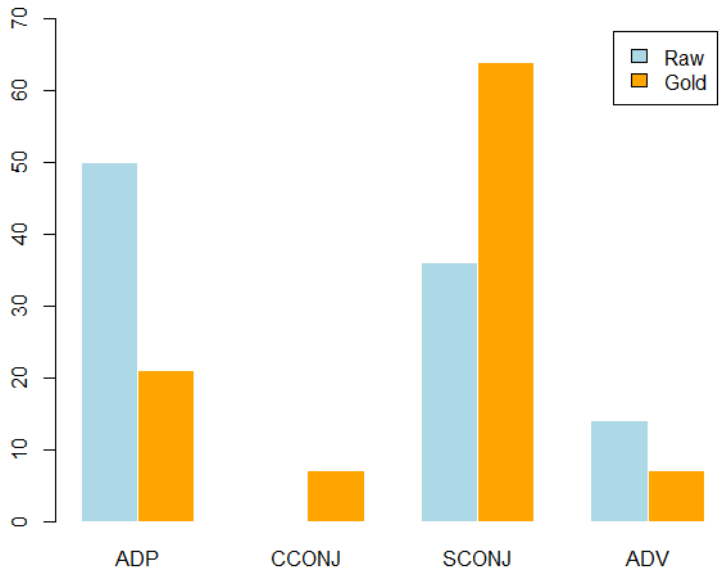

Figura 7: Rappresentazione grafica della distribuzione dei POS nel token come nell'automatica e nel gold corpus

É possibile spiegare l'errore nel POS tagging del token come nella frase 1 guardando il modo in cui è stata analizzata sintatticamente la frase: il sistema ha infatti considerato ma congiunzione coordinante di dice, quindi il token come è

stato considerato come semplice avverbio; questo errore è probabilmente dovuto alla mancanza della punteggiatura a indicare la presenza dell'inciso tra come e legge:

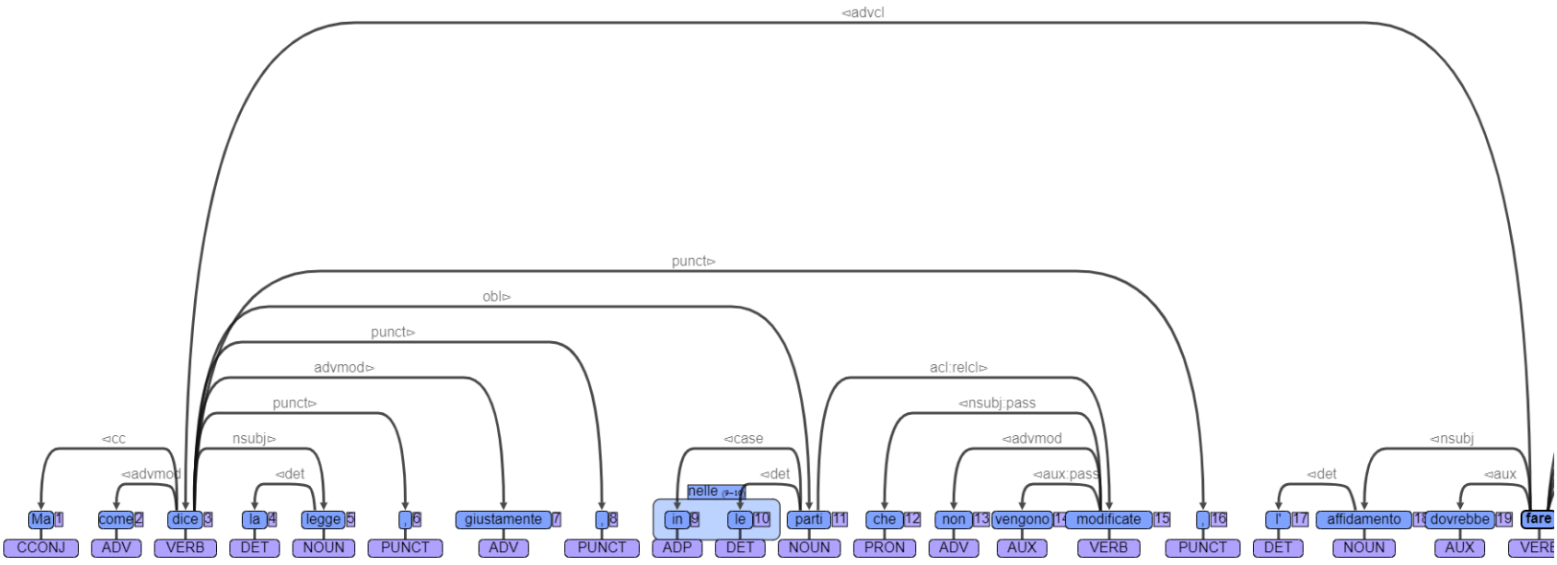

Figura 8: Annotazione a dipendenze automatica

Nella seconda frase invece l'errore di POS è probabilmente da imputare al fatto che il tagger ha etichettato il token intese come ADJ quindi come, non essendo preceduto da nessun verbo, è stato considerato come ADP.

Dal punto di vista dell'annotazione a dipendenze sono stati fatti in totale 442 errori, di cui 209 (circa il 47%) nel testo della seduta del 2015 e 233 (circa il 53%) nel testo della seduta del 2020. Gli errori sono stati molti, ma in generale è possibile notare alcune tendenze:

	Freq assoluta		Percentuale $\frac{f}{tot_{Seduta}}$	
	Seduta 2015	Seduta 2020	Seduta 2015	Seduta 2020
Passivo	16	6	8 %	3 %
Soggetto - oggetto	6	10	3 %	4 %
Radice	6	8	3 %	3 %
Testa congiunzione	11	11	5 %	5 %
Vocative	1	10	0.5 %	4 %
Paratassi	1	4	0.5 %	2 %
Apposizione	3	3	1 %	1 %

Tabella 4: Tabella contenente **alcuni** errori relativi alle relazioni di dipendenza

Un errore che è stato fatto piuttosto frequentemente è non riconoscere correttamente la radice della frase: si consideri infatti che nel testo della Seduta del 2015 su 42 frasi sono state sbagliate 6 radici, quindi il 14% delle volte, mentre nel testo del 2020 su 51 frasi sono state sbagliate 8 radici, ovvero il 16%. Questo tipo di errore ha spesso portato a una modifica totale della frase annotata automaticamente come nelle seguenti frasi:

- Frase 18: Forse sarebbe meglio avere un po' più di cautela e guardare all'insieme del provvedimento senza limitarci a dire:
  In questo caso il parser ha riconosciuto come radice avere, quando in realtà la radice dovrebbe essere meglio, questo porta ad una scorretta analisi di una buona parte della frase:

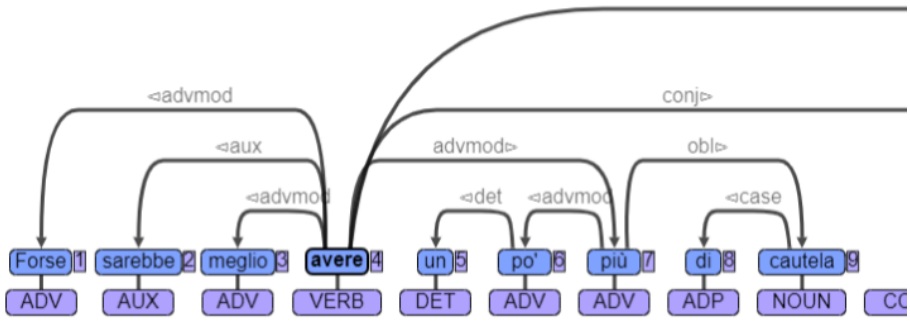

Figura 9: Annotazione automatica di una porzione della frase 18

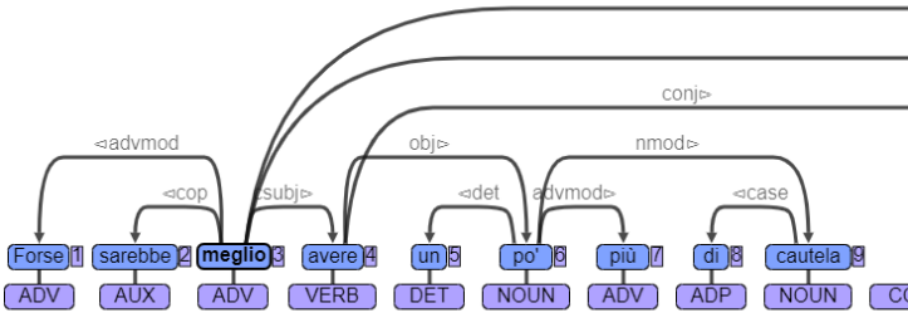

Figura 10: Revisione Gold corpus di una porzione della frase 18

- Frase 21: Signor Presidente, l'emendamento che vado a illustrare è unico e anche relativamente semplice. (il vocativo è stato riconosciuto come root). In questo frase è presente un errore abbastanza comune, ovvero riconoscere il vocativo come radice della frase, ed anche questo porta ad una modifica abbastanza massiccia della frase:

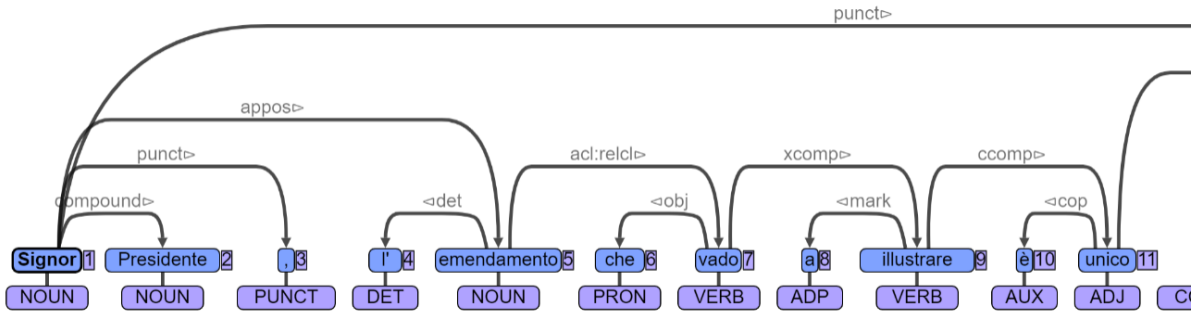

Figura 11: Annotazione automatica di una porzione della frase 21

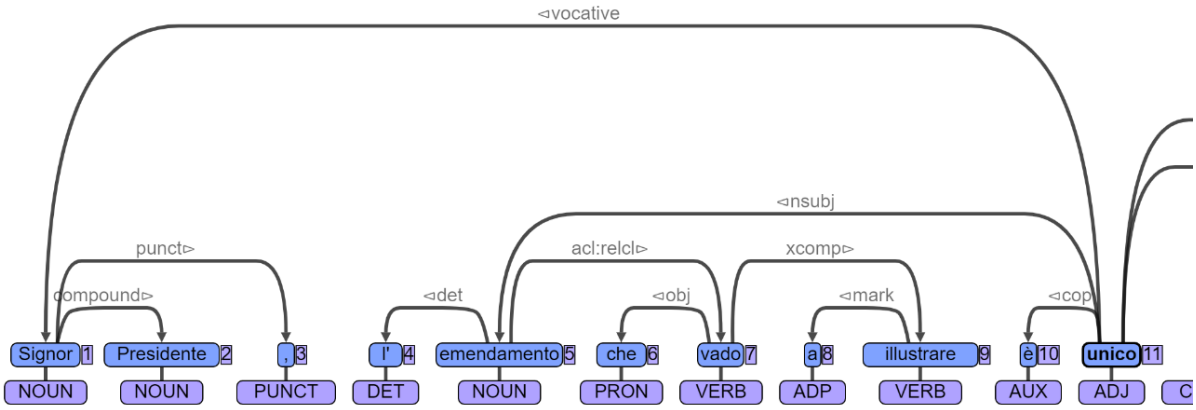

Figura 12: Revisione Gold corpus di una porzione della frase 21

- Frase 24: Anche il fatto di prevedere una data e di lasciare quella che, con una brutta parola, si può chiamare la via preferenziale o la prelazione della famiglia affidataria nella successiva adozione dei minori, crea non pochi rischi.
  In questo caso l'errato riconoscimento della radice potrebbe essere dovuto al fatto che il token crea, ROOT della frase, è posto in fondo alla frase, preceduto da una lunga sequenza di frasi subordinati e incisi; inoltre, un altro fatto probabilmente determinante potrebbe essere stato l'errato riconoscimento dell'etichetta di crea (indicata dal tagger come ADJ).
- Frase 1: Tutto questo, tra l'altro, quando l'indice di delinquenza - per fortuna, grazie alle Forze dell'ordine e grazie alle azioni messe in campo in questi anni - sta calando.

> Qui, l'errore dovrebbe essere conseguenza della mancanza di un vero verbo principale, infatti il parser ha considerato calando come radice, quando invece è questo.

Una altro errore che si è ripetuto con una certa costanza riguarda il passivo: infatti nel corpus su 39 passivi (tra verbi, ausiliari e soggetti) sono 7 i casi in cui il parser non ha riconosciuto il passivo, quindi il 17%; mentre per quanto riguarda il riconoscimento errato del passivo, il parser ha individuato erroneamente 15 passivi. Seguono alcuni esempi tratti dal primo testo relativi al mancato riconoscimento del passivo:

1. Ci troviamo in una situazione per la quale, visto che la **legge** oggi non **è** applicata (e sarà applicata ancora meno, se approveremo queste norme), finiamo col prendere atto di questa mancata applicazione.

2. L'insegnante deve prendersi la massima cura dei ragazzi **che** gli **sono** affidati, come deve fare l'affidatario, ma non sono suoi figli e non lo diventeranno.

Per quanto riguarda invece il riconoscimento errato del passivo troviamo:

1. L'**affidamento** dovrebbe essere totalmente diverso **dall'adozione**, e non perché sia una misura di serie B.
   in questa frase una possibile spiegazione dell'errato riconoscimento del passivo potrebbe essere data dal fatto che il parse ha indicato erroneamente come obl:agent dall'adozione.

2. Dunque, questi **figli sono** tornati con la propria madre originaria.

3. Il fatto che questi bambini vengono portati via, rapiti in una scuola, senza che i **genitori vengano** neanche informati su dove si trovino.

Un caso particolare da considerare riguarda la frase 32 del primo testo, infatti questa risulta essere mal formata poiché mancante di un verbo; durante il confronto la mia collega ed io abbiamo deciso di realizzare un'annotazione che dia quanto più possibile un senso alla frase:

> Questa dovrebbe essere l'estrema ratio, ma poiché nella stragrande maggioranza dei casi ciò avviene per ragioni economiche, ovvero per il fatto che la famiglia si trova a perdere il lavoro - come capita sempre più facilmente - e di conseguenza a non riuscire a pagare l'affitto dell'immobile in cui vive, a subire uno sfratto e a non trovare un luogo adatto in cui vivere.

La mal formazione della frase sta nella mancanza di un verbo coordinante avente come testa ratio e come congiunzione coordinante ma; per poter comunque fornire alla frase un'annotazione completa, abbiamo deciso di legare ma a perché e di considerare quest'ultimo come mark avente testa avviene.

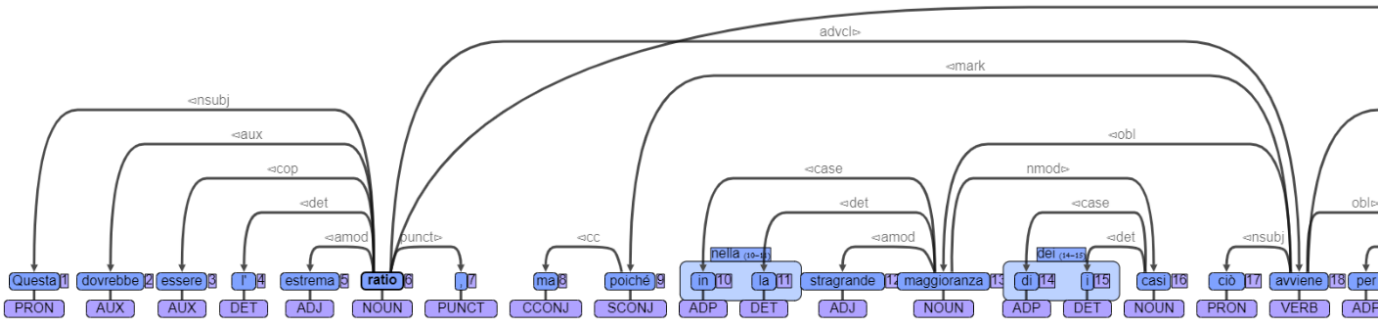

Figura 13: Revisione Gold corpus di una porzione della frase 32

A tal proposito vorrei proporre l'aggiunta di una possibile etichetta alle relazioni di dipendenza, volta a evidenziare casi di frasi mal formate in particolare nel caso in cui manca la testa a cui un determinato token si sarebbe dovuto attaccare. La nuova etichetta, chiamata ad esempio missing, potrebbe essere collocata nella sezione Special. Segue la possibile nuova annotazione della frase precedente a seguito dell'introduzione dell'etichetta missing:

- ma dipendente del verbo a cui la frase mancante avrebbe dovuto dipendere, in questo caso ratio
- relazione tra ma e ratio etichettata come missing.
- ma viene considerato come se fosse il verbo mancante e quindi avviene diventa dipendente di ma

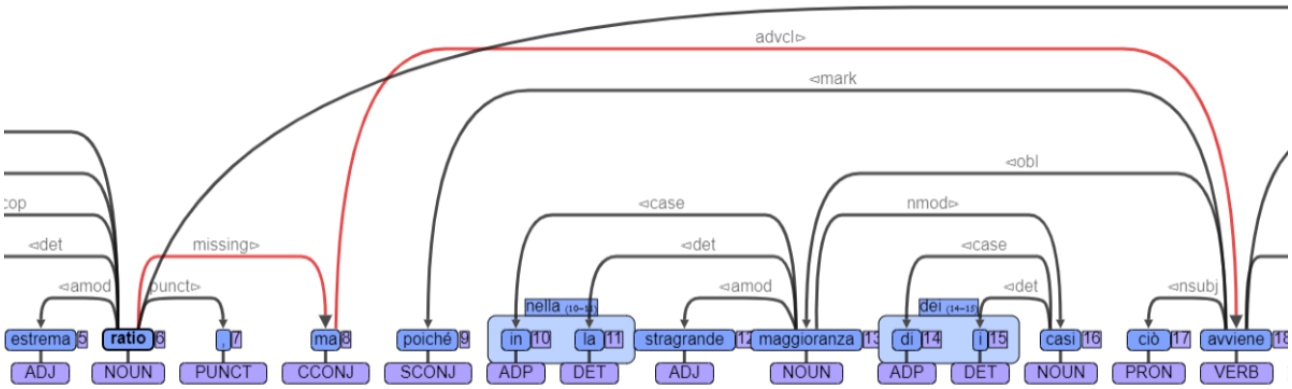

Figura 14: Possibile annotazione della frase mal formata con introduzione della relazione missing

Un tipo di etichetta simile e già presente in Universal Dependecies è orphan, tuttavia tratta in maniera specifica di ellissi e non di frasi mal formate.

Gli ultimi due errori da analizzare riguardano l'ordine non canonico del soggetto e oggetto e l'errato riconoscimento della testa delle congiunzioni. Il primo caso spesso riguarda il soggetto posposto o l'inversione di posizione tra soggetto-oggetto; di seguito alcuni esempi:

- Frase 24: Vogliamo, quindi, sapere cosa ha intenzione di fare [...]
  Cosa in questo caso è stato indicato dal parser come soggetto mentre si tratta in realtà di un oggetto; questo errore è probabilmente dovuto al fatto che il token si trova nella posizione canonicamente occupata dal soggetto, il quale oltretutto, in questa frase, è omesso
- Frase 7: L'insegnante deve prendersi la massima cura dei ragazzi che gli sono affidati, come deve fare l'affidatario, ma non sono suoi figli e non lo diventeranno.
  Questo è il caso opposto a ciò che è successo nella frase precedente, ovvero il parser ha etichettato la relazione tra affidatario e fare come obj, quando invece dovrebbe essere nsubj; ciò è appunto dovuto al fatto che il soggetto è posposto rispetto al predicato a cui si riferisce.

Mentre il secondo caso riguarda sia congiunzioni tra frasi coordinanti che tra parole:

- Frase 1: Ci sono delle supplenze, dei trasferimenti, degli spostamenti e degli avanzamenti di carriera.
  In questa frase non è stata riconosciuta correttamente la relazione di congiunzione tra spsotamenti e supplenze; il token spostamenti infatti è stato riconosciuto dal parser come soggetto avente come testa sono.

- Frase 31: Se andassimo a guardare l'adeguatezza di tutti i genitori del Paese o addirittura la salubrità di tantissimi siti [...]
  In questo caso è stata riconosciuta correttamente la relazione di congiunzione del dipendente salubrità, tuttavia è stata individuata la testa scorretta poiché non si tratta del token guardare, ma del token adeguatezza

- Frase 18: Ho ascoltato la sua relazione, ma nonostante i miei sforzi, mi perdoni, Ministro, sicuramente per un mio limite personale, non sono riuscito a capire bene quale sia la sua linea e cosa voglia fare effettivamente per risolvere il problema della situazione carceraria.
  In questa frase invece si tratta di frasi coordinanti, infatti il parser non è riuscito a riconoscere la relazione di congiunzione tra riuscito e ascoltato, probabilmente a causa della presenza di molti incisi tra la due frasi

6 Inter-Annotator Agreement

In questa sezione verrà trattata la verifica dell'accordo in fase di revisione dell'annotazione automatica e verranno discusse le maggiori differenze tra le due revisioni legate al livello di annotazione.

Seguono le tabelle relativa all'Inter-Annotator Agreemment del POS tagging (tab. 5) e delle Depencies (tab. 6) confrontando la revisione effettuata da me e la revisione della mia collega:

	Avarage obs. Agreement	**Kappa**
Seduta 2015	0.9936183790682833	0.9928448664770807
Seduta 2020	0.9910941475826972	0.9900297503768064

Tabella 5: Tabella Inter-Annotato Agreement per POS

	Avarage obs. Agreement	**Kappa**
Seduta 2015	0.9387364390555201	0.9383744137802204
Seduta 2020	0.9217557251908397	0.9213631749502206

Tabella 6: Tabella Inter-Annotato Agreement per Depencies

Osservando le tabelle, è possibile notare che in generale le revisioni da noi effettuate superano di gran lunga la soglia richiesta (Kappa > 0.8), quindi esiste una coerenza interna dell'annotazione; andando nello specifico si nota che i valori relativi al Part-Of-Speech tagging risultano più alti rispetto a quelli delle dipendenze: per quanto riguarda il POS la percentuale sia per l'AOA che il Kappa sfiorano il 100%, mentre con le depencies i valori si abbassano leggermente fino ad arrivare al 94% nel testo della Seduta del 2015 e il 92% nella Seduta del 2020.

Mettendo a confronto le nostre revisioni possiamo notare che è presente un numero maggiore di differenze nelle relazioni di dipendenza, 96 nel testo del 2015 e 123 nel testo del 2020, rispetto alle etichette, 10 differenze nel testo del 2015 e 14 nel testo del 2020.

	POS	**Depencies**
Seduta 2015	10	96
Seduta 2020	14	123

Tabella 7: Tabella differenze di annotazione tra due annotatori

Segue una tabella di alcune differenze a livello di Depencies tra le due revisioni:

	Seduta 2015	**Seduta 2020**
Testa punct	28	31
Tipo/testa subordinata	12	14
Tipo/testa obl|nmod	10	20

In fase di confronto e costruzione del gold coprus ci siamo accorte che è possibile riscontrare molte differenze per quanto riguarda la testa della punteggiatura, infatti nel testo del 2015 circa il 29% delle differenze sono relative a una diversa testa della punteggiatura, mentre nel testo del 2020 si parla del 25%; questo si pensa sia dovuto alla poca chiarezza presente nelle linee guida della Universal Dependecies, infatti è possibile vedere come c'è un diverso tipo di annotazione delle virgole, in quanto incisi, se si considerano lingue differenti:

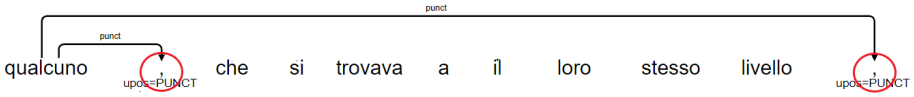

Figura 15: Annotazione di virgole in presenza di una frase relativa in italiano

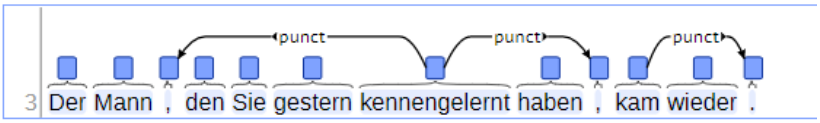

Figura 16: Annotazione di virgole in presenza di una frase relativa in tedesco

Un'altra differenza è relativa al tipo e/o alla testa della frase subordinata (quindi se ccomp, xcomp, advcl, acl o acl:relcl); è possibile incontrare questo genere di discordanza soprattutto in presenza di frasi piuttosto lunghe costituite da molti incisi e frasi subordinate, un esempio è il seguente:

> Ci sono delle supplenze, dei trasferimenti, degli spostamenti e degli avanzamenti di carriera, per cui in questo arco di tempo, che dovrebbe durare al massimo due anni, se il bambino e la famiglia affidataria avessero rapporti con i servizi sociali lo avrebbero con una serie di persone diverse, la cui efficienza e buona volontà (che peraltro non possiamo dare per scontate, perché sono esseri umani anche loro) sarebbero comunque insufficienti a supplire al continuo ricambio.

In questa frase il token avrebbero è stato annotato da me come acl:relcl avente testa il token supplenze, mentre la mia collega lo ha annotato come advcl avente come testa il token sono.

L'ultima differenza interessante da trattare è il modo in cui abbiamo etichettato relazioni relative a complementi di specificazione; spesso la differenza di etichetta della relazione è dovuta a una scelta della testa differente:

> Ci troviamo in una situazione per la quale, visto che la legge oggi non è applicata (e sarà applicata ancora meno, se approveremo queste norme), finiamo col prendere atto di questa mancata applicazione.

In questa frase la mia collega ha annotato la relazione tra applicazione e prendere come obl, mentre io ho indicato atto come testa di applicazione e dunque ho etichettato la relazione come nmod.
Ci sono stati diversi casi in cui la mia collega ed io ci siamo trovate a discutere sulla testa effettiva della relazione obl/nmod, questo è dovuto al fatto che questo tipo di complemento è spesso soggetto all'interpretazione data dal lettore, e quindi all'arbitrarietà.

7 Confronto con modelli ISDT e POSTWITA

Passiamo ora alla fase di confronto tra il gold corpus, costruito sulla base delle revisioni manuali, e il corpus annotato automaticamente utilizzando i modelli ISDT e Postwita. Come già anticipato nella sezione, per verificare la correttezza di dell'analisi automatica è stato impiegato lo script di valutazione distribuito in occasione dello *CoNLL 2018 Shared task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Per quanto riguarda il testo relativo alla seduta del 2015 il confronto con l'annotazione con modello ISDT ha portato ai risultati seguenti:

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	98.04	98.04	98.04	98.04
XPOS	97.50	97.50	97.50	97.50
UFeats	98.51	98.51	98.51	98.51
AllTags	97.09	97.09	97.09	97.09
Lemmas	98.85	98.85	98.85	98.85
UAS	89.53	89.53	89.53	89.53
LAS	87.03	87.03	87.03	87.03
CLAS	82.81	82.70	82.75	82.70
MLAS	81.04	80.93	80.98	80.93
BLEX	81.86	81.74	81.80	81.74

Mentre per quanto riguarda il testo della seduta del 2020 abbiamo:

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	96.15	96.15	96.15	96.15
XPOS	95.42	95.42	95.42	95.42
UFeats	96.88	96.88	96.88	96.88
AllTags	94.95	94.95	94.95	94.95
Lemmas	97.28	97.28	97.28	97.28
UAS	88.37	88.37	88.37	88.37
LAS	85.05	85.05	85.05	85.05
CLAS	79.65	79.65	79.65	79.65
MLAS	77.39	77.39	77.39	77.39
BLEX	77.79	77.79	77.79	77.79

Le tabelle mostrano come il testo relativo alla seduta del 2015 abbia un punteggio per ogni metrica più alto (circa 4 punti percentuali) rispetto al testo della Seduta del 2020, e questo sembra rispecchiare quanto osservato nel paragrafo 5 riguardo alla percentuale di errori sia nei POS che nelle dipendenze. Per quanto riguarda la metrica considerata più completa, ovvero LAS, la quale indica la percentuale di dipendenze identificate ed etichettate correttamente, il punteggio è di 87% per il primo testo e 85% per il secondo, quindi piuttosto alto. I

valori risultano invece essere superiori quando la valutazione è meno restrittiva, dunque considerando la metrica UAS.

Per quanto riguarda il testo relativo alla seduta del 2015 il confronto con l'annotazione con modello Postwita ha portato ai risultati seguenti:

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	94.32	94.32	94.32	94.32
XPOS	93.24	93.24	93.24	93.24
UFeats	91.62	91.62	91.62	91.62
AllTags	89.59	89.59	89.59	89.59
Lemmas	95.54	95.54	95.54	95.54
UAS	78.58	78.58	78.58	78.58
LAS	73.04	73.04	73.04	73.04
CLAS	64.79	64.44	64.62	64.44
MLAS	57.67	57.36	57.51	57.36
BLEX	61.23	60.90	61.07	60.90

Mentre per quanto riguarda il testo della seduta del 2020 abbiamo:

Metric	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	93.75	93.75	93.75	93.75
XPOS	92.69	92.69	92.69	92.69
UFeats	88.90	88.90	88.90	88.90
AllTags	86.64	86.64	86.64	86.64
Lemmas	94.09	94.09	94.09	94.09
UAS	80.27	80.27	80.27	80.27
LAS	74.42	74.42	74.42	74.42
CLAS	66.09	65.56	65.82	65.56
MLAS	53.89	53.46	53.67	53.46
BLEX	60.46	59.97	60.21	59.97

Come ci si poteva aspettare il parser ha prestazioni migliori sul corpus annotato utilizzando il modello addestrato sulla stessa varietà di lingua, cioè ISDT, mentre nel caso del modello Postwita i valori delle metriche cadono drasticamente per entrambi i testi. Questo è probabilmente dovuto all'utilizzo di un modello relativo ad una varietà della lingua molto differente. Postwita infatti è un modello addestrato sulla base di tweets, piccoli testi costituiti da un registro molto colloquiale e informale.
In particolare, la differenza tra il valore di LAS ottenuto con il modello ISDT (87% per il primo testo e 85% per il secondo) e quello ottenuto con il modello Postwita (73% per il primo testo e 74% per il secondo) è di circa 14 punti percentuali per il primo testo e 11 per il secondo.