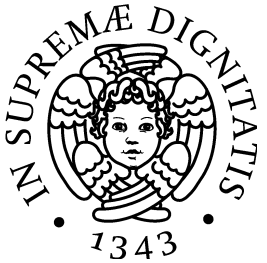

UNIVERSITÀ DI PISA

LAUREA MAGISTRALE IN
INFORMATICA UMANISTICA

LINGUISTICA COMPUTAZIONALE II A.A. 2019/20

Relazione

Resoconto della realizzazione del progetto

John Bianchi

Matricola*: 517601*

Docenti:

Prof.ssa Simonetta Montemagni

Prof.ssa Giulia Venturi

1. Introduzione **3**

2. Schema di annotazione e Software **4**

2.1 Universal Dependencies 4

2.2 Modello della lingua 5

2.3 UDPipe 5

3. Approccio **6**

3.1 Corpus in input 6

3.2 Formazione allo sviluppo del progetto 7

3.3 Tokenizzazione e sentence splitting 8

3.4 Revisione manuale dell'annotazione automatica 10

4. Confronto e analisi **12**

4.1 Calcolo dell'interannotator agreement 13

4.2 Stesura revisione condivisa del corpus 14

4.3 Verifica della correttezza delle revisione 15

4. Conclusione **18**

5. Bibliografia **19**

6. Sitografia **19**

1. Introduzione

“In LC , l’annotazione ha acquisito un ruolo centrale ormai indiscusso, in quanto permette di rendere esplicita, interpretabile ed esplorabile dal computer la struttura linguistica implicita nel testo” [1]

Padroneggiare strumenti specifici per l’annotazione del testo risulta essere un requisito fondamentale per chiunque operi in questo capo di studi.

Il presente elaborato si propone di fornire un dettagliato resoconto sul progetto, al quale il candidato ha lavorato in collaborazione con il collega Lorenzo di Gianvittorio, mirato alla realizzazione di un test corpus per valutare l’accuratezza di analisi di una catena di annotazione linguistica automatica.

Questo elaborato è stato strutturato in modo da essere conforme a quelli che sono state le varie fasi realizzative del progetto.

Nella prima parte verranno discussi nel dettaglio quale schema di annotazione e quali software sono stati utilizzati e di come essi sono stati configurati per eseguire l’indagine obiettivo del progetto.

Nella seconda parte verrà illustrata la struttura dei file analizzati rivolgendo particolare attenzione alle loro caratteristiche intrinseche proprie della specifica varietà della lingua. Verranno poi anche trattate le fasi di tokenizzazione e di revisione manuale dell’annotazione automatica.

Nella terza parte verrà messo a confronto il lavoro di entrambi i candidati attraverso il calcolo del Inter Annotator agreement. Successivamente l’elaborato verrà confrontato anche con i risultati prodotti da altri modelli della lingua e infine esaminato con uno specifico script di valutazione con l’obiettivo di esplicitare al meglio tutte le possibili divergenze.

---

[1] *Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, 2005, Testo e computer*

2. Schema di annotazione e Software

Nella prima parte di questo capitolo verrà esplicato nel dettaglio lo schema di annotazione scelto mentre nella seconda parte verranno discussi quelli che sono stati i software che hanno reso possibile la realizzazione del progetto. È opportuno chiarire che a differenza di quanto indicato nelle linee guida ufficiali del progetto, non si è utilizzato né DgAnnotatoor né UD Annotatrix.

2.1 Universal Dependencies

Universal Dependencies è lo schema di annotazione utilizzato in questo progetto. Esso si pone come obiettivo quello di costruire uno schema di annotazione valido per tutte le lingue in modo da facilitare il confronto reciproco e di standardizzare gli studi e progetti in merito. La rappresentazione si basa sull'assunzione che le teste delle relazioni sono parole di contenuto che portano quindi un maggiore contenuto semantico alla frasi in questione.

Se paragonato ad altri schemi di annotazione Universal Dependencies è meno dettagliato e le sua relazioni sono spesso di quantità inferiore. Se da un lato ciò provoca una perdita di informazione dall'altro semplifica il parsing facilita la portabilità dell'annotazione rendendola completamente cross-linguistica.

2.2 Modello della lingua

Il modello della lingua utilizzato per le fasi iniziali del progetto è stato quello addestrato sulla treebank PoSTWITA-UD realizzata, come indicato nella documentazione presente nel sito ufficiale di Universal Dependencies[2], da Cristina Bosc e Manuela Sanguinetti. Questa è stata creata arricchendo il dataset utilizzato per il compito di EVALITA 2016 di Part-of-Speech tagging dei Social Media, quindi risulta preferibile ad altre treebank se ci si occupa di varietà della lingua italiana propria dei social media.

---

[2] *sito* ufficiale di Universal Dependencies sezione PoSTWITA-UD https://universaldependencies.org/treebanks/it_postwita/index.html

2.3 UDPipe

È la catena di annotazione linguistica realizzata appositamente per il progetto Universal Dependencies. Per questo progetto è stata utilizzata fornendo come parametro il modello della lingua PoSTWITA precedentemente citato nel punto 2.2.

La versione di questo software utilizzata per questo progetto è quella 1.2.0 [3] ed è stata utilizzata su di un computer con sistema operativo basato sul kernel Linux.

[3] *sito* ufficiale di UDPipe http://ufal.mff.cuni.cz/udpipe

3. Approccio

In questo capitolo verranno inquadrati, in linea generale, quelle che sono le caratteristiche proprie della varietà della lingua che caratterizzano il corpus utilizzato come base di partenza per la realizzazione della seguente indagine.

Nella sezione successiva verranno descritte in maniera sequenziale le fasi realizzative del progetto delineando per ciascuna di esse quelle che sono state le decisioni metodologiche di entrambi i candidati.

3.1 Corpus in input

Il corpus utilizzato per le varie fasi di analisi è costituito da un insieme di tweet scritti da utenti non specificati riguardanti tre specifici macro argomenti: Coronavirus, Festival di Sanremo 2020, Fridays for future. Le successive fasi realizzative sono state condotte sia considerando questi tre insiemi uniti sia separati.

Per quanto riguarda l'argomento Coronavirus sono presenti 27 tweet composti per la maggior parte da più di una frase, questi sono formati da 845 tokens di cui 524 unici.

Per l'argomento Fridays for future sono presenti 23 tweet composti per la maggior parte da più di una frase, formati da 842 tokens di cui 472 unici.

In Sanremo 2020 ci sono 29 tweet formati per la maggior parte da più di una frase, questi sono composti da 832 tokens di cui 458 unici.

Considerando i dati nella totalità ci sono quindi 79 tweet, questi sono formati da 2519 tokens di cui 1279 unici.

3.2 Formazione allo sviluppo del progetto

Il primo passaggio che ha portato allo svolgimento del progetto è stato di carattere formativo. I vari task necessari al compimento del progetto risultavano essere impossibili da svolgere con le sole conoscenze pregresse o con quelle apprese durante le lezioni del corso da parte dei candidati. In particolare la parte della revisione dell'annotazione automatica richiede specifiche conoscenze che possono essere maturate solo mediante lo studio di determinate aree tematiche nello specifico che riguardano non solo la

conoscenza dei software ma anche e soprattutto della singolare codifica di annotazione proposta dal progetto Universal Dependencies. In questo primo approccio al progetto entrambi i candidati infatti hanno deciso di dedicarsi allo studio di questo schema presso la documentazione disponibile presso il sito web ufficiale[4]. In particolare ci si è dovuti soffermare sull'apprendimento delle relazioni semantiche fondamentali per il corretto svolgimento di molti dei task del progetto.

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Figura 1: Screenshot che mostra le Universal dependencies relations - https://universaldependencies.org/u/dep/

Oltre a ciò si è deciso di consultare il gold standard di PoSTWITA-UD nel quale è possibile consultare come nel concreto sono state applicate queste regole dai professionisti di questa materia. È opportuno però fare presente che in questo documento ci sono diversi errori che risultano essere dovuti probabilmente ad una mera distrazione. Questi sono ad esempio quello di associare la parola "lucia" come lemma della parola "luce" o di confondere il verbo essere con il verbo stare.

---

[4] Linee guida ufficiali Universal Dependencies
https://universaldependencies.org/guidelines.html

3.3 Tokenizzazione e sentence splitting

Il modello addestrato sulla treebank PoSTWITA-UD è stato utilizzato per eseguire il processo di divisione in token e frasi sul corpus attraverso il software UDPipe. Il file generato è stato preso in analisi dai candidati i quali hanno constatato le seguenti problematiche in svariate sezioni:

- Tweet accavallati a vicenda
- Token accorpati alla punteggiatura

In un primo passaggio sono state quindi corrette queste incongruenze e nel successivo sono stati elaborati una serie di criteri mirati a migliorare l'analisi del corpus da parte del parser che verrà utilizzato nel successivo task del progetto.

I criteri adottati a livello di divisione del testo in token sono i seguenti:

- Token come nomi considerati come separati ("Mc" "Donalds' " "s")
- Punteggiatura equivalente ad un token unico
- Apostrofo inteso come accento e considerato insieme al token
- Hashtag come unico token
- Mention come unico token
- Link come unico token
- Verbi clitici come due tokens

Per quanto concerne la fase di sentence splitting i criteri adottati sono stati i seguenti:

- Considerare il Tweet come frase assenstante (a meno che i criteri successivi non lo permettano)
- Se all'interno del Tweet vi è una frase più lunga di 20 token che termina con un segno di punteggiatura di fine frase allora da considerare il token separato.
- Il segno di punteggiatura dei due punti così come quelli di esclamazione e interrogazione non sono stati considerati come punti di segmentazione.
- Segmentare sequenze di più di 40 token anche senza i segni di punteggiatura

queste ultime meritano senz'altro ulteriori chiarimenti. Sono infatti state decise perché si è pensato che considerare il tweet come unità di analisi assestante avrebbe giovato

all’interpretazione del modello durante il successivo task ma anche che una sequenza di token troppo lunga, per lo più scritta in una forma della lingua italiana molto distorta, potesse essere risultare ardua da interpretare sia in per il software automatico che per i candidati. Questo è dovuto al fatto che si è il criterio di analisi utilizzato considera una sola root per unità frasale. Il fatto di avere sequenze di token più lunghe di 40 unità avrebbe creato parecchie situazioni arbitrarie nell’analisi a dipendenze ed è per questo motivo infatti che si è optato per inserire il criterio che stabilisce che le sequenze più lunghe di 40 tokens vadano suddivise in sentence diverse.

3.4 Revisione manuale dell’annotazione automatica

Una volta completata la procedura esplicata nel punto 3.3 si è passati al task successivo che impone che il corpus precedentemente modificato venga adesso annotato automaticamente sempre mediante il modello PoSTWITA ma stavolta a livello di Parts-of-speech tagging e syntactic parsing. Il file generato è stato quindi revisionato manualmente dai singoli candidati che hanno eseguito questo passaggio individualmente.

Da entrambe le parti sono state effettuate revisioni che si basano sempre su criteri stabiliti in maniera condivisa ma che stavolta sono stati applicati individualmente e concordati non caso per caso ma in linea generale.

A livello di lemmatizzazione sono stati corretti i grossolani errori che il software ha frequentemente generato. Questi sono per lo più causati da nomi molto spesso riportati come verbi ma anche da nomi propri che sono stati modificati in forme errate in maniera ricorrente come ad esempio nel caso del nome “Achille” che è stato riportato più volte in “Achilla” o nel caso della parola “luce” riportata come “lucia”.

Inoltre gli errori propri di battitura del corpus sono stati corretti a livello di lemma e non a livello di token come ad esempio:

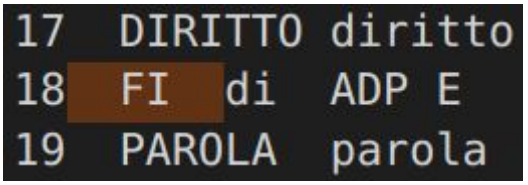

Figura 2: Esempio di correzione nel lemma nella fase di revisione

Per quanto concerne la categoria UPosTag le correzioni hanno riguardato i seguenti casi:

- Il tag SYM erroneamente scambiato con PROPN
- Il token “ L’ ” identificato come NUM e non come DET
- Errori nell’identificazione tra participi e aggettivi
- Token che compongono i marchi registrati considerati come a sé stanti e non come nomi propri ciascuno (come mostrato dalla figura 3)

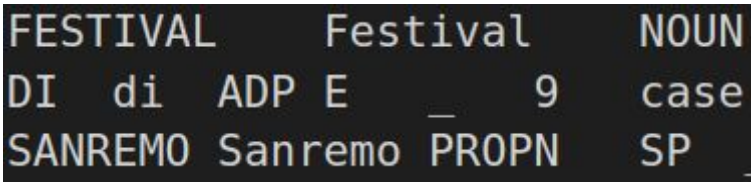

Figura 3: Esempio di correzione dell'identificazione dei nomi di marchi registrati

Esaminando la sezione dell’annotazione dedicata alle dipendenze sintattiche gli errori più tipici - oltre a quelli che vi si generano a partire da questi - sono stati i seguenti:

1. Identificazione errata della root della frase come nel caso illustrato dalla figura 4

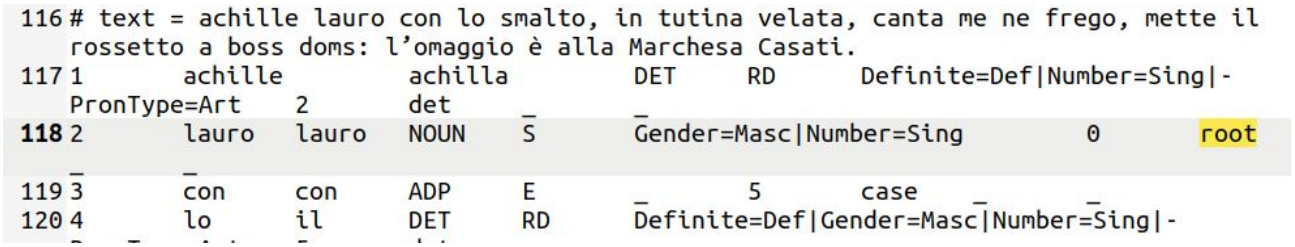

Figura 4: Esempio di assegnazione della root errata

2. Erronea identificazione dei link a fine frase classificati non come “dep” come mostrato nella figura immediatamente sottostante:

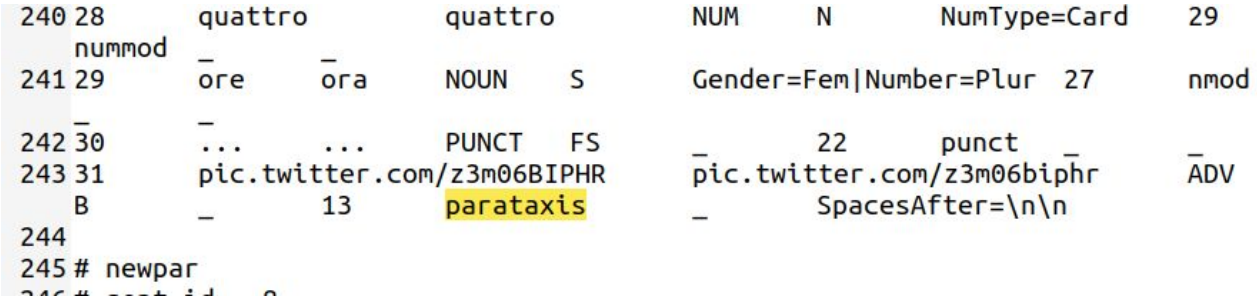

Figura 5: Esempio di assegnazione erronea di un indirizzo link

3. Uso del tag “obj” al posto di “obl”

4. Scarso utilizzo dei tag di tipologia “vocative”.

4. Confronto e analisi

In questo capitolo verranno illustrate, con i relativi dati, le fasi di verifica dell'accordo tra i due annotatori tramite il calcolo dell'interannotator agreement. Successivamente mediante uno specifico script, fornito da Universal Dependencies, verrà verificata la correttezza di analisi del corpus usando modelli di UDPipe addestrati su diverse verità della lingua.

4.1 Calcolo dell'interannotator agreement

I dati che si sono generati dalla revisione manuale dei due candidati necessitano di essere valutati attraverso una misura che certifica la loro affidabilità. Questa si calcola comparando a coppia i file revisionati e verificando il loro grado di Interannotator agreement.

Ai fini dell'indagine è stato calcolato utilizzando due diversi tipi di indice:

1. Observed agreement che corrisponde al semplice rapporto basato sul numero di elementi sui quali gli annotatori si sono trovato in disaccordo

2. K di Cohen ovvero una misura più complessa che considera gli item indipendenti gli uni dagli altri così come gli annotatori e le categorie utilizzate vengono considerate come mutualmente esclusive ed esaustive.

Questo indice è stato applicato sia sulle part of speech che sulle dipendenze e nella sezione successiva è possibile consultare i risultati

	Average observed agreement	K di Cohen
Sanremo	0,828	0,827
Fridays	0,893	0,893
Coronavirus	0,842	0,841

Tabella 1: Interannotator agreement dipendenze

	Average observed agreement	K di Cohen
Sanremo	0,957	0,953
Fridays	0,979	0,977
Coronavirus	0,961	0,957

Tabella 2: Interannotator agreement part of speech

È interessante notare come la divergenza tra gli annotatori sia stata maggiore per quanto riguarda l'analisi a dipendenze. Se rapportati in media i valori osservati di average observed agreement risultano avere una differenza sostanziale. Il valore medio di average observed agreement per part of speech è di 0,965 mentre quello per le dipendenze è di 0,854. Questa differenza è constatabile anche per i valori di K di Cohen che, se rapportati alla media, risultano possedere valori molto simili a quelli precedenti con 0,965 per le part of speech e 0,853 per le dipendenze.

Questa sostanziale differenza è causata dal fatto che le part of speech sono categorie molto spesso fisse che non lasciano spazio all'interpretazione del revisore mentre nel caso delle dipendenze sintattiche l'approccio, seppure preventivamente concordato, rimane di carattere più personale ed interpretativo. Inoltre la revisione delle dipendenze comporta molto spesso un inevitabile effetto "a cascata" che si può verificare anche banalmente qualora il revisore decida ad esempio di cambiare la radice di una frase.

4.2 Stesura revisione condivisa del corpus

Una volta completata la fase precedente i due candidati hanno revisionato il corpus di partenza in modo collaborativo lavorando quindi su di un unico file ed utilizzando regole e parametri condivisi.

Questi sono stati decisi a priori del processo e comprendono perlopiù quelli già elencati al punto 3.4 con l'aggiunta di qualche regola fissa come quella di considerare il token "Coronavirus" sempre come "PROPN". C'è stato invece una grande attenzione per casi specifici come ad esempio quello di considerare o meno il token "cosplay" come

“flat:foreign” o meno. Questo è infatti un caso borderline dove si ha a che fare con una parola anglosassone che viene utilizzata come una italiana.

Oltre a questo passaggio che ha riguardato i casi specifici c’è stata anche una grande correzione di piccoli errori che allo sguardo del singolo annotatore non erano stati corretti mentre in questo passaggio, grazie all’attenzione di una persona in più, è stato possibile individuare con maggiore facilità.

4.3 Verifica della correttezza delle revisione

I dati revisionati dai candidati sono stati quindi comparati con modelli di UDPipe addestrati su di un’altra varietà della lingua, Il modello utilizzato è stato infatti “UD Italian ISDT” che, come indicato nella pagina dedicata[5]

“The Italian corpus annotated according to the UD annotation scheme was obtained by conversion from ISDT (Italian Stanford Dependency Treebank), released for the dependency parsing shared task of Evalita-2014 (Bosco et al. 2014).”

risulta essere il prodotto di una conversione del modello ISDT.

Questa operazione di verifica della correttezza è stata ottenuta grazie allo script di valutazione distribuito in occasione dello CoNLL 2018 Shared task: Multilingual Parsing from Raw Text to Universal Dependencies ed è liberamente accessibile presso la pagina dedicata all’interno del sito di Universal Dependencies[6].

I risultati di questa operazione sono riportati in seguito:

---

[5] Sito ufficialeUD Italian ISDT: https://universaldependencies.org/treebanks/it_isdt/index.html

[6] CoNLL 2018 Shared task: Multilingual Parsing from Raw Text to Universal Dependencies:

https://universaldependencies.org/conll18/evaluation.html

Metric	Precision	Recall	F1Score	Align succ
Tokens	99,93	99,85	99,89	
Sentences	100	100	100	
Words	99,93	99,89	99,91	
UPOS	82,72	82,7	82,71	82,78
XPOS	80,9	80,87	80,88	80,96
UFeats	87,75	87,71	87,73	87,81
AllTags	76,86	76,83	76,85	76,91
Lemmas	90,24	90,21	90,22	90,3
UAS	68,93	68,9	68,91	68,97
LAS	61,73	61,71	61,72	61,77
CLAS	49,66	50,56	50,11	50,65
MLAS	40,88	41,61	41,24	41,69
BLEX	46,06	46,89	46,48	46,98

Tabella 3: Corpus revisionato comparato con il modello ISDT

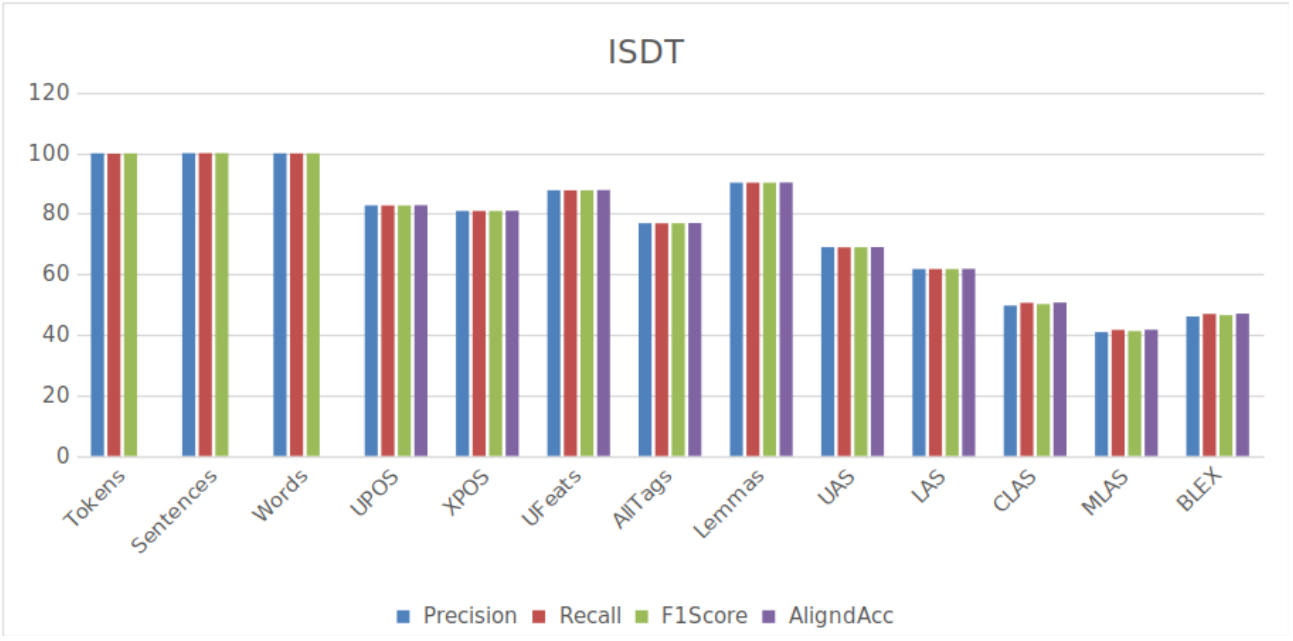

Grafico 1: Corpus revisionato comparato con il modello ISDT

Dai risultati prodotti possiamo immediatamente notare come il livello di accordo per quanto concerne le categorie di “Tokens”, “Sentences” e “Words” sia di livello massimo. Questo è dovuto al fatto che il file analizzato dal modello ISDT è stato precedentemente sincronizzato a questo livello per garantire una corretta esecuzione dello script che li ha confrontati. Per quanto concerne le altre categorie è interessante notare come le maggiori differenze siano state registrate in quelle di “UAS”, “LAS”, “CLAS”, “MLAS”, e “BLEX”, mentre per le categorie di “Lemmas” e “UFeats” non ci siano state significative differenze.

Lo stesso file revisionato è stato comparato anche con il medesimo modello che si è utilizzato per svolgere i principali task di questo progetto ovvero PoSTWITA-UD; i risultati sono mostrati in seguito:

Metric	Precision	Recall	F1 Score	Align succ
Tokens	99,93	99,85	99,89	
Sentences	100	100	100	

Words	99,93	99,89	99,91	
UPOS	93,64	93,61	93,63	93,71
XPOS	93,61	93,58	93,59	93,68
UFeats	97,19	97,16	97,17	97,26
AllTags	93,22	93,19	93,21	93,29
Lemmas	96,7	96,67	96,68	96,77
UAS	90,45	90,42	90,43	90,51
LAS	85,74	85,71	85,73	85,8
CLAS	78,19	78,14	78,16	78,28
MLAS	72,65	72,61	72,63	72,74
BLEX	75,39	75,34	75,37	75,48

Tabella 4: Corpus revisionato comparato con il modello PoSTWITA-UD

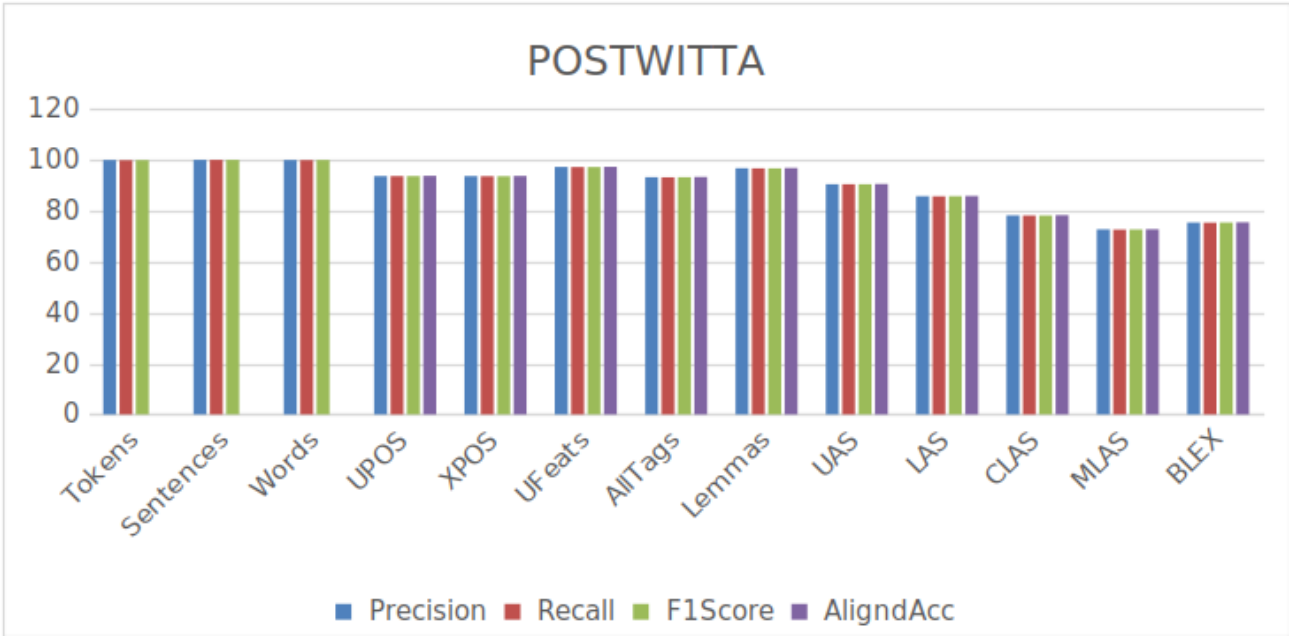

Grafico 2: Corpus revisionato comparato con il modello PoSTWITA-UD

Questo secondo confronto rende perfettamente esplicito come le i risultati prodotti dal confronto con ISDT siano perfettamente analoghi a questi ultimi con la differenza che in questo caso si registrano in quantità minore con valori quasi dimezzati.

4. Conclusione

Il task di revisione manuale dell'annotazione è fondamentale per il raggiungimento di un'annotazione priva di errori. Questo lo si deve principalmente al fatto che l'attuale stato dell'analisi automatica su questa specifica varietà della lingua da parte di strumenti automatici è ancora ben lontana dallo stato dell'arte. Essi infatti, come ampiamente dimostrato, riscontrano una grande difficoltà nell'analisi di particolari fenomeni linguistici come ad esempio nel compito di disambiguazione linguaggio. Nonostante ciò è grazie proprio al lavoro di revisione che è possibile affinare l'abilità di questa tipologia di strumenti.

5. Bibliografia

[1] *Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, 2005, Testo e computer*

6. Sitografia

[2] *sito* ufficiale di Universal Dependencies sezione PoSTWITA-UD
https://universaldependencies.org/treebanks/it_postwita/index.html

[3] *sito* ufficiale di UDPipe http://ufal.mff.cuni.cz/udpipe

[4] Linee guida ufficiali Universal Dependencies
https://universaldependencies.org/guidelines.html

[5] Sito ufficialeUD Italian ISDT: https://universaldependencies.org/treebanks/it_isdt/index.html

[6] CoNLL 2018 Shared task: Multilingual Parsing from Raw Text to Universal Dependencies:

https://universaldependencies.org/conll18/evaluation.html