UNIVERSITÀ DI PISA

LAUREA MAGISTRALE IN
INFORMATICA UMANISTICA

CORSO DI LINGUISTICA COMPUTAZIONALE II

Costruzione di un *Gold* corpus per la valutazione dell'accuratezza della catena di annotazione linguistica automatica *UDPipe*

Aldo Cerulli

(Matricola 533566)

Anno Accademico 2021/2022

Indice

1 Introduzione

In questa relazione verranno ripercorse e analizzate le fasi che hanno portato alla costruzione di un corpus *Gold* per valutare l'accuratezza della catena di annotazione linguistica automatica *UDPipe* rispetto a due modelli addestrati su varietà di italiano diverse. *UDPipe* è disponibile sia come demo online[1] sia come versione scaricabile[2]. In questo contesto si è usata la seconda. I modelli statistici di riferimento sono *italian-isdt-ud-2.5-191206.udpipe* e *talian-postwita-ud-2.5-191206.udpipe*: il primo è addestrato su ISDT, una *treebank* (14167 frasi, 278429 *token*[3]) di testi eterogenei rispetto al genere (*Wikipedia*, articoli di giornale, testi giuridici, etc.); il secondo è addestrato, invece, su PoSTWITA, un'ampia raccolta di *tweet* (6712 frasi, 119342 *token*[4]). I due corpora sono annotati secondo il formalismo delle *Universal Dependencies*.

Il lavoro è stato realizzato insieme ai colleghi Giacomo Cerretini e Luca Poggianti come progetto di esame per il corso di *Linguistica Computazionale II*, tenuto dalle docenti Simonetta Montemagni e Giulia Venturi. Si tratterebbe di un progetto pensato per due studenti; tuttavia, in via eccezionale, ci è stato permesso di lavorare in tre in quanto allo scrivente mancava un compagno.

Le professoresse ci hanno fornito un corpus di cinque testi, rappresentativi di due domini testuali, gastronomia e scienza, prodotti in momenti diversi del '900. Il *capitolo 2* sarà dedicato alla loro descrizione. Il corpus è stato annotato da *UDPipe* con il modello addestrato su ISDT limitatamente ai livelli di base e l'*output* è stato corretto dai tre annotatori insieme (*capitolo 3*). Si è poi proceduto alle analisi morfosintattiche e sintattiche, revisionate questa volta individualmente (*capitoli 4-5*). Terminate le correzioni, il gruppo si è riunito per effettuare calcoli incrociati del grado di accordo raggiunto da coppie di corpora revisionati (*capitolo 6*) e per definire, a partire da questi, un *Gold standard* comune che risultasse dalla scelta delle soluzioni di annotazione considerate più corrette (*capitolo 7*). Tale corpus è stato infine usato per valutare l'accuratezza dei due modelli a cui si faceva riferimento poc'anzi.

Fondamentali si sono rivelati il ricevimento di gruppo di mercoledì 18 maggio e i seminari tenuti dagli studenti nelle ultime tre lezioni del corso: il primo è stata l'occasione per chiarire dubbi fondanti nel momento delicato in cui si stavano iniziando le operazioni di revisione manuale. I secondi hanno rappresentato un momento di confronto a progetto inoltrato, utili per valutare quanto fatto fino a quel momento e, dove necessario, correggere il tiro.

[1] La demo online di *UDPipe* è disponibile al sito https://lindat.mff.cuni.cz/services/udpipe/
[2] La versione completa di *UDPipe* è scaricabile da https://ufal.mff.cuni.cz/udpipe/1
[3] Composizione di ISDT su https://universaldependencies.org/treebanks/it_isdt/index.html
[4] Composizione di PoSTWITA su https://universaldependencies.org/treebanks/it_postwita/index.html

2 Descrizione del corpus

I testi da annotare in modo semiautomatico sono stati selezionati nell'ambito del progetto TrAVaSI (*Trattamento Automatico di Varietà Storiche di Italiano*), iniziativa che vede la collaborazione tra l'Accademia della Crusca e l'Istituto di Linguistica Computazionale "Antonio Zampolli" nella realizzazione di risorse per il trattamento automatico di varietà storiche dell'italiano (Favaro et al., 2021). Il punto di partenza di TrAVaSI è rappresentato da un corpus di 41 testi – compresi quelli oggetto di questa analisi – rappresentativi di sette domini (arte, cucina, diritto, giornali, letteratura, paraletteratura, e scienze), che è a sua volta un sotto-corpus di quello raccolto per la realizzazione del VoDIM (*Vocabolario Dinamico dell'Italiano Moderno*) (Marazzini e Maconi, 2018)[5].

Il corpus fornito dalle docenti si compone di cinque testi appartenenti a due dei domini menzionati, cucina e scienza, rappresentati rispettivamente nella misura di tre e due testi. Da un punto di vista cronologico, c'è una significativa differenza tra i due generi: i testi gastronomici risalgono al primo trentennio del '900, mentre quelli scientifici vanno dai primi anni Sessanta a oggi.
Analizziamo brevemente la composizione del corpus. I testi di argomento culinario sono: *cucina-sample_centosessantamaniere_1907*, tratto dal libro *160 Maniere di cucinare gli Erbaggi e i Legumi*, illustra cinque pietanze a base di cavolo e cavolfiore; *cucina-sample_lazzariturco_1947*, contenuto ne *Il Piccolo Focolare* della baronessa Giulia Turco Turcati Lazzari, espone cinque modi di cucinare la polenta; infine, *cucina-sample_boni_1927*, estratto dal *Talismano della felicità* di Ada Boni, riporta le ricette della besciamella e della salsa agro-dolce. Tra i testi di scienza abbiamo *scienze-sample_fermi2_1962*, dissertazione di Enrico Fermi sugli atomi dai toni formali e sintassi articolata, e *scienze-sample_bianucci_astronomia_2015*, dalla *Storia sentimentale dell'astronomia* di Piero Bianucci, che racconta in modo leggero e divulgativo la scoperta dei neutrini. Quanto detto è riassunto in *tabella 1*. Per ogni testo è riportato il codice identificativo che si userà da qui in avanti al posto del nome del file per rendere la trattazione più snella.

Dominio	Testo	Id	Anno (1. ed)
Cucina	*cucina-sample_centosessantamaniere_1907*	*C*	1907
	cucina-sample_lazzariturco_1947	*L*	1908
	cucina-sample_boni_1927	*B*	1927
Scienza	*scienze-sample_fermi2_1962*	*F*	1962
	scienze-sample_bianucci_astronomia_2015	*BI*	2012

Tabella 1. Composizione del corpus rispetto al dominio e l'anno di pubblicazione dei testi.

[5] La banca dati del VoDIM è digitalmente consultabile al sito https://vodim.accademiadellacrusca.org/

Nei capitoli che seguono si avrà modo di approfondire i testi da un punto di vista morfosintattico e sintattico. Per adesso ci limitiamo a fornire alcune caratteristiche generali. I dati riportati in *tabella 2* vanno intesi in seguito agli interventi di revisione rispetto ai livelli di base che verranno descritti nel prossimo capitolo.

Testo	**N° frasi**	**Lunghezza frase**			**N° parole**		**TTR**
		Media	**Min**	**Max**	**Token**	**Tipo**	
C	31	20,87	2	68	647	249	0,38
L	35	18,45	2	75	646	267	0,41
B	26	22,46	3	77	584	257	0,44
F	13	44,15	3	76	574	254	0,44
BI	30	22,23	3	52	667	321	0,48

Tabella 2. Caratteristiche generali e formali dei cinque testi.

Si consideri che nei testi di cucina *C* e *L* i titoli delle ricette sono preceduti da un numero puntato, che – come si vedrà meglio in *capitolo 3* – costituisce una frase a sé stante. È il motivo per cui (1) essi hanno un numero di frasi più elevato, specialmente *L*, e (2) la loro frase più breve è costituita da due *token*, il numero e il punto. Nel complesso, quindi, i testi sono abbastanza bilanciati tranne *F*, che però è costituito da frasi lunghe – in termini di numero di *token* – mediamente il doppio rispetto a quelle degli altri testi. Anche a livello di dimensione totale – sempre rispetto ai *token* – si registra un discreto equilibrio, con una differenza massima di 93 unità tra *B* e *F*.
Oltre ai *token*, si è ricavato anche il numero di parole tipo. Ciò ha permesso di dare una valutazione, seppur approssimativa, della ricchezza lessicale tramite l'indice TTR (*Type/Token Ratio*). Sebbene ci siano delle differenze, in particolare tra *BI* e *C*, in generale i testi sono abbastanza allineati tra loro e non sembrano avere vocabolari molto variegati.

Si è anche condotto uno studio sulla leggibilità dei testi con la demo online[6] di READ-IT, il primo strumento per la lingua italiana che misura la leggibilità combinando caratteristiche linguistiche di base, lessicali, morfosintattiche e sintattiche (Dell'Orletta et al., 2011). Per ciascun testo, il *tool* ha restituito quattro indici di leggibilità (*tabella 3*), calcolati da altrettanti modelli basandosi su gruppi diversi di tratti linguistici. Ogni valore indica con quale probabilità il testo appartiene alla classe dei testi «di difficile leggibilità» in riferimento alle *features* considerate. READ-IT calcola anche il GulpEase, un indice abbastanza datato che considera solo il numero medio di caratteri per parola e quello di parole per frase (Lucisano e Piemontese, 1988). Il risultato è un valore compreso tra 0 e 100, interpretabile come segue: (1) se inferiore a 80, il testo sarebbe di difficile lettura per chi ha la licenza elementare; (2) se inferiore a 60 per chi ha la licenza media; (3) se inferiore a 40 per chi

[6] La demo di READ-IT è disponibile al sito http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest

ha un diploma di scuola superiore. Si tenga presente, tuttavia, che, proprio per lo scarso numero di tratti considerati, diversi studi recenti hanno dimostrato l'inaffidabilità di questa ultima metrica.

Indice	**Testo**				
	C	***L***	***B***	***F***	***BI***
READ-IT Base	31,6%	17,3%	45,8%	96,3%	44,2%
READ-IT Lessicale	95,6%	82,2%	61,8%	57,4%	100,0%
READ-IT Sintattico	97,7%	57,0%	51,8%	100,0%	34,2%
READ-IT Globale	100,0%	99,2%	94,9%	100,0%	99,1%
GulpEase	52,4	58,5	55,0	44,0	53,2

Tabella 3. Punteggi di leggibilità calcolati da READ-IT per i quattro modelli e l'indice GulpEase.

Dai risultati emerge che le valutazione della leggibilità per uno stesso testo possono essere molto diverse in base al modello che le esegue: *F* e, in misura poco minore, *C* rientrano nei testi «difficili» con probabilità massime – o comunque superiori al 95% – per tre modelli su quattro. *F* parrebbe il testo più difficile da leggere. I punteggi calcolati per gli altri tre testi variano maggiormente rispetto ai modelli. Tuttavia, i punteggi di READ-IT Globale – che, combinando tratti generali, lessicali e sintattici, dovrebbe fornire un giudizio complessivo – classificherebbero tutti i testi come «difficili da leggere» con probabilità che arrivano (o sfiorano) il 100%. Solo *B* è inferiore di poco al 95%. Infine, i valori di GulpEase – tutti < 60 e > 40 – indicherebbero che i cinque testi siano facilmente leggibili solo da chi ha conseguito il diploma, anche se *L* è prossimo al 60. Inoltre, con un punteggio di 44, *F* risulterebbe di nuovo il testo più difficile da leggere.

3 Livelli linguistici di base

La prima fase del progetto è consistita nell'annotazione semiautomatica del corpus rispetto ai livelli di base della descrizione linguistica. Specificando l'opzione «--tokenize», i testi sono stati passati alla catena *UDPipe*, che li ha divisi in frasi (*sentence splitting*), le quali a loro volta sono state segmentate in *token* (tokenizzazione). Gli *output* sono stati poi corretti manualmente dai tre membri del gruppo congiuntamente. Il risultato delle revisioni ha costituito il punto di partenza comune per le successive analisi individuali.

A questo livello, l'*output* generato è un file in formato CoNLL strutturato come in *esempio 1*. Le frasi, così come i *token* al loro interno, sono identificate da un valore numerico progressivo. Ciascun *token* occupa una riga. Gli *underscore* che seguono la forma del *token* rappresentano le colonne in cui saranno aggiunte le annotazioni morfosintattiche e sintattiche nelle analisi successive. L'ultima colonna è dedicata ai *MISC attributes*, annotazioni ulteriori, perlopiù facoltative, aggiunte durante la *tokenizzazione*[7]. Nell'esempio se ne vedono due: «SpaceAfter=No» del *token* «5», che specifica che «cuoco» è seguito da un segno di punteggiatura senza uno spazio tra i due, e «SpacesAfter=\n» del punto finale, che indica che nel testo analizzato quel token precede un ritorno a capo.

```
# sent_id = 30
# text = Buon fuoco fa buon cuoco.
1 Buon _     _     _     _     _     _     _     _
2 fuoco      _     _     _     _     _     _     _     _
3 fa   _     _     _     _     _     _     _     _
4 buon _     _     _     _     _     _     _     _
5 cuoco      _     _     _     _     _     _     _     SpaceAfter=No
6 .    _     _     _     _     _     _     _     SpacesAfter=\n
```

Esempio 1. *Output* dell'annotazione ai livelli di base di [*L*, 35]

In generale, la correttezza dell'analisi automatica è stata altissima. Ma ciò era prevedibile. Tuttavia, ci sono stati dei problemi, che, occorrendo svariate volte – specialmente nei testi di cucina – hanno reso la revisione a tratti inaspettatamente faticosa. Li analizziamo nelle prossime sottosezioni.

3.1 *Sentence splitting*

Tutti e cinque i testi hanno un titolo iniziale. In più, quelli di cucina hanno anche dei titoli interni che marcano il passaggio tra una ricetta e quella successiva. In fase di *sentence splitting*, *UDPipe* ha avuto difficoltà rispetto ad ambo le tipologie. Partiamo dalla prima.

[7] La documentazione completa sui *MISC attributes* è disponibile qui: https://universaldependencies.org/misc.html

Nel file *.txt* del testo *BI*, il titolo «Fantasmi chiamati neutrini» è separato dal corpo del testo da un ritorno a capo. L'annotatore umano capisce immediatamente che si tratta di una frase a sé stante, ma quello automatico non ne è stato in grado, considerandolo parte della frase successiva:

```
# sent_id = 1
# text = Fantasmi chiamati neutrini
I neutrini sono
1     Fantasmi
2     chiamati
3     neutrini
4     I
5     neutrini
6     sono
[…]
```

```
# sent_id = 1
# text = Fantasmi chiamati neutrini
1   Fantasmi
2   chiamati
3   neutrini

# sent_id = 2
# text = I neutrini sono…
1   I
2   neutrini
3   sono
[…]
```

Esempio 2. Divisione errata (a *sx*) e correzione (a *dx*) del titolo di *BI*.

Con ogni probabilità, l'errore va imputato all'assenza di un segno di interpunzione forte a indicare esplicitamente la fine del titolo. Ciò troverebbe conferma nel fatto che i titoli dei testi di cucina *B* e *L*, rispettivamente «Salsa besciamella.» e «PIATTI DI FARINA.», terminano con un punto e, guarda caso, sono stati correttamente trattati come frasi indipendenti.

Passiamo al testo *F*. Di nuovo abbiamo un titolo «ATOMI E STELLE» senza punto e separato dal resto da un accapo. Essendo la stessa conformazione del titolo di *BI*, ci aspetteremmo un simile trattamento. E invece il *tool* ha attuato una divisione a dir poco originale, trattando inspiegabilmente il *token* «ATOMI» come frase indipendente e attaccando «E STELLE» alla frase successiva:

```
# sent_id = 1
# text = ATOMI
1       ATOMI

# sent_id = 2
# text = E STELLE In questa
conferenza […]
1   E
2   STELLE
3   In
4   questa
5 conferenza
[…]
```

```
# sent_id = 1
# text = ATOMI E STELLE
1   ATOMI
2   E
3   STELLE

# sent_id = 2
# text = In questa conferenza […]
1   In
2   questa
3   conferenza
[…]
```

Esempio 3. Divisione errata (a *sx*) e correzione (a *dx*) del titolo di *F*.

Passiamo ai titoli intermedi dei testi di cucina *C* e *L*. Si tratta di intestazioni costituite da un numero puntato, seguito dal nome di una ricetta e un punto. La loro analisi corretta richiede la divisione in

due frasi in prossimità del primo punto. Ad esempio, «52. Cavolo cappuccio in umido» deve essere scisso in «52.» e «Cavolo cappuccio in umido».
Data la medesima struttura nei due testi, anche in questo caso ci aspetteremmo lo stesso trattamento. E invece il modello risulta nuovamente incoerente, in quanto divide – correttamente – i titoli nel testo *L*, e li tiene uniti – a eccezione del primo – in *C*. Può darsi che il modello segua una sua logica nel prendere tali scelte, ma, visti i contesti di frase identici, è difficile trovare una spiegazione.

Altra fonte di errori per il *sentence splitter* automatico, questa volta limitatamente al testo *L*, sono state le indicazioni delle unità di misura che accompagnano le quantità di alcuni ingredienti. Ci si riferisce alle parole «chilogr.», «chilog.» e «gr.», che si ripetono rispettivamente quattro, una e tre volte. Nelle occorrenze delle prime due, il punto è stato interpretato come terminatore di frase:

```
# sent_id = 15
# text = […] un chilogr.
5     un
6     chilogr
7     .

# sent_id = 16
# text = circa di farina[…]
1     circa
2     di
3     farina
[…]
```

```
# sent_id = 10
# text = […] un chilogr. circa di
farina […]
5     un
6     chilogr.
7     circa
8     di
9     farina
[…]
```

Esempio 4. Divisone errata (a *sx*) e correzione (a *dx*) di un'occorrenza di «chilogr.» in *L*.

In prossimità delle occorrenze di «gr.», invece, non è stata realizzata una divisione in due frasi, ma il punto è stato trattato come *token* indipendente. È chiaramente un errore separare il punto quando il *token* in questione è una abbreviazione:

«[…] 100-150 **gr.** di formaggio grasso tagliato a dadolini […]»

```
[…]
25    150
26    gr
27    .
28    di
29    formaggio
[…]
```

```
[…]
33    150
34    gr.
35    di
36    formaggio
37    grasso
[…]
```

Esempio 5. *Tokenizzazione* errata (a *sx*) e correzione (a *dx*) di un'occorrenza di «gr.» in *L*.

Sono analisi scorrette da imputare al mancato riconoscimento da parte del modello di abbreviazioni arbitrarie e non standard, evidentemente mai incontrate in fase di addestramento.

Si tenga presente che questi *token* hanno generato problemi anche a livello morfosintattico. Il *tool* è riuscito a POS-taggarli correttamente come nomi, ma ha assegnato loro lemmi uguali alle forme. In quanto ai tratti morfologici, l'analisi automatica è stata integrata con l'informazione relativa al genere (maschile), ma si è evitato di specificare il numero. Dal contesto di frase sembrerebbe che «chilogr.» sia la forma singolare, poiché sempre preceduto da «un», e «chilog.» il plurale («chilog. 1 1/2 circa»). Ma servirebbero più dati per dirlo con certezza.

Finora si sono esposte le due tipologie di errori commessi dal modello durante il *sentence splitting*. Ma le revisioni hanno coinvolto anche un cospicuo numero di suddivisioni in frasi in prossimità dei segni «:» e «;» di per sé non errate, bensì incoerenti con le convenzioni adottate dal gruppo.
Al pari del punto, *UDPipe* interpreta i due punti come terminatori di frase pressoché sempre. Ma, dal momento che essi spesso sono seguiti da un chiarimento o un'integrazione di quanto affermato nella frase che li precede, spesso può avere senso evitare suddivisioni. Pertanto, si è deciso di tenere unite le frasi intramezzate dai due punti quando questi introducono (1) un elenco – es. «Nel 1930 si conoscevano soltanto due particelle**:** l'elettrone e il protone.» [*B*, 10] – o (2) una frase con valore di apposizione – es. «…principio-base della fisica**:** la conservazione dell'energia.» [*B*, 12].
Relativamente al punto e virgola, dobbiamo distinguere tra testi di cucina e scientifici. Nei primi, sembra che esso venga utilizzato con la virgola e le congiunzioni coordinanti per scandire i passaggi delle ricette descritte. Per questo motivo, si è deciso di dividere le frasi solo in prossimità del punto, tenendo unite le frasi separate «;» attaccate a quelle che lo precedono. A livello pratico, si è intervenuti nei casi in cui il *sentence splitter* aveva eseguito la suddivisione dopo i due punti.
Nei testi scientifici, invece, si è optato per tenere unite le frasi in prossimità di «;» solo qualora ci esistesse tra loro un legame esplicito, realizzato per esempio da una congiunzione – es. «…la mia esposizione dovrà per necessità limitarsi a pochi fatti essenziali**;** ed è questo che mi dà l'ardire…».
Va detto che nella quasi totalità dei casi, *UDPipe* aveva già provveduto a non dividere le frasi.

3.2 *Tokenizzazione*

Sul versante *tokenizzazione*, il modello ha riscontrato difficoltà nel trattamento di due sole strutture linguistiche: le preposizioni articolate e i costrutti formati da un verbo e un pronome clitico.
Costituendo un gruppo piuttosto contenuto e comparendo copiose in qualsiasi testo, le preposizioni articolate sono sempre riconosciute e divise correttamente nei loro elementi costituitivi. Tuttavia, nei testi di cucina sono presenti alcuni articoli partitivi che esprimono quantità generiche di alcuni ingredienti. Sono *token* unitari, ma, avendo frequenze molto basse, l'algoritmo, che ragiona su basi probabilistiche, li scambia per preposizioni articolate e li divide. Questi casi sono stati corretti:

«[…] se fosse troppo dolce aggiungerete ancora **dell'**aceto […]» [*B*, 18]

[…]	
7-8	**dell'**
7	**di**
8	**l'**
9	aceto
[…]	

[…]	
13	ancora
14	**dell'**
15	aceto
[…]	

Esempio 6. *Tokenizzazione* errata (a *sx*) e correzione (a *dx*) di un articolo partitivo in *B*.

Un altro problema legato al discorso precedente è consistito nel mancato riconoscimento e divisone delle preposizioni articolate «colla» e «degl'»: la prima, forma univerbata di «con la», è più comune nella lingua parlata che nell'uso scritto; l'altra è la forma elisa di «degli». In entrambi i casi, l'errore deriva dall'assenza di queste forme insolite dal *training* corpus. Va detto che «colla» è attestata due volte in ISDT, ma ovviamente si riferisce al nome. Questi errori sono stati corretti come segue:

«Cavolfiore **colla** besciamella» [*C*, 8]

3	Cavolfiore
4-5	**colla**
4	**con**
5	**la**
6	besciamella

Esempio 7. *Tokenizzazione* corretta di «colla».

«[…] le dosi **degl'**ingredienti […]» [*L*, 19]

5	dosi
6-7	**degl'**
6	**di**
7	**gl'**
8	ingredienti

Esempio 8. *Tokenizzazione* corretta di «degl'».

Il modello ha avuto difficoltà anche nella divisione dei costrutti verbo-clitico, frequentissimi nei testi di cucina. Il comportamento del *tokenizzatore* è difficilmente spiegabile, in quanto si attestano molte divisioni mancate – per l'esattezza: 4 in *B*, 20 in *C* e 13 in *L* – così come un discreto numero di analisi corrette. L'esempio seguente illustra un caso di univerbazione con la relativa correzione:

«[…] **strofinatele** con l'aglio […]» [*C*, 31]

[…]	
10	**strofinatele**
11	con
12	l'
13	aglio
[…]	

[…]	
10-11	**strofinatele**
10	**strofinate**
11	**le**
12	con
[…]	

Esempio 9. *Tokenizzazione* errata (a *sx*) e correzione (a *dx*) di un costrutto pronome-clitico in *C*.

Infine, si segnala la decisione unanime di considerare l'espressione «agro–dolce», divisa dal *tool* in «agro», «–» e «dolce», un *token* unico, considerandola come variante meno comune di agrodolce e più vicina all'originale francese *aigre-doux*.

4 Morfologia e morfosintassi

Il secondo passaggio del lavoro è consistito nell'annotazione automatica e nella revisione manuale rispetto al livello morfologico e morfosintattico del corpus risultato dalla fase precedente. L'analisi è stata eseguita dalla catena *UDPipe*, questa volta specificando l'opzione «--tag». A ogni *token* è stato assegnato (1) un lemma, ossia la sua forma base, quella riportata nei dizionari; (2) una *part-of-speech* dal *tagset* delle *Universal Dependencies*[8]; (3) una *part-of-speech* specifica per l'italiano dall'*ISST-TANL morpho-syntactic tagset*[9] e (4) un insieme, anche vuoto, di proprietà grammaticali e lessicali. Diversamente dalla precedente, in questa fase le analisi sono state corrette in autonomia da ogni membro del gruppo.

Di seguito si mostra la frase dell'*esempio 10* arricchita con le informazioni suddette, inserite nelle quattro colonne del CoNLL successive a quella relativa alla forma:

```
#  sent_id = 35
#  text = Buon fuoco fa buon cuoco.
1  Buon  buono  ADJ  A  Gender=Masc|Number=Sing  _  _  _  _
2  fuoco  fuoco  NOUN  S  Gender=Masc|Number=Sing  _  _  _  _
3  fa  fare  VERB  V  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
4  buon  buono  ADJ  A  Gender=Masc|Number=Sing  _  _  _  _
5  cuoco  cuoco  NOUN  S  Gender=Masc|Number=Sing  _  _  _ SpaceAfter=No
6  .  .  PUNCT  FS  _  _  _  _  SpacesAfter=\n
```

Esempio 10. *Output* dell'annotazione morfologica e morfosintattica di [*L*, 35]

Gli errori sono stati decisamente più consistenti sia qualitativamente che quantitativamente rispetto al livello precedente. Le operazioni di revisione hanno perciò richiesto tempo e un buon livello di attenzione, anche se la maggior parte dei problemi riscontrati non è stata difficile da correggere.

Ovviamente, il grado di accuratezza del modello non è il medesimo per tutti e quattro i dati associati ai *token*. Per averne un'evidenza, si è tenuto traccia, per ciascun *token* oggetto di correzione, della/e categoria/e di errore a cui appartenesse. Dati alla mano, si è studiata la composizione del campione complessivo di errori/interventi, resa graficamente dalla *figura 1*. Si mostrano le percentuali di errori/interventi rispetto ai parametri «LEMMA», «UPOS» e «FEATS» sia presi singolarmente, sia in tutte le loro combinazioni possibili. La categoria complessiva «LEMMA+UPOS+FEATS» comprende i *token* che hanno ricevuto un'analisi totalmente sbagliata.

Non indugiamo sulla descrizione della figura poiché già di per sé chiara. Ci concentriamo invece sulla fetta più ampia, riguardante i *token* analizzati correttamente a eccezione del lemma. Il fatto

[8] Il *tagset* di UD è disponibile a sito https://universaldependencies.org/u/pos/index.html
[9] Il *tagset* ISST-TANL è disponibile qui: http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

che questa categoria da sola copra quasi un terzo dei casi è perfettamente coerente con quanto già rilevato da Favaro et al. (2021), che avevano motivato questo fenomeno «nel fatto che "UDPipe […] utilizza un dizionario costruito a partire dal corpus di addestramento integrato da euristiche di analisi morfologica utilizzate per trattare forme sconosciute».

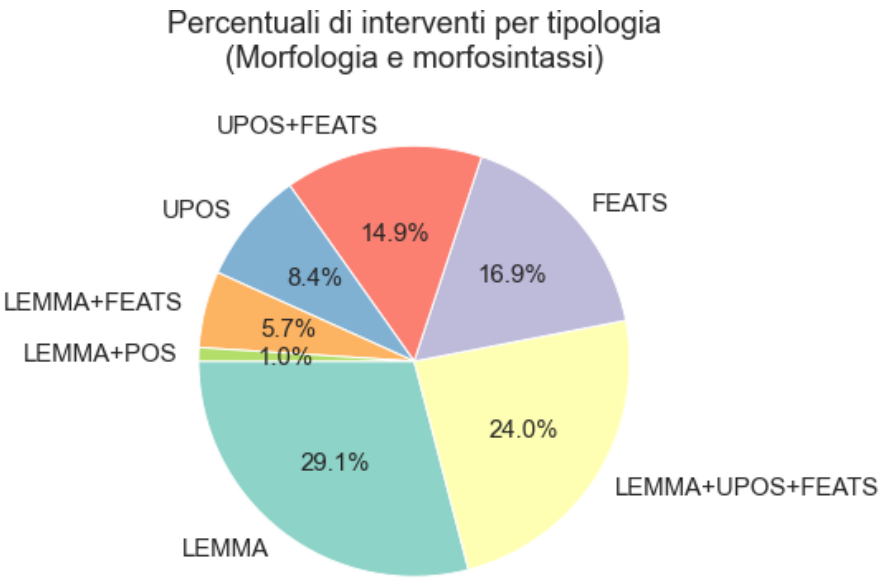


Figura 1. Percentuali delle tipologie di errori morfosintattici nel corpus globale.

Se poi a questa percentuale aggiungiamo quelle in cui «LEMMA» si combina con altre categorie, arriviamo al 59,8% di interventi che hanno riguardato – solo o anche – la correzione del lemma. Considerando insieme queste tipologie, la *tabella 5* mostra il numero di lemmi sbagliati per testo e la percentuale rispetto al numero totale dei *token* che lo compongono: se *C*, *B* e *BI* sono abbastanza allineati tra loro, *F* e, in particolare, *L* se ne distaccano vistosamente, raggiungendo i due estremi opposti. La differenza tra i due è notevole: *L* ha ben 58 lemmi errati in più di *F*.

Testo	**Lemmi errati**	
	N°	**%**
C	39	6,0%
L	73	11,3%
B	32	5,5%
F	15	2,6%
BI	30	4,5%

Tabella 4. Numeri e percentuali di *token* per testo a cui è stato assegnato un lemma sbagliato.

Da un punto di vista qualitativo, la casistica di errori rispetto al lemma è piuttosto variegata: si va da (1) lemmi inventati ma tutto sommato legittimi e coerenti rispetto al resto dell'analisi – es. «**rimestiate**» [*L*, 11] è ricondotto a «rimestiare» anziché «rimestare» – a (2) lemmi esistenti ma incoerenti con un'analisi corretta assegnata allo stesso *token* – es. «Cavolfiore **pasticciato**» [*C*, 2]

giustamente annotato come aggettivo ma associato al lemma verbale «pasticciare» –. E ancora da (3) casi in cui lemma e analisi sono coerenti ma sbagliati rispetto al *token* a (4) casi in cui lemma e analisi sono incoerenti tra loro ed entrambi errati rispetto al *token*. Infine, registriamo casi in cui (5) più occorrenze di una stessa parola ricevono lemmi diversi pur trovandosi in contesti simili e ricevendo analisi corrette.

Rispetto alla tipologia (4), è paradigmatico il caso di «Cominciate poi a sollevarla **adagio** […]» [*L*, 10]: «adagio» in italiano può essere l'avverbio sinonimo di «lentamente», un nome riferito a un movimento musicale e la I persona singolare dell'indicativo presente di «adagiare». L'analisi ha in qualche modo racchiuso tutti e tre i significati, in quanto il lemma è stato quello verbale, l'analisi (UPOS+XPOS+FEATS) quella di un nome, quando in realtà si tratta della forma avverbiale.

Rispetto alla tipologia (5), invece, segnaliamo il caso di «polenta». La sue 15 occorrenze in *L* hanno avuto tre lemmi distinti, tutti errati: «polento» in sei casi, «polentire» (verbo di terza coniugazione) in altri cinque e «polentare» (verbo di prima coniugazione) nei restanti quattro.

Torniamo al discorso iniziale. Si è vista la distribuzione delle categorie di errori a livello generale; *ma come varia rispetto ai singoli testi e domini?* Si osservi la figura sottostante:

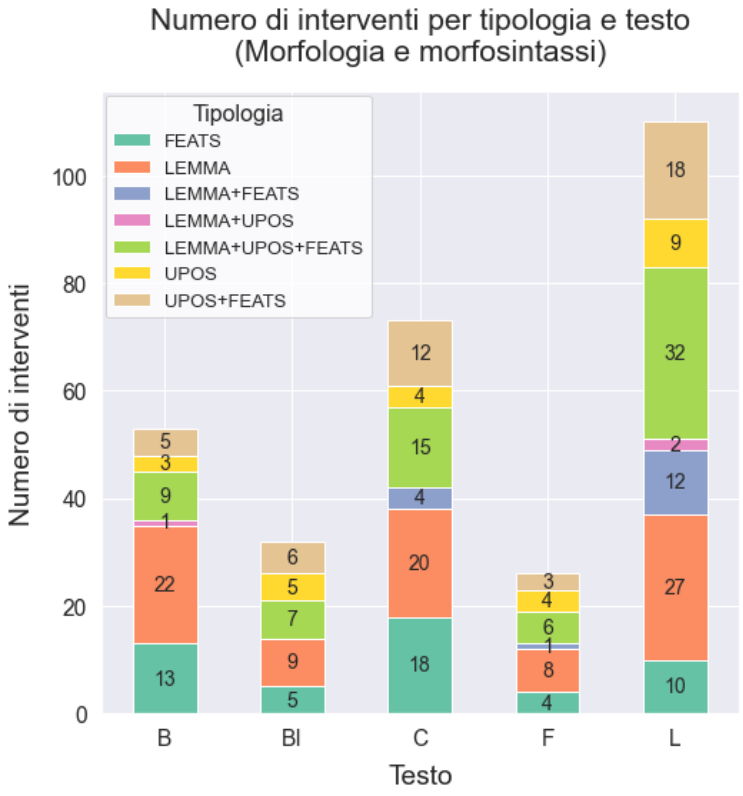


Figura 2. Numero di errori morfosintattici per tipologia e testo.

Dal grafico emerge chiaramente il divario tra domini: con 110 errori, *L* è il testo più problematico, l'unico in cui sono rappresentate tutte e sette le categorie considerate; si distanzia notevolmente

anche dagli altri due testi gastronomici (in particolare da *B*), nei quali, comunque, si registrano numeri di errori molto più alti dei testi scientifici. Basti vedere che *C* ne ha più del doppio sia di *BI* che di *F*, che è il testo rispetto al quale il modello ha raggiunto il grado di accuratezza massimo.

Nel corpus sono stati individuati due errori tipografici: uno in *F* (*esempio 11*) e l'altro in *L* (esempio 12). In fase di revisione si è optato per lasciare invariata la forma, in modo da non alterare il testo, e correggere il lemma e l'analisi morfosintattica. In più, tra i tratti morfologici si è aggiunto l'attributo «Typo=Yes», definito proprio per marcare una forma tipograficamente errata.

«[…] e da queste si poté solo in seguito, per **meno** di laboriose indagini arrivare a riconoscere […]»

```
meno    meno    ADV    B    _
        mano    NOUN   S    Gender=Fem|Number=Sing|Typo=Yes
```

Esempio 11. Errore tipografico riscontrato in [*F*, 5] con relativa correzione.

«[…] studiandovi di portare in alto la pasta ch'è in fondo al paiolo, **mai** sempre con mano leggera […]»

```
mai     mai     ADV       B       _
        ma      CCONJ     CC      Typo=Yes
```

Esempio 12. Errore tipografico riscontrato in [*L*, 10] con relativa correzione.

Dal contesto si capisce bene che quel «meno» in realtà sarebbe un «mano» e quel «mai» un «ma».

4.1 *Open class words:* verbi, aggettivi e nomi

Al livello morfosintattico le parole appartenenti alle classi chiuse sono tendenzialmente analizzate con accuratezze massime. I problemi maggiori derivano ovviamente dall'annotazione delle classi aperte. In questa sezione, si traccia quindi un quadro degli interventi principali in merito a «verbi», «aggettivi» e «nomi».

Si è detto in *capitolo 2* che i testi di cucina descrivono ricette. Le indicazioni sui passaggi da seguire vengono date ai lettori sia in costruzioni con il *si* – perlopiù passivante – sia attraverso verbi alla II persona plurale dell'imperativo. Costrutti del primo tipo si registrano quasi esclusivamente nel testo *B*. A titolo esemplificativo, si riporta un estratto della frase n°6:

> «[…] **si versa** nella casseruola il latte, **si stempera** bene il composto di farina e burro, **si condisce** con sale, e un nonnulla di noce moscata […]» [*B*, 6]

Decisamente diverso è, invece, il caso degli imperativi: complessivamente, nei testi di cucina se ne contano ben 90: 18 in *B*, 51 in *C* e 21 in *L*. Il totale sale a 92 se si aggiungono le due occorrenze della frase 18 del testo *BI*: «Così, caro popolo radioattivo, **esaminate** e **giudicate**.».

Si tratta di una quantità notevole se si considera che all'interno della *treebank* ISDT – su cui, ricordiamo, il modello utilizzato per l'annotazione automatica è stato addestrato – si registrano solo 248 verbi al modo imperativo, di cui 24 (meno del 10%) sono coniugati alla II persona plurale. Quanto appena detto, dunque, rivela che i tre testi di cucina (92 frasi e 1877 *token*) contengono più del triplo degli imperativi di II persona plurale di tutta ISDT (14167 frasi, 278429 *token*).

Quanto più un fenomeno è attestato in un *training set*, tanto maggiore è la probabilità che il modello addestrato con quei dati riesca a riconoscerne correttamente nuovi casi. Considerato lo scarso numero di imperativi contenuti in ISDT, si può immaginare che le predizioni effettuate dal modello rispetto a questa categoria di verbi non debbano essere troppo accurate. E infatti, se si osservano le differenze tra numero di imperativi riconosciuti e di quelli non riconosciuti per testo (*tabella 5*), ci si accorge che nei casi migliori – testi *B* e *C* – le interpretazioni corrette raggiungono il 50%, mentre in *L* solo 6 imperativi su 21 (28,6%) sono riconosciuti.

Testo	N° Imperativi	Non Riconosciuti	Riconosciuti
B	18	9	9
C	51	26	25
L	21	15	6
BI	2	2	0
Totale	92	52	40

Tabella 5. Imperativi di II persona plurale nel corpus: numero totale, di quelli riconosciuti e non.

Si tenga presente che nei «Riconosciuti» sono inclusi anche i nove imperativi – 1 in *L*, 7 in *C* e 1 in *B* – con lemma sbagliato, ma analisi morfologica e morfosintattica corretta.

Analizziamo il comportamento del modello nella disambiguazione degli imperativi, fornendo un ventaglio delle analisi restituite e riflettendo su possibili motivazioni alla base di tanta varietà.
Va detto che nell'80,8% dei casi non riconosciuti (42 su 52), l'errore è compiuto a livello dei tratti morfologici associati alle forma. In particolare, 28 imperativi – 5 in *B*, 13 in *C*, 8 in *L* e 2 in *BI* – sono stati annotati come «participio passato femminile plurale» (*esempio 13*); gli altri 14 – 2 in *L*, 8 in *C* e 4 in *B* – come verbo «indicativo presente di seconda persona plurale» (*esempio 14*).

```
mescolate mescolare VERB V Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part
                            Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
```

Esempio 13. Imperativo scambiato per participio passato femminile plurale [*B*, 15].

```
fate fare VERB V Mood=Ind|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
                 Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
```

Esempio 14. Imperativo scambiato per indicativo presente di II persona plurale [*L*, 28].

Si fa notare che ad alcune occorrenze di entrambe le tipologie era stato assegnato anche un lemma sbagliato, talvolta incoerente rispetto alla POS «VERB», altre volte legittimo ma inesatto dal punto di vista morfologico. Ne riportiamo un esempio per categoria:

```
versate  versato  VERB  V  Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part
         versare     Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
```

Esempio 15. Imperativo non riconosciuto e lemmatizzato in modo incoerente [*L*, 18]

```
disponete disponere  VERB  V  Mood=Ind|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
          disporre     Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
```

Esempio 16. Imperativo non riconosciuto e lemmatizzato in modo legittimo ma errato [*L*, 31]

Nel primo esempio, «versato» è il lemma di un aggettivo. Poiché «versato» esiste come aggettivo, l'analisi potrebbe essere legittima in un altro contesto, ma non lo è in riferimento a un verbo. Da notare che una forma identica in una frase precedente [*L*, 9] era stata lemmatizzata correttamente. Nel secondo esempio, invece, «disponere» è a tutti gli effetti il lemma di un verbo, ma non è corretto nella forma. Non avendo mai incontrato il *token* «disponete», il modello avrà sicuramente provato a lemmatizzarlo seguendo qualcuna delle euristiche a cui già si è fatto cenno.

Il restante 19,2% dei casi non riconosciuti è costituito da dieci imperativi POS-taggati in maniera errata, con conseguente analisi morfologica, e talvolta anche lemma, sbagliata. Di questi, tre – 2 in *L* e 1 in *C* – sono stati annotati come «aggettivo», quattro – 1 in *L* e 3 in *C* – come «nome», due in *L* come «nome proprio» e uno in *C* come «aggettivo indefinito femminile plurale». Vediamo un esempio per ognuna di queste tipologie:

```
cessate cessato ADJ   A Gender=Fem|Number=Plur
        cessare VERB  V Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
```

Esempio 17. Imperativo scambiato per aggettivo femminile plurale [*L*, 9].

```
pulite pulire NOUN S Gender=Masc|Number=Sing
              VERB V Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
```

Esempio 18. Imperativo scambiato per nome ma lemmatizzato correttamente [*C*, 20].

Si noti l'incoerenza tra il lemma verbale «pulire» e un'analisi di natura nominale.

```
Spremete spremete PROPN SP
         spremere VERB V Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
```

Esempio 19. Imperativo scambiato per nome proprio [*L*, 27].

L'annotazione di «Spremete» come «nome proprio» è, almeno in parte, legata alla maiuscola. Tuttavia, non è l'unico caso collocato a inizio frase e, anzi, molti imperativi in questa posizione

vengono analizzati correttamente. Il motivo più probabile sarebbe di nuovo l'assenza su ISDT di form riconducibili al lemma «spremere». Nella completa «ignoranza», il modello si sarà basato sull'unico indizio esplicito, la maiuscola per l'appunto.

```
1 Bagnate bagnate DET DI Gender=Fem|Number=Plur|PronType=Ind
          bagnare VERB V Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin

2 poscia posciare NOUN S Gender=Fem|Number=Plur
         poscia  ADV  B _
```

Esempio 20. Imperativo scambiato per aggettivo indefinito femminile plurale [*C*, 16].

In quest'ultimo caso, oltre al *token* «incriminato», si è riportato anche quello successivo, poiché funzionale alla comprensione del comportamento del modello: «poscia», avverbio di uso letterario che sta per «dopo», «poi», non è attestato nel corpus di addestramento. *UDPipe* lo ha analizzato – chissà perché – come «nome femminile plurale», assegnando un lemma verbale, e, nell'incertezza generale, ha deciso che «Bagnate» dovesse essere il suo determinante indefinito.
È un esempio eloquente di come il modello, addestrato sull'italiano giornalistico contemporaneo, «inciampi» nel confrontarsi con forme antiche o desuete.

Nel capitolo precedente si è visto come le costruzioni verbo-pronome clitico, copiose nei testi di cucina, siano state oggetto di numerosi errori in fase di *sentence splitting*. Aggiungiamo adesso che anche a livello morfosintattico ci sono stati molti problemi. Considerando che, nella maggior parte dei casi, queste strutture coinvolgono verbi imperativi, alla luce di quanto esposto finora potremmo pensare che il mancato riconoscimento di questi ultimi possa riflettersi e avere un qualche ruolo nell'analisi erronea del pronome che segue.
Nei casi più semplici, l'analisi è corretta a eccezione del lemma, che, nel caso specifico, deve essere uguale alla forma. È il caso, per esempio, di «stratificate**la**» [*L*, 28], a cui è stato assegnato il lemma «il» (al posto di «la»), o di «condite**le**», a cui è stato associato «lo» (invece di «le»). Sono anche frequenti i casi in cui i clitici sono stati scambiati per gli omografi – e molto più comuni – articoli determinativi: tipicamente questo si verifica in contesti in cui il costrutto è seguito da un nome; il modello interpreta il clitico come articolo del nome che lo segue. Vediamo un esempio:

«Tagliatela poi a spicchiettini e **datele sapore** con burro, pepe e sale.»

```
7   date  dare  VERB  V  Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part
                        Mood=Imp|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin

8   le  il  DET  RD  Definite=Def|Gender=Fem|Number=Plur|PronType=Art
        le  PRON  PC Clitic=Yes|Gender=Fem|Number=Sing|Person=3|PronType=Prs
```

```
9   sapore  sapore  NOUN  S  Gender=Fem|Number=Plur
                             Gender=Masc|Number=Plur
```

Esempio 21. Contesto di frase in cui un clitico è scambiato per articolo determinativo [*C*, 10].

L'imperativo "date" è annotato come «participio passato», «le» come «articolo definito femminile plurale» e «sapore» – complice il fatto che termina con la «e» come la maggioranza dei sostantivi femminili plurali in italiano – concorda per genere e numero con esso.
Casi simili sono «facendove**la piovere** adagio» e «lavorate**la fortemente** col mestolo»: «piovere» e «fortemente», senza un'apparente logica, sono stati trattati come nomi femminili singolari e i due «la» come loro articoli definiti.

Si è approfondito lo studio degli imperativi perché hanno rappresentato una delle maggiori – se non la maggiore – fonti di errori relativamente al livello morfosintattico. Ma anche altri modi e tempi verbali *hanno dato del filo da torcere* al modello, in particolare i verbi al futuro. Nel corpus se ne contano 26, così suddivisi: 11 in *B*, 4 in *C*, 5 in *L*, 2 in *BI* e 4 in *F*. Dieci di questi, coniugati alla II persona plurale, sono utilizzati nei testi di cucina per indicare al lettore cosa fare una volta compiuto un certo passaggio di una ricetta – es. «[…] quindi passate il sugo ottenuto, nel quale **getterete** poi il cavolo […]» [*L*, 27]. Bene, nessuno di questi è analizzato in modo completamente corretto: in sei casi l'errore è circoscritto al lemma, che in tre occasioni è realizzato da pseudo-infiniti – «diluiere», «metteere» e «spremeere» al posto di «diluire», «mettere» e «spremere» – e in tre coincide con la forma. Gli altri quattro, invece, hanno ricevuto un lemma identico alla forma e sono stati analizzati come «nome femminile singolare». Vediamo un esempio:

«[…] nel caso contrario la correggerete con qualche altro cucchiaino di zucchero.»

```
correggerete  correggerete  NOUN  S  Gender=Fem|Number=Sing
              correggere    VERB  V
                                    Mood=Ind|Number=Plur|Person=2|Tense=Fut|VerbForm=Fin
```

Esempio 22. Futuro di II persona plurale scambiato per nome femminile singolare [*B*,18].

Si tratta chiaramente di un errore sistematico compiuto dal modello, la cui spiegazione sembrerebbe ancora una volta legata alla scarsità di dati all'interno nella *treebank*: da un'interrogazione condotta tramite *Grew-match* si è infatti scoperto che essa ha solo cinque occorrenze di verbi alla II persona plurale del futuro, tre dei quali sono voci del verbo «vedere».
Sempre rispetto al tempo futuro, si segnala solamente l'analisi errata di «cercherò» [*F*, 2], annotato come verbo alla III persona singolare dell'indicativo passato, invece che alla I singolare del futuro. Gli altri 15 futuri – ausiliari e verbi alla III persona singolare – sono stati annotati correttamente.

Soffermiamoci sugli ausiliari e i modali, che nel tagset di UD vengono ricondotti alla POS AUX. Costituendo una classe chiusa di parole, è possibile rappresentarli in modo abbastanza consistente

all'interno del *training corpus* di un modello di analisi automatica. Dal canto suo, il nostro modello addestrato su ISDT ha analizzato propriamente la maggior parte di essi trasversalmente ai modi e ai tempi verbali. Tuttavia, in due situazioni i testi sono riusciti a *fare lo sgambetto* al modello.
La prima riguarda la difficoltà nel riconoscere una verbo come ausiliare (o modale) di un participio passato quando i due sono intramezzati da un certo numero di *token*. Ciò è capitato nella frase «[…] e da queste si **poté** solo in seguito, per meno di laboriose indagini **arrivare** a riconoscere […]» [*F*, 5] – che, se vogliamo, è un caso estremo – ma anche in contesti più semplici.

Altro caso: «avere» tende a non essere riconosciuto come ausiliare quando il participio passato a cui si riferisce è concordato con il complemento oggetto rispetto a genere e numero. Si tratta di un aspetto della lingua letteraria del primo '900 – epoca di composizione dei testi di cucina –, che lo standard contemporaneo ha accantonato in favore di un uso invariato del participio, tranne nei casi in cui il l'oggetto è rappresentato da un pronome diretto. Vediamo un esempio:

«[…] unitevi 4 acciughe, che **avrete** prima **pulite** e **disfatte** nell'olio caldo, […]»

```
18 avrete avere VERB V Mood=Ind|Number=Plur|Person=2|Tense=Fut|VerbForm=Fin
                AUX  VA

20 pulite pulito ADJ  A Gender=Fem|Number=Plur
          pulire VERB V Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part

22 disfatte disfatto ADJ  A Gender=Fem|Number=Plur
            disfare  VERB V Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part
```

Esempio 23. Mancato riconosciuto di un ausiliare e dei suoi due participi [*C*, 16]

Dal momento che «pulite» e «disfatte» concordano per genere e numero l'oggetto «che» (= *le quali*, riferito ad «acciughe»), il modello li ha annotati come aggettivi. In assenza di participi passati, «avrete» non poteva che essere analizzato come verbo semplice.

Quest'ultimo esempio introduce un problema che si è presentato di continuo, specialmente nei testi gastronomici, sia nella revisione dell'analisi morfosintattica sia nella creazione del *Gold standard*: l'eterno dilemma tra participio passato e aggettivo. Si è cercato di attuare il consiglio delle docenti di annotare la forma ambigua come «verbo» se essa avesse dei dipendenti che corrispondessero formalmente ai dipendenti di un verbo; come «aggettivo» altrimenti.

L'unica eccezione rispetto al criterio di riferimento suddetto è rappresentata dal seguente caso:

«[…] tre manate di fagioli **sgranati** e **cotti** bene nell'acqua salata, […]»

```
46 sgranati sgranare ADJ A Gender=Masc|Number=Plur
                     VERB V Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part
```

```
48 cotti cotto    ADJ  A Gender=Masc|Number=Plur
         cuocere  VERB  V Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part
```

Esempio 24. Unico caso in cui si è fatta eccezione per dare coerenza a due participi coordinati [*L*, 16].

Il modello aveva considerato «sgranati» e «cotti» aggettivi. Ma, se la prima analisi può aver senso, dal momento che il *token* non ha dipendenti, la seconda sembrerebbe inesatta: da «cotti», infatti, dipende un complemento obliquo – di luogo – che ha «acqua» come testa. Si è dunque corretta quest'ultima annotazione. Poi, trattandosi di due *token* coordinati sintatticamente da una congiunzione esplicita, si è cambiata anche la prima in modo da renderli due verbi.
Oltre a queste situazioni ambigue, si registrano anche casi di annotazioni sbagliate che coinvolgono le due parti del discorso in oggetto. Se ne segnala uno per tutti:

«[…] la cui osservazione è resa **incerta** dalla enorme grandezza e lontananza.»

```
Incerta  incrire  VERB  V  Gender=Fem|Number=Sing|Tense=Past|VerbForm=Part
         incerto  ADJ   A  Gender=Fem|Number=Sing
```

Esempio 25. Aggettivo erroneamente scambiato per participio passato femminile singolare [*F*, 2].

La forma «incerta» è chiaramente un «aggettivo femminile singolare». L'analisi automatica, che l'ha ricondotta a un presunto verbo «incrire», è pertanto sbagliata senza nessuna ambiguità.

Sono frequenti anche i casi in cui un nome è scambiato per aggettivo (*esempio 26*) e quelli in cui, viceversa, un aggettivo è annotato come nome (*esempio 27*). Ciò può essere più o meno legittimo, perché (1) da un punto di vista di analisi morfologica si tratta di due parti discorso simili, in quanto il lemma è tipicamente la forma maschile singolare e i tratti assegnati sono «genere» e «numero»; e (2) alcune forme effettivamente appartengono all'una o all'altra categoria a seconda del contesto.

«Salsa **besciamella**.»

```
besciamella   besciamella   ADJ   A   Gender=Fem|Number=Sing
                            NOUN  S
```

Esempio 26. Nome scambiato per aggettivo a causa poiché preceduto da un nome [*B*, 1].

«Fatelo bollire alcuni istanti e servitelo **caldo**.»

```
caldo   caldo   NOUN   S   Gender=Masc|Number=Sing
                ADJ    A
```

Esempio 27. Aggettivo scambiato per nome poiché preceduto da un presunto articolo [*C*, 28].

Nel primo caso, «besciamella» è stato analizzato come aggettivo probabilmente in riferimento al nome «salsa» che lo precede. Nel secondo esempio, invece, l'errore potrebbe derivare dal fatto che il pronome clitico «lo» sia stato in origine trattato come «articolo determinativo».

5 Sintassi

Il terzo *step* del lavoro ha riguardato l'analisi semiautomatica rispetto al livello sintattico del corpus risultante dai due passaggi precedenti. I testi, annotati fino alla morfosintassi, sono stati passati a *UDPipe*, digitando l'opzione «--parse», che ha ricostruito – o almeno ci ha provato – le relazioni binarie di dipendenza tra parole. Nello specifico, a ogni *token* è stato assegnato (1) l'identificatore univoco della testa da cui dipende e (2) l'etichetta del tipo di dipendenza secondo il formalismo UD. Tali informazioni sono inserite rispettivamente nella settima e nell'ottava colonna del CoNLL (*esempio 28*). Anche in questa fase le correzioni sono state eseguite autonomamente.

```
#  sent_id = 35
#  text = Buon fuoco fa buon cuoco.
1  Buon  buon  ADJ  A  Gender=Masc|Number=Sing  2  amod  _  _
2  fuoco  fuoco  NOUN  S  Gender=Masc|Number=Sing  3  nsubj  _  _
3  fa fare VERB V  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 0 root
4  buon  buono  ADJ  A  Gender=Masc|Number=Sing 5  amod  _  _
5  cuoco  cuoco  NOUN  S  Gender=Masc|Number=Sing  3  obj  _  SpaceAfter=No
6  .  .  PUNCT  FS  _  3  punct  _  SpacesAfter=\n
```

Esempio 28. *Output* dell'annotazione sintattica di [*L*, 35]

Le operazioni di revisione hanno rilevato un cospicuo numero di errori trasversalmente ai cinque testi. La sintassi è un livello complesso, che considera le parole non come elementi a sé stanti, bensì in relazioni su cui si costruisce la struttura della frase. Questo fa sì che un errore possa propagarsi e arrivare a invalidare anche interi periodi. Pertanto, le correzioni sono state molto impegnative, specialmente in relazione ai testi scientifici. Vediamo come si sono distribuiti gli errori:

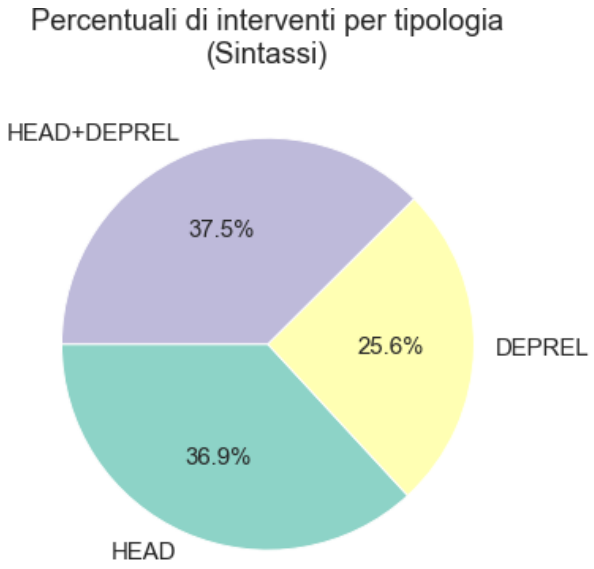


Figura 3. Percentuali delle tipologie di errori sintattici nel corpus globale.

Si può vedere che il numero di casi in cui una relazione corretta viene collegata alla testa sbagliata quasi coincide con il totale di quelli in cui sia testa che etichetta sono errati. Si tratta rispettivamente di 108 e 110 interventi. Sensibilmente minore – 75 errori – è la quantità delle parole collegate alla testa corretta ma etichettate in modo inesatto.

Ma è l'osservazione della *figura 4* che fornisce i dati più interessanti: con 70 errori, *L* è ancora il testo più problematico. Ma siamo ben lontani dalla situazione vista al livello precedente. Il numero di errori nei testi scientifici aumenta al punto che *F* (59 errori) supera *B* (56 errori) e si colloca al pari di *C*. *BI* è adesso il testo analizzato più accuratamente, anche se si riduce molto la differenza tra numero massimo e numero minimo di interventi: 21 contro gli 84 della morfosintassi. Si tratta in realtà di qualcosa che ci aspettavamo: come si vedrà nella prossima sezione, la sintassi dei testi scientifici ha dato problemi strutturali enormi al *parser* e, di conseguenza, anche ai revisori.

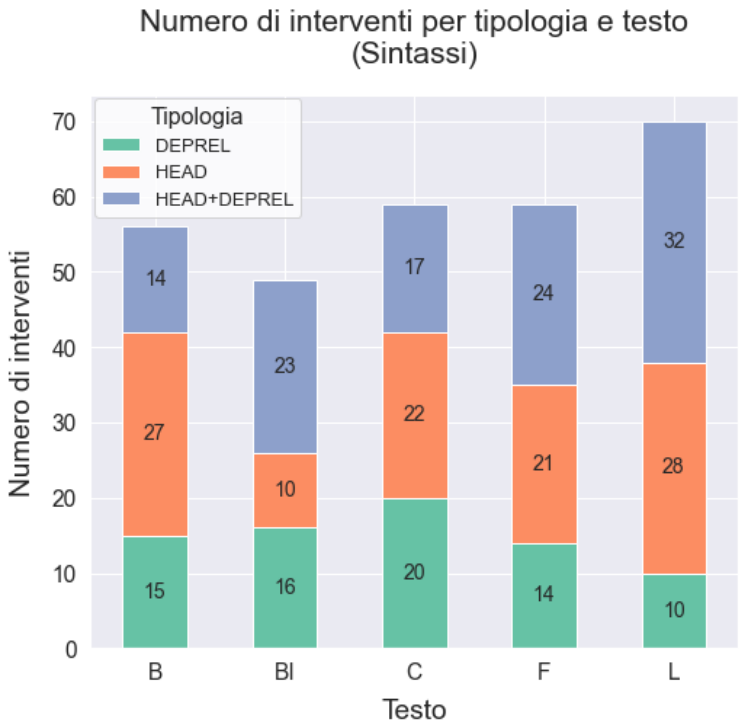


Figura 4. Numero di errori sintattici per tipologia e testo.

5.1 Domini diversi: sintassi diversa

In più occasioni si è fatto cenno alle differenze sostanziali tra i due domini testuali considerati. In effetti, una ricetta gastronomica e una dissertazione scientifica sono generi letterari lontanissimi, così come diversi sono gli ambiti e i pubblici a cui si riferiscono. In questa sezione ci soffermiamo sulla struttura sintattica dei testi, evidenziando per i due domini i principali tipi di problemi avuti dal *parser* nel ricostruire le relazioni tra proposizioni.

Una ricetta è pensata e si rivolge a un pubblico eterogeneo, potenzialmente popolato da persone di ogni età e *background* culturale. Per essere funzionale, dunque, deve (1) comunicare tutte e sole le informazioni strettamente necessarie alla riuscita del procedimento e (2) avere una sintassi lineare. In termini pratici, questo si traduce prima di tutto in una costruzione dei periodi di tipo paratattico. E infatti, i periodi dei testi di cucina in oggetto sono costituti da sequenze di frasi coordinate, che descrivono singoli passaggi e che, come già indicato in *Sezione 3.2*, sono scandite da congiunzioni coordinanti o dai segni di punteggiatura *virgola* e *punto e virgola*. Il punto marca la fine dei periodi. Ci sono poche subordinate, perlopiù di 1° grado, ossia direttamente dipendenti dalla principale o da una sua coordinata. La radice è (quasi) sempre il primo, o al massimo il secondo, verbo (o parte nominale nel caso di un predicato nominale) incontrato – dunque non è difficile da individuare – e tutte le proposizioni coordinate dipendono da essa tramite la relazione «conj».
È una sintassi piuttosto semplice. Non ci stupisce, dunque, che molti periodi siano stati analizzati perfettamente. Ne riportiamo due casi a titolo esemplificativo:

> «Da ultimo la **ridurrete** leggermente a una palla, **scuotendo**la contro le pareti del paiolo, e poi la **verserete** sopra un tagliere bianco e la **servirete fumante** con un filo per **staccar**ne le fette.»

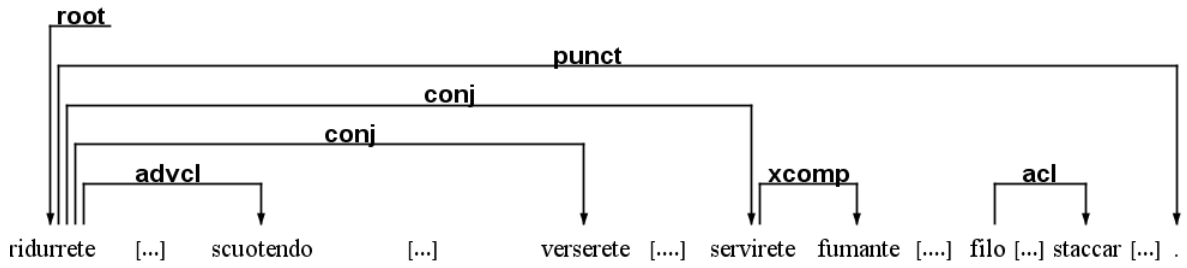


Figura 5. Periodo ricostruito correttamente dal modello [*L*, 13].

«Da ultimo […] a una palla» è la proposizione principale, con «ridurrete» come radice. Le frasi «e poi la verserete […] bianco» e «e la servirete […] con un filo» sono coordinate della principale. Le subordinate «scuotendola […] paiolo» e «per staccarne le fette» dipendono rispettivamente dalla principale e dall'ultima coordinata – in altre parole, sono di 1° grado.

> «**Bagnate**lo allora con mezzo bicchiere scarso di aceto e **mescolate** con un cucchiaio di legno per **sciogliere** bene lo zucchero.»

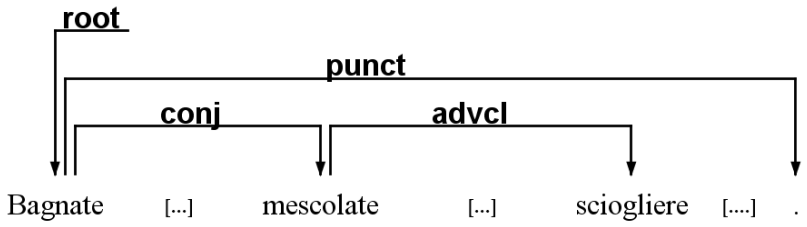


Figura 6. Periodo ricostruito correttamente dal modello [*B*, 15].

Questo secondo caso è ancora più semplice del precedente. La radice è di nuovo rappresentata dal primo verbo, «Bagnate»; «mescolate» è la testa di una proposizione coordinata alla principale, da cui dipende la subordinata di 1° grado «per sciogliere bene lo zucchero».

Ma *non è tutto oro ciò che luccica*. Molte indicazioni sono espresse con i verbi causativi «fare» e «lasciare»: al lettore è richiesto di *fare* o *lasciare* che il cibo in preparazione «compia» una certa azione – espressa dal verbo della subordinata – prima di proseguire. Si tratta di casi come «Lasciate sobbollire mezz'oretta il composto […]» [*L*, 28] e «[…] fate liquefare lo zucchero […]» [*B*, 14]. Alcuni di questi verbi si trovano a inizio frase e ne sono la radice. È capitato che una coordinata della principale (con radice nel verbo causativo) non sia stata collegata ad essa, ma alla subordinata retta dal causativo (*figura 7*). Si è provveduto ad apportare le dovute correzioni.

> «**Fate** cuocere due o tre minuti e poi **aggiungete** un ramaiolo di brodo di carne e un ramaiolo di brodo di pesce.»

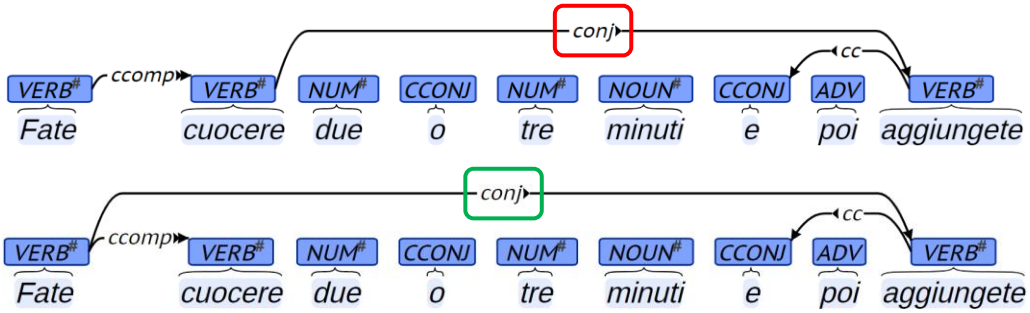
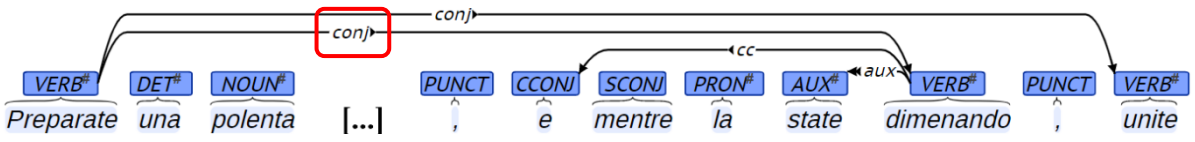


Figura 7. Coordinata della radice ricondotta erroneamente a una sua subordinata.

Il *parser* aveva collegato «aggiungete» – testa della proposizione coordinata – a «cuocere», anziché a «Fate». Si fa anche presente che in casi del genere, la subordinata dipendente dal verbo causativo era annotata con la relazione «xcomp». Interrogando la *treebank*, si è invece visto che in questi contesti si usa «ccomp». Perciò, si sono apportate le dovute correzioni.

A livello ipotattico, il modello ha avuto difficoltà nel ricostruire adeguatamente le dipendenze tra proposizioni nei casi in cui tra due coordinate è inserita una subordinata di 1° grado che si riferisce alla seconda di esse. Si osservi la seguente analisi:

> «**Preparate** una polenta di un chilogr. di farina, e mentre la state **dimenando**, **unite**vi un pezzo di burro della grossezza d'un uovo circa, […]»



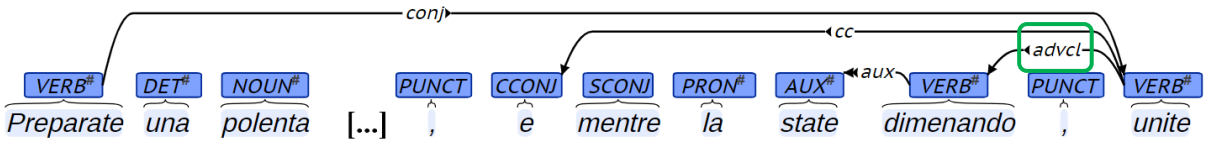


Figura 8. Mancato riconoscimento della subordinata retta dalla coordinata della principale [*L*, 16].

Il modello aveva considerato «e mentre […] dimenando» e «unitevi […] burro» coordinate della principale; non aveva, cioè, riconosciuto la frase intermedia come subordinata della successiva. Si è pensato che l'errore potesse avere un legame con la punteggiatura: introducendo la coordinata «unitevi», la congiunzione «e» avrebbe dovuto essere seguita da una virgola a fare il paio con quella dopo «dimenando», racchiudendo la subordinata. In altre parole, una scrittura più corretta sarebbe stata: «[…] un chilogr. di farina, e**,** mentre la state dimenando**,** unitevi […]».
Lo stesso comportamento è stato attuato anche rispetto alla frase «fatela sgocciolare, e quando è ancora ben calda, mettetela […]» [*C*, 6], che, guarda caso, presenta una punteggiatura analoga. Si è dunque fatto il confronto tra questi due esempi e una terza frase con la stessa conformazione, ma che ha la virgola nel posto indicato:

> «[…] Dopo un paio di minuti si **versa** nella casseruola il latte, si stempera bene il composto di farina e burro, si condisce con sale, e un nonnulla di noce moscata, e, senza **smettere** mai di mescolare, si **fa** addensare la salsa.» [*B*, 6]

In questo caso, «smettere» è stato annotato come «conj» della radice «versa» e «fa» come «advcl» di «smettere». Dunque, il modello ha distinto tra una coordinata e una subordinata, ma ha invertito le analisi. Continua, dunque, a essere un'analisi sbagliata, ma il fatto che ci sia una differenza con quelle precedenti potrebbe rivelare un coinvolgimento di qualche tipo della virgola.

I testi di scienza, invece, sono rivolti a pubblici più o meno competenti, a cui forniscono spiegazioni dettagliate e tecniche dell'argomento trattato. La loro struttura sintattica è ciò che di più lontano ci sia da quella delineata per i testi gastronomici. I periodi, generalmente più lunghi, sono costruiti ricorrendo abbondantemente alla ipotassi, le subordinate sono annidate tra loro e spesso precedono la proposizione principale. Riportiamo un caso (forse) estremo, ma paradigmatico, tratto da *F*:

> «Anche non volendo **ricordare** le speculazioni dei Greci, Democrito e i suoi seguaci, che **giunsero** all'ipotesi che la materia fosse **costituita** da atomi, e cioè da tante particelle **staccate** una dall'altra, poiché non **arrivavano** a **comprendere** come una materia continua potesse essere **compressibile**, la nozione dell'esistenza degli atomi e delle molecole venne **introdotta** nella scienza moderna per due vie differenti**.**» [*F*, 6]

Il primo dato che colpisce è certamente l'estensione del periodo: ben 76 *token*. In *capitolo 2* si era già visto che *F* ha circa la metà delle frasi degli altri, lunghe in media quasi il doppio.

Ci imbattiamo in una serie di subordinate incassate tra loro prima di individuare nell'ultima frase la proposizione principale. L'analizzatore ha avuto difficoltà tremende nell'analizzare il periodo: ha individuato la radice in «ricordare», che in realtà è testa della subordinata concessiva («advcl») legata dalla radice «introdotta», che a sua volta è stata annotata come «xcomp» di «comprendere». Ha riconosciuto «che giunsero all'ipotesi» come proposizione relativa, riconducendola a «seguaci» anziché a «Greci»; «poiché non arrivavano» è stata considerata giustamente «advcl» (subordinata causale), ma erroneamente collegata a «staccate» invece che a «giunsero». Per chiudere in bellezza, «come una […] compressibile» non è stata riconosciuta come «ccomp» (subordinata oggettiva) di «comprendere», bensì come «advcl» di «introdotta». Insomma, il povero *parser* ha scoperchiato *un vaso di Pandora*.

A parte le macroscopiche differenze strutturali, questo esempio mette in luce un problema comune ai testi di scienza e mai riscontrato nei gastronomici: l'individuazione della radice sbagliata. Se ne registrano almeno altri tre casi, nei confronti dei quali il *parser* ha generato dei *mostri*.

Altro aspetto emerso dall'esempio – questa volta trasversale ai due domini – riguarda le subordinate relative. Il *parser* riesce praticamente sempre a riconoscerle, ma spesso le collega all'antecedente sbagliato. Si mostra un esempio per ciascun dominio:

«I neutrini sono una componente dei raggi cosmici che **richiede** un discorso a sé»

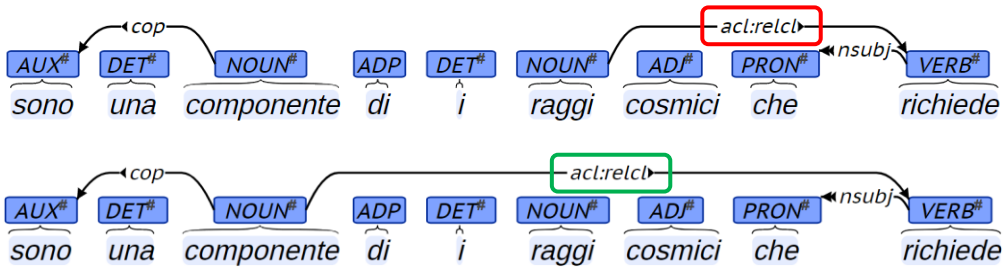


Figura 9. Subordinata relativa ricondotta a un antecedente sbagliato [*B*, 2].

«[…] unitevi un po' di verde di spinaci che le **comunicherà** un grazioso color verde pallido»

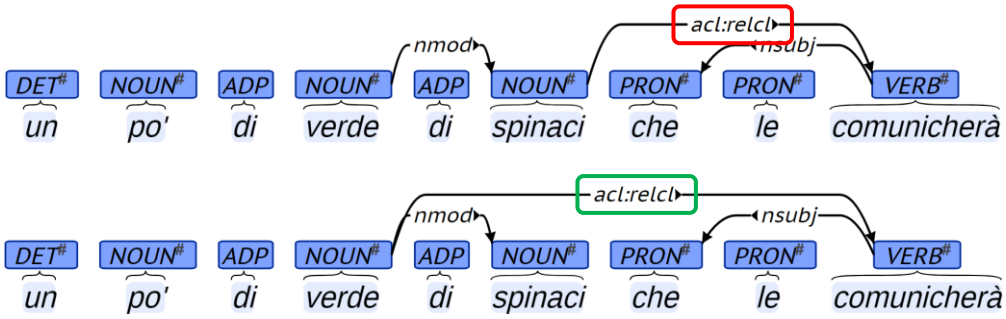


Figura 10. Subordinata relativa ricondotta a un antecedente sbagliato [*B*, 24].

Nel primo caso «richiede» era stato ricondotto a «raggi» anziché a «componente», così come nel secondo «comunicherà» a «spinaci» invece che a «verde».
Un discorso analogo vale anche per la relazione «amod», sempre riconosciuta ma spesso riferita al nome sbagliato. Se ne riporta un caso per tutti:

«La farina grossa detta «franta» esige una quantità d'acqua **maggiore** di quella in cui sia […]»

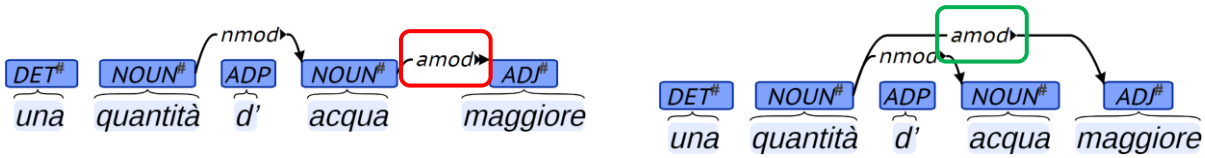


Figura 11. Assegnazione errata della testa della dipendenza «amod» [*L*, 7]

L'aggettivo «maggiore» è ricondotto ad «acqua» anziché a «quantità». Dagli esempi emergerebbe una tendenza a individuare l'antecedente nel nome più vicino. Può darsi che sia un comportamento sistematico del modello, ma andrebbe fatta un'analisi più approfondita per poterlo affermare.

Si segnala infine la difficoltà, comune ai due domini, nell'assegnazione delle sotto-tipologie della relazione «expl», specialmente nei casi di si passivante.

5.2 *Core arguments*

In quest'ultima sezione si espongono alcuni problemi abbastanza frequenti legati a quel gruppo di relazioni di dipendenza che l'iniziativa UD racchiude sotto l'etichetta di *core arguments.*

Il *parser* ha avuto difficoltà nel riconoscere le sequenze di complementi oggetto coordinati, presenti in numero discreto nei testi gastronomici. Si osservi il seguente periodo:

> «Preparate una polenta di un chilogr. di farina, e mentre la state dimenando, unitevi un **pezzo** di burro della grossezza d'un uovo circa, 100-150 **gr.** di formaggio grasso tagliato a dadolini, tre **manate** di fagioli sgranati e cotti bene nell'acqua salata, 3-6 **cucchiai** di formaggio grattato e 100-150 **gr.** di salame o di lucanica tagliati a dadolini.» [*L*, 16]

È la stessa frase che, in *sezione 5.1*, era stata presa come esempio della errata ricostruzione delle relazioni di coordinazione e subordinazione: lunga 75 *token*, rappresenta un'eccezione nel testo *L*, le cui frasi, ricordiamo, hanno lunghezza media di 18,45 *token*.
Dei cinque complementi oggetto (in grassetto), solo «pezzo» è riconosciuto; «gr.» è annotato come modificatore nominale di «uovo», «manate» come soggetto della radice «Preparate», «cucchiai» e «gr.» come «nmod» rispettivamente di «manate» e «formaggio». A discolpa del modello, va detto che il periodo mostrato è sicuramente un caso limite, in quanto i complementi oggetto sono lontani

tra loro e hanno un certo numero di dipendenti. Tuttavia, errori di questo tipo sono stati compiuti anche in contesti meno complessi.

Sappiamo che in italiano l'ordine canonico seguito dalle frasi è Soggetto Verbo Oggetto (SVO). Ma è possibile formulare anche strutture marcate, invertendo le posizioni dei due argomenti per scopi comunicativi particolari. Più volte, soprattutto nel corpus culinario, soggetti postverbali sono stati scambiati per oggetti (*figura 12*). Ciò è comprensibile, in quanto le predizioni si basano sulle evidenze statistiche ricavate dal *training corpus* ed effettivamente quella posizione è occupata dall'oggetto in un numero molto più alto di casi.

«Per tre litri d'acqua v'occorrerà un **chilogr.** abbondante di farina franta […]»

Figura 12. Soggetto in posizione postverbale scambiato per complemento oggetto [*L*, 8].

Lo stesso avviene, ad esempio, in «variano le **proporzioni** dei componenti» [*B*, 2].
Si sono riscontrate anche situazioni in cui l'oggetto in posizione canonica è scambiato per soggetto marcato. Alla base di tali predizioni, però, c'è sempre un motivo evidente. Si consideri il caso:

«Tagliatela poi a spicchiettini e date**le sapore** con burro, pepe e sale»

Figura 13. Complemento oggetto in posizione canonica scambiato per soggetto postverbale [*C*, 10].

Avendo già individuato l'oggetto in «le», il modello ha annotato «sapore» come soggetto di «date». Ciò è possibile anche perché il modo imperativo non prevede l'espressione esplicita del soggetto.
Esponiamo un altro problema legato alla posizione canonica dell'oggetto. Piccola premessa: quasi tutti i complementi indiretti in italiano sono retti da una preposizione, ma in alcuni di questi essa può essere omessa. Ebbene, il corpus di cucina presenta numerosi complementi di tempo continuato che il *parser* ha scambiato per complementi oggetto. Si veda il seguente esempio:

«Lasciate sobbollire mezz'**oretta** il composto […]»

Figura 14. Complemento di tempo continuato scambiato per complemento oggetto [*L*, 28].

Casi simili sono «Fate cuocere due o tre **minuti**» [*B*, 21] e «Fatelo bollire alcuni **istanti**» [*C*, 28]. Il motivo sembra essere banale: il *parser* vede una sequenza «determinante + nome» in posizione postverbale e annota quel nome con «obj».

A proposito di obliqui, un errore frequente – in particolare *BI* – riguarda il mancato riconoscimento dei complementi di agente: spesso sono annotati con «obl» anziché con «obl:agent», il sottotipo specificamente definito per l'analisi di questi complementi. Se ne riporta un esempio per tutti:

> «In ogni istante della nostra vita, ogni centimetro quadrato del nostro corpo è attraversato da 60 **miliardi** di neutrini […]»

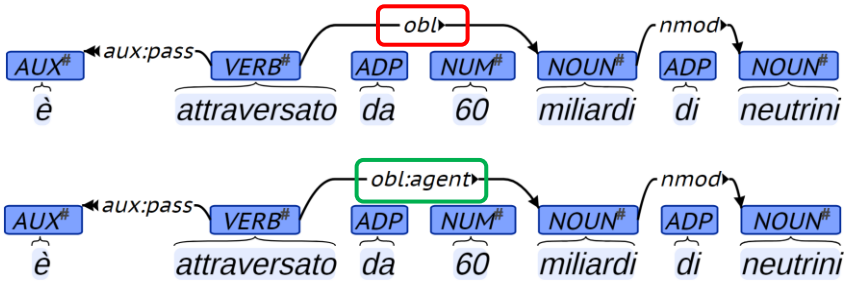


Figura 15. Complemento di agente non riconosciuto [*BI*, 4].

Per concludere, si segnala che i pronomi relativi «che» vengono sempre analizzati come soggetto della subordinata che introducono, anche quando ricoprono la funzione di oggetto. Essendo una tendenza sistematica del modello, si potrebbe pensare che ciò derivi dalle evidenze contenute nella *treebank* di addestramento, ma servirebbero studi più approfonditi per affermarlo.

6 Inter-Annotator Agreement

Le operazioni individuali di correzione dell'annotazione morfosintattica e sintattica hanno portato ciascun annotatore a costruire il proprio corpus in modo semiautomatico.
Terminate le revisioni, i membri del gruppo si sono riuniti per valutare il grado di accordo raggiunto dai loro risultati. È infatti buona prassi valutare l'affidabilità di un processo di annotazione manuale o semiautomatica con un punteggio di accordo complessivo.

Tracciando il panorama delle metriche di misurazione dell'I.A.A., Artstein (2017) espone dapprima l'*observed agreement* – letteralmente «accordo osservato» –, un indice rudimentale che esprime l'accordo in termini di proporzione dei casi rispetto a cui gli annotatori sono in accordo sull numero totale di quelli da annotare. Non tenendo in conto il fatto che parte dell'accordo può essere accidentale, il calcolo è di fatto poco attendibile.
Più affidabili risultano, invece, le metriche della famiglia *kappa/alpha*, che valutano la quantità di accordo attestata al netto di quella che potrebbe essere dovuta al caso. Uno di questi coefficienti – di riferimento per il progetto in oggetto – è il Kappa di Cohen, definito e descritto dallo statistico americano Jacob Cohen (Cohen, 1960). Matematicamente, è espresso come:

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

Dove, $P(A)$ è la proporzione del numero di volte in cui gli annotatori sono in accordo e $P(E)$ è la proporzione del numero di volte in cui si aspetta che essi lo siano per caso.
In quanto al risultato, Artstein e Poesio (2008) suggeriscono di interpretare un punteggio di almeno 0,8 come indicatore (1) di adeguatezza dello schema di annotazione rispetto al *task* specifico e (2) di buona comprensione di esso da parte degli annotatori coinvolti.

Dal punto di vista pratico, il calcolo è stato effettuato ricorrendo a uno script in Python fornito dalle docenti. Si tratta di un programma eseguibile da terminale, che prende in input due file contenenti testo annotato in formato CoNLL e calcola l'accordo tra gli annotatori sia in termini di *observed agreement* che di Kappa di Cohen. Specificando l'opzione «pos», l'accordo è misurato rispetto alle parti del discorso, mentre l'opzione «dep» effettua il calcolo a livello di relazioni di dipendenza. In più, aggiungendo l'opzione «-u» è possibile visualizzare la lista delle righe in cui si è attestato disaccordo. Questa funzionalità è stata sfruttata per lo studio delle tendenze di disaccordo che descriveremo più avanti.

Va specificato che *k,* differentemente da altri coefficienti, si limita a valutare l'accordo tra due sole annotazioni. Pertanto, essendo il nostro gruppo formato da tre studenti, si è optato per effettuare

misurazioni incrociate: per ognuna delle tre possibili coppie di annotatori – formate rispettivamente dagli studenti 1-2, 1-3 e 2-3 – si è calcolato l'accordo in relazione ai due livelli di descrizione linguistica, prima sul corpus globale e poi sui sotto-corpora rappresentativi dei due generi testuali considerati. In questo modo, è stato possibile valutare il grado di allineamento del lavoro di ogni annotatore con quelli degli altri due e osservare come questo vari a seconda del dominio.

Procedendo con ordine, commentiamo innanzitutto gli I.A.A. calcolati sui corpora generali per il livello morfosintattico. Si tenga presente che, dei due valori restituiti dallo script, in quanto segue si terrà in considerazione solo *k*, poiché, come affermato a inizio capitolo, esprime una metrica più affidabile rispetto al mero accordo osservato. Inoltre, si è scelto di mantenere tre cifre significative dopo la virgola per rendere le differenze tra valori leggermente più evidenti.
Le misurazioni circa le POS (*figura 16*) sono molto soddisfacenti: i *k* superano abbondantemente la soglia di 0,8, arrivando quasi a coincidere con il valore massimo. Va notato che esistono lievi differenze: l'accordo più alto è raggiunto dalla coppia 2-3 (*k* = 0,991, 27 *token* diversamente POS-taggati), quello meno alto dagli annotatori 1-3 (*k* = 0,979, 61 *token* diversamente POS-taggati); tuttavia, l'intervallo tra i due estremi è abbastanza ridotto (0,012). Con *k* = 0.988, infine, la coppia 1-2 occupa una posizione intermedia, avvicinandosi di più all'estremo più alto.

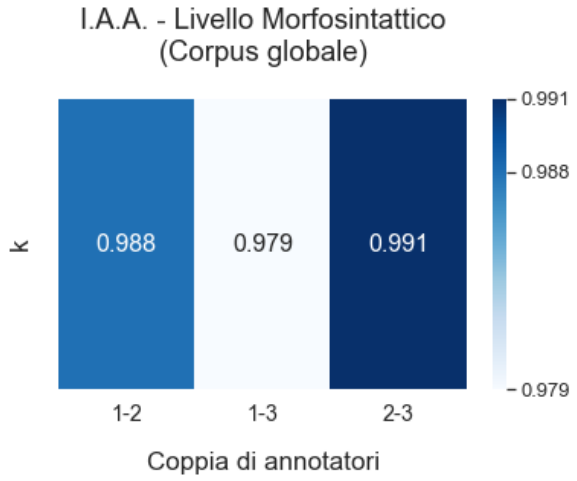


Figura 16. I.A.A (*k*) per coppie di annotatori rispetto alle POS del corpus globale.

È noto che i disaccordi tendono a distribuirsi in modo non casuale. In un'analisi empirica condotta su *dataset* POS-taggati, Plank et al. (2014) hanno dimostrato che spesso i disaccordi tra annotatori sono sistematici e non dipendono da errori casuali, bensì dai cosiddetti *hard cases*. Si tratta di casi difficili, rispetto a cui le teorie linguistiche non forniscono risposte chiarissime e, di conseguenza, i criteri di applicazione degli schemi di annotazione risultano più difficili da seguire.

Detto questo, ci soffermiamo sulla coppia 1-2 per provare a capire se i disaccordi rilevati siano effettivamente da imputare a degli *hard cases* oppure siano legati al *background* linguistico degli

annotatori o a eventuali disattenzioni in fase di revisione. Ci concentriamo su questa coppia perché, oltre a registrare il valore di accordo intermedio, riassume piuttosto bene le tendenze di disaccordo globalmente riscontrate tra le tre coppie.

Nei corpora degli annotatori 1 e 2, solo 34 *token* (dei 3118 totali) sono annotati con POS diverse. Nel complesso, i 34 disaccordi si sono generati in relazione a dieci coppie di POS. Escludendo quelle che occorrono una volta – e che, quindi, riguardano un solo *token* – otteniamo le sette coppie di *tag* diversi con cui sono stati annotati almeno due *token*. La *figura 17* mostra la distribuzione dei 31 *token* rispetto alle coppie di POS alternative con cui sono stati taggati:

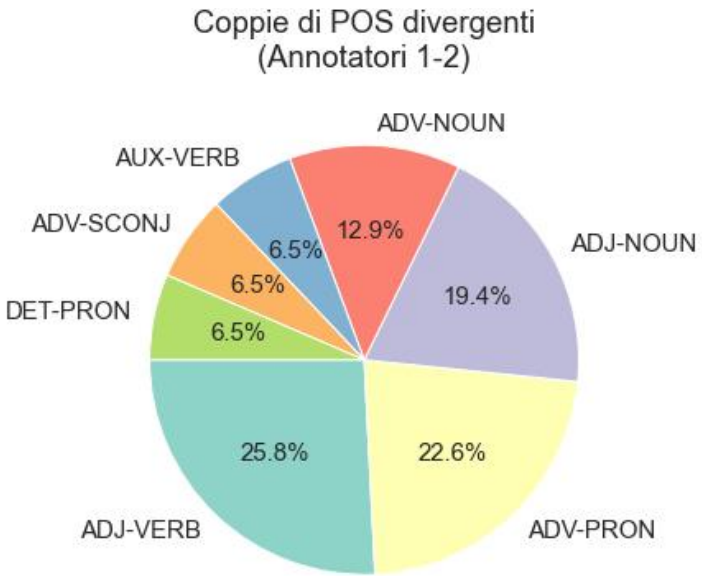


Figura 17. Coppie di POS su cui gli annotatori 1-2 sono risultati in disaccordo.

La fetta più ampia è etichettata con la coppia ADJ-VERB: vi rientrano otto *token* rispetto ai quali il disaccordo è correlato al dilemma ben noto – e affrontato in *sezione 4.1* – della distinzione tra aggettivo e participio passato. Alcuni esempi sono: «e quando sarà **sciolta** unite» [*B*, 20], «Quando la salsa sarà ben **montata** unitevi» [*B*, 24], «passateli nell'uovo **sbattuto**» [*C*, 21] e «parmigiano **grattato**» [*C*, 22]. La fetta successiva, poco più piccola, reca l'etichetta ADV-PRON e comprende sette token, tra cui «aggiungete**ci**» [*B*, 23], «unendo**vi**» [*C*, 27] e «aggiunger**vi**» [*L*, 17], che un annotatore ha considerato pronomi clitici (cioè, «aggiungete **a esso**»), l'altro avverbi di luogo (cioè, «aggiungete **qui**»). La porzione in rosso riguarda quattro occorrenze di «un poco» (e varianti), su cui si è generato disaccordo rispetto alla coppia di POS ADV-NOUN. Seguendo il suggerimento di un membro del gruppo e controllando sulla *treebank*, si è deciso, in vista della costruzione del *Gold standard*, di considerarlo *nome* quando è seguito da un complemento di specificazione – es. «un **po'** d'acqua», «un **po'** di salsa» [*L*, 27] –, *avverbio* altrimenti – es. «fate cuocere ancora un **pochino**» [*B*, 16], «bagnatele un **poco** nell'acqua» [*C*, 31]. Trattandosi di situazioni per cui più interpretazioni sono – più o meno – legittime, possiamo considerarli degli *hard cases*.

Concentriamoci adesso sulla fetta viola, terza in ordine di grandezza: ne fanno parte sei sostantivi che uno dei due annotatori ha scambiato per aggettivi (ad esempio «un po' di **verde** di spinaci» [*B*, 24] , «Per avere il **verde** di spinaci» [*B*, 25] e «col **composto** suddetto» [*L*,28]). Probabilmente si tratta di token che UDPipe ha erroneamente etichettato come aggettivi – perché, di fatto, possono esserlo – e che il revisore umano non ha corretto. Il disaccordo che ne consegue va dunque imputato a una svista dell'annotatore stesso. Lo stesso discorso vale per la fetta verde: un revisore non si è accorto che i pronomi «**la** correggerete» [*B*, 18] e «dar**gli** torto» [*BI*, 15] erano stati annotati come articoli definiti. Anche questi disaccordi, dunque, sono scaturiti dalla disattenzione umana.
Si osservi infine la fetta azzurra: i due token che ne fanno parte – «**Volendo** cuocere» [*C*, 17] e «**dev'**essere» [*L*, 10] – non sono stati identificati da uno studente come ausiliari. Sembrerebbe un errore riconducibile alla comprensione non ottimale dello schema di annotazione delle Universal Dependencies, che impone che i verbi modali (o servili) vengano annotati con il tag AUX.

Passiamo al livello sintattico. In *figura 18* sono mostrati i *k* calcolati sui corpora interi rispetto alle relazioni di dipendenza. Come prevedibile, i risultati sono più bassi, ma si confermano molto buoni e abbondantemente superiori a 0,8. Sono confermate le tendenze emerse al livello di POS-tagging: la coppia 2-3 raggiunge il grado di accordo più alto (*k* = 0,956, 147 *token* annotati diversamente) e la coppia 1-3 il meno alto (*k* = 0,895, 348 *token* annotati diversamente), anche se in questo caso la differenza tra i due estremi è più ampia (0,061); gli annotatori 1 e 2 sono di nuovo in una posizione intermedia con *k* = 0,939 e un totale di 201 *token* analizzati sintatticamente in modo differente.

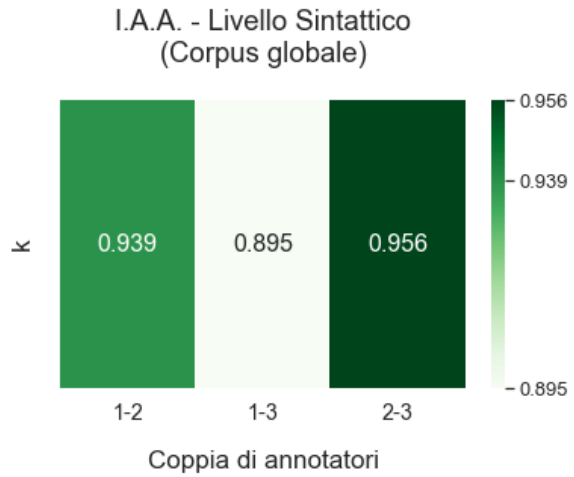


Figura 18. I.A.A (*k*) per coppie di annotatori rispetto alle dipendenze del corpus globale.

Data la maggiore complessità della sintassi rispetto alla morfosintassi e i numeri nettamente più alti di disaccordi, a questo livello si preferisce non aprire una digressione sulle divergenze – casuali o sistematiche – tra annotatori. Tuttavia, nel prossimo capitolo si riporteranno alcuni esempi di frasi che hanno richiesto particolare attenzione nella costruzione del *Gold standard*.

Vediamo adesso come variano gli I.A.A. tra i due domini testuali rappresentati nel corpus, partendo dal livello morfosintattico. Come si vede in *figure 19-20*, i *k*, calcolati per coppie di revisori, relativi al sotto-corpus di scienza sono sempre più alti di quelli ottenuti per i testi gastronomici. Le differenze vanno dal minimo di 0,003 al massimo di 0,015 rispettivamente delle coppie 2-3 e 1-3. Se gli accordi relativi alla cucina ripropongono un quadro simile a quello del corpus globale – con accordo massimo tra gli annotatori 2-3 e minimo tra 1-3 –, quelli scientifici tracciano una situazione diversa. Questa volta è la coppia 1-2 che raggiunge l'accordo massimo (*k* = 0,96), superando di 0,4 2-3, che a loro volta si collocano esattamente a metà. Di nuovo, il *k* minimo è tra i revisori 1-3.

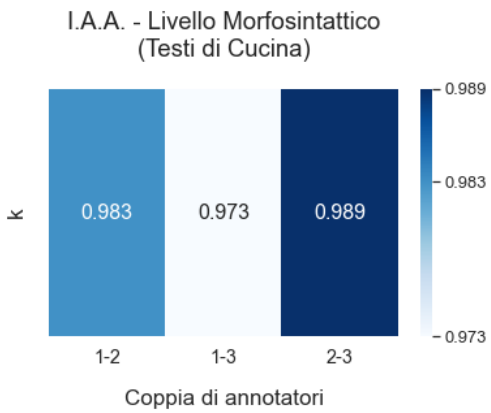


Figura 19. I.A.A (*k*) per coppie di annotatori rispetto alle POS del corpus di cucina.

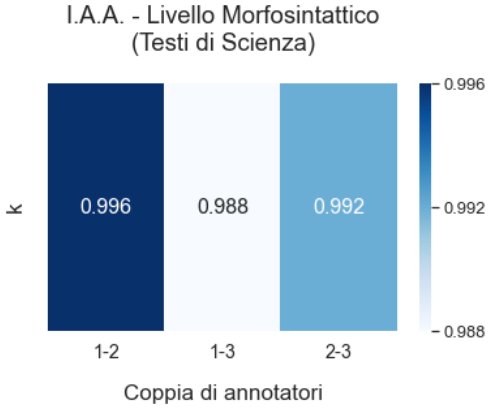


Figura 20. I.A.A (*k*) per coppie di annotatori rispetto alle POS del corpus di scienza.

Sul piano sintattico, riscontriamo differenze interessanti rispetto a quanto visto finora: se da un lato *k* è di nuovo maggiore nei testi scientifici per le coppie 1-2 e 1-3, i revisori 2-3 hanno raggiunto un accordo più alto (+0,006) nel sotto-corpus di cucina. In più, gli I.A.A. delle coppie 1-2 e 2-3 relativi ai testi scientifici sono identici (*k* = 0,952) e si distaccano significativamente (+0,048) dalla coppia 1-3, che di nuovo, e per entrambi i domini, raggiunge l'accordo minimo.

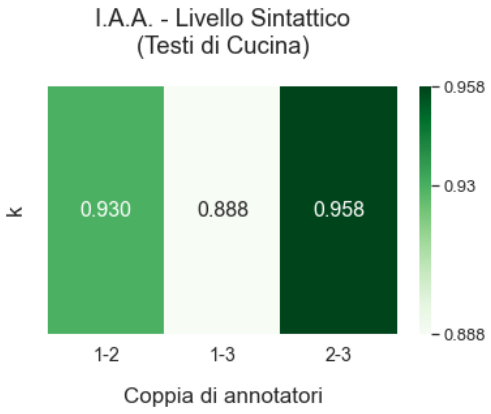


Figura 21. I.A.A (*k*) per coppie di annotatori rispetto alle dipendenze del corpus di cucina.

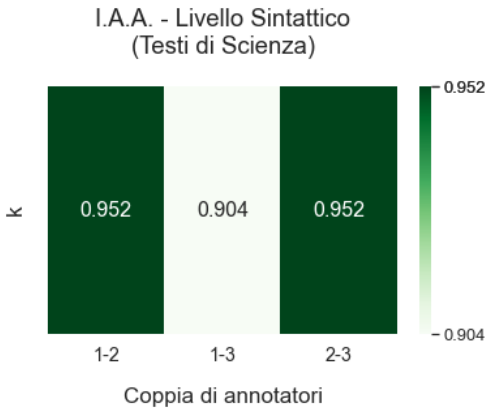


Figura 22. I.A.A (*k*) per coppie di annotatori rispetto alle dipendenze del corpus di scienza.

7 Costruzione del *Gold standard*

Gli alti gradi di accordo raggiunti hanno costituito un punto di partenza promettente per procedere alla creazione di un *Gold standard* unico, risultante dal confronto tra i corpora individuali e dalla scelta, frase per frase, delle soluzioni di annotazione ritenute più adeguate.
Come base è stato selezionato il corpus che, di comune accordo, è sembrato quello più corretto ed è stato confrontato direttamente con uno degli altri due usando il *plugin* «Compare» di Notepad++. Questo consente di affiancare due testi ed evidenzia in giallo le differenze. Il terzo revisore seguiva le operazioni scorrendo il proprio corpus e proponendo le proprie soluzioni quando necessario.
In questa fase è risultato fondamentale il confronto continuo con la *treebank* – attraverso il servizio *Grew-match* – per raggiungere un punto comune in casi ardui, o semplicemente per avere conferme. Non ci focalizziamo sul livello morfosintattico, perché (1) abbiamo già esposto una casistica dei disaccordi più comuni e (2) non ci si è imbattuti in situazioni troppo difficoltose. Ci concentriamo, invece, sul livello sintattico, riportando tre esempi di analisi significativamente divergenti.

Il primo caso che esaminiamo rientra nel discorso più ampio del trattamento degli elementi di una lista che segue i due punti. Questi, essendo casi specifici di un *token* prima del segno «:», devono dipendere sintatticamente da esso. *Ma qual è la relazione giusta per etichettare una dipendenza di questo tipo?* Questo aspetto è stato oggetto di un intenso confronto. Si osservi la seguente frase:

> «Nel 1930 si conoscevano soltanto due **particelle:** l'**elettrone** e il **protone**.» [*BI*, 10]

Un revisore ha trattato «elettrone» e «protone» come apposizioni di «particelle». La sua soluzione, pertanto, prevede una dipendenza di tipo «appos» tra «elettrone» e «particelle» e una di tipo «conj» tra «protone» ed «elettrone». Gli altri due, invece, hanno interpretato i nomi dopo i due punti come modificatori nominali di «particelle». Il fatto che le loro analisi sono identiche e che entrambi non hanno riconosciuto la costruzione con si passivante – con conseguente mancata annotazione di «si» come «expl:pass» e «particelle» come «nsubj:pass» – farebbe pensare che sia stata mantenuta l'analisi iniziale. Se così fosse, la scelta di «nmod» non deriverebbe da loro, bensì dal *parser*.
Si è dunque interrogato la *treebank*: per situazioni simili sono state trovate annotazioni eterogenee. Oltre ad «appos» e «nmod», ci sono casi di «parataxis» e addirittura «dislocated». Infine, abbiamo inserito nel *Gold* la prima soluzione esposta:

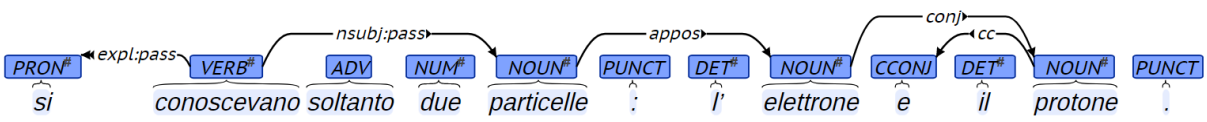


Figura 23. Rappresentazione grafica di quanto esposto circa la frase [*BI*, 10].

Si ponga ora l'attenzione sulla seguente frase:

> «Il tre per cento dell'energia del Sole, infatti, non viene emesso sotto forma di luce ma di neutrini, **e poiché questi elusivi mattoncini dell'universo attraversano indisturbati la Terra, il Sole ci illumina di neutrini giorno e notte**:» [*BI*, 8]

In grassetto è indicata la porzione su cui si sono generati i disaccordi: due revisori hanno annotato correttamente «il Sole ci illumina […] e notte» come coordinata della principale, facendo dipendere «illumina» da «emesso», e «poiché questi […] la Terra» come subordinata («advcl») – causale per l'esattezza – di tale coordinata. Il terzo annotazione, forse per disattenzione o per troppa fiducia nei confronti del *parser* automatico, ha invertito le dipendenze tra proposizioni: «poiché questi […] la Terra», nonostante la congiunzione subordinante «poiché», è stata considerata una coordinata della principale e «il Sole […] giorno e notte» inspiegabilmente una subordinata relativa dipendente da tale coordinata. In *figura 24* si mostra la rappresentazione grafica della soluzione inserita nel *Gold*:

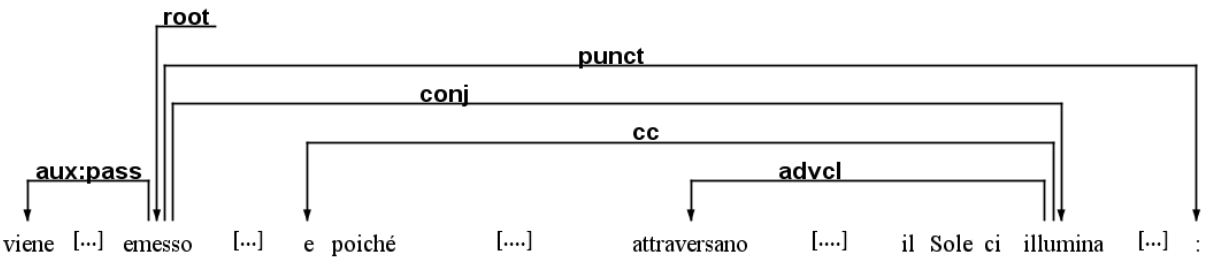


Figura 24. Rappresentazione grafica di quanto esposto circa il periodo [*BI*, 8].

Di per sé, questa frase non ha richiesto troppo sforzo per arrivare a una soluzione comune, poiché di fatto due annotatori su tre erano già d'accordo. Si è comunque ritenuto utile riportarla in questa sede per evidenziare che la tendenza del *parser* a invertire le relazioni tra proposizioni – di cui si è già parlato in *sezione 5.1* – è stata alla base di disaccordi tra annotatori.

Concludiamo commentando una delle frasi che in assoluto hanno richiesto più attenzione:

> «Se si tratta d'un chilogr. circa di farina, basterà che **dal momento ch'essa comincia a cuocere la rimestiate una ventina di minuti, per una polenta più grande v'occorrerebbe mezz'ora**.» [*L*, 11]

La radice è «basterà», «Se si tratta […] di farina» è una subordinata condizionale («advcl») che dipende da essa. E fino a qui i tre annotatori si sono trovati d'accordo.
Proseguiamo: «che la rimestiate una ventina di minuti» è la subordinata soggettiva della principale. Pertanto, «rimestiate» dipende da «basterà» tramite la relazione «csubj». Questo ha causato il primo problema, poiché due revisori l'hanno interpretata come subordinata oggettiva dipendente sempre da «basterà», annotandola con «ccomp».

Per ricostruire la corretta analisi di «dal momento ch'essa comincia a cuocere» è stata interrogata la *treebank* immettendo la sequenza di lemmi «da + il + momento + che». In base ai risultati, si è considerato «dal momento» un complemento obliquo («obl») di «rimestiate», «ch'essa comincia» una frase che lo modifica («acl») e «a cuocere» una subordinata oggettiva implicita dipendente da «comincia» tramite la relazione «xcomp». Si è deciso di usare «xcomp» e non «ccomp» – come invece il *parser* automatico aveva fatto e due revisori avevano mantenuto – perché, come indicato dalle istruzioni fornite delle *Universal Dependencies*[10], un uso di «xcomp» è consigliato quando la subordinata ha lo stesso soggetto della principale. In questo caso, è la «polenta» che comincia ed è la stessa che cuoce.

Terminiamo con la proposizione «per una polenta più grande v'occorrerebbe mezz'ora». Si è scelto di collegarne la testa «occorrerebbe» alla radice «basterà» tramite «parataxis» su suggerimento di un revisore. Gli altri due l'avevano annotata come «advcl» dipendente dalla radice. Tuttavia, non ci sono indizi di subordinazione, bensì sembra proprio una coordinazione non esplicita. Infine, un revisore, forse per distrazione, non aveva corretto l'annotazione automatica relativa a «v'» e «ora», che erano stati considerati rispettivamente un complemento oggetto (al posto di oggetto indiretto) e un complemento obliquo (al posto di soggetto posposto). La *figura 25* riassume graficamente le soluzioni riportare nel *Gold* rispetto a periodo analizzato.

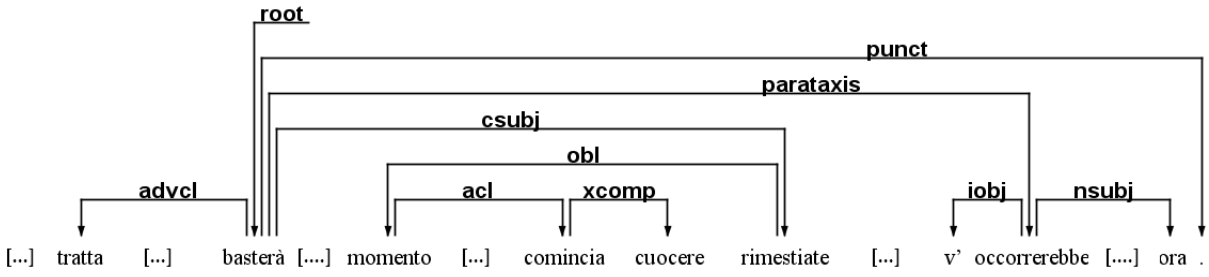


Figura 25. Rappresentazione grafica di quanto esposto circa il periodo [*L*, 11].

[10] Linee guida delle UD riguardo alla relazione «xcomp»: https://universaldependencies.org/it/dep/xcomp.html

8 Valutazione dell'accuratezza di analisi automatiche

L'ultima fase del progetto è consistita nell'uso del *Gold standard* per valutare l'accuratezza delle analisi automatiche eseguite da due modelli di *UDPipe* addestrati su varietà diverse dell'italiano: lo stesso *italian-isdt-ud-2.5-191206.udpipe*, con cui si erano eseguite le annotazioni da revisionare, e *talian-postwita-ud-2.5-191206.udpipe*[11], addestrato sulla raccolta di *tweet* PoSTWITA.

Il corpus revisionato congiuntamente dai membri del gruppo rispetto ai soli livelli di base, già punto di partenza per produrre le analisi morfosintattiche da correggere, è stato annotato con entrambi i modelli. In questo modo, si sono ottenuti due corpora annotati automaticamente da confrontare con le analisi riviste manualmente del *Gold*.
Dopodiché, si sono effettuate le valutazioni applicando lo script distribuito in occasione del *CoNLL 2018 Shared task: Multilingual Parsing from Raw Text to Universal Dependencies*[12]. Si tratta di una altro programma eseguibile da riga di comando, che prende in input due file – il *Gold* e il file annotato automaticamente dal modello che si intende valutare – e restituisce una tabella con i valori calcolati rispetto a tredici metriche relative ai diversi livelli della descrizione linguistica.
Per quanto concerne l'annotazione di base, vengono valutate le percentuali di *token* (*Token*), frasi (*Sentences*) e parole sintattiche (*Words*) del corpus *Gold* riconosciute dal modello; sul versante morfologico e morfosintattico, lo script stima l'accuratezza rispetto alle parti del discorso – *coarse-grained* (*UPOS*) e *fine-grained* (*XPOS*) –, ai tratti morfologici (*Ufeats*) sia singolarmente sia complessivamente (*AllTags*), e ai lemmi (*Lemmas*). Gli ultimi cinque parametri si riferiscono alla sintassi e sono: *UAS*, che valuta l'accuratezza nella ricostruzione delle relazioni di dipendenza tra *token*; *LAS*, che analizza anche le etichette associate a tali relazioni; *CLAS*, estensione di *LAS* che considera solo le dipendenze di tipo contenuto e che a sua volta ha due estensioni, *MLAS* e *BLEX*: oltre a dipendenze ed etichette, la prima tiene conto delle UPOS e dei tratti morfologici; la seconda incorpora nella valutazione anche la lemmatizzazione.

8.1 Modello *italian-isdt-ud-2.5-191206.udpipe*

In questa sezione esponiamo i risultati della valutazione dell'accuratezza del modello addestrato sulla *treebank* ISDT (*tabella 6*). In quanto ai primi tre parametri non c'è molto da dire. Si è già specificato che i file confrontati – il *Gold* e il corpus annotato automaticamente – partono da una base comune.

[11] I modelli sono stati scaricati da https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131
[12] Lo script è stato scaricato da https://universaldependencies.org/conll18/evaluation.html

Sul versante morfosintattico i risultati sono molto buoni: l'accuratezza rispetto alle quattro metriche riferite a singoli aspetti dell'analisi – *UPOS*, *XPOS*, *UFeats* e *Lemmas* – oscilla tra il 93,39% di *UFeats* e il 94,93% di *UPOS* e supera il 90% anche in relazione al parametro complessivo *AllTags*. Tuttavia, emergono due tendenze, piuttosto prevedibili in realtà: l'accuratezza del modello tende a calare (1) man mano che si entra in livelli più profondi e granulari e (2) rispetto a quelle metriche che riuniscono insieme più componenti dell'annotazione – come appena visto per *AllTags*.
Tali tendenze diventano molto evidenti a livello sintattico: il valore più alto, 84%, è minore del 10% circa rispetto a quelli morfosintattici e si riferisce – non a caso – alla metrica più semplice, *UAS*. Si assiste poi a un decremento continuo e vistoso che culmina nel poco più del 65% di accuratezza rispetto alle due metriche più complesse, *MLAS* e *BLEX*. Evidenziamo, infine, i circa otto punti percentuali di differenza tra le metriche *LAS* e *CLAS*: dal momento che i legami tra una parola funzionale e la sua testa sono più facili da ricostruire di quelli tra due parole contenuto, il calo notevole nelle prestazioni del modello non stupisce.

Metrica	**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	94,93	94,93	94,93	94,93
XPOS	94,42	94,42	94,42	94,42
UFeats	93,39	93,39	93,39	93,39
AllTags	91,76	91,76	91,76	91,76
Lemmas	94,03	94,03	94,03	94,03
UAS	84,00	84,00	84,00	84,00
LAS	79,38	79,38	79,38	79,38
CLAS	71,69	70,99	71,34	70,99
MLAS	66,52	66,87	66,20	66,87
BLEX	65,91	65,27	65,58	65,27

Tabella 6. Valutazione del modello addestrato su ISDT rispetto al corpus globale.

Finora si è parlato del corpus globale. *Ma come cambia la qualità delle prestazioni del modello tra i due domini considerati?* Per rispondere alla domanda sono stati valutati separatamente prima i soli testi di cucina e poi quelli scientifici. Dall'osservazione dei risultati (*tabelle 7 e 8*) risultano confermate le tendenze emerse in precedenza. Ma ciò che salta agli occhi è la maggiore accuratezza del modello nell'analisi dei testi scientifici rispetto a tutti i parametri considerati con differenze che vanno da un minimo di 4,13 (*UPOS*) a un massimo di 13,72 punti percentuali (*BLEX*). Volendo generalizzare, si può affermare che il divario cresce all'aumentare di granularità e profondità di analisi, in particolare nel passaggio dalle metriche morfosintattiche a quello sintattiche.

Metrica	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	93,29	93,29	93,29	93,29
XPOS	92,70	92,70	92,70	92,70
UFeats	90,89	90,89	90,89	90,89
AllTags	89,03	89,03	89,03	89,03
Lemmas	91,80	91,80	91,80	91,80
UAS	82,10	82,10	82,10	82,10
LAS	76,24	76,24	76,24	76,24
CLAS	68,38	67,25	67,81	67,25
MLAS	62,25	61,22	61,73	61,22
BLEX	60,97	59,96	60,46	59,96

Tabella 7. Valutazione del modello addestrato su ISDT rispetto al dominio gastronomico.

Metrica	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	97,42	97,42	97,42	97,42
XPOS	97,02	97,02	97,02	97,02
UFeats	97,26	97,26	97,26	97,26
AllTags	95,97	95,97	95,97	95,97
Lemmas	97,42	97,42	97,42	97,42
UAS	86,87	86,87	86,87	86,87
LAS	84,13	84,13	84,13	84,13
CLAS	77,16	77,41	77,29	77,41
MLAS	73,74	73,98	73,86	73,98
BLEX	74,06	74,30	74,18	74,30

Tabella 8. Valutazione del modello addestrato su ISDT rispetto al dominio scientifico.

Si è visto a lezione che l'annotazione morfosintattica automatica tipicamente ha un'accuratezza del 96-97%. Nell'ambito della campagna di valutazione *CoNLL 2018 Shared Task* è emerso che le accuratezze di sistemi diversi addestrati su ISDT oscillano tra il 96,31% e il 98,13% per le *UPOS* e tra 95,89% e 97,99% per le *XPOS*. Al livello sintattico, la campagna ha mostrato un'accuratezza media – rispetto alla metrica *LAS* – dell'87,61%, con una deviazione standard di ±4,12.
Osservando i nostri risultati, possiamo dire che, per il dominio scientifico, essi sono perfettamente in linea con lo stato dell'arte; per quello gastronomico, invece, siamo nettamente al di sotto della media, specialmente a livello di *LAS* (–11,37%). Ciò è, almeno in parte, imputabile alla dimensione diacronica e rivelerebbe quanto ci sia ancora da fare in materia di adattamento degli strumenti di annotazione a nuovi domini e varietà, che, non a caso, è uno degli obiettivi del progetto TrAVaSI.

8.2 Modello *talian-postwita-ud-2.5-191206.udpipe*

In ultima battuta, vediamo i risultati della valutazione della correttezza delle annotazioni eseguite dal modello addestrato su PoSTWITA. I valori in *tabella 9* rivelano accuratezze sensibilmente inferiori di quelle raggiunte dall'altro modello per tutte le metriche. Tra i parametri morfosintattici, l'unico che supera di pochissimo il 90% di accuratezza è *UPOS*; gli altri vanno dall'85,34% di *AllTags* all'89,35% di *XPOS*. Dal punto di vista sintattico, come ci si poteva aspettare, la situazione peggiora: *UAS* non arriva, seppur di poco, all'80% di accuratezza, *LAS* supera appena la soglia del 70% e le metriche più complesse, *MLAS* e *BLEX*, non raggiungono neppure il 55%. La scarsa capacità analitica e predittiva del modello è perfettamente comprensibile, dal momento che è stato

addestrato su un genere testuale – il *tweet* – tanto lontano per vocabolario e strutture a una ricetta culinaria o una dissertazione scientifica.

Al di là dei confronti tra i due modelli, i risultati della valutazione sembrano seguire le tendenze evidenziate nella sezione precedente. Ciò vale anche in relazione alle differenze tra generi testuali (*tabelle 10 e 11*): di nuovo le prestazioni relative ai testi scientifici sono migliori di quelle relative al corpus di cucina. Tuttavia, i divari sono meno marcati, andando da un minimo di 0,38 (*UAS*) a un massimo di 5,48 punti percentuali (*MLAS*).

Metrica	**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	90,54	90,54	90,54	90,54
XPOS	89,35	89,35	89,35	89,35
UFeats	88,49	88,49	88,49	88,49
AllTags	85,34	85,34	85,34	85,34
Lemmas	89,19	89,19	89,19	89,19
UAS	78,32	78,32	78,32	78,32
LAS	71,26	71,26	71,26	71,26
CLAS	61,68	62,28	61,98	62,28
MLAS	53,59	54,11	53,85	54,11
BLEX	52,99	53,50	53,24	53,50

Tabella 9. Valutazione del modello addestrato su PoSTWITA rispetto al corpus globale.

Metrica	**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	89,45	89,45	89,45	89,45
XPOS	88,33	88,33	88,33	88,33
UFeats	86,36	86,36	86,36	86,36
AllTags	83,32	83,32	83,32	83,32
Lemmas	88,39	88,39	88,39	88,39
UAS	78,05	78,05	78,05	78,05
LAS	69,95	69,95	69,95	69,95
CLAS	60,49	60,54	60,51	60,54
MLAS	51,84	51,90	51,87	51,90
BLEX	51,65	51,70	51,68	51,70

Tabella 10. Valutazione del modello addestrato su PoSTWITA rispetto al dominio gastronomico.

Metrica	**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	92,18	92,18	92,18	92,18
XPOS	90,89	90,89	90,89	90,89
UFeats	91,78	91,78	91,78	91,78
AllTags	88,48	88,48	88,48	88,48
Lemmas	90,41	90,41	90,41	90,41
UAS	78,73	78,73	78,73	78,73
LAS	73,25	73,25	73,25	73,25
CLAS	63,64	63,30	64,46	65,30
MLAS	56,62	58,10	57,35	58,10
BLEX	55,18	56,63	55,90	56,63

Tabella 11. Valutazione del modello addestrato su PoSTWITA rispetto al dominio scientifico.

Bibliografia

Artstein, Ron. 2017. «Inter-Annotator Agreement». In *Handbook of Linguistic Annotation*, a cura di Nancy Ide e James Pustejovsky, 297–313. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2_11.

Artstein, Ron, e Massimo Poesio. 2008. «Survey Article: Inter-Coder Agreement for Computational Linguistics». *Computational Linguistics* 34 (4): 555–96. https://doi.org/10.1162/coli.07-034-R2.

Cohen, Jacob. 1960. «A Coefficient of Agreement for Nominal Scales». *Educational and Psychological Measurement* 20 (1): 37–46. https://doi.org/10.1177/001316446002000104.

Dell'Orletta, Felice, Simonetta Montemagni, e Giulia Venturi. 2011. «READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification». In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, 73–83. Edinburgh, Scotland, UK: Association for Computational Linguistics. https://aclanthology.org/W11-2308.

Favaro, Manuel, Marco Biffi, e Simonetta Montemagni. 2021. «Risorse linguistiche di varietà storiche di italiano: il progetto TrAVaSI». In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021*, a cura di Johanna Monti, Fabio Tamburini, e Felice Dell'Orletta, 178–86. Collana dell'Associazione Italiana di Linguistica Computazionale. Torino: Accademia University Press. http://books.openedition.org/aaccademia/8515.

Lucisano, Pietro, e Maria Emanuela Piemontese. 1988. «Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana». Scuola e Città 3: 57–68.

Marazzini, Claudio, e Ludovica Maconi. 2018. «Il Vocabolario dinamico dell'italiano moderno rispetto ai linguaggi settoriali. Proposta di voce lessicografica per il redigendo VoDIM». *Italiano digitale* 7 (4): 100.

Plank, Barbara, Dirk Hovy, e Anders Søgaard. 2014. «Linguistically debatable or just plain wrong?» In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 507–11. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2083.