DOI: 10.17469/O2106SLI000016

RACHELE SPRUGNOLI, MATTEO PELLEGRINI,
MARCO PASSAROTTI, FLAVIO M. CECCHINI

EvaLatin 1.0: un Corpus per la Valutazione delle Tecnologie del Linguaggio Applicate al Latino

Questo articolo presenta il corpus EvaLatin 1.0, sviluppato per la prima campagna di valutazione di strumenti di Trattamento Automatico del Linguaggio per il latino. La campagna si è concentrata su due analisi linguistiche, ovvero la lemmatizzazione e l'annotazione delle parti del discorso. Particolare attenzione è stata rivolta alla costruzione del corpus in modo da affrontare problematiche di variabilità di genere e diacronica del latino.[1]

Parole chiave: latino, risorse linguistiche, corpus, annotazione linguistica, trattamento automatico del linguaggio.

1. *Introduzione*

Negli ultimi anni, in seguito alla crescente disponibilità di testi in formato digitale per le lingue antiche, e soprattutto per il greco antico e il latino, si sta assistendo anche a un incremento delle risorse linguistiche e degli strumenti di Trattamento Automatico del Linguaggio (TAL) ad esse relativi (ad es. Bouma & Adesam 2017; Sprugnoli & Passarotti 2020). Dato che l'affidabilità dei risultati delle ricerche condotte con l'ausilio di tali risorse e strumenti dipende in maniera cruciale dalla loro qualità, emerge l'esigenza di una valutazione sistematica.

È a questa necessità che si propone di far fronte EvaLatin, la prima campagna di valutazione di strumenti di TAL interamente dedicata alla lingua latina, che si inserisce in un'ampia tradizione di eventi di

[1] La responsabilità principale delle Sezioni dell'articolo va attribuita come segue. Rachele Sprugnoli: §2., §3., §3.1; Flavio M. Cecchini: §3.2; Marco Passarotti: §5. Le Sezioni §1 e §4 sono da ascrivere a tutti gli autori.

valutazione tramite *shared task*; si veda ad es. SemEval (https://alt.qcri.org/semeval2018/), CoNLL (https://www.conll.org/) e – specificamente sull'italiano – EVALITA (http://www.evalita.it/). In uno *shared task*, strumenti diversi di TAL devono risolvere uno specifico compito, come la lemmatizzazione, utilizzando dati comuni a tutti i partecipanti sia in fase di addestramento che di valutazione.

Per rendere possibile una campagna di valutazione, è necessario innanzitutto mettere a disposizione dei partecipanti i dati annotati su cui addestrare i propri strumenti di TAL (*training set*), le cui prestazioni verranno poi valutate su dati diversi (*test set*). La risorsa linguistica presentata in questo articolo, denominata EvaLatin 1.0, contiene sia i testi in prosa classica del *training set*, sia quelli del *test set* – relativi in parte agli stessi autori classici del *training set*, in parte a testi di epoca medievale e di poesia classica.

L'articolo è strutturato come segue: la Sezione 2 descriverà la prima edizione della campagna EvaLatin, mentre la Sezione 3 fornirà dettagli circa i dati soffermandosi sulla loro composizione e annotazione linguistica. I risultati della campagna di valutazione saranno presentati nella Sezione 4 insieme a un'analisi di alcune importanti caratteristiche dei dati. La Sezione 5 raccoglierà una discussione finale e presenterà un breve sguardo sulla successiva edizione di EvaLatin.

2. *EvaLatin*

I risultati della prima edizione sono stati presentati al "1st Workshop on Language Technologies for Historical and Ancient LAnguages" (LT4HALA 2020: https://circse.github.io/LT4HALA/2020/), nell'ambito della conferenza internazionale "Language Resources and Evaluation".

Nell'organizzare una campagna di valutazione di strumenti di TAL per il latino, va tenuto conto del fatto che sotto la comune etichetta di "latino" ricadono testi relativi a epoche e generi letterari diversi, risultando quindi in una notevole variazione diacronica e stilistica. Ciò tende ad impattare negativamente sui risultati dell'applicazione di un modello addestrato a testi relativi ad un'altra epoca o ad un altro genere rispetto a quelli del *training set* (Ponti & Passarotti 2016). Per poter valutare l'impatto di questo problema, i due task proposti nella prima edizione di EvaLatin – lemmatizzazione e Part-of-Speech

(PoS) tagging, cioè riconoscimento delle parti del discorso – sono divisi ciascuno in tre sotto-task, denominati rispettivamente "Classical" (i dati del *test set* sono della stessa epoca e dello stesso genere letterario di quelli del *training set*), "Cross-genre" (i dati del *test set* sono di un altro genere letterario) e "Cross-time" (i dati del *test set* sono relativi a un'altra epoca).

I testi di epoca classica sono stati tratti dalla Perseus Digital Library (Smith *et al.* 2000). L'annotazione dei lemmi e delle parti del discorso è stata ottenuta applicando ai testi alcuni modelli automatici addestrati usando UDPipe (Straka & Straková 2017) sui dati annotati manualmente del corpus sviluppato presso il centro di ricerca LASLA (Verkerk *et al.* 2020); il risultato di questa analisi automatica è stato quindi controllato e corretto manualmente da due annotatori, con eventuali dubbi risolti da un terzo annotatore. I testi di epoca medievale sono stati invece annotati manualmente nell'ambito del progetto Index Thomisticus Treebank (Passarotti 2019).

3. *Dati*

I testi forniti come dati di addestramento sono di cinque autori classici: Cesare, Cicerone, Seneca, Plinio il Giovane e Tacito. Per ogni autore abbiamo selezionato circa 50.000 token annotati, per un totale di quasi 260.000 token. Ogni autore è rappresentato da testi in prosa: trattati nel caso di Cesare, Seneca e Tacito, discorsi pubblici per Cicerone e lettere per Plinio il Giovane. La Tabella 1 riassume la composizione dei dati di addestramento.

Tabella 1 - *Composizione dei dati di addestramento*

Autori	*Testi*	*# token*
Cesare	De Bello Gallico	44.818
Cesare	De Bello Civili (libro II)	6.389
Cicerone	Philippicae (libri I-XIV)	52.563
Seneca	De Beneficiis	45.457
Seneca	De Clementia	8.172
Plinio il Giovane	Epistulae (libri I-VIII)	50.827
Tacito	Historiae	51.420
TOTALE TOKEN		259.646

Per quanto riguarda i dati di test, nel sotto-task Classical abbiamo distribuito testi in prosa degli stessi autori presenti nei dati di addestramento, selezionando circa 11.000 token per ciascuno. Per il sotto-task Cross-genre, invece, abbiamo usato i Carmina di Orazio, mentre per quello Cross-time, una parte del libro IV della Summa Contra Gentiles di Tommaso d'Aquino. In altre parole, abbiamo un testo in poesia e uno di epoca medievale. La composizione dei dati del sotto-task Classical è presentata nella Tabella 2, mentre dettagli sui dati relativi agli altri due sotto-task sono riportati nelle Tabelle 3 e 4, rispettivamente.

Tabella 2 - *Composizione dei dati di test: sotto-task Classical*

Autori	*Testi*	*# token*
Cesare	De Bello Civili (libro I)	10.898
Cicerone	In Catilinam	12.564
Seneca	De Vita Beata	7.270
Seneca	De Providentia	4.077
Plinio il Giovane	Epistulae	9.868
Tacito	Agricola	6.737
Tacito	Germania	5.513
TOTALE TOKEN		56.927

Tabella 3 - *Composizione dei dati di test: sotto-task Cross-genre*

Autore	*Testo*	*# token*
Orazio	Carmina	13.290

Tabella 4 - *Composizione dei dati di test: sotto-task Cross-time*

Autore	*Testo*	*# token*
Tommaso d'Aquino	Summa Contra Gentiles (parte del IV libro)	11.556

Il corpus è rilasciato nel formato standard CoNLL-U (https://universaldependencies.org/docs/format.html), adottato nel progetto Universal Dependencies (Nivre *et al.* 2016; de Marneffe *et al.* 2021). Secondo il formato CoNLL-U, ogni testo corrisponde ad un file in cui ogni frase è rappresentata da una struttura a 10 colonne separate da tabulazioni. Una riga vuota segna la divisione tra una frase e l'altra. Nei dati di EvaLatin 2020, solo le prime 4 colonne (di

seguito dettagliate) presentano del contenuto, mentre le altre sono riempite da un trattino basso:

1. identificatore numerico per ogni token, ovvero un numero intero che riparte da 1 ad ogni frase;
2. forma della parola così come appare nel testo;
3. lemma, ovvero forma di citazione;
4. etichetta della parte del discorso "universale" (Universal Part Of Speech, o UPOS; vedi Sezione 3.2).

Inoltre, ogni frase è preceduta da due linee di commento che iniziano con un carattere di cancelletto: una linea riporta il numero identificativo della frase, l'altra il testo. Un esempio è fornito nella Figura 1.

Figura 1 - *Il formato dei dati: una frase di esempio*

```
# sent_id = 306
# text = Debere se suspicari simulata Caesarem amicitia quod
exercitum in Gallia habeat sui opprimendi causa habere
1   Debere      debeo     VERB  _  _  _  _  _  _
2   se          sui       PRON  _  _  _  _  _  _
3   suspicari   suspicor  VERB  _  _  _  _  _  _
4   simulata    simulo    VERB  _  _  _  _  _  _
5   Caesarem    Caesar    PROPN _  _  _  _  _  _
6   amicitia    amicitia  NOUN  _  _  _  _  _  _
7   quod        quod      SCONJ _  _  _  _  _  _
8   exercitum   exercitus NOUN  _  _  _  _  _  _
9   in          in        ADP   _  _  _  _  _  _
10  Gallia      Gallia    PROPN _  _  _  _  _  _
11  habeat      habeo     VERB  _  _  _  _  _  _
12  sui         sui       PRON  _  _  _  _  _  _
13  opprimendi  opprimo   VERB  _  _  _  _  _  _
14  causa       causa     NOUN  _  _  _  _  _  _
15  habere      habeo     VERB  _  _  _  _  _  _
```

3.1 Dettagli sulla lemmatizzazione

La lemmatizzazione è il processo di riconduzione di ogni forma di parola alla sua forma di citazione corrispondente all'entrata del dizionario (cioè al lemma). Le convenzioni che abbiamo seguito sono riassunte di seguito.

– I verbi sono lemmatizzati sotto la prima persona singolare del presente indicativo attivo (o passivo, nel caso dei deponenti): es. PRS.ACT.INF *accingere* 'cingere' → *accingo*; FUT.PASS.IND.1.SG/PRS.PASS.SBJV.1.SG *sequar* 'seguirò/(che io) segua' → *sequor*.
– Le abbreviazioni sono espanse: es. *L.* → *Lucius*; *s.* → *salus*.
– Il lemma dei numeri romani (es. *ccc*, *CCCXVIII*) è *numerus_romanus*. Il lemma dei numeri arabi (es. *12*, *53*) è *num_arab*. Il lemma delle parole greche (es. *εἰσήλασαν*) è *uox_graeca*.

– Le espressioni multi-parola non sono combinate in un singolo token ma ogni loro parte è analizzata separatamente: *res publica* 'repubblica/stato' è formata da due token con due lemmi e due categorie grammaticali.
– I clitici non sono separati dal token: *exercitumque* ('esercito.ACC.SG=e') ha come lemma *exercitus* e il clitico *-que* non viene analizzato separatamente.

3.2 Dettagli sull'annotazione delle parti del discorso

Nel corpus, ogni token è annotato con la propria parte del discorso. Le etichette adottate per l'annotazione sono quelle dello schema di annotazione di UD, per cui si rimanda alle linee guida del progetto.[2] Si noti che sono usate tutte le etichette a eccezione di PUNCT e SYM, indicanti rispettivamente punteggiatura e simboli, che non sono presenti nel corpus.

Qui ci limitiamo a segnalare la caratteristica di UD per cui le categorie si presentano "in coppia", distinguendo classi funzionali e lessicali: per esempio, alla classe lessicale ADJ degli aggettivi (es. *agrestis* 'agreste') corrisponde quella funzionale DET dei determinanti (es. *hic* 'questo'), che a sua volta contempla la sottoclasse dei numerali NUM (es. *mille* 'mille'). Si hanno così NOUN/PROPN (es. *mater* 'madre', *Hercules* 'Ercole') rispetto a PRON (es. *ego* 'io'), e VERB (es. *rapiebat* 'rapiva') rispetto a AUX (che include anche la copula: nel corpus solo SUM 'essere' e, solo in alcune particolari costruzioni, EO 'andare', come in *datum iri … facultatem* 'si sarebbe data la possibilità', da Cesare, Bellum Civile, fr. 448). L'unica classe lessicale invariabile in latino è quella ADV degli avverbi (es. *certe* 'certamente'); fra le restanti, tutte funzionali, notiamo la differenza fra congiunzioni coordinanti CCONJ (es. *et* 'e') e subordinanti SCONJ (es. *quod* 'che'), e che le negazioni *non*, *ne* e *haud* ricevono la classe PART (e quindi non ADV). Infine, X è una classe residuale usata per quei token a cui per vari motivi (parole straniere, passaggi lacunosi, …) non è possibile assegnare un'analisi nel sistema latino.

Segnaliamo inoltre che l'annotazione nel nostro corpus differisce in modo rilevante dagli standard di UD riguardo al verbo *sum*: anziché essere annotato uniformemente come AUX anche laddove si com-

[2] https://universaldependencies.org/u/pos/index.html.

porta come verbo copula, questa etichetta è usata solamente quando *sum* compare come ausiliare in tempi perifrastici o composti (es. ***sunt*** *demonstratae* 'sono state dimostrate'), mentre in tutti gli altri casi (es. *paratiores* ***essent*** 'fossero più preparati', *mihi animi* ***sit*** '(io) abbia l'intenzione' lett. 'mi sia d'animo di') viene analizzato come VERB.

4. *Risultati e analisi*

EvaLatin 2020 ha visto la partecipazione di 5 gruppi, di cui 3 hanno preso parte sia alla lemmatizzazione che al riconoscimento delle parti del discorso mentre 2 al solo riconoscimento delle parti del discorso. Tutti i gruppi partecipanti erano affiliati a istituzioni di ricerca non italiane (canadesi, ceche, tedesche e statunitensi). La Tabella 5 mostra l'accuratezza raggiunta dal sistema risultato migliore, ovvero una versione appositamente creata per EvaLatin di UDPipe, strumento di analisi linguistica automatica sviluppato dall'Università Carolina di Praga (Straka & Straková 2020).[3]

In generale, i testi più semplici da elaborare per tutti i sistemi automatici partecipanti sono stati "In Catilinam" di Cicerone e "De Bello Civili" di Cesare. Al contrario, i testi che hanno registrato più errori sono "Germania" di Tacito e "De Vita Beata" di Seneca. Come previsto, tutti i sistemi, compreso UDPipe, subiscono un calo nelle prestazioni quando applicati a un genere o a un periodo temporale diverso da quello dei dati di addestramento: tale calo può arrivare anche a 10 punti percentuali. In particolare, si nota una migliore accuratezza sui testi medievali di Tommaso d'Aquino che sulla poesia classica per quanto riguarda la lemmatizzazione, mentre il contrario avviene per il riconoscimento delle parti del discorso: si ha un'accuratezza maggiore sulla poesia di Orazio che sui testi di Tommaso d'Aquino.

Tabella 5 - *Accuratezza del sistema con le migliori prestazioni*

	Classical	*Cross-genre*	*Cross-time*
Lemma	96,19 %	87,13 %	91,01 %
PoS	96,74 %	91,11 %	87,69 %

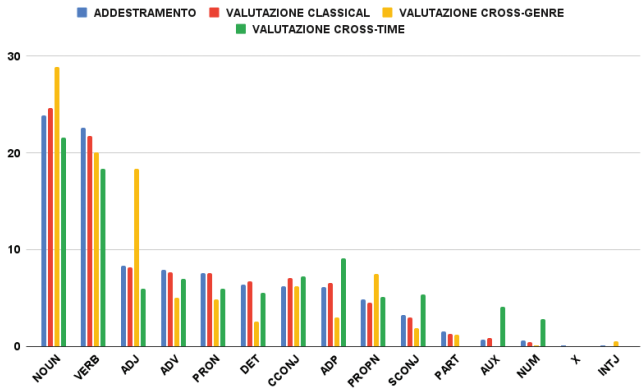
[3] L'accuratezza è il rapporto tra predizioni corrette del sistema e predizioni.

Le ragioni alla base di queste differenze possono essere ricercate in alcune caratteristiche linguistiche dei testi presi in esame. Nello specifico, per quanto riguarda la lemmatizzazione, Tommaso d'Aquino presenta un vocabolario meno ricco e vario rispetto a Orazio: il rapporto tra lemmi e token in Orazio è 0,26, mentre in Tommaso d'Aquino e nei dati di valutazione Classical lo stesso rapporto è di 0,09. Un'altra difficoltà dei Carmina risiede nel fatto che contengono un maggior numero di lemmi non presenti nei dati di addestramento (i cosiddetti lemmi *out-of-vocabulary*). Tali lemmi (ad es. *arcus* 'arco') coprono il 29% dei lemmi totali nel testo di Orazio: i lemmi *out-of-vocabulary* di Tommaso d'Aquino sono, invece, il 26% (ad es. *Christus* 'Cristo') e il 14% nei dati di valutazione del sotto-task Classical (ad es. *Archippus* 'Archippo').

Anche dal punto di vista delle parti del discorso, notiamo una distribuzione differente delle etichette nei dati, come rappresentato graficamente nella Figura 2. Se la distribuzione percentuale non varia sensibilmente tra dati di addestramento (in blu) e dati di valutazione Classical (in rosso), lo stesso non si può dire per i dati di valutazione Cross-genre (in giallo) e Cross-time (in verde). Nel testo di Orazio, infatti, abbiamo meno verbi (-3%) ma più nomi comuni (+5%), nomi propri (+3%) e aggettivi (+10%). L'alto numero di nomi propri in Orazio è stato osservato in vari studi e in letteratura non mancano indagini di natura etimologico-onomastica (Bo 1967; Roncali 2013; Paradisi 2019). Per quanto riguarda gli aggettivi, un'analisi quantitativa nei testi poetici presenti nel corpus LASLA ha rivelato che la percentuale di aggettivi rispetto al numero totale di parole è maggiore in poesia rispetto che in prosa. Infatti, se in prosa la loro percentuale media si aggira intorno all'8%, nei testi poetici del periodo classico questa è sempre superiore al 10% con evidenti punte in due delle opere di Orazio: 15,8% negli Epodi e 18,3% nei Carmina. Più preposizioni (+3%), ausiliari (+3%), congiunzioni subordinanti (+2%) e numeri (+2%) rispetto ai dati di addestramento sono invece presenti nel testo di valutazione Cross-time. Queste differenze sono da collegarsi sia a caratteristiche specifiche della prosa medievale di Tommaso d'Aquino che ad alcune discrepanze nei criteri di annotazione. Per quanto riguarda la prima motivazione, Tommaso d'Aquino tende: (i) ad usare sintagmi preposizionali laddove il latino classico userebbe la flessione dei casi per indicare il ruolo sintattico di

un sintagma nominale (Palmer 1954); (ii) a sostituire la costruzione *accusativus cum infinitivo* con proposizioni subordinate introdotte da congiunzioni subordinanti come *quia*/*quod*/*ut*; (iii) a citare frequentemente la Bibbia riportando il numero dei versi. Il maggior numero di ausiliari è, invece, da attribuire ad una discrepanza nei criteri di annotazione in quanto nel testo di Tommaso d'Aquino, tratto da un corpus annotato in altro contesto (vedi Sezione 3), sono annotate come ausiliari anche le copule verbali, cosa che non avviene negli altri testi usati in EvaLatin 2020 (Bamman 2008).

Figura 2 - *Confronto della distribuzione percentuale delle etichette relative alle parti del discorso nei dati*



È infine utile soffermarsi rapidamente sulle tipologie di errori che il sistema risultato vincitore della competizione commette nei due task proposti.

Per quanto riguarda la lemmatizzazione, la Tabella 6 mostra i 5 lemmi che presentano il maggior numero di discrepanze tra l'annotazione proposta dal sistema di Straka & Straková (2020) e quella dei dati del test set relativamente al sotto-task Classical.

Tabella 6 - *Errori di lemmatizzazione più frequenti in Straka & Straková (2020)*

Lemma (Gold)	*N. Errori*
qui	78
quod	58
quis	40
numerus_romanus	26
bonum	24

Si può notare che tra i lemmi che presentano un maggior numero di errori compaiono il pronome relativo *qui* e il pronome indefinito/interrogativo *quis*, che in molte forme flesse sono indistinguibili dal punto di vista formale e vengono dunque spesso confusi l'uno con l'altro. Inoltre, alcune forme sono identiche ad elementi invariabili e dunque lemmatizzate come tali, in particolare il nominativo/accusativo neutro *quod* come SCONJ, l'ablativo maschile/neutro *quo* come SCONJ o come ADV, e l'accusativo femminile *quam* come ADV. Queste stesse ragioni motivano, nel senso opposto, la frequente confusione del lemma *quod* come SCONJ con l'omonima forma flessa del pronome relativo (PRON).

Compare poi il nome *bonum* 'bene', che altro non è che il neutro sostantivato dell'aggettivo *bonus* 'buono', con cui condivide molte forme flesse e con cui è dunque sovente confuso. Infine, molti numeri romani non vengono riconosciuti come tali e lemmatizzati di conseguenza, in alcuni casi anche per via di omografie con altre forme (ad es. *ii* confuso con la forma di nominativo plurale maschile del pronome dimostrativo *is*).

I problemi evidenziati riguardo alla lemmatizzazione si intrecciano con l'analisi dal punto di vista morfosintattico delle parti del discorso. Basandoci sulla capacità del sistema di Straka & Straková (2020) di riconoscerle correttamente e misurandone la precisione relativa a ciascuna parte del discorso, rileviamo quattro principali nuclei di confusione, tutti con punteggi inferiori al 95%: uno che coinvolge i numerali (NUM; precisione del 66,9%), uno le congiunzioni subordinanti (SCONJ; 90,9%), uno gli ausiliari (AUX; 92,3%), e infine uno gli aggettivi (ADJ; 92,3%).

Per quanto riguarda le SCONJ (e in misura minore anche gli avverbi, che si attestano sotto il 96% di precisione, mentre ADP, CCONJ, PART e INTJ sono tutte sopra il 99%), trattandosi di una categoria composta da parole morfologicamente invariabili, l'assegnazione della parte del discorso va di pari passo con la lemmatizzazione, e si ripresentano così le problematiche appena discusse per quest'ultima.

Osserviamo tuttavia altri due fattori che hanno apparentemente influito sugli errori del modello di UDPipe in oggetto: da una parte alcune oscillazioni nell'annotazione dell'insieme di addestramento e/o test, e dall'altra la non contiguità di token appartenenti allo stesso sintagma.

Il primo fattore è ben illustrato dagli errori nella classe dei numerali (NUM): qui la precisione molto bassa è fortemente influenzata dalle molte occorrenze di forme del lemma *unus* 'uno', che, mentre nell'insieme di addestramento sono uniformemente annotati come NUM, nel test compaiono (93 occorrenze) solo come DET, rimanendo però altrimenti indistinguibili: il sistema li etichetta tutti come NUM, tranne 1 come ADV, così sbagliando.

Il secondo fattore può essere invece esemplificato da alcuni errori rispetto alla classe degli aggettivi (ADJ). Qui osserviamo come siano coinvolte soprattutto le categorie NOUN e VERB, sia in un verso che nell'altro, cioè queste tre etichette vengono (relativamente) di frequente scambiate fra loro. Infatti, se, su 4.636 ADJ, 169 sono stati etichettati come NOUN e 111 come VERB, fra i NOUN le maggiori incertezze insorgono soprattutto con ADJ e VERB (rispettivamente 202 e 93 su 14.019 token), e fra i VERB con ADJ e NOUN (90 e 80 su 12.359). In latino la flessione aggettivale è quasi indistinguibile dalla prima, seconda o terza declinazione dei nomi (a seconda di classe e/o genere), e i contesti sintattici in cui queste due categorie lessicali compaiono sono estremamente simili (ad es., sia un aggettivo che un nome possono fungere da testa di un sintagma nominale). Il sistema si trova così a dover operare scelte su criteri molto sottili e a volte contrastanti con quello che ha appreso durante l'addestramento. Per esempio, osserviamo che tre volte il sistema etichetta una forma *medio* o *medium* afferente a un ADJ *medius* 'medio' come un NOUN con lemma *medium* (2 volte) o *medius* (1 volta): nei dati d'addestramento osserviamo che effettivamente le forme *medio* e *medium* si dividono fra ADJ (lemma *medius*; 13 e 11 occorrenze a testa) e NOUN (lemma

medium 'il mezzo'; 15 e 10 occorrenze), con solo una lieve preferenza per l'aggettivo, e capiamo che il sistema propende per la seconda opzione quando interpreta tale forma come testa di un sintagma in quanto non adiacente a un nome di cui potrebbe essere attributo, ad es. *in medium reciperent agmen* '[che] li salvaguardavano nel mezzo dello schieramento [lett. nel medio schieramento]' (Cesare, Bellum Civile, fr. 503), dove *medium* modifica *agmen*, ma ne è separato dal predicato. Similmente, nella categoria delle SCONJ troviamo vari *cum* 'quando' erroneamente etichettati come adposizioni (ADP), e in quella degli ADV *modo* 'soltanto' analizzato come una forma del NOUN *modus* 'maniera'.

5. *Conclusioni e lavori futuri*

In questo articolo è stato presentato il corpus EvaLatin 1.0, sviluppato per la prima campagna di valutazione di strumenti di TAL per il latino e concentrata sulla lemmatizzazione e l'annotazione delle parti del discorso.

La campagna EvaLatin è stata avviata innanzitutto per fare il punto in merito allo stato delle prestazioni degli strumenti di TAL per la lingua latina; tale esigenza riflette il fatto che il latino sia da considerarsi una lingua per cui sono ormai disponibili numerose risorse linguistiche di diverso tipo, che iniziano a garantire una copertura testuale e lessicale sufficiente dell'ampio spettro diacronico e diatopico lungo cui questa lingua si estende. In un circolo virtuoso, proprio la disponibilità crescente di corpora annotati metalinguisticamente che raccolgono testi latini di diversa epoca, provenienza e genere consente di addestrare strumenti di TAL di tipo probabilistico capaci di garantire buoni valori di accuratezza e, quindi, di facilitare lo sviluppo di ulteriori corpora annotati.

Certamente restano molte le questioni da risolvere a livello di TAL del latino e una campagna di valutazione come EvaLatin serve proprio a farle emergere. Su tutte si impone la difficoltà di portabilità dei modelli addestrati lungo l'arco diacronico e stilistico dei testi latini. Se questa è una sfida aperta per il mondo del TAL, essa si configura anche come una solida fonte d'informazioni per chi si occupa di questioni linguistiche e letterarie del latino. A tal proposito, un'indagine dettagliata non solo della quantità e distribuzione, ma anche del tipo

di errori di lemmatizzazione e attribuzione delle parti del discorso commessi dagli strumenti di TAL può fornire indizi in merito alle differenze lessicali e morfosintattiche tra i testi usati in fase di addestramento di un modello e quelli su cui il modello è stato applicato.

La svolta empirista nel mondo del TAL, il cui stato dell'arte consiste in strumenti probabilistici addestrati sulla base di evidenza empirica, si accosta dunque fertilmente alla grande disponibilità di testi latini su supporto digitale che oggi è nelle mani dei ricercatori. Da sempre gli studiosi di lingue classiche hanno avuto un rapporto stretto con il dato testuale, unica voce che ancora risuona di lingue che non hanno più parlanti nativi, ma mai come ora quel dato è stato disponibile tanto facilmente, velocemente e ampiamente: ciò solleva la necessità di strumenti che lo possano analizzare in modo automatico, valorizzando così una svolta massivamente empirista anche negli studi classici.

Il corpus EvaLatin 1.0 aspira ad essere il primo di una serie, dal momento che è prevista l'organizzazione di nuovi *shared task* per gli strumenti di TAL del latino, di volta in volta dedicati a diversi livelli di annotazione metalinguistica. La seconda edizione di EvaLatin si è tenuta il 25 giugno del 2022 a Marsiglia nell'ambito del "2nd Workshop on Language Technologies for Historical and Ancient LAnguages" (LT4HALA 2022: https://circse.github.io/LT4HALA/2022/). Oltre alla lemmatizzazione e all'annotazione delle parti del discorso, la campagna di valutazione si è concentrata sul trattamento automatico dei tratti morfologici.

Ringraziamenti

Il progetto "LiLa: Linking Latin" è finanziato dal Consiglio Europeo della Ricerca (ERC) nell'ambito del programma di ricerca e innovazione European Union's Horizon 2020 – Grant Agreement No. 769994.

Riferimenti bibliografici

Bamman, David & Passarotti, Marco & Crane, Gregory. 2008. A Case Study in Treebank Collaboration and Comparison: Accusativus cum Infinitivo and Subordination in Latin. *The Prague Bulletin of Mathematical Linguistics* 90, 109–122.

Bo, Domenico. 1967. *L'uso dei nomi propri greci come parametro del progresso artistico di Orazio*. Torino: Giappichelli.

Bouma, Gerlof & Adesam, Yvonne (a cura di). 2017. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Gothenburg, 22-24 maggio 2017*. Gothenburg: Linköping University Electronic Press.

Chiari, Isabella & De Mauro, Tullio. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In Basili, Roberto & Lenci, Alessandro & Magnini, Bernardo (a cura di), *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, 113–116. Pisa: Pisa University Press.

de Marneffe, Marie-Catherine & Manning, Christopher D. & Nivre, Joakim & Zeman, Daniel. 2021. Universal dependencies. *Computational linguistics* 47(2): 255–308.

Palmer, Leonard Robert. 1954. *The Latin language*. London: Faber and Faber.

Paradisi, Patrizia. 2019. Donne oraziane: onomastica e identità. *il Nome nel testo. Rivista internazionale di onomastica letteraria*, 155–167.

Passarotti, Marco. 2019. The Project of the Index Thomisticus Treebank. In Berti, Monica (a cura di), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, 299–319. Berlino-Boston, De Gruyter GmbH: 299-319.

Petrov, Slav & Das, Dipanjan & McDonald, Ryan. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* 2089–2096. Istanbul: European Language Resources Association (ELRA).

Ponti, Edoardo Maria & Passarotti, Marco. 2016. Differentia compositionem facit. A slower-paced and reliable parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 683–688. Portorož: European Language Resources Association (ELRA).

Roncali, Renata. 2013. Orazio. In Roncali, Renata (a cura di), I *classici nella storia della letteratura latina. I poeti*, 269–340. Bari: Edizioni di Pagina.

Smith, David A. & Rydberg-Cox, Jeffrey A. & Crane, Gregory R. 2000. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing* 15(1): 15–25.

Sprugnoli, Rachele & Passarotti, Marco (a cura di). 2020. *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages.* Marsiglia, European Language Resources Association (ELRA).

Straka, Milan & Straková, Jana. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver: Association for Computational Linguistics.

Straka, Milan & Straková, Jana. 2020. UDPipe at EvaLatin 2020: Contextualized Embeddings and Treebank Embeddings. In Sprugnoli & Passarotti (a cura di), 124–129.

Verkerk, Philippe & Ouvrard, Yves & Fantoli, Margherita & Longrée, Dominique. 2020. LASLA and Collatinus: a convergence in lexica. *Studi e Saggi Linguistici*, 58(1), 95–120.