4 Gli strumenti informatici. Sviluppo e risultati

di Sara Tonelli, Rachele Sprugnoli, Alessio Palmero Aprosio, Giovanni Moretti, Stefano Menini (FBK)

4.1 Introduzione

Per riconoscere i tratti linguistici di interesse su un corpus composto da quasi tremila temi e per annotarli in modo coerente si è reso necessario lo sviluppo di diversi strumenti informatici. Tali software appartengono a due tipologie: da un lato, si sono sviluppati alcuni moduli per l'analisi del testo, che in modo automatico riconoscono dei tratti o estraggono delle informazioni parziali utili a riconoscere i tratti in modo manuale. Dall'altro, si è adattata al progetto una piattaforma online che permette di effettuare annotazione linguistica multilivello con diversi annotatori al lavoro in parallelo su porzioni diverse del corpus di temi. Per la prima tipologia di attività, si è esteso lo strumento per l'analisi del testo TINT (Palmero Aprosio e Moretti, 2019), un sistema sviluppato dalla Fondazione Bruno Kessler basato sulla architettura open-source StanfordCoreNLP (Manning *et al.*, 2014). TINT includeva alcuni moduli di base per l'analisi dell'italiano contemporaneo che sono stati estesi con numerosi componenti ad hoc per il riconoscimento dei tratti di interesse. I dettagli dell'implementazione sono illustrati nella sezione seguente. Per quanto riguarda la piattaforma di annotazione, si è optato per un software online, in quanto consente di gestire centralmente diversi annotatori in parallelo, senza la necessità di installazioni di applicativi. La piattaforma adottata è CAT (Bartalesi Lenzi *et al.,* 2012), descritta nel dettaglio in questo capitolo, che ha permesso di creare un account utente per ogni annotatore, di assegnare da un account di amministratore i temi da annotare a ogni utente e di svolgere un controllo qualità su alcuni temi presi a campione annotati in parallelo da più utenti.

Alcune delle analisi di TINT sono state utilizzate come input per CAT, affinché gli annotatori potessero vedere alcuni tratti come pre-annotati e intervenire con delle correzioni oppure specificando degli attributi. La combinazione dei due tool, insieme alla versatilità dell'interfaccia online, hanno consentito di velocizzare notevolmente la annotazione dei tratti linguistici di interesse. Uno studio preliminare

aveva infatti stimato in un'ora per tema il tempo necessario per annotare tutti i tratti manualmente utilizzando un editor di testo come Word. Con CAT in combinazione con TINT, invece, i tempi medi per tema si sono dimezzati. A questo vantaggio si aggiunge anche la possibilità di esportare da CAT i temi annotati in formato XML, da cui partire per estrarre le statistiche sui tratti. Affidandosi a un'annotazione manuale su Word, sarebbe venuto meno il controllo sull'omogeneità delle annotazioni, rendendo più difficoltoso anche il calcolo di tali statistiche.

4.2 Annotazione automatica e semi-automatica dei tratti linguistici

Vista la crescente disponibilità di dati in formato digitale, i mondi della ricerca e dell'industria avvertono sempre più la necessità di poter analizzare automaticamente grandi quantità di dati. Siccome non è possibile effettuare queste analisi manualmente, ha avuto grande impulso la creazione di software in grado di esaminare rapidamente testi, immagini e video per trasformarli in dati strutturati. La maggior parte di questi sforzi e dei progetti in questa direzione, però, si concentra sulla lingua inglese, più diffusa a livello mondiale. Su molte altre lingue il numero di risorse è sempre stato piuttosto limitato, soprattutto per quanto riguarda lo sviluppo di software liberamente disponibili e open source. Per questo motivo, nel 2016 si è intrapreso alla Fondazione Bruno Kessler lo sviluppo di TINT, un software di analisi testuale pensato appositamente per l'italiano. Il cuore dell'applicazione si basa su CoreNLP, una libreria sviluppata dall'Università di Stanford per l'analisi della lingua inglese pensata per essere facilmente estesa anche ad altre lingue.

Il funzionamento di TINT è basato sul concetto di *data pipeline*, un insieme di moduli connessi in serie per cui l'output di uno è utilizzato come input del successivo.

Il primo passo per l'analisi testuale (per quasi tutti i sistemi attualmente disponibili in qualsiasi lingua) è la *tokenizzazione*, ovvero la separazione del testo in *token*, elementi atomici che essenzialmente sono parole e segni di interpunzione (più di recente, anche le emoticon sono entrate a far parte della categoria, ma nel caso dei temi delle scuole superiori questo aspetto non è considerato). Il secondo passo consiste nel raggruppare i token in periodi.

I successivi passaggi, che corrispondono a diversi moduli di TINT, vengono attivati in sequenza:

- *Analisi morfologica* - A ciascun token vengono associate tutte le possibili forme conosciute della lingua italiana, scorrendo un dizionario pre-caricato nel programma. La parola «cancello» per

esempio, può essere classificata come nome maschile singolare o come prima persona presente del verbo «cancellare».

- *Assegnazione delle categorie grammaticali* - Ogni token viene analizzato (anche e soprattutto in relazione agli altri del medesimo periodo) e classificato in base alla sua funzione grammaticale: nome, aggettivo, pronome, verbo, ecc.
- *Lemmatizzazione* - Usando le informazioni ottenute dai due passi precedenti (secondo il principio precedentemente illustrato di *pipeline*), il modulo di lemmatizzazione cerca di assegnare un lemma, ovvero il termine base del token. Nel caso di «cancello», se il modulo di analisi grammaticale ha stabilito che si tratta di un verbo, la lemmatizzazione assegnerà come lemma «cancellare». Nel caso in cui l'analisi morfologica non avesse trovato nulla nel dizionario (come nel caso di un neologismo), il sistema proverà a risalire al lemma originale in base al risultato dell'analisi grammaticale. Per esempio se «insalatiamo» viene classificato come verbo presente in prima persona plurale, il lemma di origine sarà «insalatare», anche se nella lingua italiana non esiste.
- *Derivatario* - Questo modulo utilizza il derIvaTario (Talamo *et al*., 2016), una risorsa elettronica che contiene circa 11.000 derivativi italiani,[1] e associa a ciascun token un insieme di informazioni che riguardano la sua derivazione. La parola è segmentata nella radice e nei suoi affissi (per esempio da «visione» viene estratta la parola radice «vedere», l'affisso «-zione» e l'allomorfo «-ione»).
- *Classificazione dei tempi verbali* - I token che rappresentano forme verbali (informazione raccolta dall'analisi grammaticale) vengono raggruppati e classificati per modo, tempo e forma (attiva/passiva). Questo passaggio è particolarmente complesso poiché il sistema deve capire quando due parole etichettate come verbo nei passaggi precedenti sono parte di una stessa forma verbale, anche se non adiacenti (per esempio «ho subito riconosciuto»). Anche la distinzione tra forme passive e passato prossimo è complessa da riconoscere automaticamente, vista la somiglianza tra le due strutture sintattiche (cfr. «è mangiato» - «è andato»).

Oltre ai moduli descritti sopra, TINT ne include altri, come il riconoscimento delle entità geografiche e dei nomi di persona menzionati nel testo, o l'analisi dell'albero sintattico, che però non sono stati utilizzati per l'annotazione dei temi poiché non erano necessari per riconoscere i tratti. Ulteriori dettagli su questi moduli sono riportati in Palmero Aprosio e Moretti (2018).

[1] http://derivatario.sns.it/

Infine, dopo l'analisi automatica di un testo, TINT fornisce anche informazioni statistiche sul documento, come il numero di token, frasi, periodi, parole, sillabe, che possono essere utilizzati per analizzare - tra le altre cose - il livello di complessità del documento.

La versione base di TINT è rilasciata come open source[2] con la versione 3 della licenza GPL, che permette a chiunque di usarlo liberamente, anche in contesti commerciali (si veda la pagina ufficiale del tool o della licenza per ulteriori dettagli). I moduli aggiuntivi descritti in queste pagine sono disponibili, con la stessa licenza, sulla pagina dei progetti sperimentali dell'unità Digital Humanities di FBK[3].

4.2.1 Descrizione dei tratti annotati con TINT

Come precedentemente illustrato, TINT è stato utilizzato sia per annotare in modo automatico alcuni tratti linguistici tra quelli selezionati per il presente studio, sia per effettuare una pre-annotazione, da sottoporre agli annotatori per una verifica o arricchimento manuale. Per chiarezza, riportiamo nella tabella seguente la lista dei tratti, indicando con «A» i tratti identificati con TINT in modo automatico, con «S» quelli che sono poi stati verificati, corretti o arricchiti manualmente (semi-automatici), e con «M» quelli che sono stati annotati in modo completamente manuale utilizzando la piattaforma CAT, descritta nella sezione seguente.

Tipo	ID	Tratto	Descrizione
S	1	Monosillabi	Vocaboli monosillabi con l'accento sbagliato
A	2	Apostrofi	Uso scorretto dell'apostrofo per l'articolo «un»
S	3	Maiuscole	Uso scorretto delle lettere maiuscole
A	4	«il»	Uso scorretto di «il» (davanti a w-j-pn-ps)
S	5	Pronomi personali	Annotazione di tutti i pronomi personali e dell'uso di «loro» con significato di «a loro»
S	6	«gli»	Differenti usi di «gli», inclusi quelli scorretti
S	7	«questo»	Annotazione di «quest*» quando si riferisce al contesto del discorso
A	8	Parole comuni	Parole generiche, come «bello», «brutto», «fare», «dire», «cosa»
S	9	Indicativo imperfetto	Diversi tipi di imperfetto (al posto del congiuntivo, nelle frasi ipotetiche, ecc.)
S	10	Gerundio	Diversi tipi di gerundio

[2] Tutte le informazioni e il codice sorgente possono essere trovati sul sito http://tint.fbk.eu/
[3] Disponibile sul sito https://github.com/dhfbk/dh-utils

S	11	Indicativo presente	Diversi tipi di indicativo presente
A	12	«stare», «andare»	Uso di «stare» e «andare» nella costruzione della frase
S	13	Affissi	Uso di affissi come «anti-», «-ismo», «trans-», ecc.
S	14	Statistiche	Conteggio di parole, frasi, periodi e frasi nominali
S	15	Connettivi 1	Annotazione dei connettivi generici «che», «dove», «allora» e il loro uso corretto o scorretto
S	16	Connettivi 2	Conteggio dei connettivi complessi come «nondimeno», «sebbene», «qualora», ecc.
S	17	Punteggiatura	Annotazione dei segni di punteggiatura, con il relativo utilizzo (corretto o scorretto)
S	18	Connettivi a inizio frase	Identifica «perché» e «quando» a inizio frase, con l'annotazione del loro utilizzo
S	19	Registro informale	Estrazione di un elenco di forme tipiche del parlato («della serie», «tipo», «troppo forte», ecc.)
S	20	Anglicismi	Estrazione dei termini derivati dalla lingua inglese, adattati e non
S	21	Politicamente corretto	Uso di termini politicamente corretti («ministra», «sindaca», ecc.)
S	22	Polirematiche	Estrazione delle espressioni formate da più parole (polirematiche)
S	23	Plastismi	Estrazione delle espressioni abusate e prive di originalità
M	24	Dislocazioni	Annotazione delle dislocazioni destre e sinistre
S	25	Frasi scisse	Annotazione delle frasi scisse
S	26	«li»	Annotazione della parola «li» e relativo uso corretto/scorretto dell'accento
A	27	«d» eufonica	Annotazione della «d» eufonica prima di una parola che inizia per vocale
M	28	Altri tratti	Altri fenomeni linguistici rilevanti non definiti dai precedenti tratti

Tabella 1 - Lista dei tratti linguistici annotati, divisi in automatici (A), semi-automatici (S) e manuali (M)

Di seguiti descriviamo nel dettaglio l'uso di TINT per ogni tratto identificato in modo automatico o semi-automatico.

T1 - Monosillabi

Tra le varie informazioni estratte con TINT c'è anche quella relativa al numero di sillabe. Le parole analizzate nei temi che ne contengono solo una vengono estratte, confrontate con un elenco di tutti i monosillabi noti in italiano e classificate in «da verificare» (quando non sono incluse nell'elenco), «corretto», «scorretto» e «ambiguo». L'ultimo caso si verifica quando il termine è presente nell'elenco in più forme (con/senza accento/apostrofo) e quindi il sistema non può stabilirne la correttezza. Dopo la prima classificazione, l'informazione viene vagliata manualmente (cfr. Sezione 3) per stabilire se i casi da verificare e quelli ambigui ricadano tra i corretti o gli scorretti.

T2 - Apostrofi

L'analisi morfologica può stabilire se un determinato nome/aggettivo è maschile o femminile, e quindi verificare se «un» deve essere seguito da un apostrofo. In TINT è stata quindi inserita una regola che verifica se una parola che inizia per vocale è preceduta dall'articolo «un» e, in base al genere del sostantivo, stabilisce se l'uso dell'apostrofo è corretto.

T3 - Maiuscole

Tutti i termini che iniziano con lettera maiuscola (e che non si trovano all'inizio della frase) vengono identificati dal sistema in modo automatico. Confrontando poi questi termini con una lista di parole che solitamente sono erroneamente scritte con la maiuscola iniziale, per esempio i nomi di giorni e mesi, il sistema è in grado di riconoscere gli errori in modo completamente automatico.

T4 - «il»

Le parole che iniziano per w-, j-, pn- e ps- e sono precedute dall'articolo «il» sono individuate automaticamente e conteggiate.

T5 - Pronomi personali

Le occorrenze di tutti i pronomi personali vengono selezionate e conteggiate. Poiché non è possibile identificare automaticamente se «loro» è utilizzato al posto di «a loro», tutte le menzioni di «loro» vengono evidenziate per la verifica manuale da parte degli annotatori.

T6 - «gli»

Tutte le occorrenze della parola «gli» sono selezionate per gli annotatori che poi dovranno stabilire se l'utilizzo è corretto.

T7 - «questo»

Le occorrenze di «questo», «questa», «questi» e «queste» sono selezionate per l'annotazione manuale, in cui gli annotatori dovranno stabilire il tipo di utilizzo del termine in base al contesto della frase.

T8 - Parole comuni

Le parole di uso più comune («bello», «brutto», «grande», «piccolo», «buono», «cattivo», «vecchio», «nuovo», «cosa», «fare», «dire») sono identificate e conteggiate in modo completamente automatico.

T9 - Indicativo imperfetto, T10 - Gerundio, T11 - Indicativo presente

I verbi nei modi/tempi in oggetto sono raccolti e selezionati per una successiva verifica, in cui verranno annotati manualmente la tipologia e l'uso più o meno corretto del modo e del tempo verbale.

T12 - «stare»/«andare»

Le occorrenze delle parole «stare» e «andare» sono raccolte e conteggiate distinguendo i casi in cui reggono la costruzione con il modo infinito («stare per» o «andare a» seguito da infinito).

T13 - Affissi

Vengono conteggiate le parole che includono uno dei seguenti affissi: «anti-», «dopo-», «trans-», «iper-», «super-», «-ista», «-tore»,

«-zione», «-mento», «-tura», «-aggio», «-ità», «-ismo», «-izzare», «-ale», «-iano», «-istico», «-ato», «-izzazione».

T14 - Statistiche

TINT raccoglie dati relativi al numero di parole, frasi e periodi, etichettando anche le frasi che non contengono il verbo. Poiché le frasi nominali sono molto frequenti nei titoli, è stato implementato un modulo aggiuntivo che esclude i titoli dai conteggi, in modo da non inficiare la validità delle statistiche.

T15 - Connettivi 1

Vengono conteggiati e raccolti i seguenti connettivi: «che», «dove», «infatti», «cioè», «cioé», «allora», «dunque», «quindi», «siccome».

T16 - Connettivi 2

Vengono conteggiati e raccolti i seguenti connettivi: «qualora», «nondimeno», «sebbene», «quantunque», «affinché», «sicché», «talché».

T17 - Punteggiatura

Vengono conteggiati e annotati i segni di punteggiatura, considerando in particolare i casi in cui più segni sono usati per aggiungere enfasi (come l'unione di più punti esclamativi o interrogativi).

T18 - Connettivi a inizio frase

I casi di «perché» e «quando» a inizio frase vengono identificati e conteggiati.

T19 - Registro informale

Le espressioni gergali («troppo forte», «della serie», ecc.) vengono identificate e conteggiate.

T20 - Anglicismi

Le espressioni derivanti dall'inglese vengono individuate automaticamente e conteggiate. L'elenco degli anglicismi è stato stilato a partire da fonti online. Per ciascun termine è anche aggiunta in automatico una annotazione per stabilire se è stato adattato (ovvero coniugato/declinato in italiano) oppure se è stato utilizzato nella sua forma originale.

T21 - Politicamente corretto

A partire da un elenco stilato manualmente, TINT estrae tutte le occorrenze di termini politicamente corretti, come per esempio «operatore ecologico», «non vedente», «ministra», ecc.

T22 - Polirematiche

Le espressioni formate da più parole sono estratte automaticamente, a partire da un elenco stilato utilizzando varie fonti online. TINT identifica anche le forme discontinue, intervallate da un aggettivo o da un avverbio (per esempio da «Giovanni andava sempre cauto» si estrae l'espressione «andare cauto»).

T23 - Plastismi

Le espressioni abusate e prive di originalità sono estratte a partire da un elenco stilato manualmente (e contenente, per esempio, «offrire uno spaccato», «poteva essere una strage», ecc.).

T25 - Frasi scisse

Le frasi contenenti le strutture tipiche delle frasi scisse (come per esempio «È Maria che parla alla radio») sono identificate in modo automatico ma vengono successivamente controllate dagli annotatori per evitare che siano inclusi nei conteggi anche costrutti dalla struttura simile ma che non sono frasi scisse.

T26 - «li»

Le occorrenze della parola «li» (con o senza accento) vengono identificate e successivamente passate al vaglio degli annotatori che devono specificare manualmente se la presenza o assenza dell'accento è corretta.

T27 - «d» eufonica

Una recente regola grammaticale, discussa anche dall'Accademia della Crusca, stabilisce che la «d» eufonica dovrebbe essere introdotta solamente quando la congiunzione «e» e la preposizione «a» sono seguite da una parola che inizia con la medesima vocale (per esempio «ed ecco», «ad andare»). Il sistema identifica automaticamente tutti i casi trovati, classificandone l'uso corretto e scorretto.

Oltre alle componenti necessarie per l'annotazione dei tratti, TINT è stato arricchito con un ulteriore modulo finalizzato a individuare passaggi testuali nei temi che riprendono il testo della traccia. Infatti, tutte le tracce dei temi analizzati sono accompagnate da un elenco di fonti bibliografiche che gli studenti possono utilizzare come riferimento nello svolgimento del compito. È alquanto probabile, quindi, che nel manoscritto siano presenti citazioni tratte dai documenti allegati alla traccia. Analizzando manualmente alcuni lavori, si è anche notato che non sempre la citazione è alla lettera, ma può essere modificata per farla rientrare nel contesto della frase in cui è inserita.

Se il conteggio dei tratti viene effettuato includendo anche i passaggi di testo ripresi dalle fonti bibliografiche, le statistiche potrebbero non essere affidabili. Per esempio, termini aulici o desueti (cfr. i connettivi nel tratto 16) potrebbero essere ripresi dalle fonti, e non essere presenti invece nelle parti originali dell'elaborato scritto dallo studente. Per evitare questo problema, TINT è stato esteso in modo che, mettendo a confronto due testi, stabilisca quali parti del primo sono state incluse nel secondo. In particolare, i testi della traccia e delle fonti del compito scelto dallo studente sono stati messi a confronto con il tema svolto, e le frasi trovate sono state escluse da tutte le analisi dei tratti sopra descritti. Con questa strategia, è stato eliminato circa il 2% del testo complessivo, con variazioni molto rilevanti tra i vari temi (in alcuni casi il 46% del singolo tema è risultato essere sovrapponibile alle fonti).

Per riuscire a intercettare anche quelle citazioni modificate dallo studente, nel modulo di TINT è stato utilizzato anche il pacchetto software FuzzyWuzzy[4], che permette di associare anche stringhe di testo non perfettamente identiche. Il codice che permette di svolgere questo compito è liberamente disponibile nel pacchetto software di TINT e può avere anche utilizzi che esulano dagli scopi delle analisi svolte in questo documento, come per esempio il riconoscimento di casi di plagio.

4.3 Analisi manuale dei tratti linguistici

L'annotazione manuale dei tratti si è svolta usando la piattaforma CAT (Bartalesi Lenzi *et al*, 2012), un software basato su tecnologia web e sviluppato per l'annotazione manuale di corpora. Grazie a CAT è stato possibile creare uno schema di annotazione specifico per il progetto codificando i vari tratti e i loro attributi. Per annotazione si intende la selezione dei segmenti di testo, detti *markable*, che realizzano linguisticamente il tratto da analizzare e la scelta del valore giusto per ogni attributo. Per i tratti semi-automatici, i markable erano già annotati ed evidenziati nei testi e gli attributi avevano i valori già selezionati in base al risultato dell'analisi automatica fornita da TINT. Markable e attributi potevano essere modificati o rimossi se errati ma anche aggiunti se mancanti. I markable potevano essere singole parole (come i monosillabi) ma anche espressioni di lunghezza variabile (come nel caso del plastismo «teatrino della politica») e discontinue (come per le polirematiche verbali che presentano una rigidità strutturale variabile e quindi possono apparire interrotte dall'interposizione di un'altra parola, ad esempio «prestare molta attenzione»).

L'interfaccia di CAT, nello specifico con un esempio di annotazione del tratto *anglicismi*, è mostrato nella Figura 1. Il markable «slogan» è evidenziato in verde; cliccandoci sopra si apre una finestra denominata «Markable Attributes» in cui aggiungere le informazioni relative agli attributi che possono essere espressi in vari modi: campo di testo libero, menù a tendina, pulsanti di opzioni (radio button), caselle di spunta (check box). Si noti che, ad ogni markable è associato un campo di commento in cui segnalare eventuali dubbi o problemi, ad esempio la presenza di errori di ortografia (e.g. «La banalizzazione e la pubblicità della vita sono collegate ai social networck»).

[4] Informazioni sul pacchetto sul sito https://github.com/xdrop/fuzzywuzzy

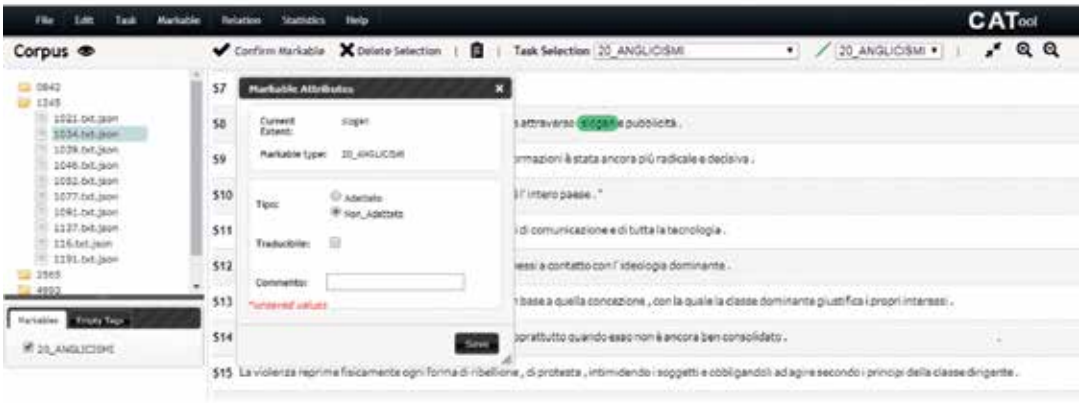

Figura 1 - Interfaccia di CAT

Per il funzionamento di CAT non è stata necessaria alcuna installazione ma solo la connessione Internet e l'uso di un normale browser: ciò ha ridotto il rischio che si presentassero problemi tecnici dovuti alla presenza di diversi sistemi operativi. Inoltre, la tecnologia web su cui si basa CAT ha permesso di controllare da remoto il lavoro degli annotatori, di assegnare loro i temi caricandoli online e di salvare le annotazioni su un database centrale.

Nella prima fase di lavoro, quindici tra professori e professoresse di italiano delle scuole secondarie di secondo grado della Provincia Autonoma di Trento hanno partecipato all'annotazione dei temi. Per insegnare loro ad usare la piattaforma sono state organizzati degli incontri ed è stato compilato un manuale in cui erano descritte le funzioni principali di CAT e le caratteristiche da annotare per ogni tratto. Ad ogni insegnante sono stati assegnati da 30 a 150 temi, in base alla disponibilità concordata con ciascuno di loro ad inizio progetto. Una volta conclusa questa prima fase di annotazione, sono stati selezionati in maniera casuale 150 temi (dieci per ciascun insegnante) che sono stati fatti controllare da altri tre professori.

Qui di seguito si descrivono i 22 tratti manuali o semi-automatici così come codificati nello schema di annotazione sulla piattaforma. La numerazione riprende quella riportata nella lista dei tratti in Tabella 1.

T1 - Monosillabi

I markable già annotati corrispondono a monosillabi che, secondo l'analisi automatica, sono ambigui (ad esempio, «lì»/«li») o errati relativamente alla presenza dell'accento (ad esempio, «fà» invece di «fa»). Un attributo specifica se l'errore è dovuto a una mancata accentazione o a un'accentazione erroneamente aggiunta.

T3 - Maiuscole

I markable già annotati corrispondono a parole non ad inizio periodo aventi l'iniziale maiuscola. Un attributo specifica se si tratta di aggettivi etnici, glottonimi e nomi di mesi/giorni.

T5 - Loro

Tutte le occorrenze della parola «loro» sono evidenziate nel testo come markable. Una casella di spunta è da selezionare nel caso in cui «loro» sia usato al posto di «a loro».

T6 - Gli

Tutte le occorrenze della parola «gli» sono già evidenziate nel testo. L'annotatore deve indicare attraverso un attributo se è usato col significato di «a lui», «a lei», «a loro» oppure se ha la funzione di articolo. Un'opzione viene selezionata per indicare se l'uso di «gli» non segue le regole grammaticali.

T7 - Questo

Tutti i lemmi di «questo» che il sistema automatico riconosce essere pronomi sono evidenziati nel testo come markable. Un attributo è usato per specificare i casi in cui il pronome dimostrativo è usato in funzione anaforica.

T9 - Indicativo imperfetto

I verbi all'indicativo imperfetto identificati dal sistema automatico sono già evidenziati come markable. Per ogni markable un attributo viene usato per indicarne l'uso: proprio, di cortesia, all'interno di un periodo ipotetico, al posto del congiuntivo.

T10 - Gerundio

I verbi al gerundio identificati dal sistema automatico sono già evidenziati come markable. Per ogni markable un attributo viene usato per indicarne l'uso: corretto, scorretto o testuale.

T11 - Indicativo presente

I verbi all'indicativo presente identificati dal sistema automatico sono già evidenziati come markable nel testo. Un attributo serve a distinguere i vari possibili usi: contemporaneità rispetto al momento dell'enunciazione, presente storico, con valore temporale di futuro, imperativo, al posto del congiuntivo.

T13 - Affissi

I markable già colorati nel testo corrispondono a parole che hanno un affisso tra quelli oggetto di monitoraggio, ovvero i prefissi anti-, dopo-, trans-, iper-, super- o i suffissi: -ista, -tore, -zione, -mento, -tura, -aggio, -ità, -ismo, -izzare, -izzazione, -ale, -iano, -istico, -ato. L'analisi automatica dei suffissi si è basata sul derIvaTario, una risorsa della Scuola Normale di Pisa contenente 11.000 derivati italiani[5] (Talamo et al., 2016). Un attributo serve a distinguere le parole con un suffisso, con un prefisso o con entrambi gli affissi. Attraverso dei menù a tendina viene specificato quale prefisso e/o suffisso è presente nella parola evidenziata.

[5] http://derivatario.sns.it/

T14 - Frasi nominali

Le frasi nominali automaticamente identificate sono già evidenziate nel testo. Un attributo viene usato per specificare se la frase nominale fa parte del titolo del tema.

T15 e 16 - Connettivi

I connettivi scelti dal progetto (per esempio «che», «dove», «infatti», «cioè», «allora», «dunque», «quindi», «siccome») sono evidenziati come markable nel testo. Connettivi diversi richiedono attributi diversi: nel caso di «che» e «dove» si specifica se sono polivalenti, si indica se «siccome» è usato al posto di «poiché» e per gli altri se il loro utilizzo all'interno della frase è corretto, scorretto o ha valore di riempitivo.

T17 - Punteggiatura

Tutti i seguenti segni di punteggiatura sono evidenziati come markable: punto e virgola, due punti, punto esclamativo, virgolette alte, puntini di sospensione, virgola, punto. Un attributo è usato per segnalare casi in cui la punteggiatura è usata in maniera scorretta, ad esempio quando le virgolette non sono chiuse o quando indicano lessico improprio.

T18 - Perché/Quando

Tutte le occorrenze di «perché» e «quando» ad inizio frase sono evidenziate come markable nel testo. I casi di uso testuale sono segnalati con un apposito attributo.

T19 - Registro informale

Le parole o espressioni indicate dal progetto per questo tratto (per esempio «troppo forte», «della serie», «tipo», «praticamente», «assolutamente», «ovviamente», «per così dire», «voglio dire») sono evidenziate come markable nel testo. L'annotatore deve controllare che tali parole o espressioni appartengano ad un registro informale nel contesto della frase.

T20 - Anglicismi

Nel testo sono evidenziati come markable gli anglicismi non adattati tratti dal dizionario del De Mauro e quelli adattati tratti dalla lista indicata dal progetto (per esempio «gol», «chattare», «skillato», «stoppare», «mixare», «demo», «app», «info», «spoilerare», «buggato», «flammare», «killare», «whatsappare», «twittare», «stalkerare»). Per ogni anglicismo si segnala se si tratta di forma adattata o meno e si seleziona una specifica opzione solo quando esso si trova in un contesto di traducibilità.

T21 - Politicamente corretto

Le parole o espressioni indicate dal progetto per questo tratto (per esempio «ministra», «sindaca», «assessora», «magistrata», «operatore ecologico», «operatore socio assistenziale», «non udente», «non vedente», «diversamente abile», «danno collaterale», «esubero») sono evidenziate come markable nel testo anche nella loro versione al plurale. L'annotatore deve controllare la correttezza di questa annotazione automatica.

T22 - Polirematiche

Nel testo sono evidenziate come markable le espressioni polirematiche riconosciute preliminarmente da TINT. L'annotatore deve controllare la correttezza di questa annotazione automatica.

T23 - Plastismi

Le parole o espressioni indicate per questo tratto dal progetto (per esempio «è giallo», «il cerchio magico», «la morsa del gelo») sono automaticamente identificate ed evidenziate nel testo. L'annotatore deve controllare la correttezza di questa annotazione automatica.

T24 - Dislocazioni

Questo tratto richiede un'annotazione totalmente manuale. Per ogni markable un attributo specifica se si tratta di dislocazione a destra, dislocazione a sinistra o tema sospeso.

T25 - Frasi scisse

Grazie ad un'analisi automatica del testo, sono state evidenziate nel testo le frasi potenzialmente scisse. L'annotatore deve controllare la correttezza di questa annotazione automatica e aggiungere eventuali frasi scisse mancanti.

T26 - Li

Tutti i «li» che sono parole autonome nel testo sono evidenziati come markable. Non sono invece evidenziati gli enclitici che devono essere manualmente annotati. I casi scorretti in cui «li» è usato al posto di «gli» vengono segnalati con un apposito attributo.

T28 - Altri tratti

Questo markable è stato aggiunto per permettere agli insegnanti di indicare ulteriori tratti significativi presenti a livello di tema, diversi da quelli scelti dal progetto, su cui potrebbe essere interessante lavorare in futuro. Tra questi tratti rientrano diversi errori di ortografia, terminologia e sintassi.

4.4 Prime analisi del corpus annotato

In questa Sezione introduciamo i dati su cui sono state svolte le analisi automatiche, semi-automatiche e manuali presentate nelle Sezioni 2 e 3. Per lo scopo sono stati raccolti un totale di 2.928 elaborati che coprono la produzione di diversi tipi di istituti e un arco temporale di 15 anni. Nelle tabelle che seguono si riportano come sono suddivisi i dati sia in base alla loro provenienza che alla distribuzione nel corso degli anni.

Tipologia di scuola	Numero Temi
Liceo classico	150
Liceo scientifico	659
Istruzione professionale	228
Altro (liceo linguistico, liceo SU e LES, istituti tecnici)	1.891
Totale	2.928

Tabella 2 - Suddivisione dei temi per tipologia di scuola

Nelle Tabelle 2 e 3 si riporta il numero di temi usati per le analisi, aggregati in base alla tipologia dell'istituto di provenienza seguendo due diversi criteri. La prima tiene separati liceo classico, scientifico e istituti professionali, aggregando le rimanenti provenienze in una singola categoria, mentre la seconda mostra una suddivisione in base alla tipologia d'istituto (liceo - istituto tecnico - istituto professionale). Si può osservare come le categorie maggiormente rappresentate siano gli istituti tecnici e i licei scientifici che coprono rispettivamente il 38,8% e il 22,2% dei dati a disposizione.

Tipologia di scuola	Numero Temi
Licei	1.564
Tecnici	1.136
Istruzione professionale	228
Totale	2.928

Tabella 3 - Suddivisione dei temi per tipologia di scuola (macro-categorie)

I temi usati per la ricerca sono stati raccolti ad intervalli di 3 anni a partire dal 2000-2001 fino al 2015-2016. La Tabella 4 illustra il numero di temi per ognuno degli anni presi in esame divisi per tipologia di scuola. Sebbene certe tipologie d'istituto siano maggiormente rappresentate, come già evidenziato dalle precedenti tabelle, la proporzione dei temi per tipologia di scuola nell'arco temporale considerato resta pressoché invariata.

Tipologia scuola	Numero di temi per anno					
	2000-01	03-04	06-07	09-10	12-13	15-16
liceo classico	26	26	26	27	24	21
liceo scientifico	109	109	109	106	113	113
altri licei	108	114	119	130	153	131
tecnici	187	189	189	174	196	201
istruzione professionale	40	40	39	53	27	29
Totale	470	478	482	490	513	495

Tabella 4: Distribuzione dei temi in base a tipologia di scuola e anno scolastico preso in analisi

Gli elaborati raccolti si suddividono inoltre nelle tipologie di *Saggio Breve* e *Articolo.* Nonostante la traccia chiedesse esplicitamente agli studenti di indicare in calce all'elaborato la tipologia, numerosi temi non presentano questa informazione[6]. La distribuzione delle tre classi (*Saggio breve*, *Articolo* e *Non disponibile*) in base al tipo di istituto è riportata nella Tabella 5 e in Figura 2. L'informazione riguardante la tipologia di tema svolto è disponibile solo per una porzione degli scritti (il 64,6%), mentre le restanti prove sono conteggiate come N/D.

In generale si osserva come gli studenti manifestino una netta preferenza, indifferentemente dalla tipologia di istituto, in favore del saggio breve. L'unico caso in cui la differenza tra le due categorie risulta meno marcata è rappresentata dagli istituti professionali, mentre negli altri casi il saggio breve può arrivare ad avere il doppio delle preferenze rispetto alla scrittura di un articolo.

Tipologia scuola	Tipologia Tema		
	Saggio Breve	Articolo	N/D
liceo classico	60	39	51
liceo scientifico	278	156	225
altri licei	321	192	242
tecnici	485	223	428
istruzione professionale	75	63	90

Tabella 5: Distribuzione della tipologia di tema tra i diversi tipi di istituti scolastici

[6] In un secondo momento, da parte del gruppo di lavoro si è risaliti a una definizione della tipologia scelta (articolo o saggio breve), anche laddove non era indicata dallo studente, con la lettura diretta dei file e la valutazione di alcuni requisiti: presenza/assenza della destinazione editoriale, caratteristiche testuali. Ne risultano 2030 saggi e 898 articoli. Per i dettagli e per le fasi della ricerca, vedere l'*Introduzione* in questo volume e il capitolo 3.

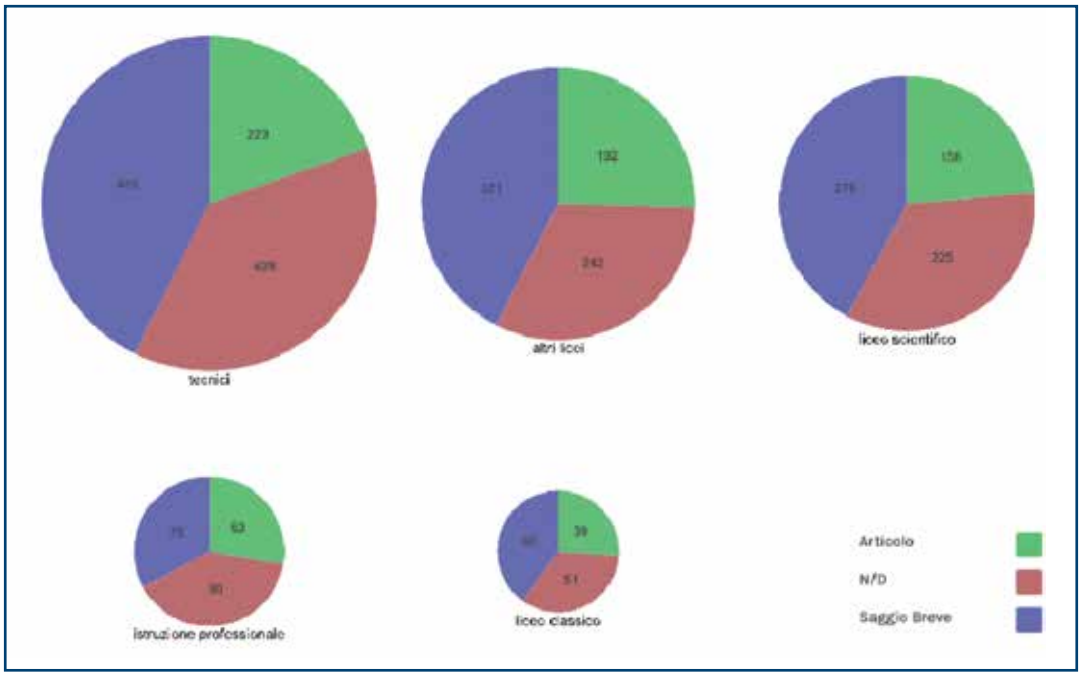

Figura 2: Distribuzione della tipologia di tema tra i diversi tipi di istituto scolastico

Nella Tabella 6 viene riportata un'analisi simile alla precedente, riguardante la distribuzione dei temi tra le tipologie di *Saggio Breve* e *Articolo* nei diversi anni presi in analisi. Anche in questo caso viene confermata la preferenza per il saggio breve, con un divario rispetto a chi sceglie l'articolo che aumenta progressivamente nel corso degli anni (da uno scarto di 52 scritti nel 2000-2001 ai 145 del 2015-2016). L'aumento del divario tra la scelta di saggio breve e articolo è ancora più visibile osservando la Figura 3. Resta presente come nelle precedenti analisi una porzione di temi nei quali l'informazione sulla tipologia non è stata specificata.

Anno Scolastico	Tipologia Tema		
	Saggio Breve	Articolo	N/D
2000-01	192	140	161
03-04	150	121	207
06-07	183	83	216
09-10	225	118	147
12-13	250	115	149
15-16	242	97	157

Tabella 6: Distribuzione della tipologia di tema nel corso degli anni

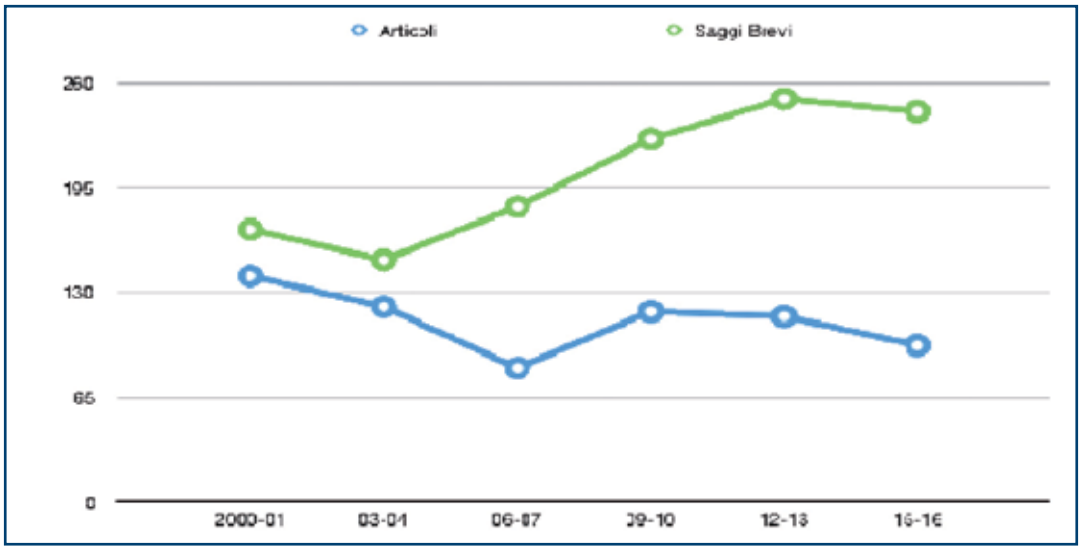

Figura 3: Distribuzione della tipologia di tema nel corso degli anni

Si riportano infine in Tabella 7 due misure, raggruppate per tipologia di scritto e anno scolastico. Il primo valore indica il numero totale di token presenti per le varie categorie, mentre il secondo, tra parentesi, rappresenta la lunghezza media dei singoli temi. Si può notare come non sia presente un andamento crescente o decrescente relativo alla lunghezza degli elaborati nel corso degli anni, sebbene in alcuni le produzioni scritte risultino avere una lunghezza sensibilmente superiore alla media. Allo stesso modo risulta minimo anche il divario di lunghezza tra *saggi brevi* e *articoli*.

	2000-01	03-04	06-07	09-10	12-13	15-16	Totale
Token Articoli	92.454 (660)	85.869 (710)	57.405 (692)	79.780 (676)	76.597 (666)	62.801 (647)	454.906 (675)
Token Saggi Brevi	121.087 (631)	116.599 (777)	136.426 (745)	152.727 (679)	180.823 (723)	167.914 (694)	875.576 (705)
Token N/D	106.666 (663)	152.569 (737)	151.459 (701)	104.363 (710)	97.721 (656)	102.780 (655)	715.558 (690)
Totale Token	320.207 (650)	355.037 (743)	345.290 (716)	336.870 (687)	355.141 (691)	333.495 (672)	2046040 (693)

Tabella 7: Numero di *token* in tutti gli elaborati per anno e per tipologia di tema. Il valore tra parentesi riporta la lunghezza media degli elaborati.

Bibliografia

Bartalesi Lenzi, V., Moretti, G., & Sprugnoli, R. (2012). *CAT: the CELCT Annotation Tool.* Negli Atti di *LREC 2012* (pp. 333-338).

Manning, C., Surdenau, M., Bauer, J., Finkel, J., Bethard, S., McClosky (2014). *The Stanford CoreNLP Natural Language Processing Toolkit.* Negli Atti di ACL 2014 (pp. 55-60).

Palmero Aprosio, A. & Moretti, G. (2018). *Tint 2.0: An All-Inclusive Suite for NLP in Italian.* Negli Atti di *CLIC-it 2018*.

Talamo, L., Celata, C., & Bertinetto, P. M. (2016). *DerIvaTario: An annotated lexicon of Italian derivatives.* In *Word Structure*, *9*(1), 72-102.