
UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

# Analisi linguistico-computazionale del linguaggio di Twitter, un confronto interno ed esterno.

**Candidato:** *Chiara Scarpitta*

**Relatore:** *Felice Dell'Orletta*

**Correlatore:** *Mirko Tavosanis*

Anno Accademico 2014-2015

## Sommario

# 1. Introduzione

L'ampia diffusione di Internet ha rilanciato l'uso della lingua scritta, creando nuove occasioni per disporre e fruire di testi sempre più brevi e vicini a uno scambio di battute tra utenti simile a quello che si avrebbe nel parlato, anche se questo fenomeno di fatto si osserva solo in determinati settori e fenomeni specifici. In particolare, l'allontanamento *dell'italiano digitato*[1] dall'italiano scritto si ha nelle *chat*, dove lo scambio di battute è rapido, e nelle reti sociali (o *social network*), servizi web che permettono agli utenti di condividere brevi messaggi contenenti informazioni e messaggi personali. La scrittura, in questi settori, si è dotata di funzioni informali in precedenza svolte solo dal parlato, attraverso il tentativo degli utenti di riprodurre tratti del parlato con gli strumenti messi a disposizione dallo scritto, ad esempio sfruttando tratti grafici a fini espressivi, come accade con le *emoticons*[2], nel tentativo di rimpiazzare l'intonazione, i segni paralinguistici che accompagnano i discorsi e i tratti indicali che identificano il singolo parlante.

Mirko Tavosanis, ne *L'italiano del web* (2012), individua tre macro-categorie di generi testuali presenti nel Web 2.0 che fanno uso di un linguaggio più vicino a quello parlato che a quello scritto: i *blog* , i *forum* e le reti sociali. Tutte e tre le categorie sono accomunate dall'essere (quasi sempre) luoghi di ritrovo per persone con interessi affini che scrivono testi "linguisticamente poco sorvegliati"[3].

[1] Per *italiano digitato* o *e-italiano*, Giuseppe Antonelli intende la varietà diamesica che nasce dalla comunicazione telematica, per la quale identifica come caratteristiche peculiari l'imitazione del parlato e la telecompresenza.

[2] Le *emoticons* sono anche dette "smiley" (faccine sorridenti) perché la più usata è quella che rappresenta un sorriso, formata da due punti, trattino e parentesi chiusa :-). Vengono utilizzate per esprimere stati d'animo ed emozioni; il termine "emoticon" nasce infatti dalla fusione tra "emotional" (emotiva) e "icon" (icona), quindi "icona emotiva", usata con lo scopo di simulare con la punteggiatura un volto umano.

Le *emoticons* possono essere distinte in *emoticons* occidentali ed *emoticons* orientali o giapponesi. Le *emoticons* occidentali vengono lette orizzontalmente da sinistra verso destra, ad esempio ;-), mentre le *emoticons* orientali vengono lette verticalmente, con gli occhi in alto e la bocca in basso, come O_O.

[3] Tavosanis, 2011, p. 148

Il termine *blog* è stato coniato nel 1999 da Peter Merholz, per abbreviazione di *weblog* ("diario di rete"), nome con cui John Barger definì nel 1997 la lista di link presente nel suo sito.

Nella loro prima fase di vita, i *blog* erano infatti liste di link commentati e recentemente aggiornati ed hanno goduto di molta fortuna comunicativa tra il 2002 e il 2007. Godono tutt'ora di buona salute quei *blog* che difficilmente potrebbero comunicare in altro modo i loro contenuti, ad esempio attraverso le reti sociali; si parla di *blog* che condividono esperienze personali o condividono informazioni pratiche e teoriche. Attualmente, il panorama dei *blog* offre varie sottocategorie, tra cui il *blog* tematico, il diario e il *blog* letterario. Il registro linguistico è variabile in base al tipo di sottocategoria di appartenenza; ad esempio, i *blog* di attualità presentano un linguaggio vicino a quello giornalistico, mentre i diari divergono maggiormente dall'italiano standard.

I *forum* sono gruppi di discussione in cui gli utenti si scambiano informazioni su argomenti più o meno generici ma solitamente a carattere tematico, avviando discussioni che possono essere regolamentate dai moderatori. Il prodotto dei *forum* è collettivo, non essendo presente un singolo autore; questo porta gli utenti ad adeguarsi al linguaggio specifico del *forum*. Spesso gli interventi dei visitatori presentano l'uso di *emoticons*, parolacce o sostituzioni eufemistiche di esse, e si riscontrano errori di battitura dovuti alla rapidità di scrittura, abbreviazioni, uso non standard delle maiuscole, della punteggiatura e della sintassi, parole straniere non adattate, in particolare dall'inglese.

Le reti sociali, o *social network*, nacquero nel 2002 con Friendster (attualmente inattiva). Nel 2004 nacque Facebook, creata da Mark Zuckenberg per gli studenti di Harvard, considerata attualmente la principale rete sociale. Una quantità sempre crescente di persone fa uso di questi servizi, che permettono di rimanere in contatto con altri utenti, condividendo informazioni personali quali foto, pensieri o anche soltanto informazioni meno riservate. Un'altra rete sociale che si sta diffondendo anche in Italia negli ultimi anni, anche se non con l'intensità di Facebook, è Twitter. La concentrazione del presente contributo verterà sul linguaggio degli utenti di Twitter.

Twitter è un servizio lanciato nel 2006 da Jack Dorsery, *social network* e piattaforma di micro blogging che fornisce agli utenti registrati la possibilità di registrarsi con un

proprio *account* e di scrivere un breve messaggio di testo, chiamato in gergo *tweet*, "cinguettio", composto da un massimo di 140 caratteri, che "rappresenta probabilmente la forma più estrema del parlar spedito[4]".

Twitter è strettamente integrato con tutti i dispositivi mobili, telefono incluso, ma le interazioni possibili sono limitate: l'utente può solo scegliere se seguire i *tweets* inviati da altri utenti, rispondere ad essi o rilanciare (*ritweettare*) il *tweet*.

La comunicazione tra utenti di Twitter (definiti da loro stessi *twitteri*) si fonda su convenzioni veloci da codificare per l'utente abituale, costituendo una sorta di grammatica, della quale gli elementi caratterizzanti sono *retweet*, *@risposta*, *@menzione*, *hashtag*.

*Ritwettare* significa rilanciare un *tweet* scritto da un altro utente perché lo si ritiene interessante, in modo da aumentarne la diffusione; rispondere a un *tweet* implica invece il riferimento specifico all'utente destinatario attraverso la formula *@nomeutente*, che condivide con la menzione, usata per citare il singolo utente o segnalare qualcosa ad un gruppo.


**Figura 1 Esempio di *retweet*; in questo caso, l'utente Alessandra Sciullo ha rilanciato un *tweet* dell'utente Lilli Mandara.**

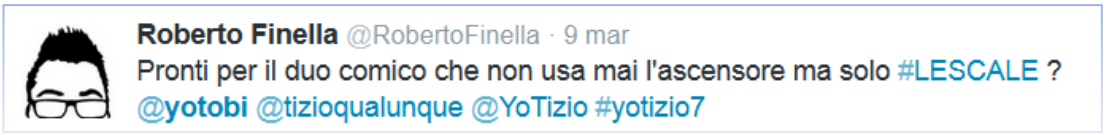


**Figura 2 Esempio di menzione; l'utente Roberto Finella ha menzionato con il segno "@" gli utenti yotobi, tizioqualunque e YoTizio.**


**Figura 3 Esempio di risposta; l'utente Marco Gattuso risponde all'utente yotobi tramite il segno "@".**

[4] Cfr. Roncaglia, 2010, p. 5.

Ad esempio, in Figura 1 si mostra un esempio di *retweet*; ovvero l'utente Alessandra Sciullo ha *ritwittato* un *tweet* dell'utente Lilli Mandara. In Figura 2 un *tweet* mostra l'uso della menzione, usata dall'utente Roberto Finella in riferimento agli utenti yotobi, tizioqualunque e YoTizio attraverso il simbolo "@", scrivendo infatti "@yotobi", "@tizioqualunque" e "@YoTizio".
L'esempio in Figura 3 mostra un *tweet* facente parte di una conversazione: si nota che l'utente Marco Gattuso risponde all'utente yotobi scrivendo "@yotobi".

L'*hashtag*, di cui è mostrato un esempio in Figura 4, è una parola chiave preceduta da "#", ad esempio "#terremoto", divenuta un link di rimando ad una pagina che contiene soltanto i *tweets* con quell'*hasthag*.
Se l'*hashtag* è composto da più parole, esse vengono scritte di filato, senza tradizionali trattini separatori ("-").
Un approfondimento sull'uso dell'*hashtag* è stato effettuato da Francesca Chiusaroli, che lo definisce parte de "la struttura artefatta del testi di Twitter rispetto alla scrittura ordinaria e convenzionale, poiché la stringa frasale risulta concretamente alterata da tali figure tradizionalmente non contemplate nelle regole ortografiche della lingua standard[5]". L'*hashtag* viene classificato come esempio di "scrittura breve[6]", indispensabile per l'innesco della conversazione per la sua capacità, sopracitata, di generare *link* in cui sono raccolti tutti i *tweets* caratterizzati dall'*hashtag* in questione.

[5] Cfr. Chiusaroli, 2014, p. 117.

[6] Francesca Chiusaroli gestisce un sito, visitabile all'indirizzo http://www.scritturebrevi.it/, in cui definisce la categoria di "Scritture Brevi", in cui è incluso anche l'*hashtag*: "L'etichetta "Scritture Brevi" è proposta come categoria concettuale e metalinguistica per la classificazione di forme grafiche come abbreviazioni, acronimi, sigle, punteggiatura, segni, icone, indici e simboli, elementi figurativi, espressioni testuali e codici visivi per i quali risulti dirimente il principio della "brevità" connesso al criterio dell'"economia".
In particolare sono comprese nella categoria "scritture brevi" tutte le manifestazioni grafiche che, nella dimensione sintagmatica, si sottraggono al principio della linearità del significante, alterano le regole morfotattiche convenzionali della lingua scritta, e intervengono nella costruzione del messaggio nei termini di "riduzione, contenimento, sintesi" indotti dai supporti e dai contesti.
La categoria ha applicazione nella sincronia e nella diacronia linguistica, nei sistemi standard e non standard, negli ambiti generali e specialistici."

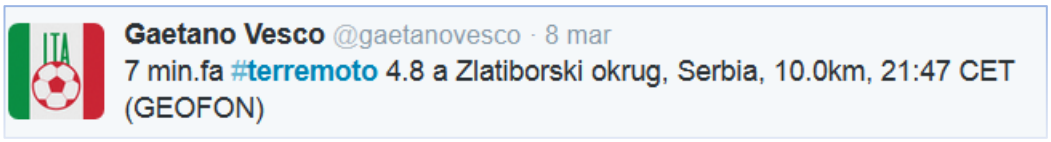


**Figura 4 Esempio di *hashtag*, in questo caso *#terremoto*.**

Diversi studi sono stati effettuati sull'uso di Twitter, sia da parte della psicologia e della sociologia che della linguistica. Tra i contributi che hanno sondato i *tweets* dal punto di vista linguistico, ricordiamo la tesi di laurea triennale di Cristina Zaga "Twitter un'analisi dell'italiano nel *microblogging*", in cui Twitter viene valutato dal punto di vista qualitativo, suddividendo i *tweets* in categorie di appartenenza ed esaminando le deviazioni dalla grammatica standard che avvicinano il linguaggio di Twitter al parlato. Inoltre, vengono osservati i meccanismi di alterazione presenti nella scrittura; tra di essi ricordiamo le abbreviazioni, le variazioni grafiche, l'uso non standard delle maiuscole e della punteggiatura e l'uso delle *emoticons*.

In questa sede verrà effettuato un confronto tra il linguaggio di testi giornalistici su argomenti generici che fa uso di grammatica standard e il linguaggio di Twitter.
Inoltre, verrà effettuata un'ulteriore distinzione all'interno di Twitter tra messaggi scritti in momenti casuali e scritti durante una catastrofe naturale (ad esempio il terremoto de L'Aquila del 2009) che descrivono o trattano dei danni causati da tali eventi, quali danneggiamenti a persone, animali, edifici, confrontando due modi diversi di usare Twitter.
E' la prima volta che viene affrontata la questione di come Twitter possa essere un mezzo che cambia il modo d'uso in base allo scopo per cui viene usato, nonostante il vincolo tecnico che impone ai messaggi un limite di 140 caratteri. Nonostante questo vincolo, si vuole osservare il cambiamento del registro linguistico a seconda delle occasioni in cui Twitter viene utilizzato.

Nel capitolo 2 verranno descritti nel dettaglio i *corpora* usati per l'analisi testuale, nel capitolo 3 si parlerà delle tecnologie linguistiche attualmente a disposizione per effettuare analisi linguistiche, nel capitolo 4 verranno confrontati i risultati delle analisi linguistiche dei *corpora*.
Se l'utente, con l'abitudine, riesce a comprendere l'uso di *hashtag*, menzioni, risposte, link esterni, *retweet*, ciò diventa problematico per uno strumento di analisi

che deve analizzare il testo, pertanto nel capitolo 5 verranno discusse le prestazioni del sistema per l'annotazione, addestrato su *corpora* differenti rispetto ai due *corpora* di *tweets* analizzati in questa sede, nel capitolo 6 si trarranno le conclusioni dalle osservazioni effettuate. Il capitolo 7 conterrà la bibliografia e il capitolo 8 una piccola appendice riguardante i dati utilizzati per le analisi.

# 2. Descrizione dei *corpora*

Un *corpus* è una collezione di testi selezionati e organizzati in maniera tale da soddisfare specifici criteri che li rendono funzionali per le analisi linguistiche.[7]

I *corpora* testuali rappresentano la principale fonte di dati in linguistica computazionale; la creazione di un *corpus* funzionale è stata resa più semplice grazie allo sviluppo della tecnologia informatica, che permette di creare *corpora* sempre più grandi, ottimizzare la ricerca di dati linguistici interessanti e sviluppare modelli computazionali della lingua.

Ogni *corpus* proviene da un'opera di selezione di contenuti appropriati per la successiva analisi linguistica; tale selezione viene effettuata sulla base di alcuni parametri: generalità, modalità, cronologia, lingua, integrità dei testi e codifica dei testi.

Il grado di generalità dipende da quanto i testi che compongono il *corpus* sono trasversali rispetto a diverse varietà di una lingua, distinguendo tra *corpora* specialistici, che si occupano di una specifica varietà linguistica, o generali, che tentano di descrivere una lingua nel suo complesso.

La modalità distingue in base all'appartenenza dei testi alla lingua scritta, parlata o a entrambe in proporzioni variabili.

La cronologia riguarda l'appartenenza dei testi alla stessa finestra temporale (sincronia) o a periodi diversi (diacronia), a seconda che lo scopo sia monitorare una fase o più fasi di una lingua.

La lingua di un *corpus* può essere la stessa per tutti i testi, generando *corpora* monolingue, oppure possono essere presenti due o più lingue, come accade nei *corpora* bilingue e multilingue.

Per integrità dei testi si intende se il *corpus* contiene testi interi o porzioni di essi.

La codifica digitale dei testi contraddistingue i *corpora* digitali in base al grado di annotazione.

In questa sede verranno analizzati tre *corpora*: due costituiti da *tweets* e uno rappresentante il linguaggio giornalistico estratto da *La Repubblica*.

[7] Lenci, Montemagni, Pirrelli, 2005, p. 26

I corpora di *tweets* sono *TOrdinario*, composto da 5239 *tweets* estratti in maniera casuale, e *TCatastrofe*, composto da 7645 *tweets* scritti durante eventi catastrofici. Sia *TOrdinario* che *TCatastrofe* sono in lingua italiana ed appartenenti ad una fascia temporale ridotta. Per rendere consistenti i confronti abbiamo creato *corpora* che hanno lo stesso numero di *tokens*[8]; ovvero *TOrdinario* contiene 65667 tokens, mentre *TCatastrofe* 65675.
Entrambi i *corpora* sono da considerarsi specialistici, sincronici, monolingue; specialistici perché si occupano di una precisa varietà linguistica (il linguaggio di Twitter), sincronici poiché i testi che ne fanno parte appartengono ad una ristretta finestra temporale, monolingue perché interamente in italiano. I *corpora* sono inoltre composti esclusivamente da testi appartenenti alla lingua scritta.

L'altro *corpus*, che chiameremo *Rep*, ha le stesse dimensioni in termini di *tokens* dei primi due *corpora* (ovvero contiene 65688 *tokens*) ed è stato estratto dal *corpus La Repubblica*, *corpus* in lingua italiana che include le annate de "La Repubblica" dal 1985 al 2000. Contiene circa 175 milioni di parole ma è in via di ampliamento fino a 400 milioni; è codificato in XML secondo lo standard TEI e annotato a livello morfosintattico.
Anche *Rep*, come *TOrdinario* e *TCatastrofico*, è specialistico, sincronico, e monolingue.

In questa sede verranno effettuati due confronti tra i tre *corpora*: il primo è da considerarsi un confronto "esterno", poiché prevede il paragone tra i risultati dell'analisi linguistica di *Rep* e di *TOrdinario*, composti da testi appartenenti ad ambiti differenti. Il secondo è invece un confronto "interno" a due varietà di linguaggio presenti in Twitter, ovvero il linguaggio usato correntemente, *TOrdinario,* e quello caratteristico dei periodi di allerta, *TCatastrofe*.

[8] Per quanto la nozione di *token* includa quella di parola, il *token* non si identifica solo con essa: limitatamente a quanto accade nelle lingua con ortografia segmentata come l'italiano, il *token* è una sequenza di caratteri delimitata sa spazi, pur con le dovute eccezioni date da punteggiatura (non preceduti da spazi), acronimi (U.S.A.) , abbreviazioni (Sig. Rossi), cifre (9,45), date (25-12-2014), indirizzi internet o e-mail (www.google.it), *tokens* graficamente complessi (La Spezia).

Se i testi che compongono *Rep* e *TOrdinario* sono stati estratti casualmente, diverso è il discorso per i *tweets* che compongono *TCatastrofe*.

Il *corpus* rappresenta una collezione di *tweets* scritti nel periodo immediatamente successivo ai disastri naturali di cui parlano. L'estrazione dei *tweets* è stata effettuata grazie a "The Streaming APIs[9]" per gli eventi più recenti, a "Historical Twitter Data Access[10]" per quelli avventi da più tempo.

I *tweets* sono stati filtrati attraverso delle parole chiave, come "terremoto[11]" e "scossa", e degli *#hastag* specifici dell'avvenimento, ad esempio *#allertameteoSAR*[12].

[9] https://dev.twitter.com/streaming/overview

[10] https://gnip.com/sources/twitter/historical/

[11] Uno dei terremoti a cui si fa riferimento è quello che colpì L'Aquila il 6 aprile del 2009.

[12] Il *tag* fa riferimento all'alluvione che colpì la Sardegna nel novembre del 2013; su Twitter, molti dei messaggi che ne parlavano furono contraddistinti *da*ll'*hastag #allertameteoSAR.*

# 3. Tecnologie linguistiche

Le tecnologie linguistico-computazionali permettono l'accesso al contenuto informativo dei testi attraverso l'individuazione della struttura linguistica sottostante e la sua rappresentazione esplicita.

Gli strumenti che abbiamo analizzato all'interno di questa tesi sono LinguA[13], (Linguistic Annotation Pipeline ), una catena di strumenti statistici di Trattamento Automatico del Linguaggio sviluppati dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa che permette l'annotazione automatica del testo, e READ-IT[14], che permette di creare il profilo linguistico dello stesso e definire la leggibilità sulla base di parametri linguistici complessi estratti dallo strumento che spaziano tra i vari livelli di analisi linguistica effettuati da LinguA. Tali strumenti rappresentano lo stato dell'arte per la lingua italiana, essendo i più precisi e affidabili[15].

READ-IT è stato sviluppato dall'*Italian Natural Language Processing Laboratory* (ItaliaNLP Lab)[16] dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa.

READ-IT è in grado di implementare un indice di leggibilità avanzato basato su analisti linguistica multi-livello; in particolare, è basato su una combinazione di tratti linguistici che spaziano tra più livelli di descrizione linguistica: lessicale, morfo-sintattico e sintattico. In base a tali caratteristiche linguistiche è in grado di calcolare la leggibilità dei testi classificandoli come testi di *facile* o *difficile* lettura, in base ad un classificatore statistico che associa i testi in input a due classi di lettura definite a priori. Le due classi sono state formate attraverso l'addestramento su due *corpora* rappresentativi di testi complessi e semplificati appartenenti al genere testuale di

[13] Una demo di LinguA è disponibile all'indirizzo http://linguistic-annotation-tool.italianlp.it/

[14] Una demo on-line di READ-IT è disponibile alla pagina http://www.italianlp.it/demo/read-it/

[15] Gli strumenti di cui fa uso Linguistic Annotation Pipeline sono risultati i più precisi e affidabili nell'ambito delle campagne di valutazione di strumenti per l'analisi automatica dell'italiano, EVALITA. (http://www.evalita.it)

[16] www.italianlp.it

prosa giornalistica, ovvero *La Repubblica* (Rep) e *Due Parole* (2Par)[17], l'uno rappresentativo della *difficile* lettura e l'altro della *facile*.

L'appartenenza ad una delle due classi è stabilita in base al grado di similarità tra la distribuzione di alcune delle caratteristiche linguistiche monitorate.

I corpus *Rep*, *TOrdinario* e *TCatastrofe* verranno presi come testi in input e analizzati tramite READ-IT e confrontati in base ad alcuni parametri, ovvero il profilo di base, il profilo lessicale, il profilo sintattico e infine l'analisi globale della leggibilità.

Descriveremo ora cosa si intende per annotazione linguistica automatica, creazione del profilo linguistico di un testo e di analisi della leggibilità.

## 3.1 Analisi linguistica automatica

L'identificazione della struttura linguistica del testo avviene in modo incrementale attraverso analisi linguistiche a livelli di complessità crescente. Inizialmente il testo viene *tokenizzato*[18], ovvero segmentato in *tokens*, poi viene analizzato morfo-sintatticamente[19] (*POS-tagging*) e lemmatizzato[20]. Infine viene analizzata la struttura sintattica in base alle relazioni di dipendenza (*parsing* a dipendenze).

[17] I *corpora La Repubblica* e *DueParole* hanno svolto la funzione di *corpus* di addestramento, o *training corpus*, perché i dati estratti da essi sono stati usati per addestrare il modello, ovvero stimare la probabilità del verificarsi degli eventi prodotti dal sistema. Attraverso essi, READ-IT è stato reso capace di annotare nella maniera (più possibile) corretta *corpora* successivi.

[18] Il processo di segmentazione del testo in *token* è detto "tokenizzazione".

[19] L'annotazione morfo-sintattica consiste nell'assegnazione a ogni parola (o *token*) del testo l'informazione relativa alla categoria grammaticale della parola nel contesto specifico. Ad essa si aggiunge la lemmatizzazione.

[20] La lemmatizzazione consiste nel ricondurre ogni parola del testo al proprio lemma, ovvero ossia quella parola che per convenzione è scelta per rappresentare tutte le forme di una flessione.

| | ID | Token | Lemma | C-POS | F-POS | Morphosyntactic feats | HEAD | DEP |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Le | il | R | RD | num=p\|gen=f | 2 | det |
| | 2 | immagini | immagine | S | S | num=p\|gen=f | 7 | subj |
| | 3 | del | di | E | EA | num=s\|gen=m | 2 | comp |
| | 4 | maltempo | maltempo | S | S | num=s\|gen=m | 3 | prep |
| | 5 | in | in | E | E | _ | 4 | comp_loc |
| | 6 | Sardegna | Sardegna | S | SP | _ | 5 | prep |
| | 7 | fanno | fare | V | V | num=p\|per=3\|mod=i\|ten=p | 0 | ROOT |
| | 8 | male | male | B | B | _ | 7 | mod |
| | 9 | al | a | E | EA | num=s\|gen=m | 7 | comp |
| | 10 | cuore | cuore | S | S | num=s\|gen=m | 9 | prep |
| | 11 | . | . | F | FS | _ | 7 | punc |

**Figura 5: Esempio di rappresentazione tabellare di un *tweet* annotato linguisticamente**

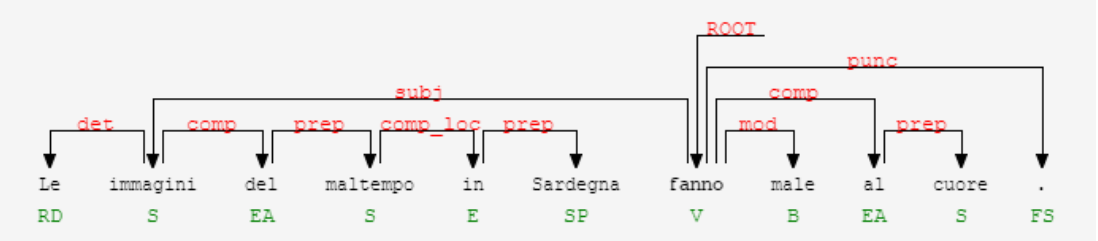


**Figura 6: Rappresentazione grafica dell'annotazione linguistica dell'esempio in Figura 5 attraverso l'albero sintattico**

In Figura 5 viene mostrato il risultato del processo di annotazione linguistica del *tweet* "Le immagini della Sardegna fanno male al cuore.", effettuato grazie allo strumento "LinguA". Il testo è stato tokenizzato (colonna *Token*), ovvero segmentato in parole ortografiche, poi è stato analizzato morfo-sintatticamente ed è stato lemmatizzato (colonna *Lemma*), cioè è stata assegnata a ciascun *token* la propria categoria grammaticale nel contesto (colonna *C-POS*[21] e *F-POS*[22]) ed è stato assegnato ad ogni forma il lemma corrispondente.

Tale informazione è integrata da ulteriori specificazioni morfologiche nella colonna *Morphosyntactic feats* (riguardanti ad esempio categorie flessionali come persona, genere, numero, ecc.). Infine il testo viene annotato sintatticamente (colonna *HEAD* e *DEP*) e vengono fornite le informazioni riguardo le relazioni di dipendenza tra parole, in genere relazioni binarie asimmetriche tra una testa e un dipendente, come "soggetto", "oggetto diretto", "modificatore", eccetera. In particolare, la colonna

[21] Nella colonna C-POS, V = verbo; R = articolo; S = sostantivo, A = aggettivo, E= preposizione.

[22] Nella colonna F-POS vengono identificate le eventuali sottocategorie di C-POS, ad esempio EA = preposizione articolata, RD-RI = articolo definito-indefiniti, SP = nome proprio.

*HEAD* riporta l'identificatore univoco della forma che costituisce la testa da cui dipende (0 per il verbo della proposizione principale, radice dell'albero sintattico (contrassegnato in Figura 2 dal tag "ROOT" assegnato al verbo "fanno"), mentre la colonna *DEP* specifica il tipo di dipendenza.
In Figura 6 viene mostrato che la parola *immagini* è il soggetto del verbo *fanno* e che *male* è un modificatore del verbo. La Figura 6 è una rappresentazione grafica dell'albero di dipendenze sintattiche esplicitate nella Figura 5 dalle colonne *HEAD* e *DEP*.

### 3.1.2 Profilo linguistico del testo: profilo di base, profilo lessicale, profilo sintattico

Lo strumento READ-IT permette di fare il monitoraggio linguistico di un documento permettendo di definirne il profilo linguistico a tre livelli: il profilo di base, il profilo lessicale e il profilo sintattico.
Il profilo di base tiene conto del numero totale di periodi in cui si articola il testo, del numero totale di parole (*tokens*), della lunghezza media dei periodi in termini di *tokens* e della lunghezza media delle parole in caratteri.
Il profilo lessicale prende in considerazione l'insieme delle parole tipo[23] ricorrenti nel documento, confrontandone i lemmi con quelli appartenenti al *vocabolario di base* costruito sulla base dizionario di riferimento *Grande Dizionario italiano dell'uso* (GRADIT, De Mauro, 2000).
Ulteriori informazioni sulla composizione del vocabolario si ricavano dall'analisi della ripartizione della porzione del vocabolario riconducibile al vocabolario di base rispetto ai repertori d'uso quali il repertorio fondamentale (che include 2000 parole conosciute e usate da chi ha almeno la licenza elementare), il repertorio ad alto uso (3000 parole conosciute e usate da chi ha almeno la licenza media inferiore) ed il repertorio ad alta disponibilità (contenente 2000 parole presenti nell'uso comune ma usate solo all'occorrenza dai parlanti).

[23] Per parola tipo di un'unità si intende il valore dell'unità normalizzato. Per (parole) unità si intendono forme grafiche dello stesso tipo indistinguibili a prescindere dalla posizione che occupano nel testo.

Un altro indice fornito dal profilo lessicale è il *rapporto tipo/unità*[24] (calcolato rispetto alle prime 100 parole del testo), ovvero il numero di parole tipo diviso il numero di *tokens* che compongono il testo. Il risultato è compreso tra 0 e 1, dove la maggior vicinanza a 0 implica un vocabolario meno vario e la maggior vicinanza a 1 indica una maggior ricchezza lessicale.

Infine il profilo lessicale fornisce il valore della densità lessicale, ovvero il rapporto tra parole semanticamente piene[25] rispetto al totale di occorrenze di parola. Più alto è il valore, maggiore è la leggibilità.

Il profilo sintattico fa una distinzione tra analisi morfo-sintattica e sintattica.

L'analisi morfo-sintattica effettua delle misurazioni sulle categorie grammaticali, calcolando la percentuale sostantivi (e nomi propri), di aggettivi, di verbi e di congiunzioni, distinguendo tra congiunzioni coordinanti e subordinanti. La distribuzione delle categorie grammaticali è un parametro della misurazione della leggibilità del testo; ad esempio, un alto numero di costruzioni ipotattiche è correlato con una complessità maggiore.

L'analisi sintattica analizza la struttura sintattica del testo, attraverso il numero medio di proposizioni per periodo e la percentuale di coordinate e subordinate; all'aumentare delle proposizioni per periodo, in particolare se si tratta di subordinate, si ha un aumento della complessità del testo. Si ha inoltre il numero medio di parole per proposizione e il numero medio di dipendenti per testa verbale. Riguardo l'albero sintattico (vedi Figura 6), viene fornita la media delle altezze massime, la profondità media delle strutture nominali complesse e delle catene di subordinazione. Infine, sono disponibili anche la media tra le distanze in parole tra testa e dipendente e la media tra le massime distanze tra testa e dipendente, valori che aumentando indicano una complessità testuale crescente.

[24] Il *rapporto tipo/unità* viene spesso chiamato *Type/Token Ratio*, o TTR. La formula è $|V_T|/|T|$, in cui $|V_T|$ indica il numero di parole tipo presenti in $|T|$, che invece rappresenta il numero di parole unità nel testo.

[25] Per parole semanticamente piene si intendono tutte quelle parole che portano significato, quali nomi, verbi, aggettivi, avverbi, mentre le parole funzionali o vuote si intendono articoli, congiunzioni, pronomi, preposizioni.

| Tipo di caratteristica | Livello di annotazione linguistica | Caratteristica |
|---|---|---|
| Di base | Divisione in frasi | Lunghezza media dei periodi e delle parole |
| Lessicale | Lemmatizzazione e annotazione morfo-sintattica | Percentuale di lemmi appartenenti al *Vocabolario di Base* del *Grande dizionario italiano dell'uso* (De Mauro, 2000) |
| | | Distribuzione dei lemmi rispetto ai repertori di uso (Fondamentale, Alto uso, Alta disponibilità) |
| Morfo-sintattico | Annotazione morfo-sintattica | Distribuzione delle categorie morfo-sintattiche |
| | | Densità lessicale |
| Sintattico | Annotazione sintattica a dipendenze | Distribuzione dei vari tipi di relazioni di dipendenza |
| | | Rarità verbale |
| | | Caratteristiche relative alla struttura dell'albero sintattico analizzato:<br>- altezza media dell'intero albero,<br>- lunghezza media della più lunga relazione di dipendenza |
| | | Caratteristiche relative all'uso della subordinazione:<br>- distribuzione di frasi principali vs. subordinate,<br>- lunghezza media di sequenze consecutive di subordinate |
| | | Caratteristiche relative alla modificazione nominale:<br>- lunghezza media dei complementi preposizionali dipendenti in sequenza da un nome |

**Tabella 1: Riassunto delle caratteristiche considerate da READ-IT nella fase di monitoraggio linguistico**

### 3.1.3 Analisi globale della leggibilità

READ-IT sfrutta tutti i parametri misurati per assegnare un punteggio di leggibilità; per analisi globale della leggibilità si intende il risultato dell'analisi condotto in relazione al documento (in questo caso, al *corpus*). Attraverso READ-IT, la valutazione globale della leggibilità viene effettuata sulla base di diverse configurazioni di caratteristiche del testo, tramite l'uso di diversi modelli di analisi della leggibilità, ovvero READ-IT BASE, READ-IT LESSICALE, READ-IT SINTATTICO, READ-IT GLOBALE.

Il modello READ-IT BASE misura la leggibilità del testo attraverso la lunghezza delle frasi (numero medio di parole per frase) e la lunghezza delle parole (numero medio di caratteri per parola).

Il modello READ-IT LESSICALE si focalizza sulla composizione del vocabolario[26] e sulla sua ricchezza lessicale.

Il modello READ-IT SINTATTICO si serve dell'informazione grammaticale, ovvero della combinazione di tratti morfosintattici e sintattici ricavati dai corrispettivi livelli di analisi linguistica.

Il modello READ-IT GLOBALE è un modello che combina tratti di varia natura, come caratteristiche generali del testo (READ-IT BASE) e quelle lessicali e sintattiche (READ-IT LESSICALE e READ-IT SINTATTICO).

Ciascun modello esprime il proprio risultato con una percentuale che esprime il livello di difficoltà; più la percentuale è alta, più è probabile che il testo in considerazione sia di difficile leggibilità.

Un altro indice di leggibilità è fornito dall'indice Gulpease[27], tarato sulla lingua italiana: Gulpease si serve di caratteristiche di base della frase, che sono la lunghezza delle parole in caratteri e il numero di *tokens* per frase. Il suo valore va da 0 a 100, dove 0 indica la leggibilità più bassa e 100 la più alta.

[26] Il vocabolario di un testo è l'insieme di parole tipo che ricorrono in esso

[27] I testi con indice di Gulpease inferiore a 80 sono difficili da leggere per chi ha licenzia elementare, quelli con indice inferiore a 60 per chi ha la licenza media, quelli con indice inferiore a 40 per chi ha un diploma superiore.

# 4. Confronto di *corpora*

In questo paragrafo verrà effettuato il confronto esterno tra il profilo di base, lessicale e sintattico dei *corpora Rep* e *TOrdinario*, dei quali verrà effettuata anche un'analisi globale della leggibilità. In seguito verrà eseguito il confronto interno anche tra i due *corpora* composti da *tweets*, ovvero *TOrdinario* e *TCatastrofe*.

## 4.1 Confronto tra *Repubblica VS Twitter, ovvero Rep VS TOrdinario*

I due *corpora*, l'uno contenente prosa giornalistica e l'altro *tweets* raccolti in momenti casuali, *Rep* e *TOrdinario*, verranno confrontati tramite il profilo base, il profilo lessicale e il profilo sintattico al fine di apprezzarne le differenze e spiegare a cosa possano essere dovute. In seguito, verranno confrontati attraverso la funzione "Analisi globale della leggibilità".

### 4.1.1 Profili: di base, lessicale, sintattico

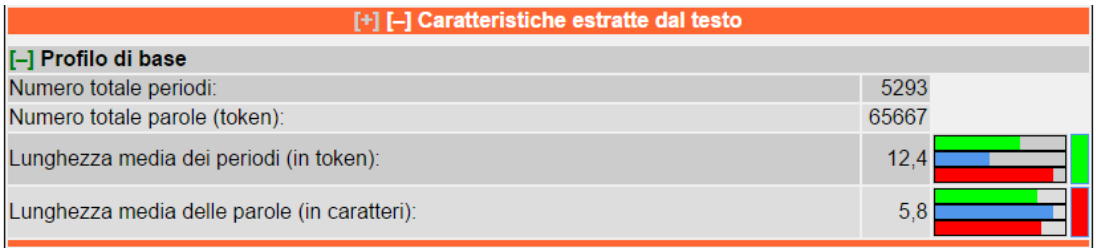


**Figura 7 Esempio di output del sistema del Profilo di base di *TOrdinario***

| **Profilo di base** | **TOrdinario** | **Rep** |
|---|---|---|
| Lunghezza media periodi | 12,4 | 22,9 |
| Lunghezza media parole | 5,8 | 5,0 |

**Tabella 2: Profilo di base di *TOrdinario* e *Rep***

Osservando la Tabella 2, la differenza che salta maggiormente all'occhio è la differenza di lunghezza media dei periodi: *Rep* contiene frasi molto più lunghe, per quanto la lunghezza media delle parole sia di poco inferiore a quella di *TOrdinario*. Questo potrebbe essere dovuto al limite di 140 caratteri imposto a Twitter, che costringe gli utenti a usare frasi brevi. Il quotidiano, invece, non risponde ad alcun tipo di limitazione.

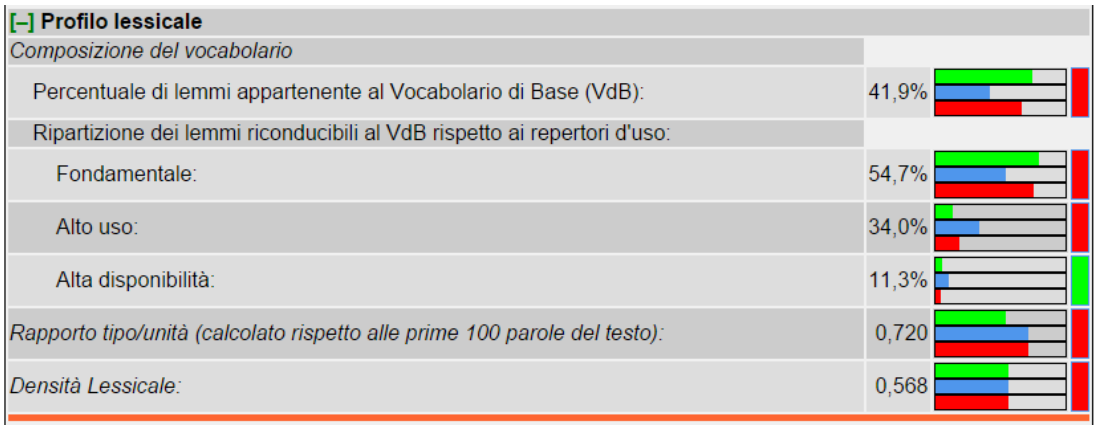


**Figura 8: Esempio di output del sistema del Profilo lessicale di *Rep***

| **Profilo lessicale** | **TOrdinario** | **Rep** |
|---|---|---|
| Percentuale lemmi VdB | 23,2% | 41,9% |
| Repertorio fondamentale | 64,5% | 54,7% |
| Repertorio ad alto uso | 26,2% | 34,0% |
| Repertorio alta disponibilità | 9,3% | 11,3% |
| Rapporto tipo/unità | 0,830 | 0,720 |
| Densità lessicale | 0,581 | 0,568 |

**Tabella 3: Profilo lessicale di *TOrdinario* e *Rep***

Osservando nella Tabella 3 la percentuale di lemmi appartenenti al vocabolario base si nota che questo valore è maggiore nel *corpus Rep*; questo suggerisce che in *TOrdinario* ci sia un'alta percentuale di elementi di uso comune in Twitter (*hashtag*, menzioni, risposte, link esterni) che non fanno parte del vocabolario di base, in quanto non fanno parte della grammatica standard.

In questo modo si abbassa la percentuale di lemmi appartenenti al vocabolario di base, facendo sembrare il testo più complicato, ma osservando i parametri successivi

si nota una discrepanza: infatti la percentuale di lemmi riconducibili al repertorio ad uso fondamentale è più alta in *TOrdinario*, mentre i lemmi riconducibili ai repertorio ad alto uso e ad alta disponibilità sono minori; questo può significare *TOrdinario* usa parole in genere più facili da comprendere, mentre *Rep* fa un discreto ricorso a termini conosciuti da meno persone.

Il rapporto tipo/unità di *Rep* è inferiore rispetto a quello di *TOrdinario*, segno di una minore, se pur di poco, ricchezza lessicale. La densità lessicale leggermente inferiore in *Rep* indica invece una leggibilità più bassa rispetto a quella di *TOrdinario*.

Si deduce che in *Rep* le parole semanticamente piene, pur essendo presenti in quantità leggermente maggiori, non rendono il testo troppo difficile da comprendere, mentre *TOrdinario* le parole semanticamente piene sono presenti in quantità inferiore ma le parole usate sono più varie.

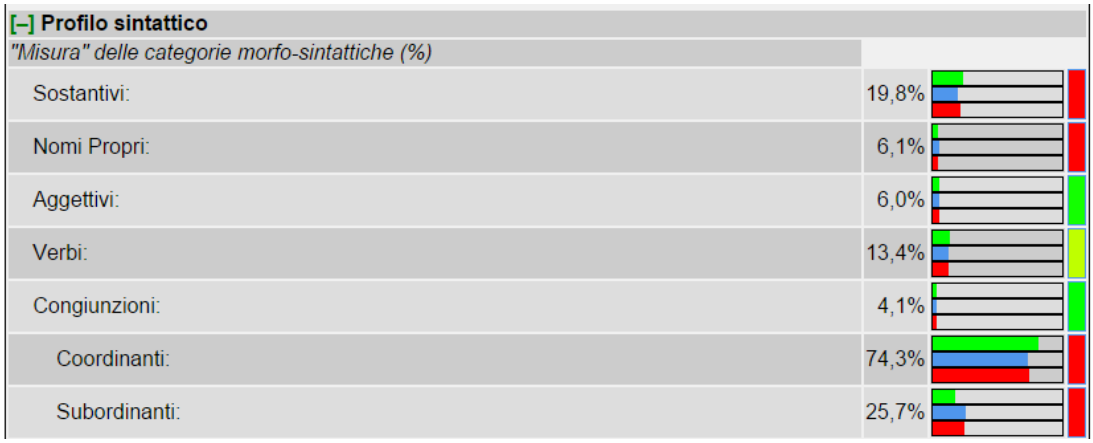


**Figura 9: Esempio di output del sistema del Profilo sintattico di *Rep***

| **Categorie morfosintattiche** | **TOrdinario** | **Rep** |
|---|---|---|
| Sostantivi | 18,1% | 19,8% |
| Nomi propri | 9,7% | 6,1% |
| Aggettivi | 6,3% | 6,0% |
| Verbi | 9,6% | 13,4% |
| Congiunzioni | 2,7% | 4,1% |
| Congiunzioni coordinanti | 73,6% | 74,3% |
| Congiunzioni subordinanti | 26,4% | 25,7% |

**Tabella 4: Profilo sintattico, proprietà morfo-sintattiche in *TOrdinario* e *Rep***

Osservando la Tabella 4, si nota che sia *TOrdinario* che *Rep* presentano un numero di sostantivi piuttosto alto e simile tra di loro, anche se *TOrdinario* contiene una quantità maggiore di nomi propri, dovuti forse alle conversazioni tra utenti, il cui nome viene ripetuto attraverso la menzione, la risposta e il *retweet*. Se in *TOrdinario* si ha un verbo ogni due nomi, questa proporzione è inferiore in *Rep*, in cui la predominanza dei nomi sui verbi non è così evidente. In *Rep* compaiono inoltre più congiunzioni, ma la proporzione tra congiunzioni coordinanti e subordinanti è simile in entrambi i *corpora*, se pur con una leggera maggioranza di congiunzioni coordinanti in *Rep*.

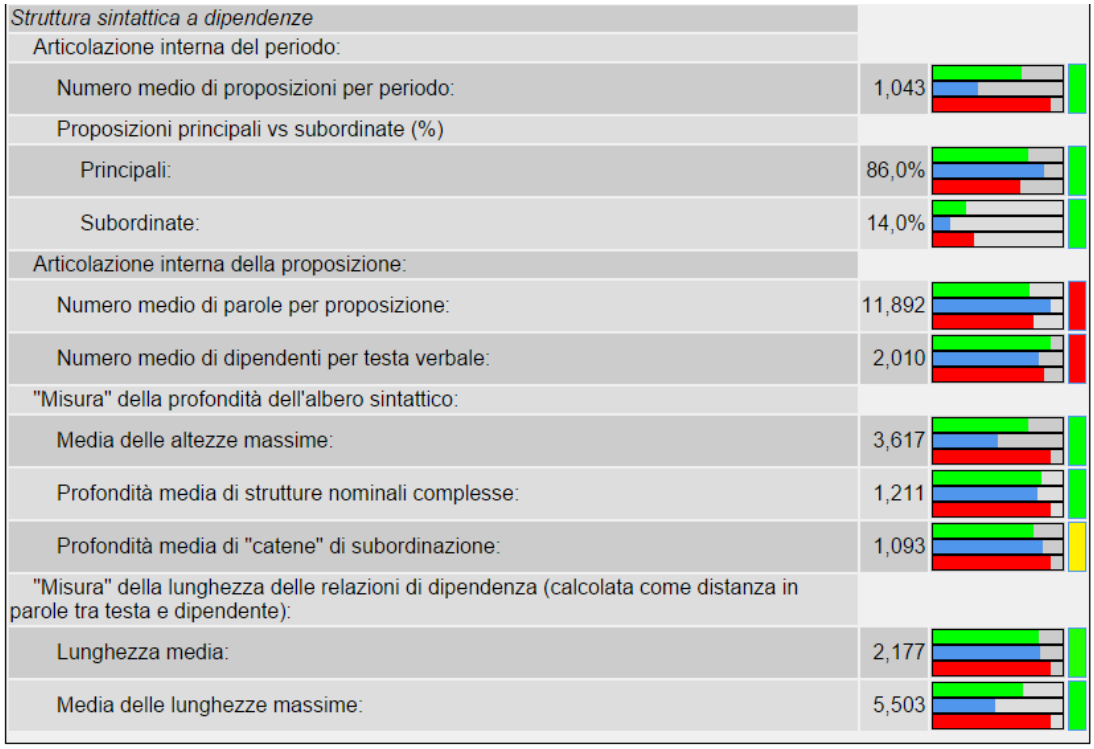


**Figura 10: Esempio di output del sistema del Profilo di base di *TOrdinario***

| **Categorie morfosintattiche** | **TOrdinario** | **Rep** |
|---|---|---|
| Proposizioni per periodo | 1,043 | 2,524 |
| Principali | 86,0% | 68,0% |
| Subordinate | 14,0% | 32,0% |
| Parole per preposizione | 11,892 | 9,080 |
| Dipendenti per testa verbale | 2,010 | 2,0370 |

| | | |
|---|---|---|
| Altezza massima albero sintattico | 3,617 | 5,673 |
| Profondità strutture nominali | 1,211 | 1,276 |
| Profondità catene di subordinazione | 1,093 | 1,233 |
| Lunghezza testa-dipendente | 2,177 | 2,413 |
| Lunghezza massima-testa dipendente | 5,503 | 8,994 |

**Tabella 5: Profilo sintattico, struttura sintattica a dipendenze in *TOrdinario* e *Rep***

Dalla Tabella 5 si evince che la quantità media di proposizioni per periodo è 1.043 in *TOrdinario*, mentre 2.524 in *Rep*; questo è probabilmente dovuto al fatto che su Twitter, quasi sempre, un *tweet* è composto da un'unica proposizione, cosa che invece non accade in *Rep*, in cui si nota un abbondante uso di subordinate rispetto a quello in voga in *TOrdinario*.

La tendenza a condensare i contenuti in una singola proposizione si nota anche dalla distribuzione di parole per preposizione, più alta in *TOrdinario*.

L'altezza media dell'albero sintattico è maggiore in *Rep*, informazione ricollegabile a un più alto numero di subordinate, così come la profondità delle strutture nominali. Anche la profondità delle catene di subordinazione è maggiore in *Rep*, così come la lunghezza- testa dipendente media e massima.

### 4.1.2 Analisi globale della leggibilità

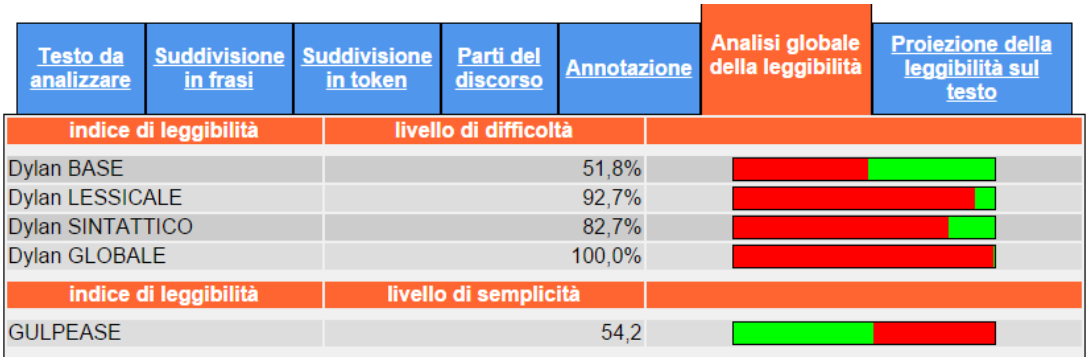


**Figura 7: Esempio di output della valutazione globale dell'analisi della leggibilità in *Rep***

| Indice di leggibilità | TOrdinario | Rep |
|---|---|---|
| READ-IT BASE | 6,3% | 51,8% |
| READ-IT LESSICALE | 99,7% | 92,7% |

| | | |
|---|---|---|
| READ-IT SINTATTICO | 91,8% | 82,7% |
| READ-IT GLOBALE | 100,0% | 100,0% |
| | | |
| GULPEASE | 62,1 | 54,2 |

**Tabella 6: Valutazione globale dell'analisi della leggibilità in *TOrdinario* e *Rep*, riepilogo**

Come si nota dalla Tabella 6, il modello READ-IT BASE ha un valore di molto maggiore in *Rep* rispetto a *TOrdinario*, come già indicato dalla maggior lunghezza delle frasi. Pertanto, i testi di Twitter sono più semplici di quelli di *Rep.* Infatti, stessa cosa si nota andando a valutare l'indice di Gulpease, inferiore per *Rep*, mostrando una maggior complessità testuale rispetto a *TOrdinario.*

I modelli READ-IT LESSICALE e READ-IT SINTATTICO hanno invece valori superiori in *TOrdinario*.

Per quanto riguarda il modello READ-IT GLOBALE, tutti i testi vengono considerati estremamente complessi, perché il READ-IT utilizzato utilizza come polo di complessità *La Repubblica*, a cui attribuisce la massima complessità, ossia 100%. Avendo *TOrdinario* una complessità sintattica e lessicale maggiore di *Rep*, verrà ugualmente considerata con complessità 100%.

### 4.2.3 Osservazioni sul confronto tra *TOrdinario* e *Rep*

Il profilo lessicale dei due corpora a confronto mostra la tendenza di *Rep* a rivolgersi a un pubblico ampio, come mostrato dall'alta percentuale di lemmi appartenenti al vocabolario di base, quindi conosciute ai più, pari al 41.9%, e dalla percentuale di lemmi riconducibili al repertorio fondamentale, presenti al 54.7%.

*Rep* fa però un uso più alto anche dei vocaboli appartenenti al repertorio ad alta disponibilità, indicando la tendenza a usare termini specifici a seconda dei contesti. In *TOrdinario* la percentuale di lemmi del vocabolario di base è più bassa (23.3%), a causa degli elementi caratteristici di Twitter presenti in alta quantità ma non interpretati correttamente dallo strumento di analisi; *TOrdinario* presenta una percentuale di vocaboli del repertorio ad uso fondamentale maggiore (64.5%) e di vocaboli del repertorio ad alta disponibilità minore (9.3%), mostrando un ampio uso di parole che rendono più agevole la comprensione.

*TOrdinario* presenta, rispetto a *Rep*, un numero inferiore di proposizioni per periodo e di subordinate. Anche la profondità delle catene di subordinazione è più bassa, così come la lunghezza massima testa-dipendente, ovvero il testo è in buona parte composto da frasi principali e gli elementi sintatticamente vicini lo sono anche graficamente.

L'indice di Gulpease è maggiore per *TOrdinario*, che quindi presenta una minor complessità, per quanto invece i modelli READ-IT LESSICALE e READ-IT SINTATTICO presentino valori maggiori rispetto a *Rep*. L'indice di Gulpease indica, per *Rep*, un valore pari a 54.3, ovvero il testo non è di facile comprensione a coloro che non dispongono di una licenza media superiore.

Il modello READ-IT DI BASE ha un valore molto più alto in *Rep*, dovuto alla maggior lunghezza dei periodi.

Se *Rep* può avvalersi di un lungo periodare ricco di subordinate (e probabilmente di proposizioni nominali, soprattutto per quanto riguarda titoli e sottotitoli, vedendo la differenza tra la percentuale di sostantivi, 19.8%, e verbi, 13.4%), *TOrdinario* comunica l'esigenza di sintesi, dovuta al poco spazio per esprimersi, in cui si usano molte parole comuni e meno parole mediamente meno conosciute, come accade in *Rep*, e dove ogni *tweet* è spesso composto da una singola frase.

## 4.2 Confronto tra *TOrdinario* e *TCatastrofe*

I due *corpora* contenenti *tweets*, *TOrdinario* e *TCatastrofe*, verranno confrontati tramite il profilo base, il profilo lessicale e il profilo sintattico al fine di apprezzarne le differenze e spiegare a cosa possano essere dovute. In seguito, verranno confrontati attraverso la funzione "Analisi globale della leggibilità".

### 4.2.1 Profili: di base, lessicale, sintattico

| **Profilo di base** | **TOrdinario** | **TCatastrofe** |
|---|---|---|
| Lunghezza media periodi | 12,4 | 8,6 |
| Lunghezza media parole | 5,8 | 5,8 |

**Tabella 7: Profilo di base di *TOrdinario* e *TCatastrofe*, riepilogo**

Osservando la Tabella 7, si può notare che sì, i *tweets* scritti durante periodi di emergenza sono più corti, ma le parole utilizzate non risultano essere di lunghezza inferiore. Entrambi i corpora hanno infatti una media di caratteri per parola pari a 5,8.

| **Profilo lessicale** | **TOrdinario** | **TCatastrofe** |
|---|---|---|
| Percentuale lemmi VdB | 23,2% | 25,3% |
| Repertorio fondamentale | 64,5% | 67,1% |
| Repertorio ad alto uso | 26,2% | 25,0% |
| Repertorio alta disponibilità | 9,3% | 7,9% |
| Rapporto tipo/unità | 0,830 | 0,700 |
| Densità lessicale | 0,581 | 0,594 |

**Tabella 8: Profilo lessicale di *TOrdinario* e *TCatastrofe*, riepilogo**

Osservando nella Tabella 8 la percentuale di lemmi appartenenti al vocabolario base si nota che questo valore è maggiore nel *corpus TCatastrofe* , ovvero il linguaggio usato è di più semplice comprensione. Anche la percentuale di lemmi riconducibili al repertorio d'uso fondamentale è maggiore in *TCatastrofe,* mentre diminuisce la percentuale di lemmi riconducibili al repertorio ad alto uso e ad alta disponibilità, maggiore invece in *TOrdinario.*

Il *rapporto tipo/unità* è inferiore in *TCatrastrofe,* che presenta quindi un vocabolario meno ricco rispetto a *TOrdinario,* e anche la densità lessicale è maggiore in *TCatastrofe,* confermando la maggior leggibilità e soprattutto la maggiore informatività (dovuta alla maggior quantità di parole piene) dei *tweets* scritti durante un'emergenza.

| **Categorie morfosintattiche** | **TOrdinario** | **TCatastrofe** |
|---|---|---|
| Sostantivi | 18,1% | 24,3% |
| Nomi propri | 9,7% | 7,2% |
| Aggettivi | 6,3% | 4,4% |
| Verbi | 9,6% | 10,8% |
| Congiunzioni | 2,7% | 3,0% |

| | | |
|---|---|---|
| Congiunzioni coordinanti | 73,6% | 78,0% |
| Congiunzioni subordinanti | 26,4% | 22,0% |

**Tabella 9: Profilo sintattico, proprietà morfo-sintattiche in *TOrdinario* e *TCatastrofe*, riepilogo**

Dall'analisi morfologica dei due *corpora*, riassunta in Tabella 9, si nota la frequenza più alta di sostantivi in *TCatastrofe*, ma allo stesso tempo si nota la minor quantità di nomi propri.

Gli aggettivi sono più numerosi in *TOrdinario*, mentre i verbi in *TCatastrofe*. L'alta percentuale di nomi e verbi, insieme a quella più bassa di aggettivi, indica che i *tweets* sono più informativi, perché descrivono più eventi.

Anche le congiunzioni sono più numerose in *TCatastrofe*, in particolare le congiunzioni coordinanti. Riguardo al rapporto nomi/verbi, *TOrdinario* presenta un valore di 1.9 e *TCatastrofe* di 2.2, riflettendo la preponderanza dei nomi rispetto ai verbi.

| **Categorie morfosintattiche** | **TOrdinario** | **TCatastrofe** |
|---|---|---|
| Proposizioni per periodo | 1,043 | 0,790 |
| Principali | 86,0% | 88,6% |
| Subordinate | 14,0% | 11,4% |
| Parole per preposizione | 11,892 | 10,870 |
| Dipendenti per testa verbale | 2,010 | 1,854 |
| Altezza massima albero sintattico | 3,617 | 2,700 |
| Profondità strutture nominali | 1,211 | 1,152 |
| Profondità catene di subordinazione | 1,093 | 1,139 |
| Lunghezza testa-dipendente | 2,177 | 1,991 |
| Lunghezza massima-testa dipendente | 5,503 | 3,692 |

**Tabella 10: Profilo sintattico, struttura sintattica a dipendenze in *TOrdinario* e *TCatastrofe*, riepilogo**

Osservando la Tabella 10, si nota che *TOrdinario* contiene in media più proposizioni per periodo, indice di una maggior complessità sintattica del testo, così come l'uso di subordinate aumenta la complessità grammaticale, di cui si da uso più ampio in *TOrdinario*. In *TOrdinario*, anche il numero di dipendenti per testa verbale è più

alto. Sempre in *TOrdinario* si osservano una media di altezze massime dell'albero sintattico e una profondità media di struttura nominali complesse maggiori. Più alti, rispetto a *TCatastrofe*, anche i valori della media delle distanze e delle distanza massime tra testa e dipendente.

### 4.2.2 Analisi globale della leggibilità

| Indice di leggibilità | TOrdinario | TCatastrofe |
|---|---|---|
| READ-IT BASE | 6,3% | 1,1% |
| READ-IT LESSICALE | 99,7% | 79,3% |
| READ-IT SINTATTICO | 91,8% | 70,2% |
| READ-IT GLOBALE | 100,0% | 100,0% |
| GULPEASE | 62,1 | 74,3 |

**Tabella 11: Valutazione globale dell'analisi della leggibilità in *TOrdinario* e *TCatastrofe*, riepilogo**

Osservando la Tabella 11, si può notare che *TOrdinario* presenta un valore del modello READ-IT BASE maggiore di *TCatastrofe*, derivante dal fatto che i *tweets* scritti al di fuori delle emergenze sono più lunghi, mentre quelli scritti durante le catastrofi naturali sono più brevi.

*TOrdinario* presenta inoltre un READ-IT LESSICALE e un READ-IT SINTATTICO maggiori rispetto a *TCatastrofe*, segno di maggior complessità lessicale e sintattica, come era già apparso nelle analisi dei profili linguistici della sezione 4.2.1.

| SID | frase | base | less. | sint. | glob. |
|---|---|---|---|---|---|
| 1. | tornare in camera e trovare l'armadio aperto, #creepy #terremoto | | | | |
| 2. | altra scossa forte. | | | | |
| 3. | #terremoto | | | | |
| 4. | e dire che non ci tenevo a fare la notte bianca, #terremoto | | | | |
| 5. | non ha sentito il terremoto. | | | | |
| 6. | #terremoto si trema ancora cazzo. | | | | |
| 7. | #terremoto grazie maya per il regalo di compleanno... | | | | |
| 8. | RaiNews24 annuncia morto nel ferrarese #terremoto | | | | |
| 9. | Svegliata ben due volte nella stessa notte dal terremoto... | | | | |
| 10. | questa era propio piccola perche non ho sentito niente .... #terremoto #milano | | | | |
| 11. | #terremoto bussa ancora a Lucca | | | | |
| 12. | Si balla ancora #terremoto | | | | |
| 13. | "danni a Ferrara" e come al solito fraintendo. | | | | |
| 14. | Ah, l'ottimismo... #terremoto #ciccione | | | | |
| 15. | I TT sono #terremoto e #chelsea, ma guai associarli | | | | |
| 16. | Ciclone #cleopatra travolge la #Sardegna morti e dispersi #allertameteoSAR - http://t.co/RmUg25psEf via @NewsLeonardo | | | | |
| 17. | Ah ecco non era suggestione, altra scossetta ora ora #terremoto , ma che cavolo | | | | |
| 18. | E cmq, grazie #terremoto :spenta luce alle 2.40, svegliato alle 4.03 ... | | | | |
| 19. | #terremoto a milano, nuova scossetta | | | | |
| 20. | Mi ha svegliato il #terremoto, sembrava crollasse tutto. | | | | |
| 21. | #bologna | | | | |
| 22. | Scossa di terremoto avvertito a Ravenna, a Bologna io non l'ho sentito | | | | |
| 23. | eh no, non è ancora finita! | | | | |
| 24. | Altra scossa ora in diretta #terremoto #Lugano | | | | |
| 25. | esprime solidarietà e sostegno ai cittadini dell'Abruzzo colpiti dal terremoto: http://htxt.it/vQ2C | | | | |
| 26. | Fanculo avevo quasi deciso di riprovare a dormire e via altre scosse #terremoto | | | | |
| 27. | #allertameteoSAR mi pare la morte del Twitter "di servizio" del 2007. | | | | |
| 28. | Fra gente che inventa notizie, numeri e dice di conoscere le vittime. | | | | |
| 29. | Terremoto in centro e nord Italia? | | | | |

**Figura 12: Leggibilità di alcuni *tweets* provenienti da *TCatastrofe***

L'indice Gulpease è invece inferiore, indicando una leggibilità più bassa per il corpus *TOrdinario.*

Dall'analisi globale emerge che, in generale, i *tweets* scritti durante periodi di emergenza, più corti, meno complessi dal punto di vista lessicale e sintattico, come a voler comunicare in fretta la situazione di pericolo ed essere di immediata comprensione.

READ-IT, oltre a dare un punteggio globale della leggibilità dei vari corpora è in grado di assegnare un grado di leggibilità anche ai singoli *tweets*, come si vede nella Figura 12: verde indica leggibilità facile, le sfumature del giallo e dell'arancione indicano una complessità sempre crescente, fino ad arrivare al rosso, che indica la complessità massima.

### 4.2.3 Osservazioni sul confronto tra *TOrdinario* e *TCatastrofe*

Il profilo di base ha messo in luce la tendenza di *TCatastrofe* verso la brevità dei periodi, scritti in fretta e leggibili in fretta; il profilo lessicale ha evidenziato la quantità maggiore di termini con lemmi riconducibili al vocabolario di base, con una

maggior quantità di termini inclusi nel repertorio ad uso fondamentale, segno che le parole usate sono semplici e accessibili a qualunque utente grazie alla scarsa complessità lessicale. Il rapporto tipo/unità più basso e il valore di densità lessicale più alto sono sintomatici, infatti, di una minor varietà linguistica.

Dal profilo sintattico è emerso un numero maggiore di sostantivi e di verbi e una preferenza per le congiunzioni, soprattutto quelle coordinanti.

Inferiore rispetto a quanto accade in *TOrdinario* anche la percentuale di proposizioni per periodo, così come la percentuale di subordinate, contribuendo ad abbassare la complessità sintattica. Le teste verbali presentano un numero inferiore di dipendenti, e gli elementi sintatticamente vicini vengono tenuti vicini anche graficamente, agevolando la lettura e la comprensione.

I *tweets* scritti durane il periodo di emergenza, con l'intento di comunicare uno stato di allerta, diffondere notizie o richiedere conferme sugli avvenimenti, presentano dunque una necessità di immediata fruibilità e comprensione da parte dell'utente che potrebbe leggere, cosa che avviene in minor parte nel resto dei *tweets*, in cui è meno forte il bisogno di comprensibilità a tutti i costi.

I modelli READ-IT BASE, READ-IT LESSICALE e READ-IT SINTATTICO confermano la tendenza di *TCatastrofe* a tendere verso una maggior semplicità testuale.

Tale ipotesi è confermata anche dall'indice di Gulpease, che in *TCatastrofe* è 74.3; tale valore indica che il testo è comprensibile anche per coloro in possesso della sola licenza elementare, mentre *TOrdinario,* con un indice di Gulpease pari a 62.1, presenta una complessità più alta comprensibile a chi ha almeno la licenza media.

| **Elementi caratteristici** | **TOrdinario** | **TCatastrofe** |
|---|---|---|
| # (hashtag) | 1.816 | 5.239 |
| @ (menzione/risposta) | 2.838 | 577 |
| *http* (link esterni) | 2.070 | 1.029 |
| totale elementi caratteristici | 6.724 | 6.845 |

**Tabella 12 Quantità di elementi tipici di Twitter riscontrati in *TOrdinario* e *TCatastrofe***

Dalla Tabella 12 si può notare la massiccia presenza di *hashtag* nel corpus *TCatastrofe*, pari a 5.277, di cui ben 2.312 di essi rappresentati da "#terremoto", 975

da “#allertameteoSAR”; questo fa pensare che gli *hashtag* presenti nel *corpus* siano perlopiù gli stessi ripetuti, in quanto sono quelli che vengono usati per parlare dell’evento specifico.

In *TOrdinario* è invece molto alta la presenza di “@” e di link esterni. I *tweets* scritti in momenti ordinari fanno parte di discussioni più ampie, e la comprensione del singolo *tweet* è possibile attraverso una lettura più globale; la presenza frequente di “@” è dovuta ai frequenti dialoghi tra utenti, più rari quando si scrive di eventi catastrofici.

Entrambi i *corpora* contengono, comunque, un’alta quantità di elementi caratteristici di Twitter.

# 5. Analisi qualitativa e quantitativa degli strumenti di analisi linguistica automatica applicati ai *tweets*

Tutte le analisi mostrate in questo lavoro si basano sui risultati dell'annotazione linguistica automatica. Essendo il sistema di annotazione addestrato su testi giornalistici, gli elementi caratterizzanti di Twitter possono diventare fonte di rumore sui dati da analizzare. In questa sezione verranno riportati i risultati di un'analisi qualitativa e quantitativa condotta per mostrare come elementi caratteristici del linguaggio di Twitter possano influire negativamente sulle analisi.

Come già detto nel capitolo 4, in cui si sono confrontati i risultati dei monitoraggi linguistici dei *corpora Rep*, *TOrdinario* e *TCatastrofe*, gli strumenti di annotazione linguistica che hanno analizzato i corpora composti da *tweets TOrdinario* e *TCatastrofe* sono stati addestrati su un *corpus* di testo giornalistico. Tale *corpus* è rappresentante delle strutture grammaticali standard, cosa che non accade nei *corpora* estratti da Twitter

Twitter presenta delle strutture caratteristiche, quali le menzioni/risposte, contrassegnate dal segno "@", gli *hashtag*, caratterizzati dal cancelletto "#"; inoltre spesso gli utenti fanno riferimento a link esterni. Tali elementi sono presenti in quantità non trascurabile e non sono di facile interpretazione per lo strumento di annotazione linguistica, essendo assenti nei *corpora* di addestramento.

La presenza di questi elementi genera errori riguardo la corretta classificazione dal punto di vista morfo-sintattico e sintattico, poiché quello che per un parlante umano è chiaramente un nome di persona, se pur scritto in forma di *hashtag* (ad esempio "#BarbaraBerlusconi") non è facilmente interpretabile dallo strumento di analisi, che sbaglia quindi anche nell'assegnare all'elemento un ruolo nella proposizione.

Negli esempi che seguono, è stato effettuato uno studio qualitativo su come lo strumento di analisi analizza gli elementi ad esso ignoti.

Per quanto riguarda l'annotazione morfo-sintattica e sintattica a dipendenze, sui testi con caratteristiche linguistiche vicine ai *corpora* di addestramento i risultati presentano un'accuratezza del 96.34% nell'identificazione simultanea della

categoria grammaticale e dei tratti morfologici associati. Riguardo l'analisi sintattica, per analizzare l'accuratezza[28] dell'analizzatore vengono usare alcune metriche, tra cui:

i. "Labelled Attachment Score" (LAS), ovvero la proporzione di parole del testo con una corretta assegnazione sia per quanto riguarda la testa sintattica sia per quanto riguarda il ruolo svolto in relazione ad essa
ii. "Unlabelled Attachment Score (UAS), ovvero la proporzione di parole nel testo con una corretta assegnazione per quanto riguarda la testa sintattica.

L'accuratezza dello strumento di analisi, in rapporto a questi parametri, è di 83,38% (LAS) e 87,71% (UAS), ovvero quasi tutti i *tokens* vengono ricondotti alla corretta testa sintattica, ma di una percentuale inferiore viene indicato il corretto ruolo svolto in relazione ad essa.

Se queste percentuali sono valide per quanto riguarda l'analisi del *corpus Rep*, non si può dire lo stesso riguardo a *TOrdinario* e *TCatastrofe*. La presenza di elementi caratteristici come "@", "#" e collegamenti esterni abbassa il livello di accuratezza dell'annotazione.

Notare, inoltre, che oltre a tali peculiarità grafiche, Twitter presenta delle differenze anche a livello di costruzione sintattica che aumentano il divario tra i *corpora* di addestramento e quelli composti da *tweets*. Tra di essi, ricordiamo ad esempio l'uso quasi esclusivo della prima e della seconda persona singolare, quasi completamente in disuso nei testi giornalistici, tipicamente in terza persona, e l'uso del futuro, quasi assente in ambito giornalistico.

## 5.1 Analisi qualitativa degli strumenti di analisi linguistica

Gli esempi che seguono forniscono una valutazione qualitativa degli errori, prendendo ad esempio i casi più ricorrenti.

[28] L'accuratezza viene calcolata come il rapporto tra il numero di *tokens* classificati correttamente e il numero totale di *tokens* analizzati.

| | ID | Token | Lemma | C-POS | F-POS | Morphosyntactic feats |
|---|---|---|---|---|---|---|
| 1 | 1 | Ciclone | ciclone | S | S | num=s\|gen=m |
| | 2 | #cleopatra | #cleopatra | S | S | num=n\|gen=n |
| | 3 | travolge | travolgere | V | V | num=s\|per=3\|mod=i\|ten=p |
| | 4 | la | il | R | RD | num=s\|gen=f |
| | 5 | #Sardegna | #Sardegna | N | N | _ |
| | 6 | morti | morto | S | S | num=p\|gen=m |
| | 7 | e | e | C | CC | _ |
| | 8 | dispersi | disperso | S | S | num=p\|gen=m |
| | 9 | #allertameteoSAR | #allertameteoSAR | N | N | _ |
| | 10 | - | - | F | FC | _ |
| | 11 | http://t.co/RmUg25psEf | http://t.co/RmUg25psEf | N | N | _ |
| | 12 | via | via | S | S | num=s\|gen=f |
| | 13 | @NewsLeonardo | @NewsLeonardo | N | N | _ |

**Figura 138 Esempio di POS tagging errato in cui "#", link esterni e "@" sono stati considerati numeri**

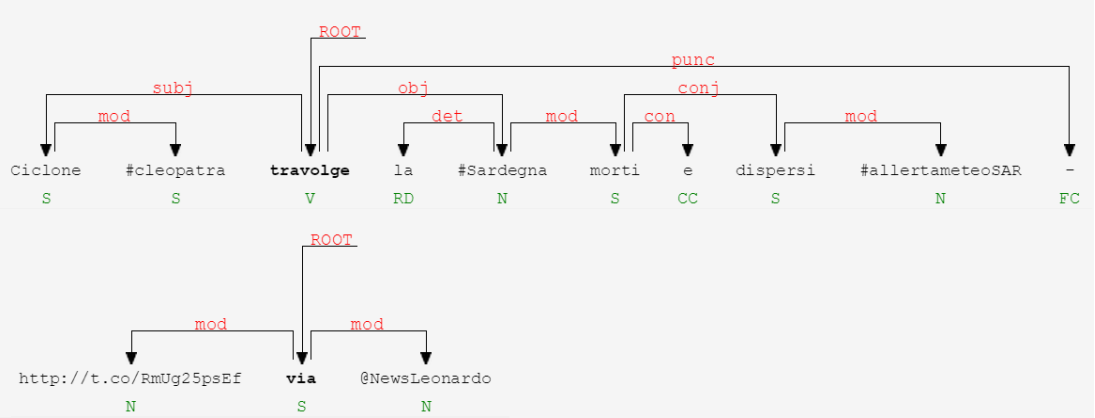


**Figura 14 Albero sintattico del *tweet* analizzato in Figura 14**

In Figura 13 si può notare che il POS tagging ha effettuato alcuni errori in corrispondenza di *hashtag*, link esterni e menzioni: ovvero, "#Sardegna", "allertameteoSAR", "http://t.co/RmUg25psEf", e "@NewsLeonardo" sono stati etichettati con "N", ovvero numeri cardinali.

Dalla Figura 14 si nota che l'albero sintattico, nonostante il POS tagging errato, rimane corretto, mostrando la robustezza dello strumento di analisi: "#Sardegna", taggato come numero, viene correttamente indicato come oggetto del verbo "travolge".

| | ID | Token | Lemma | C-POS | F-POS | Morphosyntactic feats |
|---|---|---|---|---|---|---|
| 1 | 1 | È | essere | V | VA | num=s\|per=3\|mod=i\|ten=p |
| | 2 | scontato | scontare | V | V | num=s\|mod=p\|gen=m |
| | 3 | con | con | E | E | _ |
| | 4 | @andagn, | @andagn, | S | S | num=s\|gen=m |
| | 5 | ma | ma | C | CC | _ |
| | 6 | perché | perché | C | CS | _ |
| | 7 | sono | essere | V | V | num=p\|per=3\|mod=i\|ten=p |
| | 8 | d' | di | E | E | _ |
| | 9 | accordo | accordo | S | S | num=s\|gen=m |
| | 10 | anche | anche | B | B | _ |
| | 11 | con | con | E | E | _ |
| | 12 | #BarbaraBerlusconi? | #BarbaraBerlusconi? | N | N | _ |
| 2 | 1 | Perché | perché | B | B | _ |
| | 2 | ? | ? | F | FS | _ |
| 3 | 1 | #figc | #figc | S | SW | _ |
| | 2 | #elezioni | #elezioni | S | SW | _ |
| | 3 | #voltinuovi | #voltinuovi | S | SW | _ |

**Figura 15 Esempio di POS tagging errato in cui "#" e "@" sono stati classificati talvolta come nomi, talvolta come numeri**

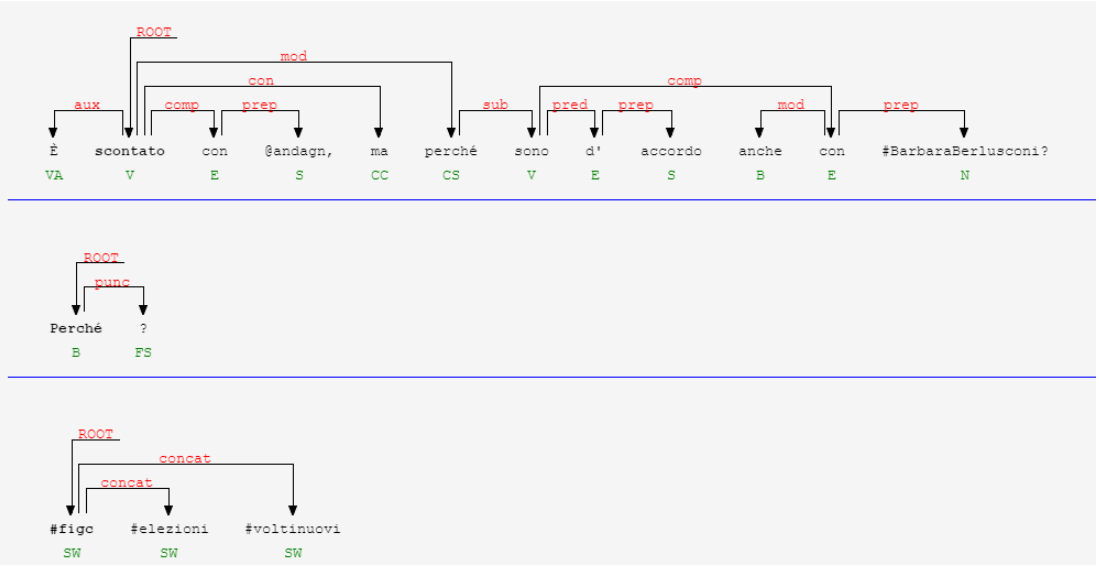


**Figura 16 Albero sintattico del *tweet* analizzato in Figura 16**

In Figura 15 gli errori sono simili a quelli presenti in Figura 13; in particolare si nota che "#BarbaraBerlusconi*"* è stato erroneamente classificato dal POS tagging come numero cardinale, mentre i tre *hashtag* "#figc", "#elezioni" e "#voltinuovi" sono stati classificati come "SW", ovvero parole straniere. Tali errori non hanno influenzato la

corretta struttura dell'albero sintattico della frase, visibile in Figura 16, almeno nella prima parte e nella seconda.

| | ID | Token | Lemma | C-POS | F-POS | Morphosyntactic feats |
|---|---|---|---|---|---|---|
| 1 | 1 | Un | uno | R | RI | num=s\|gen=m |
| | 2 | #terremoto | #terremoto | S | S | num=s\|gen=m |
| | 3 | la | il | R | RD | num=s\|gen=f |
| | 4 | scorsa | scorsa | S | S | num=s\|gen=f |
| | 5 | notte | notte | S | S | num=s\|gen=f |
| | 6 | in | in | E | E | _ |
| | 7 | Italia | Italia | S | SP | _ |
| | 8 | . | . | F | FS | _ |
| 2 | 1 | Vittime | vittima | S | S | num=p\|gen=f |
| | 2 | e | e | C | CC | _ |
| | 3 | danni | danno | S | S | num=p\|gen=m |
| | 4 | . | . | F | FS | _ |
| 3 | 1 | :-( | :-( | N | N | _ |

**Figura 17 Esempio di POS tagging di un *tweet* in cui è presente un'*emoticon***

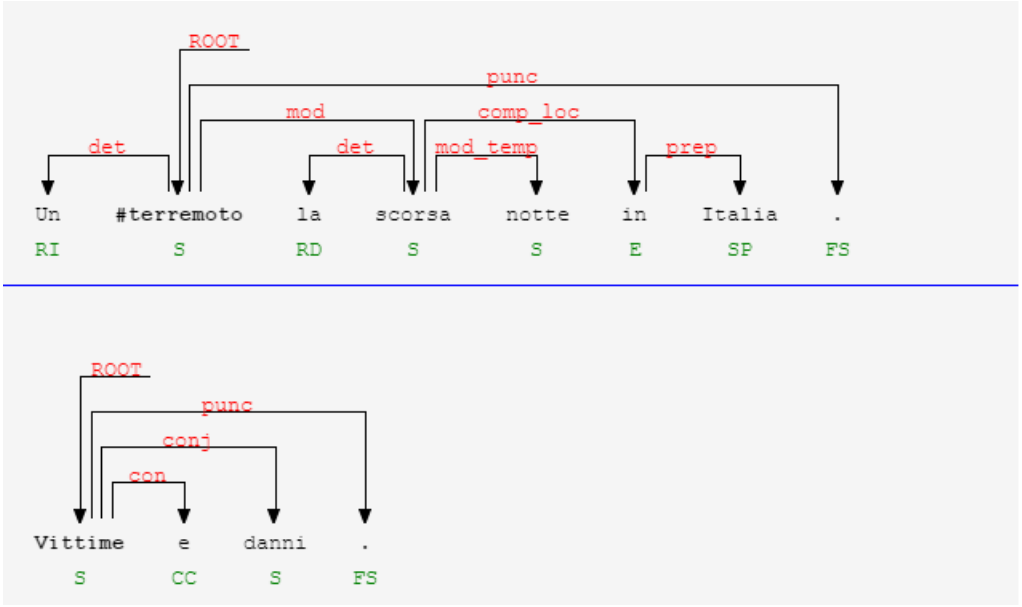


**Figura 18 Albero sintattico del *tweet* analizzato in Figura 18**

Nel *tweet* analizzato in Figura 17 è presente un'*emoticon* (":-(" ), tokenizzata correttamente ma non riconosciuta dallo strumento di annotazione ed etichettata quindi con un numero, poiché lo strumento non dispone di tag per trattare le *emoticons*.

In questo caso, "#terremoto" è stata correttamente etichettato come sostantivo.

| | ID | Token | Lemma | C-POS | F-POS | Morphosyntactic feats |
|---|---|---|---|---|---|---|
| 1 | 1 | Buondì | Buondì | S | SP | _ |
| | 2 | @ElenaBi | @ElenaBi | N | N | _ |
| | 3 | Ricordi | Ricordi | S | SP | _ |
| | 4 | il | il | R | RD | num=s\|gen=m |
| | 5 | periodo | periodo | S | S | num=s\|gen=m |
| | 6 | elezioni | elezione | S | S | num=p\|gen=f |
| | 7 | , | , | F | FF | _ |
| | 8 | ora | ora | B | B | _ |
| | 9 | basta | bastare | V | V | num=s\|per=3\|mod=i\|ten=p |
| | 10 | , | , | F | FF | _ |
| | 11 | ladri | ladro | S | S | num=p\|gen=m |
| | 12 | , | , | F | FF | _ |
| | 13 | sempre | sempre | B | B | _ |
| | 14 | uguali | uguale | A | A | num=p\|gen=n |
| | 15 | , | , | F | FF | _ |
| | 16 | via | via | S | S | num=s\|gen=f |
| | 17 | tutti | tutto | P | PI | num=p\|gen=m |

**Figura 19 Esempio di POS tagging in cui un verbo viene interpretato come un nome**

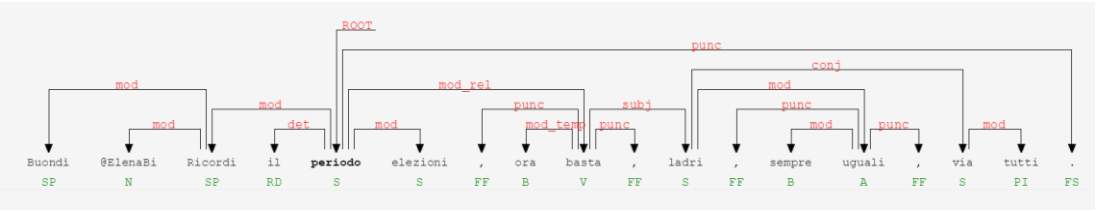


**Figura 20 Albero sintattico del *tweet* analizzato in Figura 20**

Un caso differente è rappresentato dalla Figura 19, in cui un verbo, "Ricordi", viene classificato dal POS tagging come sostantivo. Questa erronea classificazione è dovuta alla mancanza di discorsi in seconda persona nei *corpora* di addestramento, portando alla mancata individuazione della corretta costruzione verbale. Scorretta, pertanto, è anche la costruzione dell'albero sintattico in Figura 20, con l'errata assegnazione del ruolo di radice a "periodo".

| | ID | Token | Lemma | C-POS | F-POS | Morphosyntactic feats |
|---|---|---|---|---|---|---|
| 1 | 1 | Ecco | ecco | B | B | _ |
| | 2 | xke | xke | S | S | num=n\|gen=n |
| | 3 | ' | ' | F | FB | _ |
| | 4 | avete | avere | V | VA | num=p\|per=2\|mod=i\|ten=p |
| | 5 | perso | perdere | V | V | num=s\|mod=p\|gen=m |
| | 6 | milione | milione | S | S | num=s\|gen=m |
| | 7 | di | di | E | E | _ |
| | 8 | voti | voto | S | S | num=p\|gen=m |
| | 9 | . | . | F | FS | _ |

**Figura 21 Esempio di POS tagging di un *tweet* che presenta un'abbreviazione**

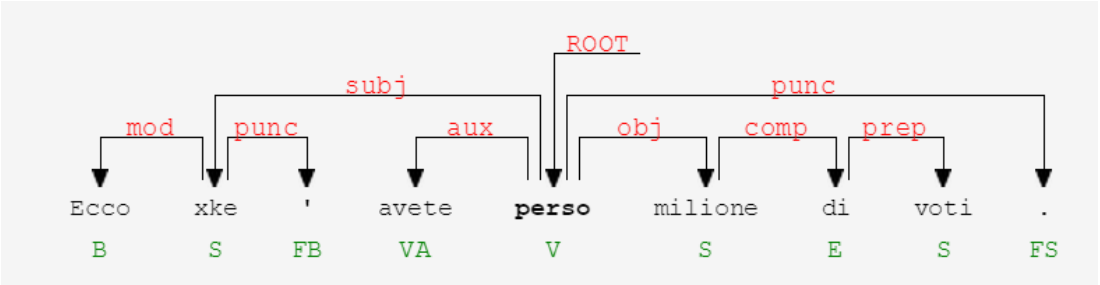


**Figura 22 Albero sintattico del *tweet* analizzato in Figura 22**

Il *tweet* analizzato nella Figura 21 mostra un'abbreviazione, ovvero "xke'" al posto di "perché". Il sistema di annotazione commette errori a livelli diversi di analisi linguistica.

Risulta errata la tokenizzazione, in quanto l'accento è considerato un apice.

Il POS tagging è scorretto perché il sistema identifica "xke'" come "S", sostantivo, in modo erroneo, non avendo mai visto prima un esempio di abbreviazione tipica del linguaggio delle *chat*. L'albero sintattico in Figura 22 colloca erroneamente "xke'" come soggetto, quando dovrebbe considerarlo modificatore del verbo.

Nonostante il sistema si sia dimostrato valido per gli scopi di questa tesi, gli esempi riportati raffigurano casi prototipici che conducono il sistema in errore, dimostrando la necessità di un adattamento degli strumenti a livelli diversi.

## 5.2 Analisi quantitativa degli strumenti di analisi linguistica

Dopo un'analisi qualitativa, osserviamo a livello quantitativo quanto è forte l'impatto degli errori nell'analisi. Abbiamo valutato gli errori che il sistema ha effettuato all'interno dei contesti tipici di Twitter: *hashtag*, menzioni e risposte.

**TCatastrofe**

| | **Valore assoluto** | **Percent.** |
|---|---|---|
| *Hashtag* erroneamente contrassegnati come "A" (aggettivi) | 142 | 3,374% |
| *Hashtag* erroneamente contrassegnati come "E" (preposizione) | 2 | 0,053% |
| *Hashtag* erroneamente contrassegnati come "N" (numeri) | 1.136 | 29,871% |
| *Hashtag* erroneamente contrassegnati come "S" (sostantivi) | 26 | 0,684% |
| *Hashtag* erroneamente contrassegnati come "V" (verbi) | 34 | 0,894% |
| | | |
| *Hashtag* correttamente contrassegnati come "S" (sostantivi) | 2.461 | 64,712% |
| *Hashtag* correttamente contrassegnati come "V" (verbi) | 2 | 0,053% |

**Tabella 133: Comportamento del sistema di fronte agli elementi caratteristici di Twitter in *TCatastrofe***

**TOrdinario**

| | **Valore assoluto** | **Percent.** |
|---|---|---|
| *Hashtag* erroneamente contrassegnati come "A" (aggettivi) | 51 | 5,095% |
| *Hashtag* erroneamente contrassegnati come "N" (numeri) | 221 | 22,079% |
| *Hashtag* erroneamente contrassegnati come "S" (sostantivi) | 58 | 5,794% |
| | | |
| *Hashtag* correttamente contrassegnati come "A" (aggettivi) | 4 | 0,399% |
| *Hashtag* correttamente contrassegnati come "N" (numeri) | 5 | 0,449% |
| *Hashtag* correttamente contrassegnati come "S" (sostantivi) | 662 | 66,134% |

**Tabella 144: Comportamento del sistema di fronte agli elementi caratteristici di Twitter in *TOrdinario***

Nella maggior parte dei casi, gli *hashtag* sono dei sostantivi, e, come si nota osservando la Tabella 13 e la Tabella 14, il sistema riesce a classificarli correttamente per più della metà della volte in entrambi i *corpora*, con un'accuratezza del 64,712% per *TCatastrofe* e del 66,134% per *TOrdinario*.
L'errore più comune commesso dal sistema di fronte all'*hashtag* è l'assegnazione della categoria grammaticale "N", *numero*, ad elementi che in realtà non sono numeri. Più raramente, il sistema confonde gli *hashtag* con aggettivi e verbi.
Dal punto di vista sintattico, l'errore commesso dal sistema quando scambia un sostantivo per un numero non è considerato grave, in quanto la struttura della frase rimane invariata. L'errore (anche se assai poco ricorrente) che maggiormente inficia la struttura sintattica si ha quando il sistema considera un sostantivo come un verbo, andando a sconvolgere l'albero sintattico della frase.

### 5.2.1 Nomi propri celati da *hashtag*, menzioni e risposte

Un'attenzione particolare è stata rivolta ai nomi propri, rappresentati dall'etichetta "SP", presenti nei *corpora TOrdinario* e *TCatastrofe*, per avere un'idea più precisa di quanto lo strumento si sia dimostrato affidabile qualora i nomi propri fossero scritti in forma di *hashtag* o di menzione/risposta. Viene inoltre fornito il tipo di errore effettuato dal sistema, ovvero l'etichetta errata da esso assegnata al nome proprio non riconosciuto.

**TCatastrofe**

| | **Valore assoluto** | **Percent.** |
|---|---|---|
| ***Hashtag (scritti in forma di "#hashtag")*** | | |
| Nomi propri correttamente etichettati come "SP" | 11 | 0,766% |
| Elementi etichettati come "SP" che in realtà non lo sono | 36 | 2,507% |
| Nomi propri scambiati per "A" (aggettivi) | 74 | 5,153% |
| Nomi propri scambiati per "E" (preposizioni) | 3 | 0,209% |
| Nomi propri scambiati per "N" (numeri) | 663 | 46,170% |
| Nomi propri scambiati per "S" (sostantivi) | 641 | 44,638% |
| Nomi propri scambiati per "V" (verbi) | 8 | 0,557% |

| *Menzioni/risposte (scritti in forma di "@**nomeutente**")* | | |
|---|---|---|
| Nomi propri correttamente etichettati come "SP" | 0 | 0,000% |
| Nomi propri scambiati per "A" (aggettivi) | 23 | 3,986% |
| Nomi propri scambiati per "N" (numeri) | 232 | 40,208% |
| Nomi propri scambiati per "S" (sostantivi) | 320 | 55,460% |
| Nomi propri scambiati per "V" (verbi) | 2 | 0,347% |

**Tabella 155: Comportamento del sistema riguardo ai nomi propri celati da elementi caratteristici di Twitter in *TCatastrofe***

**TOrdinario**

| | **Valore assoluto** | **Percent.** |
|---|---|---|
| *Hashtag (scritti in forma di "**#hashtag**")* | | |
| Nomi propri correttamente etichettati come "SP" | 10 | 1,227% |
| Elementi etichettati come "SP" che in realtà non lo sono | 5 | 0,613% |
| Nomi propri scambiati per "A" (aggettivi) | 14 | 1,718% |
| Nomi propri scambiati per "N" (numeri) | 481 | 59,018% |
| Nomi propri scambiati per "S" (sostantivi) | 305 | 37,423% |
| *Menzioni/risposte (scritti in forma di "**@nomeutente**")* | | |
| Nomi propri correttamente etichettati come "SP" | 11 | 0,387% |
| Nomi propri scambiati per "A" (aggettivi) | 24 | 0,847% |
| Nomi propri scambiati per "E" (preposizioni) | 1 | 0,035% |
| Nomi propri scambiati per "N" (numeri) | 694 | 24,454% |
| Nomi propri scambiati per "S" (sostantivi) | 2.106 | 74,207% |
| Nomi propri scambiati per "V" (verbi) | 2 | 0,070% |

**Tabella 166: Comportamento del sistema riguardo ai nomi propri celati da elementi caratteristici di Twitter in *TOrdinario***

La Tabella 15 e la Tabella 16 mostrano che il sistema, quando incontra un nome proprio sotto forma di *hashtag*, è difficilmente in grado di riconoscerlo. La

percentuale di nomi propri scritti in forma di *hashtag* correttamente individuati è molto bassa, con un valore di 0,766% per *TCatastrofe* e 1,227% per TOrdinario.
La maggior parte delle volte, il nome proprio viene scambiato per un sostantivo o per un numero; questo tipo di errore però non va a pesare sulla struttura sintattica della frase. Per quanto riguarda *TCatastrofe*, solo in pochi casi (0,557%) un nome proprio viene scambiato per un verbo, sconvolgendo la struttura sintattica della frase.

Su Twitter, ogni menzione/risposta fa riferimento ad un utente preciso, identificato da un nickname univoco e quindi da un nome proprio; il sistema di annotazione non è stato in grado di cogliere questo aspetto, e nessun nome è stato correttamente identificato in *TCatastrofe* se scritto nella forma *@nomeproprio*, solo lo 0,387% in *TOrdinario*. La maggior parte dei nomi propri è assimilata alla categoria grammaticale *sostantivo*, e un'altra buona parte alla categoria *numero*.
Come affermato in precedenza, questo tipo di errore non è rilevante per quanto riguarda la struttura sintattica della frase.

# 6. Conclusioni

Nei capitoli precedenti è stata effettuata un'analisi di monitoraggio linguistico su tre *corpora*: uno composto da testi giornalistici estratti da *La Repubblica*, chiamato *Rep*, gli altri due da *tweets* raccolti sia in momenti generici (*TOrdinario*) che caratterizzati da stati d'emergenza (*TCatastrofe*).

Oltre al confronto tra scritto formale e scritto informale (*Rep* e *TOrdinario*), l'obiettivo principale di questo contributo era analizzare per la prima volta che tipo di differenze si creano tra i *tweets* a seconda del contesto in cui vengono composti, effettuando quindi un confronto interno a Twitter in base ai risultati dell'analisi linguistica dei *corpora TOrdinario* e *TCatastrofe.*

Confrontando il profilo lessicale di *Rep* e *TOrdinario*, è emersa la tendenza del linguaggio giornalistico de *La Repubblica* a rivolgersi a un pubblico ampio, usando perlopiù parole semplici e conosciute ma anche termini specifici e meno noti quando richiesto dal contesto. Per quanto riguarda il *corpus TOrdinario*, il sistema si è trovato di fronte agli elementi caratteristici di Twitter (*hashtag*, menzioni, risposte, link esterni, *emoticons*, errori grammaticali), difficili da interpretare correttamente perché non presenti nel *corpus* di addestramento; questo ha portato ad un abbassamento del livello di leggibilità indicato dalla percentuale di lemmi appartenenti al vocabolario; la leggibilità di *TOrdinario* viene comunque suggerita dai valori assunti dagli altri parametri, quali la maggior presenza di lemmi provenienti dal repertorio d'uso fondamentale, e la minore presenza di lemmi del repertorio ad alto uso e ad alta disponibilità.

Lo strumento di annotazione ha assegnato al READ-IT BASE un valore inferiore a *TOrdinario*, basandosi sulla minor lunghezza dei *tweets*, vincolati al limite dei 140 caratteri, rispetto alle frasi di *Rep*, ma ha attribuito un valore maggiore al modello READ-IT LESSICALE.

Dal punto di vista morfo-sintattico e sintattico, lo strumento di analisi ha assegnato a *TOrdinario* un valore più alto, nonostante il profilo sintattico evidenzi come in *Rep* sia maggiore la frequenza di subordinate, il numero di dipendenti per testa verbale, l'altezza massima dell'albero sintattico, la profondità delle strutture nominali e delle catene di subordinazione, la lunghezza testa dipendente, tutte caratteristiche che rendono meno agevole la comprensione del testo. Questa discrepanza è da attribuirsi

all'uso che gli utenti generalmente fanno di Twitter, ovvero ricco di variazioni rispetto alla sintassi normale e alla grammatica dell'italiano scritto, di abbreviazioni, punteggiatura non standard, eccessive nominalizzazioni, tutti tratti caratteristici del linguaggio di Twitter.

Per quanto riguarda il confronto interno a Twitter tra *TOrdinario* e *TCatastrofe*, l'osservazione più importante riguarda la spiccata necessità palesata in *TCatastrofe* di sintesi, informatività e semplicità dal punto di vista lessicale e sintattico. Questa necessità spinge le persone a rispettare maggiormente la grammatica e la sintassi standard, deviando dallo stile tipico di Twitter.

*TCatastrofe* presenta una minore lunghezza media dei periodi (che porta ad un innalzamento dell'indice di Gulpease), una maggiore percentuale dei lemmi appartenenti al vocabolario di base e al repertorio d'uso fondamentale maggiore e una minor percentuale di quelle parole che rendono la lettura meno comprensibile, ovvero appartenenti al repertorio ad alto uso e ad alta disponibilità. La densità lessicale è maggiore, segno di maggior informatività. Il READ-IT BASE e il READ-IT LESSICALE presentano valori inferiori in *TCatastrofe*, ovvero i *tweets* raccolti durante i periodi d'emergenza sono facili da comprendere in generale, anche dal punto di vista lessicale. Anche il valore del modello READ-IT SINTATTICO è più basso in *TCatastrofe*. La presenza, più alta rispetto a *TOrdinario*, di nomi e verbi è la percentuale più bassa di aggettivi suggerisce che nei *tweets* di *TCatastrofe* vengono descritti molti eventi, ovvero c'è un'alta informatività. Inoltre, dal punto di vista sintattico, *TCatastrofe* riporta una minor quantità di proposizioni per periodo e una minor quantità di subordinate. L'altezza massima dell'albero sintattico è inferiore, così come la profondità delle strutture nominali e della lunghezza testa-dipendente.

Nel capitolo 5 abbiamo riportato un'analisi qualitativa e quantitativa sui possibili casi in cui il sistema di annotazione fornisce risultati sbagliati, ovvero quando si trova di fronte a testi che non rispettano la grammatica standard. Gli esempi prototipici e il conteggio quantitativo degli errori mostrano l'esigenza di adattare i sistemi in modo da essere in grado di trattare le caratteristiche peculiari di Twitter (link esterni, menzioni, repliche, *hasthag*, *emoticons*); per via dei limiti dello strumento che ha effettuato i monitoraggi linguistici, i dati estratti da *TOrdinario* e *TCatastrofe* non

sono esatti, perché quello che è semplice da intuire per il parlante umano non lo è altrettanto per la macchina addestrata su tutt'altro registro linguistico.

Ulteriore contributo di questa tesi è stata la creazione e annotazione manuale del *corpus* di *tweets TCatastrofe*[29], che oggi viene utilizzato all'interno del progetto "Social sensing[30]", che ha come scopo valutare le informazioni inviate dagli utenti dei *social network* durante una calamità naturale, in modo da avere un quadro in tempo reale degli eventi, delle zone coinvolte e delle conseguenze. Sistemi di annotazione addestrati su *corpora* costituiti da *tweets* possono permettere una migliore accuratezza di annotazione, permettendo di estrarre informazioni quantitativamente e qualitativamente migliori.

[29] La costruzione del *corpus* e la sua annotazione manuale, indicando se il *tweet* in esame desse o meno informazioni sui danni provocati dalla catastrofe naturale, ha costituito una parte del mio stage, effettuato presso il CNR di Pisa sotto la coordinazione del professor Felice Dell'Orletta, che è stato il mio tutor.

[30] Il progetto è portato avanti da un gruppo di ricercatori dell'Istituto di Informatica e Telematica del CNR di Pisa (IIT-CNR), coordinati da Maurizio Tesconi. Per maggiori informazioni, http://socialsensing.it/it.

# 7. Bibliografia


Brunato D., Venturi G., (2014) *Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici*. Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa. Italian Natural Language Processing Lab (ItaliaNLP Lab).

Chiusaroli F., (2014) *Sintassi e semantica dell'hashtag: studio preliminare di una forma di Scritture Brevi*. in R. Basili, A. Lenci, B. Magnini (eds.), The First Italian Conference on Computational Linguistics, CLiC-it 2014 – Proceedings, 9-10 December 2014, Pisa University Press, Pisa, vol. I, pp. 117-121.

Cresci S., Tesconi M., Cimino A., Dell'Orletta F., (2015) *A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages*. In Proceedings of the 3rd International Workshop on Social Web for Disaster Management (SWDM'15), 18 May, Firenze.

Dell'Orletta F., Montemagni S., Venturi G., (2011) READ-IT: *Assessing Readability of Italian Texts with a View to Text Simplification, in Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies* (SLPAT 2011), Edinburgh, July 30, pp. 73-83.

Lenci A., Montemagni S., Pirrelli V., (2005) *Testo e Computer. Elementi di linguistica computazionale*. Roma, Carocci.

Montemagni S., (2013) *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*. Studi Italiani di Linguistica Teorica e Applicata (SILTA) 1, Anno XLII, pp. 145-172 ". Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) – Pisa. Italian Natural Language Processing Lab (ItaliaNLP Lab).

Piemontese M.E., (1996), *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli, Tecnoid.

READ-IT , documentazione Demo Online, ILC-CNR, ItaliaNLP Lab.

Roncaglia G., (2010), *Linguaggi e tecnologia: usi della lingua e strumenti di rete*, in Libro dell'Anno 2010, Roma, Treccani.

Tavosanis M., (2011), *L'italiano del web*. Roma, Carrocci.

Zaga C., (2012). *Twitter: un'analisi dell'italiano nel micro blogging*. Italiano LinguaDue, n. 1., Milano.

# 8. Appendice

Questa appendice contiene i dati con cui sono stati effettuate le analisi linguistiche del capitolo 4.

I dati sono estratti tramite lo strumento di analisi READ-IT TextTools v2.1.9, disponibile all'indirizzo: http://www.italianlp.it/demo/read-it/.

I dati sono disponibili per la consultazione ai seguenti link:

- Dati relativi a *Rep*, *corpus* costruito con testi estratti da *La Repubblica*: http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_lang=it&tt_tmid=tm_source&tt_jid=1424254535743049_it
- Dati relativi a *TOrdinario*, *corpus* costituito da *tweets* generici: http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_lang=it&tt_tmid=tm_source&tt_jid=1423057887476373_it
- Dati relativi a *TCatastrofe*, *corpus* costituito da *tweets* scritti in corrispondenza di una catastrofe naturale, quale alluvioni o terremoti: http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_lang=it&tt_tmid=tm_source&tt_jid=1423057831819516_it