Monitoraggio linguistico di Scritture Brevi: aspetti metodologici e primi risultati

Dominique Brunato, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi
Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR)
ItaliaNLP Lab - www.italianlp.it

Abstract: Se da un lato le tecnologie del linguaggio svolgono un ruolo ormai indiscusso per l'accesso al contenuto testuale, ciò non appare scontato quando si va a considerare il loro ruolo nella valutazione delle strutture linguistiche sottostanti al testo. Questo contributo si focalizza sulla definizione di una metodologia innovativa di monitoraggio linguistico della lingua italiana che a partire dall'output di strumenti di annotazione linguistica automatica permette di ricostruire il profilo linguistico di una collezione di testi rappresentativa di una specifica varietà d'uso della lingua. Tale metodologia è stata applicata a un corpus di *tweet* e ha permesso di far luce su interrogativi aperti quali la possibilità di rintracciare tendenze lessicali, morfo-sintattiche e sintattiche peculiari all'interno di questa tipologia testuale; di studiare come queste tendenze si rapportino ai tratti caratterizzanti della lingua scritta e parlata; di individuare possibili differenze nella forma linguistica in cui si *twittano* contenuti di natura diversa.

Parole chiave: Trattamento Automatico del Linguaggio, Monitoraggio Linguistico, Varietà d'Uso della Lingua, Lingua del Web

1. Introduzione

Tra le peculiarità dell'odierna lingua del web va annoverata la presenza di nuove tipologie testuali che fanno ricorso in maniera estesa a forme di *scritture brevi*, così come definite in Chiusaroli e Zanzotto (2012). Si tratta di un fenomeno ricco di spunti per chi affronta lo studio dell'italiano contemporaneo nelle sue varietà d'uso, da prospettive e con obiettivi differenti. Lo storico della lingua, ad esempio, può voler approfondire il rapporto in diacronia tra le forme attuali e quelle attestate in epoche precedenti, così come l'analogia nei loro meccanismi di derivazione; in una prospettiva socio-linguistica, può essere interessante esplorare quanto il ricorso ad un repertorio di forme grafiche, per le quali «risulti dirimente il principio della 'brevità' connesso al criterio dell' 'economia'» (ivi), rappresenti un segnale distintivo del "gergo" giovanile o allargato ad una comunità di parlanti che si identifica nell'uso di un *medium* comune, le cui caratteristiche ne vincolerebbero anche le realizzazioni comunicative. In questo caso, si potrebbero estendere alla forma *testo* le implicazioni della celebre intuizione del sociologo canadese McLuhan, divenuta quasi slogan, secondo cui il mezzo è il messaggio. Ancora, l'uso e la diffusione di tali espedienti grafici nelle diverse tipologie di comunicazione in rete può proporsi come metrica per distinguere e classificare modelli di scrittura non canonica; è possibile valutare, ad esempio, in che misura e per quali aspetti della struttura linguistica le deviazioni dalla lingua standard siano ascrivibili alla mimesi del parlato.

In questo contributo ci proponiamo di esaminare alcune tendenze linguistiche che caratterizzano una varietà specifica di lingua del web - la lingua dei testi di *Twitter* - introducendo una innovativa metodologia di analisi basata su strumenti di annotazione linguistica automatica del testo. Oltre a fornire indicazioni utili a ricostruire il profilo linguistico di *Twitter*, tale metodologia può rappresentare il punto di partenza per approfondire alcuni dei possibili interrogativi di ricerca connessi all'uso dell'italiano nelle nuove tipologie di testi digitali. Come nota infatti Tavosanis (2012:19), nonostante l'enorme quantità di dati che il web mette a disposizione allo studioso della lingua, «a livello di studi linguistici in italiano [esso] è stato poco esaminato». E tale considerazione sembra valere soprattutto per le piattaforme di *microblogging*[1], di cui Twitter esemplifica ad oggi un rappresentante in costante crescita. Con oltre 4 milioni di utenti attivi settimanalmente in Italia[2], *Twitter* offre infatti un bacino di dati di produzione scritta di enormi dimensioni. In realtà, ad oggi, l'ottica di analisi linguistica privilegiata da cui si è guardato a queste produzioni è stata quella del contenuto, spesso con l'obiettivo di costruire un repertorio di lemmi, formule, simboli ricorrenti identificativi di stati emotivi dello scrivente, premessa a sua volta allo sviluppo di sistemi di *Sentiment Analysis e Opinion Mining*, che rappresentano trend di ricerca di estrema attualità. Non mancano inoltre contributi significativi di linguisti che hanno monitorato l'uso dello strumento da parte di una comunità di utenti ben selezionati, principalmente esponenti della scena politica italiana[3], di cui sono stati ispezionati i messaggi non solo rispetto ai temi ma anche sul piano della forma, al fine di tracciare profili di evoluzione della comunicazione politica stimolati dai nuovi paradigmi di interazione 2.0.

Molto più circoscritti sono invece gli studi campionari a più ampio raggio che hanno cercato di caratterizzare questa forma di scrittura del web nel sistema, inquadrandola all'interno del quadro di variabilità linguistica dell'italiano contemporaneo. La lingua dei *social network* è considerata infatti una tipologia testuale dell'*italiano digitato* (Gastaldi, 2002) inteso come varietà di lingua per sua natura scritta ma che, nelle più recenti rivisitazioni del noto schema di Berruto (1987), si tende a far spostare lungo l'asse diamesico verso il parlato (Antonelli, 2011).

Partendo da questi presupposti, il presente contributo intende introdurre una innovativa metodologia di monitoraggio linguistico fondata sulle tecnologie linguistico-computazionali, che permette di far luce sui seguenti interrogativi:

- È possibile rintracciare in un testo di 140 caratteri, ovvero l'estensione massima di un *tweet*, tendenze lessicali, morfo-sintattiche e sintattiche peculiari?

- In che modo si può quantificare l'influenza della componente parlata in queste forme di scrittura breve?

- Esistono differenze a livello di forma linguistica nel modo in cui si *twittano* contenuti di natura diversa, come ad esempio stati personali o contenuti legati a situazioni di calamità naturale, quali l'evento di un terremoto o di un'alluvione?

[1] Wikipedia definisce il microblogging come "una forma di pubblicazione costante di piccoli contenuti in Rete, sotto forma di messaggi di testo (normalmente fino a 140 caratteri), immagini, video, audio MP3, ma anche segnalibri, citazioni, appunti. Questi contenuti vengono pubblicati in un servizio di rete sociale, visibili a tutti o soltanto alle persone della propria comunità."

[2] http://youmark.it/wp-content/uploads/2013/01/Ricerca-Twitter-MEC.pdf

[3] Si veda ad esempio S. Spina, 2012, 2013.

A questo scopo, nel paragrafo successivo (§ 2) introdurremo gli elementi fondamentali della nostra metodologia di monitoraggio linguistico. Saranno pertanto descritti i corpora analizzati (§ 2.1.), gli strumenti di Trattamento Automatico del Linguaggio usati in questo studio (§ 2.2) e l'ampia gamma di caratteristiche linguistiche considerate in fase di ricostruzione del profilo linguistico dei testi presi in esame (§ 2.3). In questo paragrafo ci focalizzeremo anche su alcune peculiarità di Twitter che rappresentano fonti di rumore per gli strumenti di annotazione linguistica automatica del testo. Nel paragrafo 3 riporteremo i risultati del monitoraggio linguistico del corpus di testi di *Twitter*, condotto in ottica comparativa, sia rispetto ad altri generi testuali tradizionali, sia rispetto ad una distinzione interna alla tipologia di *tweet* considerati.

2. La metodologia di indagine

La metodologia di indagine qui introdotta prende le mosse dall'intuizione che «i risultati dell'annotazione linguistica automatica, sebbene includano inevitabilmente un margine di errore, se appropriatamente esplorati possono fornire indicazioni affidabili nella ricostruzione del profilo linguistico di un testo» (Montemagni, 2013). Partendo dunque da questo presupposto, abbiamo condotto un'indagine della lingua dei *tweet* basata sull'individuazione di un'ampia gamma di tratti linguistici rintracciati in corpora di *tweet* linguisticamente annotati con strumenti di annotazione linguistica automatica. In quanto segue, descriveremo gli elementi fondamentali della metodologia.

2.1. I corpora

Lo studio qui presentato ha preso in considerazione un corpus di 8.780 *tweet* di estensione pari a 131.345 parole (tokens), internamente ripartito in due categorie:

- un corpus di 3.508 *tweet* di argomento generico, selezionati in maniera casuale nei primi sei mesi del 2014, per un totale di 65.670 parole (tokens).
- un corpus di 5.272 *tweet* che fanno riferimento ad alcuni episodi di disastri naturali che hanno colpito recentemente la penisola, tra cui il terremoto dell'Aquila del 2009, quello dell'Emilia del 2012 e l'alluvione di Genova del 2014. L'estensione di questo corpus è pari a 65.675 parole (tokens).

Allo scopo di individuare le peculiarità della lingua di Twitter rispetto ad altre varietà della lingua italiana, abbiamo preso in considerazione una serie di altri corpora rappresentativi di generi testuali e varietà di lingua diversi. Tale soluzione consente infatti un'esplorazione comparativa a diversi livelli, sia rispetto a ciascun genere, sia rispetto alla loro distribuzione trasversale in tipologie più generali, quali la distinzione tra varietà di lingua scritta e orale, che assume in questo contesto una rilevanza particolare.

Come visibile dalla tabella 1, ciascun genere è ripartito internamente in due varietà di lingua, distinte rispetto ad una dimensione di complessità/semplicità linguistica per il genere in considerazione. Per quanto riguarda la lingua scritta, i generi contemplati sono la prosa giornalistica, la narrativa, i materiali didattici e la prosa scientifica. Più in dettaglio, il corpus di linguaggio giornalistico si compone di una varietà più complessa, rappresentata da una collezione di testi tratti dal quotidiano *La Repubblica* (Rep), e di una varietà più semplice, di cui sono

esemplificativi i testi pubblicati in *Due Parole* (2Par), il mensile di "facile lettura" scritto in un linguaggio controllato per lettori affetti da lievi disabilità cognitive e/o un basso livello di scolarizzazione; nell'ambito della narrativa, la distinzione è tra un corpus di letteratura per adulti e uno per bambini; per il genere dei materiali didattici, alla varietà complessa dei testi didattici per la scuola secondaria, si è affiancato un corpus di testi tratti da sussidiari per bambini; la prosa scientifica comprende una selezione di articoli scientifici più "complessi", che sono stati pubblicati su riviste specialistiche di settore e una collezione di articoli scientifici più "semplici", tratti da Wikipedia. Infine, sul versante dei corpora di lingua parlata, la distinzione è tra parlato formale primariamente monologico e parlato informale in forma prevalentemente dialogica nell'ambito familiare.

Corpus	**Abbrev.**	**n° di testi**	**n° token**
Tweet di argomento generico	TweetGen	3.508	65.670
Tweet relativi a disastri naturali	TweetDis	5.272	65.675
Totale	TweetTot	**8.780**	**131.345**

Tabella 1: I corpora di *tweet*.

Etichetta	**Genere**	**Corpus**	**n°di testi**	**n°token**
Giorn	Giornalismo	La Repubblica (Marinelli et al., 2003)	321	232.908
		Due Parole (Piemontese, 1996)	322	73.314
			Tot: 643	**Tot: 306.222**
Narr	Letteratura	Letteratura per adulti (Marinelli et al., 2003)	101	19.370
		Letteratura per bambini (Marconi et al., 1994)	327	471.421
			Tot: 428	**Tot: 490.791**
Suss	Materiali didattici	Scuola Primaria (Dell'Orletta et al. , 2011b)	127	48.036
		Scuola Secondaria (Dell'Orletta et al., 2011 b)	70	48.103
			Tot: 197	**Tot: 96.139**
Scient	Prosa scientifica	Wikipedia, sezione "Ecologia e Ambiente"	293	205.071
		Articoli scientifici specialistici	84	471.969
			Tot: 377	**Tot: 677.040**
Parlato	Parlato	Parlato formale (CO-ORAL-ROM Italia, Cresti e Moneglia, 2005)		189.382
		Parlato informale (CO-ORAL-ROM Italia, Cresti e Moneglia, 2005)		150.892
			Tot:	**Tot: 340.275**

Tabella 2: I corpora di altre varietà testuali.

2.2. Gli strumenti di Trattamento Automatico del Linguaggio

L'annotazione linguistica automatica del testo rappresenta il prerequisito indispensabile per la ricostruzione del suo profilo linguistico. Pertanto, tutti i testi presi in esame sono stati analizzati con strumenti di annotazione linguistica automatica in grado di rendere progressivamente esplicita l'informazione linguistica contenuta in un testo. Per ogni livello di descrizione linguistica uno specifico componente di analisi identifica in modo automatico la struttura del testo, utilizzando come input il risultato prodotto dal componente precedente. L'identificazione della struttura linguistica del testo, o annotazione, avviene tipicamente in modo incrementale, attraverso analisi linguistiche a livelli di complessità crescente: "tokenizzazione", ovvero segmentazione del testo in parole ortografiche (o tokens); analisi morfo-sintattica e lemmatizzazione del testo tokenizzato; analisi della struttura sintattica della frase in termini di relazioni di dipendenza.

La tabella 3 mostra un esempio del risultato dell'annotazione linguistica del seguente periodo:

L'altro giorno due uomini camminavano sul marciapiede.

		Lemmatizzazione	**Annotazione morfo-sintattica**			**Annotazione sintattica**	
Id	Forma	Lemma	CPoS	FPoS	Tratti morfologici	Testa sintattica	Relazione
1	L'	il	R	RD	num=s|gen=n	3	det
2	altro	altro	D	DI	num=s|gen=m	3	mod
3	giorno	giorno	S	S	num=s|gen=m	6	mod_temp
4	due	due	N	N	_	5	mod
5	uomini	uomo	S	S	num=p|gen=m	6	subj
6	camminavano	camminare	V	V	num=p|per=3|mod=i|ten=i	0	ROOT
7	sul	su	E	EA	num=s|gen=m	6	comp_loc
8	marciapiede	marciapiede	S	S	num=s|gen=m	7	prep
9	.	.	F	FS	_	6	punc

Tabella 3: Un esempio di annotazione linguistica.

Innanzitutto, il periodo è stato individuato grazie alla fase preliminare di segmentazione in periodi del testo. Durante la successiva fase di tokenizzazione, all'interno del periodo sono stati riconosciuti i tokens corrispondenti alle singole forme (colonna *Forma*), identificate univocamente da un numero progressivo (colonna *Id*). La fase di disambiguazione morfo-sintattica ha permesso di associare ad ogni token individuato *i)* la corretta categoria morfo-sintattica (colonna *CPoS* e *FPoS*)[4] che il token ha nel contesto specifico, *ii)* i relativi tratti morfologici (colonna *Tratti morfologici*) e

[4] Per ogni token viene riconosciuta la categoria morfo-sintattica generale (CPoS) e eventuali sottocategorie (FPoS). Ad esempio, alla forma (token) 'sul' viene associata la categoria preposizione (E) e viene ulteriormente specificato che si tratta di una preposizione articolata (EA). Allo stesso modo, il token '.' viene annotato come un segno di punteggiatura (F) di fine periodo (FS). La lista completa delle categorie morfo-sintattiche è consultabile alla pagina http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf.

iii) il lemma corrispondente (colonnna *Lemma*). Ad esempio, la forma 'uomini' (Id=5) viene annotata con la categoria sostantivo (S), viene riconosciuto che si tratta di una forma plurale (num=p) e maschile (gen=m) e ricondotta al lemma 'uomo'.

Il risultato dell'annotazione sintattica riportato nelle colonne *Testa sintattica* e *Relazione* della Tabella 3 permette inoltre di stabilire che, ad esempio, il sostantivo 'uomini' è il soggetto (subj) del verbo 'camminavano', il quale costituisce la testa sintattica della relazione.[5] Questa informazione è riportata nella colonna *Testa* dove è infatti segnalato che la testa sintattica del dipendente 'uomini' ha Id=6, l'Id cioè del token 'camminavano'. In questo caso 'camminavano' ha testa sintattica 0 dal momento che rappresenta il verbo della frase principale, radice (root) dell'albero sintattico dell'intero periodo. La fase di annotazione sintattica a dipendenze permette dunque di fornire una descrizione esplicita dell'intero albero sintattico del periodo analizzato, sotto forma di relazioni di dipendenza che legano i tokens che lo compongono. L'informazione può inoltre essere graficamente visualizzata, come mostra la figura 1 che riporta la struttura sintattica della frase annotata, rappresentata come una serie di nodi lessicali (i singoli tokens), messi in collegamento da archi di dipendenza a loro volta etichettati con il nome del tipo di relazione di dipendenza (gli archi e le etichette graficamente rappresentati).

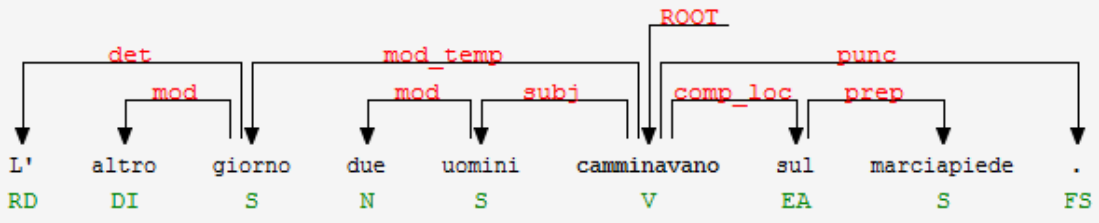

Figura 1: Un esempio di rappresentazione grafica dell'annotazione sintattica a dipendenze.

In questo studio è stata usata LinguA (Linguistic Annotation pipeline)[6], una piattaforma per l'analisi linguistica automatica di testi i cui componenti sono stati sviluppati in modo congiunto dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa e dall'Università di Pisa. Si tratta di strumenti di analisi linguistica automatica multilingue che combinano algoritmi basati su regole (*rule-based*) con metodi di apprendimento automatico (*machine learning*) e le cui prestazioni rappresentano lo stato dell'arte per la lingua italiana, come testimoniato dai risultati delle periodiche campagne di valutazione di componenti per il trattamento automatico dell'italiano EVALITA[7].

2.2.1. L'annotazione linguistica automatica di *tweet*

Come sottolineato per la prima volta da Gildea (2011), gli strumenti di Trattamento Automatico del Linguaggio hanno una drastica diminuzione di accuratezza quando sono impiegati nell'analisi di tipologie di testi rappresentativi di un dominio o di una varietà linguistica diversa da quella sui quali gli strumenti sono stati sviluppati. Hanno influenza le caratteristiche specifiche del dominio o

[5] L'inventario delle dipendenze sintattiche utilizzate con relativa descrizione è consultabile alla pagina http://www.italianlp.it/docs/ISST-TANL-DEPtagset.pdf.
[6] http://www.italianlp.it/demo/linguistic-annotation-tool/
[7] http://www.evalita.it/

varietà in esame che, differenziandosi da quelle proprie della varietà linguistica sulla quale gli strumenti sono stati addestrati, devono essere trattate in modo specifico per non influire negativamente sull'accuratezza dei risultati di annotazione. Elemento fondamentale di ogni approccio di annotazione linguistica automatica basata su algoritmi statistici è infatti il *training corpus*, l'insieme di testi linguisticamente annotati in modo manuale, dal quale gli strumenti imparano ad associare al testo in esame l'informazione linguistica corretta grazie ad un costante processo inferenziale di classificazione probabilistica, durante il quale, ad ogni passo della computazione, viene scelta l'annotazione linguistica più probabile data la parola in input. Dal training corpus gli strumenti, utilizzando algoritmi di apprendimento automatico, ricavano un modello matematico probabilistico da applicare all'annotazione linguistica di una nuova collezione di testi. Di conseguenza, tanto più i nuovi testi in esame contengono caratteristiche linguistiche diverse da quelle dei testi contenuti nel training corpus, tanto minori saranno le prestazioni degli strumenti di analisi.

La risposta al problema delineato sopra è costituita dallo sviluppo di metodi e tecniche di *Domain Adaptation* il cui fine ultimo è l'adattamento degli strumenti all'analisi di testi che appartengono a un dominio diverso da quello rispetto al quale sono stati sviluppati. Se si considera che gli strumenti di annotazione linguistica sono per lo più addestrati su corpora di testi giornalistici, considerati rappresentativi della lingua comune (standard), si può comprendere quanto la definizione di metodi automatici di adattamento a nuovi domini sia di primaria importanza. Sino ad oggi particolare interesse è stato dedicato alla definizione di metodi di adattamento di parser sintattici a domini specialistici (Sagae & Tsujii, 2007; McClosky et al., 2010; Dell'Orletta et al., 2013). Una tale attenzione è legata al fatto che l'annotazione sintattica costituisce il punto di partenza per numerose applicazioni pratiche, quali ad esempio l'estrazione automatica di informazione, la traduzione automatica, il Question Answering, ecc.

In particolare, sempre più spesso gli strumenti di Trattamento Automatico del Linguaggio si trovano a dover analizzare testi del web, passo preliminare di numerosi compiti applicativi quali ad esempio il riconoscimento di entità nominate (Named Entity Recognition) o l'individuazione di opinioni (Sentiment Analysis). Sempre di più, dunque, lo scenario d'uso reale è rappresentato dall'analisi di testi rappresentativi di *non-canonical languages*, intese come varietà devianti rispetto alla norma di una lingua. Tali varietà sono assimilate alle varietà di testi rappresentativi di *computer-mediated communication* ("comunicazione mediata dal computer", CMC), quali ad esempio chat, email e sms, alla lingua dei social media, quali blogs commenti di blogs, forum posts, microblogs, consumer review, ecc… o ancora alla lingua parlata. Le caratteristiche specifiche di queste varietà linguistiche hanno un impatto negativo sull'accuratezza degli strumenti di annotazione linguistica automatica (Petrov & McDonald, 2012). Per quanto riguarda in particolare l'analisi di Twitter, sono stati pertanto messi a punto degli approcci di adattamento rispetto a diversi livelli: a livello di definizione di uno schema di annotazione morfo-sintattica che include non solo le categorie morfo-sintattiche standard ma anche etichette specifiche usate per classificare simboli caratteristici di Twitter (Gimpel et al., 2011); a livello di adattamento del modulo di annotazione morfo-sintattica (Part-Of-Speech Tagging) (Plank et al., 2014); a livello delle applicazioni dell'annotazione linguistica come ad esempio il riconoscimento di entità nominate (Named Entity Recognition) (Plank et al., 2014).

Per valutare la portata del possibile rumore introdotto dalle caratteristiche specifiche di Twitter in fase di annotazione linguistica automatica dei *tweet*, abbiamo condotto un'indagine preliminare a campione dei *tweet* raccolti e annotati. Come si può notare dagli esempi che seguono, a marcare la distanza del *tweet* dalla norma linguistica del testo scritto *standard* non sono solo gli errori ortografici, frequenti in ogni tipo di *computer-mediated communication*[8], bensì anche una sintassi propria dell'ambiente *Twitter*, composta di simboli che anteposti ad altre parole o stringhe del testo assumono un valore semantico[9]. Questo influenza i risultati dell'annotazione linguistica automatica a diversi livelli e in diversi modi. Partendo dalla fase di annotazione morfo-sintattica gli elementi del cosiddetto «gergo di Twitter» (Chiusaroli, 2014) quali le parole precedute dai simboli # e @ sono classificati con categorie morfo-sintattiche diverse, a seconda del contesto nelle quali ricorrono. Consideriamo ad esempio il caso del *tweet* seguente tratto dal corpus *TwGen* (cfr. tabella 4 e la rappresentazione grafica nella figura 2):

(1) Il #pericolo #TTIP con i pareggi di #bilancio

Nonostante l'input non completamente conforme alla norma dell'italiano standard, in fase di annotazione morfo-sintattica, i tokens '#pericolo' e '#bilancio' sono stati correttamente classificati come sostantivi ('S'). Al contrario, il token '#TTIP', acronimo di "Transatlantic Trade and Investment Partnership" (un termine di recente introduzione nella politica economica dell'Unione Europea), non è stato classificato come sigla ma come aggettivo ('A'). Come si può chiaramente vedere nella figura 2, gli eventuali errori di classificazione morfo-sintattica non hanno tuttavia influito a livello di annotazione sintattica. Il token '#pericolo' è stato correttamente riconosciuto come radice nominale (ROOT) dell'intero *tweet*; la relazione tra '#bilancio' e la sua testa sintattica 'di' è stata classificata in modo corretto come relazione di tipo *preposition* ('prep'); il token '#TTIP', sebbene erroneamente classificato come aggettivo, è stato correttamente riconosciuto come dipendente del token '#pericolo' a cui è legato da una relazione di tipo *modifier* ('mod').

		Lemmatizzazione	**Annotazione morfo-sintattica**			**Annotazione sintattica**	
Id	Forma	Lemma	CPoS	FPoS	Tratti morfologici	Testa sintattica	Relazione
1	Il	il	R	RD	num=s | gen=m	2	det
2	#pericolo	#pericolo	S	S	num=s | gen=m	0	root
3	#TTIP	#TTIP	A	A	num=s | gen=m	2	mod
4	con	con	E	E	-	2	comp
5	i	il	R	R	num=p | gen=p	6	det
6	pareggi	pareggio	S	S	num=p | gen=p	4	prep
7	di	di	E	E	-	6	comp

[8] Per una tipologia degli 'scarti' ortografici più frequenti e della loro distribuzione nelle diverse tipologie di scrittura in rete, si veda Tavosanis (2012: 76-87).

[9] Come riportato da Zaga (2012), il più famoso di questi è il simbolo 'cancelletto' (#), che individua l'*hasthtag*, ossia la parola chiave del *tweet*, facendo sì che la stringa di parole immediatamente preceduta da questo simbolo si trasformi in un link attivo, che rimanda ad altre pagine indicizzate con lo stesso *hashtag* nel database di Twitter. Oltre alle hashtag (di cui si rimanda a Chiusaroli, 2014 per un approfondimento), sono frequenti i tag "menzione" e "risposta" (Dardi, 2011:74), entrambi identificati dal simbolo 'chiocciola' (@), che vengono preposti alla stringa di caratteri identificativa del nome di un altro utente a cui ci si vuole rivolgere.

		Lemmatizzazione	Annotazione morfo-sintattica			Annotazione sintattica	
Id	Forma	Lemma	CPoS	FPoS	Tratti morfologici	Testa sintattica	Relazione
8	#bilancio	#bilancio	S	S	num=s | gen=m	7	prep

Tabella 4: Annotazione linguistica di (1).

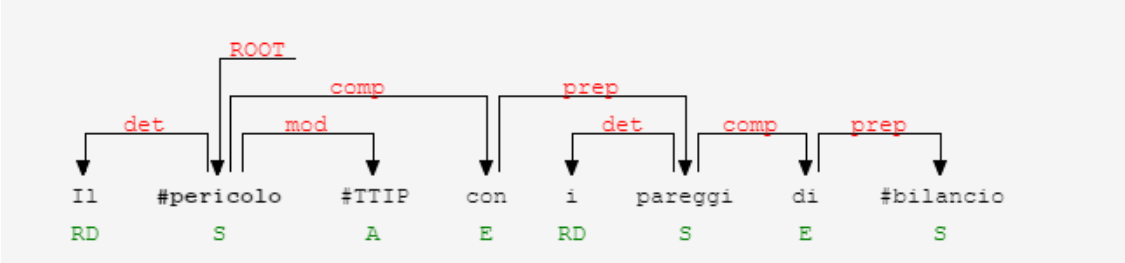

Figura 2: Rappresentazione grafica dell'annotazione linguistica di (1).

Anche nel caso dell'esempio (2) estratto dal corpus *TweetDis* l'albero sintattico risulta ricostruito in modo corretto (cfr. figura 3):

(2) @AndreaVallascas Segui il tag #allertameteoSAR per avere news.

Sebbene infatti i tokens preceduti dai simboli # e @ ('#allertameteoSAR' e '@AndreaVallascas') siano stati erroneamente classificati come numeri ('N'), tuttavia la radice sintattica dell'intero *tweet* ('Segui') è stata correttamente riconosciuta, così come le altre relazioni di dipendenza sintattica come ad esempio la relazione di tipo oggetto ('obj') che lega il token 'tag' alla sua testa sintattica 'Segui' o la relazione di tipo *modifier* ('mod') che lega '#allertameteoSAR' a 'tag'. Diverso è il caso della relazione di dipendenza che lega '@AndreaVallascas' alla corrispondente testa sintattica 'Segui'. In fase di training il parser sintattico non ha mai visto esempi da cui apprendere la funzione sintattica svolta da una menzione, in questo caso dal token '@AndreaVallascas'. Sebbene, dunque, sia stata correttamente riconosciuta l'esistenza di una relazione di dipendenza sintattica tra '@AndreaVallascas' e 'Segui', tuttavia la relazione è stata classificata con l'etichetta più probabile che potrebbe legare un token numero ('N') dipendente da un token verbale ('V'), una relazione cioè di tipo *modifier* ('mod').

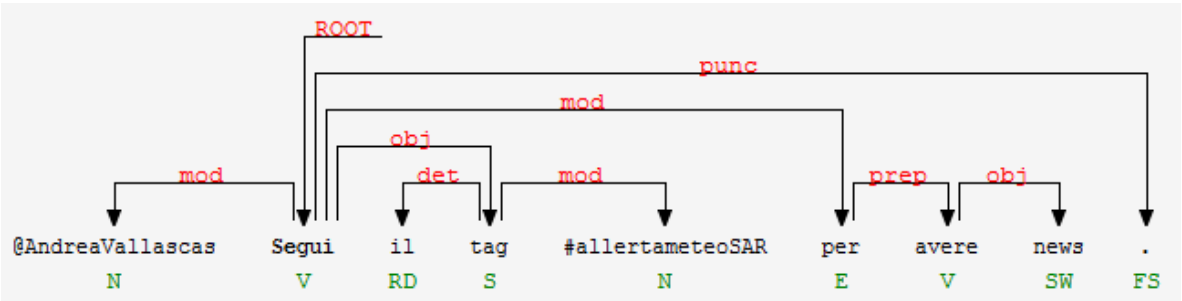

Figura 3: Rappresentazione grafica dell'annotazione linguistica di (2).

Più problematico a livello di analisi sintattica è invece il caso della frase in (3), tratta dal corpus *TwGen* (cfr. figura 4):

> (3) @hyperbros: Il candidato @LucaBraia pubblica su Fb un post su elezioni per PdBas.

Come il precedente anche questo *tweet* è introdotto dal tag "menzione", funzione svolta dal token '@hyperbros'. Questo crea un errore già in fase di tokenizzazione, i ':' non vengono correttamente tokenizzati, separati cioè dalla menzione '@hyperbros' . L'errore si propaga a cascata: a livello di annotazione morfo-sintattica i ':' non vengono dunque classificati come segni di punteggiatura. Ma è a livello dell'analisi sintattica che sono più evidenti le conseguenze della erronea tokenizzazione. Si può ad esempio notare come sia stata erroneamente riconosciuta una relazione di tipo *relative modifier* ('mod_rel'), identificativa di una frase relativa, tra il verbo 'pubblica' e la menzione '@hyperbros'. Risulta anche pregiudicata l'identificazione del ruolo sintattico di oggetto svolto dal token 'post', nonostante esso sia stato riconosciuto correttamente come sostantivo in fase di analisi morfo-sintattica. Al contrario, è stata correttamente ricostruita la struttura di alcuni dei sotto-alberi sintattici della frase come ad esempio la sequenza di complementi preposizionali dipendenti in cascata dal sostantivo 'post' ('su elezioni per PdBas').

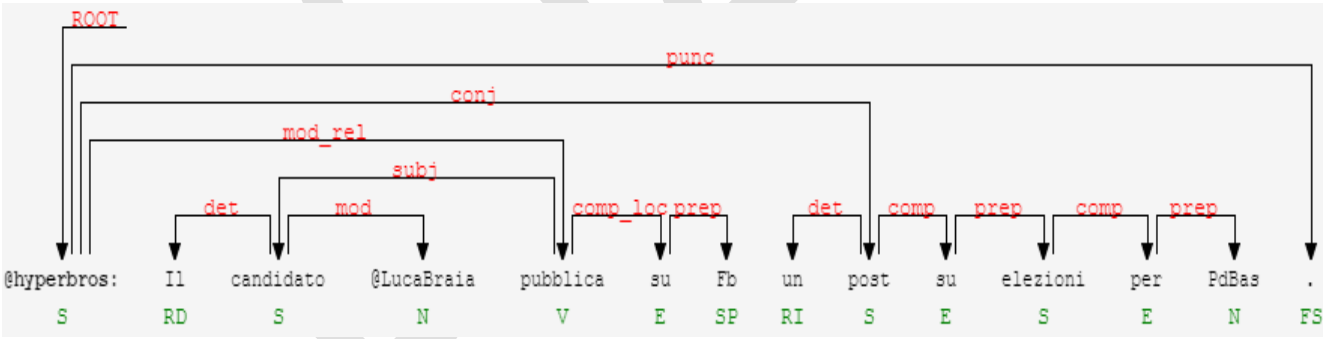

Figura 4: Rappresentazione grafica dell'annotazione linguistica di (3).

Negli esempi riportati sopra, abbiamo visto alcune delle principali difficoltà incontrate dagli strumenti di annotazione linguistica automatica nell'analisi della lingua di *Twitter*. In questa sede non è stata effettuata una correzione manuale degli errori riscontrati; un intervento di questo tipo, del resto, richiederebbe uno studio più approfondito del tipo di errore, o meglio deviazione dallo *standard* linguistico, che è possibile rintracciare in questi testi, al fine di capire in quali direzioni rivolgere le azioni di specializzazione degli strumenti di annotazione linguistica. L'obiettivo dell'analisi, piuttosto, era quello di selezionare, proprio in considerazione delle specificità rintracciate nei *tweet*, i livelli di analisi linguistica più attendibili da cui partire per monitorare la

lingua di *Twitter*, nonché la tipologia di parametri di variazione che meglio permettono di coglierne le peculiarità. A conclusione di ciò, riteniamo importante sottolineare che sebbene il livello di annotazione sintattica sia il più complesso per gli strumenti di annotazione linguistica automatica, l'analisi qualitativa condotta – per quanto ancora a uno stadio preliminare – ha dimostrato che anche analisi parziali e contenenti errori possono fornire indicazioni interessanti se considerate in relazione a corpora di vaste dimensioni. Come discusso in quanto segue, abbiamo pertanto ritenuto opportuno considerare anche questo livello di annotazione linguistica come un punto di partenza affidabile per acquisire informazioni, anche complesse, utili alla ricostruzione del profilo linguistico dei corpora di *tweet* analizzati.

2.3. Le caratteristiche linguistiche

Il testo linguisticamente annotato costituisce il punto di partenza della nostra metodologia di monitoraggio linguistico. Essa consiste nel rintracciare in modo automatico a partire dall'output degli strumenti di annotazione linguistica del testo la distribuzione di un'ampia gamma di caratteristiche linguistiche: lessicali, morfo-sintattiche e sintattiche. Come discuteremo più nel dettaglio nel paragrafo 3, l'analisi comparativa della distribuzione delle caratteristiche rintracciate nei diversi corpora esaminati ci ha permesso di ricostruire il profilo linguistico dei corpora di *tweet*, individuandone le tendenze specifiche rispetto alle altre varietà dell'italiano qui considerate.

La tabella 4 illustra le caratteristiche linguistiche prese in considerazione in fase di monitoraggio linguistico, distinte per livello di annotazione[10]. Nel paragrafo 3 discuteremo quelle che si sono rivelate più significative nella caratterizzazione dei *tweet* rispetto alle varietà d'uso della lingua che abbiamo preso in considerazione come riferimenti esterni.

Tipo di caratteristica	**Livello di annotazione linguistica**	**Caratteristica**
Di base	Divisione in frasi e Tokenizzazione	Lunghezza media dei periodi e delle parole
Lessicale	Lemmatizzazione e annotazione morfo-sintattica	Type/Token Ratio
		Densità Lessicale
		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-sintattica	Annotazione morfo-sintattica	Distribuzione delle categorie morfo-sintattiche
		Articolazione interna dei sintagmi verbali (modo, tempo, persona)
Sintattica	Annotazione sintattica a dipendenze	Articolazione interna del periodo; - numero di proposizioni per periodo; - distribuzione di frasi principali vs. subordinate; - arità verbale
		Caratteristiche relative alla struttura dell'albero sintattico

[10] Per una descrizione completa della metodologia di monitoraggio linguistico basata su caratteristiche estratte automaticamente dall'annotazione multi-livello del testo si rimanda a Montemagni (2013).

Tipo di caratteristica	Livello di annotazione linguistica	Caratteristica
		analizzato: - altezza media dell'intero albero; - lunghezza media di sotto-alberi (es. complementi preposizionali dipendenti in sequenza da un nome)
		Distribuzione dei vari tipi di relazioni di dipendenza sintattica
		Lunghezza delle relazioni di dipendenza sintattica, calcolata come la distanza tra la testa e il dipendente (in tokens)

Tabella 4: Le caratteristiche considerate in fase di monitoraggio linguistico.

Le caratteristiche *di base* fanno riferimento ai livelli più semplici di analisi automatica del testo, ovvero la segmentazione in frasi e la tokenizzazione, da cui si ricavano caratteristiche basilari della struttura del testo, quali la lunghezza media del periodo (calcolata in termini di *tokens*) e la lunghezza media delle parole (in termini di caratteri). Procedendo nell'esplorazione dell'output congiunto delle fasi di lemmatizzazione e annotazione morfo-sintattica, è possibile acquisire tratti relativi al profilo lessicale del testo, quali la sua varietà e la sua ricchezza, le cui misure quantitative di riferimento sono, rispettivamente, la Type/Token Ratio (o TTR) e la Densità Lessicale. La prima va a valutare il rapporto tra il numero di parole tipo al numeratore e il totale di occorrenze di unità del vocabolario al denominatore, restituendo un valore che oscilla tra 0 e 1, dove valori prossimi allo 0 indicano testi poco variegati dal punto di vista del vocabolario e valori vicini a 1 contraddistinguono testi lessicalmente molto vari. Si tratta ovviamente di un indice sensibile alla lunghezza del testo, per cui viene generalmente computato per porzioni testuali predefinite (ad esempio, le prime 100 parole del testo). La Densità Lessicale è calcolata come il rapporto tra il numero di parole semanticamente piene, dette anche "parole contenuto" (quindi sostantivi, verbi, aggettivi e avverbi), e il totale di tokens del testo.

La tipologia di parole usate nel testo e il loro grado di familiarità sono rintracciati sulla base della distribuzione dei lemmi presenti all'interno dei testi in esame rispetto al Vocabolario di Base (VDB) (De Mauro, 2000) e alla tripartizione interna nei repertori d'uso, dunque Lessico Fondamentale, Alto Uso e Alta Disponibilità. Inoltre, a partire dall'output dell'annotazione morfo-sintattica viene calcolata la distribuzione delle parti del discorso nel testo, sia a livello più generale (es. sostantivi, verbi, congiunzioni ecc.), sia più granulare (es. nomi propri vs nomi comuni; verbi principali, ausiliari, modali; congiunzioni coordinanti vs subordinanti). Dall'annotazione dei tratti morfologici è inoltre possibile rintracciare informazioni sulla predicazione verbale quali la distribuzione dei modi, tempi e della persona del verbo.

Infine, l'annotazione sintattica a dipendenze consente l'esplorazione di parametri più complessi e informativi della struttura grammaticale del testo, che spaziano dall'articolazione interna del periodo (ad esempio il numero di proposizioni per periodo o la proporzione tra clausole principali e subordinate), a quella della proposizione (ad esempio, il numero di parole per proposizione o l' "arità" intesa come numero medio di dipendenti per testa verbale) e, ancora, a caratteristiche relative alla struttura dell'albero sintattico analizzato, quali la profondità media dell'intero albero della frase o di alcuni sotto-alberi selezionati (ad esempio quelli identificativi di strutture nominali

complesse) e la lunghezza media delle dipendenze sintattiche (calcolata come la distanza in termini di tokens tra una testa sintattica e il suo dipendente legati da una relazione di dipendenza), e infine la distribuzione delle varie tipologie di dipendenze sintattiche (es. soggetto, oggetto diretto, oggetto indiretto).

3. Tendenze della lingua dei *tweet*

Mutuando dagli studi classici di *Register Variation* l'indicazione secondo cui «systematic differences in the relative use of core linguistic features provide the primary distinguishing characteristics among registers» (Biber, 1995: 36), ci siamo concentrati innanzitutto sul dato relativo alla distribuzione percentuale delle categorie morfo-sintattiche nei corpora considerati. Come si può vedere nella tabella 5, l'intero corpus di *tweet* (*TwGen*) si caratterizza per una maggior presenza di nomi rispetto alla distribuzione percentuale di verbi. Di conseguenza, i *tweet* si distinguono per un rapporto nomi/verbi piuttosto elevato, più vicino ai valori rintracciati nel corpus di testi scritti (*Scritto*), e in particolare a quelli della prosa scientifica (*Scient*), rispetto ai valori caratterizzanti il parlato (*Parlato*).

La prevalenza dei sostantivi sui verbi è un dato ben acquisito in letteratura rispetto al quale misurare l'opposizione scritto vs parlato (Voghera, 2005). Lo è altrettanto il fatto che, all'interno delle due varietà diamesiche, si possano riscontrano ulteriori dimensioni di variazione; nello scritto, ad esempio, è soprattutto la prosa ad alta densità informativa a ricorrere allo stile nominale rispetto ai testi di scrittura creativa (Voghera, 2005; Biber, 1995)[11]: questa tendenza è confermata dai nostri dati, che mostrano infatti una maggior affinità tra i corpora di prosa giornalistica e scientifica, da un lato, e il genere di narrativa e materiali didattici dall'altro.

Rispetto a questa distinzione più granulare, la lingua dei *tweet* oscilla molto di più verso la prosa scritta di carattere prevalentemente informativo, tendenza che molto probabilmente risente del vincolo di economia del testo posto dallo strumento, per cui chi scrive tenderà a ricorrere a quelle strutture sintattiche dotate di maggior incisività e pregnanza semantica, tipicamente quelle nominali che consentono fenomeni di ellissi e nominalizzazione. Non a caso, un altro esempio di scrittura breve che si contraddistingue per questa tendenza sono i titoli dei giornali, ovvero la sezione più esposta ad un periodare sintetico, dall'andamento tipicamente brachilogico (Bonomi, 2003: 146-148). Inoltre, come accade tipicamente nella prosa scritta informativa, i *tweet* fanno un uso limitato di pronomi (4,77%) e avverbi (4,19%), che invece predominano nel parlato. Al contrario, la lingua dei *tweet* si rivela più vicina alla lingua parlata per quanto riguarda la presenza di preposizioni e aggettivi attributivi, due categorie grammaticali la cui distribuzione tende di norma a co-variare con quella dei sostantivi (Biber, 1988, 1995). Come mostrano i valori riportati nella tabella 5, la distribuzione di aggettivi e preposizioni nel corpus *TwGen*, rispettivamente il 3,88% e il 12,38% del totale di categorie morfo-sintattiche classificate, si avvicina di più ai valori rintracciati nel corpus di parlato (4,63% e 10,50%) che in quello di testi rappresentativi della varietà scritta della lingua italiana (6,96% e 15,30%). E sempre all'influsso del parlato può essere ricondotto l'uso più consistente di interiezioni, uno dei dispositivi morfo-sintattici più usati per riprodurre caratteristiche tipiche dell'oralità.

[11] Confronta la discussione in Montemagni (2013:113-115).

	Giorn	Narr	Suss	Scient	Scritto (media)	Parlato (media)	TwGen	TwDis	TwTot
Aggettivi	6,16	6,15	7,76	8,85	6,96	**4,63**	3,88	4,35	**4,12**
Articoli	9,35	8,21	9,03	7,73	8,58	7,13	5,33	5,44	5,38
Avverbi	4,18	5,90	5,79	3,98	**4,77**	10,77	3,09	5,28	**4,19**
Congiunzioni	3,65	4,60	4,69	3,60	4,60	5,69	2,68	3,03	2,85
Interiezioni	0,03	0,14	0,06	0,05	0,07	**1,33**	0,18	0,28	**0,23**
Preposizioni	15,85	12,21	14,57	17,05	15,30	**10,50**	12,38	10,53	**11,45**
Pronomi	3,05	6,32	5,42	3,12	**4,77**	9,41	3,18	3,69	**3,44**
Sostantivi	28,29	23,63	23,25	28,53	**24,98**	17,91	35,97	32,21	**34,09**
Verbi	13,30	15,20	13,87	10,67	**13,85**	17,27	9,60	10,81	**10,21**
Rapporto Nomi/Verbi	2,13	1,55	1,68	2,67	**1,80**	0,89	3,75	2,98	**3,34**

Tabella 5: Distribuzione percentuale delle principali categorie morfo-sintattiche.

Per avanzare una possibile interpretazione di questi dati, che sembrano suggerire una minor complessità sintattica del *tweet* pur in presenza di uno stile prevalentemente nominale, ci sembra opportuno intrecciare l'output dell'analisi morfo-sintattica con il risultato dell'annotazione sintattica. Alcuni indicatori utili a questo livello, la cui rilevazione manuale sarebbe senza dubbio laboriosa per corpora di così grandi dimensioni, ci vengono dai seguenti parametri: la profondità media delle strutture nominali complesse e il numero medio di dipendenti per testa verbale.

Il primo parametro (*Lunghezza media catene nominali* nella tabella 6) permette di monitorare l'uso della modificazione nominale calcolata come profondità media di strutture nominali complesse, ovvero sotto-alberi sintattici la cui testa è un nome che viene modificato ricorsivamente da complementi preposizionali. Questa caratteristica è una spia di complessità del testo, a sua volta favorita dall'uso di nominalizzazioni e derivati nominali, ed è infatti più frequente nei testi scritti ad alta densità informativa (si veda in particolare il dato della prosa scientifica), ma poco attestata nel parlato (1,16). In questo senso, la lingua dei *tweet* ricalca la maggior leggerezza del parlato (si veda la discussione riportata in (Voghera, 2001)), mostrando catene nominali la cui profondità media è pari a circa una relazione di dipendenza (1,21).

Analogamente, il numero di dipendenti per testa verbale identifica un altro possibile tratto di complessità della struttura frasale, ovvero l'aggiunta di espansioni attorno al nucleo del predicato verbale (Shapiro et al., 1987). Va precisato che il dato qui esposto assimila la tradizionale distinzione del modello valenziale in argomenti e aggiunti, da cui l'impossibilità di discriminare se, ad esempio, un valore medio pari a tre individui una clausola semplice con predicato trivalente (es. Il presidente *ha rimandato* la riforma alle Camere) da una clausola con predicato bi-argomentale e un complemento aggiunto (es. Durante la seduta il ministro *ha illustrato* la riforma). Nonostante ciò, possiamo osservare, ancora una volta, come la tendenza all'uso di predicati con un numero maggiore di modificatori (sia argomenti sia aggiunti) sia una variabile discriminante non tanto sul piano dell'opposizione scritto vs parlato quanto piuttosto all'interno di ciascuna delle due diverse varietà diamesiche. Per quanto riguarda lo scritto, infatti, i valori medi più elevati occorrono nella prosa giornalistica e scientifica rispetto ai testi scolastici e letterari, mentre nel contesto della produzione orale è il parlato formale ad ottenere valori considerevolmente superiori a quello informale. Anche rispetto a questo parametro i *tweet* mostrano un andamento più simile a quest'ultimo corpus.

Rimanendo sul piano della dimensione sintattica, la tendenza ad un registro più informale e ad un periodare più semplice si ricava anche dalla scarsa propensione all'uso del passivo (terza riga della tabella 6) e dal ricorso limitato alla subordinazione (si veda, a questo proposito, il dato relativo al rapporto tra clausole principali e subordinate, quarta riga della tabella 6). Tuttavia, non necessariamente il limite di lunghezza del *tweet* impone la scelta di periodi monoclausali: come rivela il dato sull'occorrenza media di clausole per periodo (cfr. riga 5), questi ultimi sono più attestati nel corpus dei *tweet* relativi ai disastri rispetto a quelli di tematiche generali, che risultano anche mediamente più lunghi dei primi. Ovviamente la lunghezza media del periodo (dato qui riportato nell'ultima riga della tabella 6) è un parametro molto grezzo per valutare la complessità sintattica di un testo, oltre ad essere prevedibile la sua forte implicazione nel caratterizzare il profilo linguistico delle forme di scritture brevi qui in esame in relazione agli altri generi e varietà testuali di riferimento. Al contrario, meno scontato era che potesse agire da elemento distintivo all'interno delle due varietà di *tweet* considerati, evidenza che potrebbe segnalare la tendenza ad esprimere un unico contenuto informativo nei *tweet* legati a situazioni di allarme.

	Giorn	**Narr**	**Suss**	**Scient**	**Scritto (media)**	**Parlato (media)**	**Tw Gen**	**Tw Dis**	**Tw Tot**
Lunghezza media catene nominali	1,29	1,17	1,21	1,40	1,27	1,16	1,26	1,16	**1,21**
Dipendenti per testa verbale	2,15	1,80	1,94	2,01	1,98	**1,95**	1,88	1,75	**1,81**
Clausole con soggetto al passivo	0,26	0,18	0,40	0,54	0,36	**0,15**	0,08	0,16	**0,12**
Rapporto principali/subordinate	0,42	0,48	0,48	0,40	0,44	**0,36**	0,22	0,27	**0,25**
Clausole per periodo	2,36	2,33	3,21	2,54	2,61	**1,29**	1,11	0,86	**0,98**
Lunghezza media del periodo	22,87	17,61	27,64	28,73	24,21	**9,18**	13,16	9,36	**11,26**

Tabella 6: Distribuzione delle principali caratteristiche sintattiche.

Veniamo a esplorare più nel dettaglio il profilo lessicale dei corpora considerati. A questo livello, un ulteriore dato che segnala la maggior vicinanza della lingua dei *tweet* alla prosa scritta è il valore di densità lessicale (seconda riga della tabella 7). Va sottolineato che le oscillazioni di questo parametro si quantificano in differenze di unità centesimali, pertanto un valore medio pari allo 0,591 nel corpus considerato, dunque superiore anche alla media dello scritto, segnala in maniera piuttosto netta la pregnanza informativa che caratterizza questa tipologia di testi.

Se la densità lessicale cattura aspetti della ricchezza del vocabolario di un testo, la Type/Token Ratio è invece un indice della sua varietà (cfr. paragrafo 2.3). È interessante notare che il valore riportato dai *tweet* rispetto a questo indice è addirittura più elevato della media dello scritto, ad indicare come la connaturata sinteticità del "cinguettio" singolo (escludendo dunque la sua propagazione tramite i cosiddetti *retweet*) sia poco incline a fenomeni di ridondanza e ripetizione. Tuttavia, quando prendiamo in considerazione la distinzione interna al corpus, questo parametro registra valori più contenuti in *TwDis* (ovviamente il dato è sempre commisurato ad un campione di *tweet* pari a 100 tokens), a segnalare una maggior uniformità nella scelta delle parole con cui utenti diversi affidano ai loro *tweet* l'espressione di contenuti che ruotano attorno ad hashtag quali #terremoto #allertameteo #earthquake (alcuni dei più rappresentativi all'interno del corpus qui considerato).

	Giorn	Narr	Suss	Scient	Scritto (media)	Parlato (media)	Tw Gen	Tw Dis	Tw Tot
Type/Token Ratio	0,63	0,71	0,69	0,67	**0,67**	0,62	0,83	0,70	**0,77**
Densità lessicale	0,564	0,573	0,557	0,578	**0,568**	0,546	0,588	0,594	**0,591**

Tabella 7: Distribuzione delle principali caratteristiche lessicali.

Infine, in tabella 8, riportiamo alcuni dati sull'uso della morfologia verbale e, in particolare, sulla distribuzione dei tratti morfologici di tempo e persona nei verbi. Anche questi dati forniscono un'ulteriore evidenza di come il repertorio di forme grammaticali nei *tweet* prodotti in momenti di allerta sia distinguibile da quello di argomento generico, indipendentemente dal contenuto specifico veicolato e dalla natura comune del supporto. Quest'ultima si può rintracciare nella netta prevalenza di verbi al presente (prima riga della tabella 8), a testimonianza di come la comunicazione in *Twitter* sia per definizione legata all'attualità. Tuttavia, se si analizzano le distribuzioni dei verbi in base ai tratti di numero e persona emergono alcune differenze non trascurabili. In particolare, sebbene l'uso della prima persona, soprattutto al singolare, sia caratteristico di entrambe le varietà di *tweet*, ma con una predominanza in *TwDis* (16,02) rispetto a *TwGen* (11,69), è il ricorso alla seconda persona singolare a marcare la distinzione interna ai due corpora. Essa supera infatti il 10% nei cinguettii di argomento generico (una percentuale superiore anche alla media del parlato), ma raggiunge appena il 2% in *TwDis*. Questa variazione sembra suggerire una minor propensione all'*addressivity*, che pur è centrale in questo genere di comunicazione del web[12], quando il proposito dell'utente non è tanto quello di instaurare un dialogo con un altro interlocutore della piattaforma (esempi (2) e (4)), quanto quello di comunicare il proprio stato d'animo in una situazione di pericolo percepito (es. (5)).

	Giorn	Narr	Suss	Scient	Scritto (media)	Parlato (media)	Tw Gen	Tw Dis	Tw Tot
Verbi al presente	55,55	35,32	39,8	51,88	45,63	**83,67**	88,14	91,56	**89,95**
Verbi al futuro	3,56	2,61	1,36	1,15	2,17	2,59	6,23	2,12	4,17
Verbi all'imperfetto	4,68	16,59	15,89	1,17	9,58	11,21	5,15	5,78	5,47
Verbi al passato	1,02	11,81	7,45	1,03	5,32	2,53	0,48	0,54	0,51
Verbi alla I^a per.sing.	2,88	4,58	2,13	0,48	2,52	15,00	**11,69**	**16,02**	13,86
Verbi alla II^a per.sing.	0,45	2,03	0,71	0,67	0,96	6,29	**10,63**	**2,12**	6,37
Verbi alla III^a per.sing.	37,25	45,06	43,26	35,38	40,24	59,24	54,26	56,64	55,45
Verbi alla I^a per.pl.	2,46	2,20	1,49	1,15	1,82	5,60	4,93	4,95	4,94
Verbi alla II^a per.pl.	0,08	0,80	0,18	0,03	0,27	0,87	1,46	2,22	3,68
Verbi alla III^a per.pl.	21,68	11,67	16,72	17,53	16,90	13,00	17,04	17,94	17,49

[12] Il concetto di *addressivity* in filosofia del linguaggio e pragmatica della comunicazione esprime una proprietà costitutiva degli enunciati con cui si realizza il proposito dell'emittente di stabilire un'interazione con l'interlocutore; come osserva Zaga (2012:174), in Twitter tale funzione è assolta primariamente dal tag "@risposta", che ritroviamo infatti anche negli esempi in (2) e (3).

Tabella 8: Distribuzione della morfologia verbale rispetto ai tratti di tempo e persona.

(4) @chiarishka Mi **stai** dicendo che non hai un mutuo cointestato vero? Niente lega più di un debito a vita :D (tratto da *TwGen*)

(5) DI NUOVO scossa #terremoto a San Bovio... argh **stavo** per andare a dormire :-((tratto da *TwDis*)

4. Conclusioni

In questo contributo abbiamo illustrato le potenzialità di una innovativa metodologia di analisi e monitoraggio linguistico della lingua italiana che a partire dall'output di strumenti di annotazione linguistica automatica permette di ricostruire un profilo linguistico multi-livello di testi rappresentativi di una specifica varietà d'uso della lingua. Tale metodologia è stata applicata in particolare allo studio di una specifica tipologia di *scritture brevi*, ovvero per la ricostruzione del profilo linguistico di un corpus di testi tratti dal social network *Twitter*, contenente *tweet* di argomento generico e *tweet* che fanno riferimento a episodi di disastri naturali. I risultati raggiunti, per quanto preliminari, ci hanno permesso di cominciare a fare luce su interrogativi aperti sulla lingua del Web.

Sul versante linguistico-computazionale, abbiamo visto che è possibile rintracciare tendenze lessicali, morfo-sintattiche e sintattiche all'interno di questa tipologia testuale che, come noto, pone nuove sfide ai metodi e alle tecniche della linguistica computazionale. I risultati ottenuti hanno dimostrato come ad oggi gli strumenti di annotazione linguistica automatica siano sufficientemente affidabili e robusti per essere usati come punto di partenza per estrarre un vasto repertorio di caratteristiche lessicali, morfo-sintattiche e sintattiche dal testo linguisticamente annotato. Sebbene l'indagine qualitativa che abbiamo condotto sui *tweet* annotati a più livelli di descrizione linguistica abbia dimostrato come ci sia ampio spazio di miglioramento in termini di precisione dell'analisi linguistica automatica, tuttavia riteniamo che i risultati del monitoraggio linguistico aprano la strada a numerose prospettive di ricerca.

Sul versante linguistico, abbiamo rintracciato alcune delle caratteristiche linguistiche più idonee a caratterizzare il profilo della lingua dei *tweet* sull'asse della variazione diamesica scritto/parlato e a identificare tendenze differenti nelle due varietà di *tweet* considerate. Ovviamente, i risultati raggiunti, per quanto sicuramente promettenti, possono essere ulteriormente raffinati: ad esempio, la metodologia di monitoraggio linguistico potrebbe essere specializzata per prendere in considerazione tratti linguistici non ancora monitorati, la cui selezione dovrebbe essere guidata da studiosi esperti nello studio di *scritture brevi* in generale o della lingua del web più in particolare. Inoltre, tale metodologia potrebbe rappresentare un ausilio per questa tipologia di studi come supporto metodologico al linguista impegnato nella raccolta dei fenomeni caratterizzanti diverse forme di *scritture brevi* e nella loro generalizzazione.

Bibliografia

Antonelli, G., 2011, "Lingua", in *Modernità italiana. Cultura, lingua e letteratura dagli anni Settanta a oggi*, a cura di Andrea Afribo ed Emanuele Zinato, Roma, Carocci, pp. 15-52.

Attardi, G., Dell'Orletta, F., Simi, M., Turian, J., 2009, "Accurate Dependency Parsing with a Stacked Multilayer Perceptron", in *Evalita 2009*.

Berruto, G., 1987, *Sociolinguistica dell'italiano contemporaneo*, Carocci, Roma.

Biber, D., 1988, *Variation across speech and writing*, Cambridge & New York, Cambridge University Press.

Biber, D., 1995, *Dimensions of register variation: A cross-linguistic comparison*, Cambridge & New York, Cambridge University Press.

Bonomi, I., 2003, La lingua dei quotidiani, in Bonomi, I., Masini, A., Morgana S. (a cura di), 2003, *La lingua italiana e i mass media*, Carocci, Roma.

Chiusaroli, F., 2014, "Sintassi e semantica dell'hashtag: studio preliminare di una forma di Scritture Brevi", in *The First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa University Press, pp. 117-121.

Chiusaroli, F., Zanzotto, F.M., a cura di 2012, "Scritture brevi di oggi", Quaderni di Linguistica Zero, 1, Napoli, Università degli studi di Napoli "L'Orientale".

De Mauro, T., 2000, *Il dizionario della lingua italiana*, Torino, Paravia.

Dell'Orletta, F., 2009, "Ensemble system for Part-of-Speech tagging", in *Evalita 2009*.

Dell'Orletta, F., Venturi, G., Montemagni, S., 2013, "Unsupervised Linguistically-Driven Reliable Dependency Parses Detection and Self-Training for Adaptation to the Biomedical Domain", in *Proceedings of the 12th workshop on Biomedical Natural Language Processing (BioNLP-2013)*, Sofia, Bulgaria, August 8-9, pp. 45-53.

Gastaldi, E., 2002, "Italiano digitato", in *Italiano & oltre* 31: 134-137.

Gildea, D., 2011, "Corpus variation and parser performance", in *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, PA, pp. 167-202.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith N.A., 2011, "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, June 19-24, pp. 42-47.

McClosky, D., Charniak, E., Johnson, M., 2010, "Automatic Domain Adaptation for Parsing", in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, pp. 28-36.

M. McLuhan, Gli strumenti del comunicare, Il saggiatore, Milano, 1967.

Montemagni, S., 2013, "Tecnologie linguistico-computazionali e monitoraggio della lingua italiana", in *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, Anno XLII, Numero 1, pp. 145-172.

Petrov, S., McDonald, R., 2012, "Overview of the 2012 Shared Task on Parsing the Web", *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

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

Plank, B., Dirk, H., McDonald, R., Sögaard, A., 2014, "Adapting taggers to Twitter with not-so-distant supervision", in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pp. 1783-1792.

Sagae, K., Tsujii, J., 2007, "Dependency parsing and domain adaptation with LR models and parser ensembles", in *Proceedings of the CoNLL 2007 Shared Task in the Joint Conferences on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07 shared task)*, pp. 1044-1050.

Shapiro, L.P., Zurif, E., and Grimshaw, J. (1987). Sentence processing and the mental representation of verbs. Cognition, 27, 219-246.

Spina, S., Cancila, J. 2013, "Gender issues in the interactions of Italian politicians on Twitter: Identity, representation and flows of conversation". International Journal of Cross-Cultural Studies and Environmental Communication 2.2 (2013): 147-157.

Spina, S., 2012, *Openpolitica. Il discorso dei politici italiani nell'era di Twitter*, Milano, Franco Angeli.

Tavosanis, M., 2012, *L'italiano del web*, Carocci, Roma.

Voghera, M., 2001, *Riflessioni su semplificazione, complessità e modalità di trasmissione: sintassi e semantica*, in Dardano M., Pelo A., Stefinlongo A. (a cura di), Scritto e parlato. Metodi, testi e contesti, Atti del Colloquio internazionale di studi, Aracne, Roma, pp. 65-78.

Voghera, M., 2005, "La misura delle categorie sintattiche", in Chiari Isabella / De Mauro Tullio (a cura di) *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne, Roma, pp.125-138.

Zaga, C., 2012, "Twitter: un'analisi dell'italiano nel micro blogging", in *Italiano LinguaDue*, 1.