UNIVERSITÀ DI PISA

Corso di Laurea in Informatica

Umanistica

RELAZIONE

Applicazione di strumenti di analisi linguistica a un corpus di commenti provenienti da Facebook

Candidato: *Paolo Nasuto*

Relatore: *Mirko Tavosanis*

Correlatore: *Maria Simi*

Anno Accademico 2017-2018

Indice

Introduzione

Una nuova tipologia di campagna elettorale ha rivoluzionato il panorama politico, che ha dovuto allargare i suoi orizzonti e ritagliarsi spazio nel difficile mondo del digital marketing. I manifesti, i comizi e i megafoni non bastano più per assicurarsi il voto degli elettori. La comunicazione politica ha deciso di abbandonare sempre di più l'oscurità del suo linguaggio, spesso definito col nome di politichese[1], istituire un rapporto amichevole e quasi complice con il proprio elettore e rendere unico l'individuo suscitando in lui emozioni positive e rassicuranti. Come affermato da Tavosanis (2016, pp.677) gli studi linguistici effettuati sulla comunicazione politica italiana sono innumerevoli.

D'altra parte il linguaggio usato dai politici su Facebook possiede caratteristiche autonome interessanti, ma poco innovative rispetto a quello della comunicazione tradizionale *broadcast*.

Tra i tanti tipi di produzione linguistica offerti da Facebook, i commenti a post di personaggi pubblici (nel nostro caso politici) rientrano tra le categorie che richiedono studi più specifici. Un grande divario caratterizza l'autore del post e colui che lo commenta, comportando una forte asimmetria tra essi.

Un aspetto su cui verte principalmente il lavoro di tesi è l'analisi di un corpus di commenti provenienti dal profilo Facebook dell'ex Ministro delle politiche agricole alimentari e forestali Maurizio Martina, osservati dal 1° marzo 2016 all'11 marzo 2017, con un'attenzione particolare alla funzione assunta dal *che* all'interno delle varie frasi. Lo scopo è quello di analizzare i commenti utilizzando diversi strumenti di analisi

1 Il termine viene spesso usato dai politici stessi, in senso dispregiativo, in quanto è caratterizzato da tecnicismi e complicazioni che lo rendono incomprensibile al vasto pubblico.

linguistica per valutare la loro efficienza e accuratezza. Secondo Montemagni (2013, pp.145-172) le tecnologie linguistico-computazionali si sono dimostrate uno strumento utilissimo per il rilievo di tratti relativi a corpora sempre più estesi. Un'ulteriore riprova del grande vantaggio che si può ottenere dalla fusione di due discipline apparentemente distanti come la linguistica e l'informatica. Sarà interessante sfruttare gli strumenti di analisi linguistica e valutare quanti e quali errori commetteranno con un corpus di questo tipo.

La comunicazione politica su Facebook

2.1 Aspetti della comunicazione politica su Facebook

La politica ha compreso l'importanza della comunicazione in rete e la centralità che questa assume nella vita di tante persone. Il primo utilizzo di Internet, con una certa sistematicità, da parte dei politici italiani, coincide con le elezioni politiche del 2006; in quell'occasione, infatti, la politica italiana scopre quanto la rete, a fianco dei mezzi di comunicazione tradizionali, possa essere importante per la campagna elettorale. Inizialmente vengono utilizzati i siti Internet dei partiti e dei singoli candidati e i blog. Con il rapido sviluppo dei social network, l'aspetto sociale assume una valenza preponderante; nascono così tra la fine del 2007 e il 2008, spesso come esperimenti di singoli, i primi profili di politici su Facebook. Quella dei social media è una sperimentazione che consente da un lato di uscire dalla monodirezionalità della comunicazione televisiva, sempre meno apprezzata dal pubblico, e dall'altro di cimentarsi su una dimensione comunicativa basata sull'interazione e sul contatto diretto con i cittadini (Spina 2012, pp.62-63). Permette il passaggio, in altre parole, dalla narrazione emotiva televisiva alla narrazione relazionale della rete (Epifani *et al.*,

2011, p.20). La stampa e gli altri media iniziano a dare una certa rilevanza a questa presenza, in un momento storico caratterizzato da un forte distacco dei cittadini dalla politica. Per una figura istituzionale avere uno spazio su Facebook rende possibile costruirsi un'identità attraverso testi, immagini, video, citazioni e link condivisi. L'obiettivo primario di un esponente politico che sceglie di aprire un profilo o una pagina su piattaforme sociali quali Facebook e Twitter, è quello di allontanarsi dai mezzi di comunicazione tradizionali e recuperare un rapporto diretto con il cittadino. I politici italiani hanno mostrato una netta preferenza per Twitter in quanto, essendo un mezzo di comunicazione più elitario, arriva ad un pubblico dieci volte inferiore rispetto a quello di Facebook, consentendo loro di dialogare con i singoli utenti senza bisogno di particolari attenzioni. Su Facebook invece la comunicazione è quasi del tutto unidirezionale. Un esempio a sostegno di questa tesi è uno studio pubblicato su *Parlamento 2.0* (Bentivegna 2012, p.109) che dimostra che al momento in cui è avvenuta l'analisi il 59,9% dei parlamentari italiani in possesso di un account Facebook non aveva pubblicato un solo commento né riguardante un proprio post né riguardante lo stato di un altro utente.

In molti casi a curare i messaggi che compaiono sui profili dei politici sono i rispettivi uffici stampa. I messaggi più importanti sono redatti dai politici stessi o almeno sotto il loro diretto controllo mentre quelli di routine sono perlopiù opera degli uffici stampa. L'italiano utilizzato su Facebook dai politici vuol essere colloquiale e molto vicino al parlato (Tavosanis 2016, pp. 678-680).

D'altro canto, le numerose funzionalità offerte da Facebook, con l'integrazione di diversi formati di file (immagini e video) all'interno della pagina, rendono faticosa la ricerca delle informazioni. La difficoltà incontrata da un utente, che volesse seguire una discussione relativa ad un certo argomento su Facebook, è che un post viene immediatamente

"preso d'assalto" dagli altri, tanto da contare migliaia di commenti; reperire informazioni in modo efficiente diventa allora impossibile sia per il gestore della pagina che per l'utente stesso. I collegamenti ipertestuali che conducono ad altre risorse, come video o immagini, vengono spesso preferiti alla stesura di testi. Non è un caso il fatto che Facebook, come dimostrato da molte statistiche, sia una piattaforma sociale in cui l'attività prevalente è la lettura, e non la produzione di testi (Tavosanis 2011, p.207).

2.2 Il rapporto tra linguaggio e politica.

L'oscurità era il marchio di fabbrica del discorso politico della prima Repubblica. La perdita di credibilità dei politici dopo lo scandalo di tangentopoli (1992) ha determinato il lento declino di questa caratteristica e ha incentivato il recupero di semplicità, trasparenza e chiarezza (Spina 2012, p. 35).

In primis, si tratta di un linguaggio che mira a consolidare il consenso degli elettori, a rispondere alle loro domande e sollecitazioni e anche a difendersi dagli atteggiamenti continui di disprezzo e provocazione dei famigerati "haters", fenomeno sempre più in espansione sui social network.

In secondo luogo, viene adottata una strategia di autolegittimazione consapevole con lo scopo di mostrarsi personalità autorevoli in grado di portare a termine positivamente il proprio incarico. Uno degli obiettivi fondamentali su cui si basa il discorso politico è risultare credibili a chi ascolta, dare quindi l'impressione che chi parla abbia realmente l'intenzione e la capacità di realizzare quanto promesso (Spina 2012, p. 17). Più il consenso del pubblico è largo e convinto, più è possibile affermare che le scelte linguistiche sono state opportune.

La costruzione del consenso è un processo che si è modificato nel corso del tempo: se prima si cercava di stupire l'uditorio con il fascino delle parole difficili, ora il discorso politico deve colpire per la sua semplicità e per la sua emotività; il messaggio deve essere chiaro, ma allo stesso tempo deve suscitare delle emozioni perché in questo modo è facile che riesca a risultare credibile e venga ricordato. Il linguaggio della politica è il linguaggio del potere, della decisione e fare politica in questo senso "è un esercizio di persuasione, è una negoziazione verbale, un'interazione di natura contrattuale dove può determinarsi cooperazione oppure competizione" (Lasswell 1979).

Ne deriva che i registri utilizzati sono molto vari e informali e che il linguaggio politico punta ad innescare efficaci attività di rispecchiamento per la crescita del consenso (Desideri 1993, pp. 284-285). Il lessico viene arricchito prelevandolo da altri settori, come quello del diritto, dell'amministrazione e dell'economia. Il linguaggio filosofico, giuridico e letterario, avvertito come più prestigioso, è stato sostituito dalla terminologia finanziaria, caratterizzata dall'oggettività e dall'efficienza manageriale. Rispetto al latino, si predilige l'inglese come dimostra la fitta presenza di anglicismi nel lessico comunemente utilizzato dai politici (*governance*, *welfare*, *jobs act*). Alle citazioni d'autore subentrano le statistiche (Antonelli 2017, p.95). I numeri e la statistica infatti costituiscono una certezza di oggettività, di esattezza e di autorevolezza ed amplificano in chi ascolta quella sensazione di concretezza e di chiarezza che un numero trasmette di per sé, veritiero o meno. I rappresentanti del governo esibiscono i dati relativi al loro operato e quelli dell'opposizione presentano numeri diversi per smontare tali statistiche (Spina 2012, p.40).

Strumenti per l'estrazione dei commenti

3.1 Corpora linguistici

I corpora linguistici sono collezioni finite di informazioni, raccolte in modo rigoroso con lo scopo di riflettere la reale distribuzione (quantitativa e qualitativa) dei fenomeni linguistici presi in esame. Essendo per definizione finiti, i corpora rappresentano un sottoinsieme della *langue*. La finitezza è un requisito fondamentale per almeno due ragioni: una scientifica (a) e una pratica (b).

Per garantire l'attendibilità delle proprie affermazioni è necessario che le osservazioni fatte possano essere ripetibili (a). Il passaggio dalle schedine cartacee ai moderni corpora informatici ha permesso di realizzare operazioni statistiche sui dati (b). L'avvento del World Wide Web si è scontrato con il problema della finitezza: essendo sempre in movimento il WWW non può essere considerato un campione rappresentativo, né finito (nel senso di costituire un insieme su cui svolgere operazioni statistiche deterministiche), né definito (in grado di consentire la ripetibilità degli esperimenti). Nonostante ciò, è una risorsa inesauribile di dati linguistici, in quanto ha permesso di collezionarli sempre in quantità maggiore, ma soprattutto di facilitarne lo studio e la realizzazione di modelli computazionali (Barbera 2013, p.21). Il progresso della linguistica computazionale ha permesso la continua produzione di risorse linguistiche apportando un sostanziale vantaggio all'analisi statistica. Si è reso possibile il confronto tra caratteristiche di testi differenti, la determinazione di affinità e diversità che difficilmente si sarebbero potute analizzare senza strumenti computazionali. Questo approccio alla linguistica permette una considerevole capacità di elaborazione, sia in termini quantitativi sia qualitativi, che porta ad un rapido sviluppo di analisi.

3.2 Graph API Explorer

L'estrazione degli Status dei politici dal social network è permessa dalle API (Application Programming Interface) fornite da Facebook. L'accesso a tali interfacce è consentito attraverso la maggior parte dei linguaggi di programmazione, preceduto dalla richiesta di un Access Token (v. fig. 2), cioè una stringa di testo per l'autenticazione temporanea, nella quale viene monitorata la nostra sessione utente e riconosciuta la cattiva gestione delle risorse.

API Graph è il nome che prendono le API di Facebook. Si basano su HTML e vengono utilizzate per scaricare dati o caricare contenuti. Il nome "Graph" si riferisce all'organizzazione dei dati di Facebook, strutturati come un grafo. I nodi sono costituiti da pagine, utenti, foto ed eventi e si collegano fra loro con archi che rappresentano le possibili relazioni che possono intercorrere tra i nodi, come ad esempio amicizia, condivisione o tag.

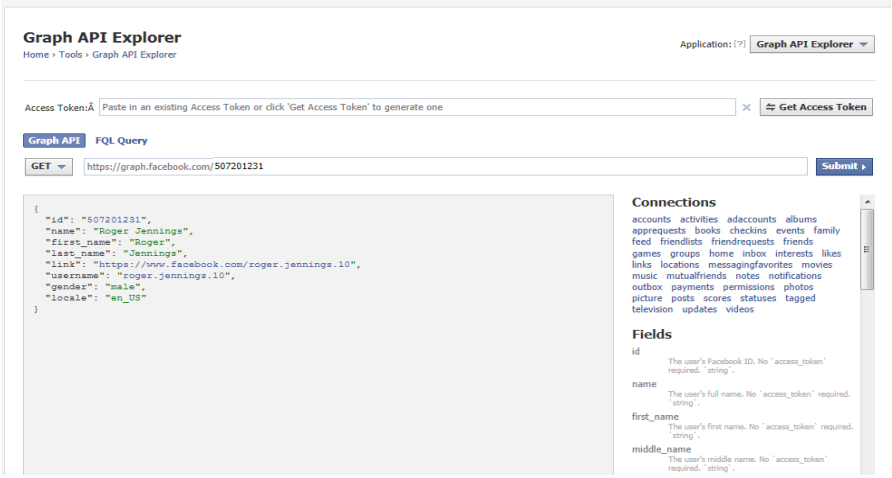

FIG. 1 API GRAPH EXPLORER

Un esempio di utilizzo del metodo GET su una pagina di un profilo Facebook viene fornito in fig.1. Il funzionamento di API Graph è chiarito da Facebook che mette a disposizione sul sito *facebook for*

developers[2] un tool di esplorazione in cui è consentito indicare l'id della risorsa, quale metodo applicare e i campi a cui essere applicato.

3.3 Descrizione del corpus

Il lavoro di analisi svolto proviene da dati già raccolti giunti dall'elaborazione della tesi magistrale della dott.ssa Tortorelli. Gli script sono stati eseguiti per tutte le pagine tra l'11 e il 12 marzo 2017.
Python[3] è stato scelto dalla dott.ssa Tortorelli come linguaggio di programmazione (dato che per quanto riguarda l'analisi di testi, e non solo, è uno dei più fruibili) per scaricare i post e i commenti, attraverso i metodi messi a disposizione da API Graph.
È necessario fare una precisazione riguardo allo scraper utilizzato per questo lavoro di analisi: dopo l'introduzione del nuovo regolamento europeo per la protezione dei dati (GDPR) e lo scandalo del caso Cambridge Analytica, Facebook ha ristretto molto la possibilità di raccogliere dati e l'attività di ricerca svolta dalla dott.ssa Tortorelli non può più essere ripetuta.
Per questa attività era stato utilizzato uno scraper scritto e realizzato da Max Woolf e distribuito sotto licenza MIT (consentito l'utilizzo, lo studio, la modifica e la ridistribuzione) sulla piattaforma GitHub[4]. In origine lo scraper era stato pensato per scaricare tutti i post e tutti i commenti disponibili. La dott.ssa Tortorelli ha modificato tale opzione e ha reso possibile scegliere il numero esatto di post e di commenti da scaricare. La nuova versione dello scraper presenta un'altra differenza importante rispetto all'originale: i commenti di risposta ai commenti (subcomments) non vengono più scaricati. Questo è un aspetto essenziale nella nostra analisi, poiché la risposta a un commento tra utenti non

2 developers.facebook.com

3 Sito ufficiale della comunità italiana Python: https://www.python.it/

4 Link per scaricare lo scraper da GitHub: https://github.com/minimaxir/facebook-page-post-scraper

rappresenta più un caso di comunicazione tra utente e personaggio pubblico.

3.4 Scelta del Ministro

Prima di partire con l'analisi linguistica vera e propria è necessario individuare il Ministro i cui post generano i commenti su cui verrà concentrato il lavoro. Su 21 componenti totali dell'esecutivo della XVII legislatura, che rimase in carica dal 12 dicembre 2016 al 1 giugno 2018, solo 14 tra ministri, sottosegretari e primo ministro avevano pagine attive. È stato deciso di prendere in considerazione l'ormai ex Ministro delle politiche agricole alimentari e forestali Maurizio Martina.

I criteri secondo il quale è stato scelto sono:

- popolarità; secondo un sondaggio Ipsos[5] risalente al 29 luglio 2017, il Ministro Martina era al quinto posto della classifica di popolarità dei ministri, con indice di gradimento di 27.
- pertinenza linguistica; i commenti generati sono tutti in lingua italiana e quelli minacciosi, offensivi e diffamatori costituiscono una minoranza, a differenza, per esempio, di quelli contro Angelino Alfano (dodicesimo posto con indice a 16) e Beatrice Lorenzin (nono posto, con 21 di indice di gradimento) maggiormente sottoposti al fenomeno mediatico.

La lingua utilizzata nei commenti, con un occhio di riguardo alla funzione del **che**, sarà l'aspetto principale di questa analisi.

5 Sondaggio consultabile all'indirizzo: https://www.corriere.it/politica/cards/popolarita-leader-ministri-classifica/tra-ministri-minniti-primo-posto.shtml

Tanl Pipeline

4.1 Introduzione a Tanl

Tanl (Text Analytics and Natural Language) è un servizio Web costituito da una suite di moduli per l'analisi del testo e l'elaborazione del linguaggio naturale. Tale pipeline è in grado di:

- dividere il testo in frasi;
- estrarre lemma, Part-of-Speech[6] e aspetti morfologici per ciascun token;
- trovare e classificare ogni singolo elemento presente in un testo all'interno di categorie predefinite (Named-entity recognition);
- costruire alberi di dipendenza.

La pipeline effettua diversi tipi di analisi ed è composta infatti da un segmentatore di frasi (Sentence Splitter), un segmentatore di token (Tokenizer), un analizzatore morfosintattico (Part-of Speech tagger), un parser a dipendenze (Dependency Parser), e un riconoscitore e classificatore di entità nominali (Named Entity Tagger).

4.2 Tanl POS Tagset

Tanl si basa sul tagset ILC / PAROLE[7] ed è conforme allo standard internazionale EAGLES[8]. Come caso di studio per la suite Tanl sono stati annotati dagli sviluppatori due sottoinsiemi significativi di Wikipedia, uno in lingua inglese e uno in lingua italiana (oltre 660.000 articoli per un totale di 5.507.225 di frasi). Il tagset è costituito da due livelli diversi di granularità: si passa dalla granularità grossolana

6 Part-of-speech, o parte del discorso in italiano, è la categoria grammaticale a cui appartiene una parola.

7 http://webilc.ilc.cnr.it/viewpage.php/sez=ricerca/id=33/vers=ita

8 http://www.ilc.cnr.it/EAGLES96/home.html

(*coarse*) per componenti relativamente grandi, alla granularità fine (*fine*) per componenti più piccoli.

La tabella 1 mostra i tag di Part-of-Speech a grana grossa (14) utilizzati per l'annotazione ISST-CoNLL[9].

Value	Description
A	adjective
B	adverb
C	conjunction
D	determiner
E	preposition
F	punctuation
I	interjection
N	numeral
P	pronoun
R	article
S	noun
T	predeterminer
V	verb
X	residual class

Tabella 1. Coarse-grained

9 Il corpus ISST-CoNLL è stato sviluppato attraverso una collaborazione tra il Dipartimento di Informatica dell'Università di Pisa e l'Istituto di Linguistica Computazionale (ILC) del Consiglio Nazionale delle Ricerche (CNR).

I tag di Part-of-Speech a grana fine, invece, presentano maggiori specificità. I verbi, per esempio, vengono suddivisi nelle seguenti categorie:

- **VA**: verbi ausiliari (essere, avere, venire).
- **VM**: verbi modali (volere, potere, dovere).
- **V**: verbi principali.

Oltre a queste categorie viene indicato anche il modo e il tempo, come possiamo notare in un estratto della tabella 2.

Vif	main verb indicative future
Vm	main verb imperative
Vcp	main verb conjunctive present
Vci	main verb conjunctive imperfect
Vd	main verb conditional
Vf	main verb infinite
Vg	main verb gerundive
Vpp	main verb participle present
Vps	main verb participle past

Tabella 2. Estratto fine-grained

Il POS tagset comprende infine una lista di morphed tags che includono informazioni morfologiche codificate come segue:

- **genere**: m (maschio), f (femmina), n (non specificato).
- **numero**: s (singolare), p (plurale), n (non specificato).
- **persona**: 1 (prima), 2 (seconda), 3 (terza).
- **modo**: i (indicativo), m (imperativo), c (congiuntivo), d (condizionale), g (gerundio), f (infinito), p (participio).

- **tempo**: p (presente), i (imperfetto), s (passato), f (futuro).
- **clitico**: c segna la presenza di clitici.

Annotazione manuale

5.1 Quadro di analisi generale

Il piano di lavoro è stato suddiviso in fasi, ognuna delle quali con un compito ben preciso per una valutazione corretta e accurata dei commenti presi in esame. Dal corpus della Dott.ssa Tortorelli, di circa 263,000 token raccolti da 14 diverse pagine Facebook, si è cominciato ad analizzare i commenti a post di Maurizio Martina.

La fase preliminare dell'analisi è focalizzata sulla selezione dei dati attraverso un'operazione manuale. In particolare dal corpus di commenti estratti da Facebook sono state selezionate le prime 251 frasi in cui è presente un *che* e in seguito il lavoro di analisi è stato ripetuto con riferimento ad altri 253 *che*. Il Gold Standard è stato unificato attraverso il comando cat del sistema Unix (descritto nella sezione 7.2) ed è composto da 504 *che*. Non sono stati presi in considerazione tutti gli altri commenti dal momento che questi non sono utili ai fini della nostra ricerca.

Nella seconda fase con l'aiuto di Tanl è stato possibile realizzare l'annotazione morfosintattica: a ogni parola è stata associata la relativa categoria grammaticale. L'annotazione consiste nell'attribuzione di un'etichetta (tag) ed è stata osservata nei vari commenti l'interpretazione del *che* fornita dalla pipeline.

La terza e ultima fase del lavoro manuale è caratterizzata dalla correzione degli errori commessi da Tanl, dalla discussione dei casi dubbi che meritano una nota di approfondimento e dalla riflessione su alcuni spunti interessanti forniti dalla pipeline.

5.2 Annotazione morfosintattica

Operazione fondamentale è stata la creazione di una tabella contenente i commenti prescelti che sono stati numerati e classificati. La classificazione è stata effettuata secondo le seguenti voci:

- **Tag Tanl**: rappresenta l'interpretazione data dalla pipeline per l'assegnazione del **che** al commento oggetto di analisi.
- **Pos corretto?** : la domanda sta ad indicare se la valutazione precedentemente effettuata è corretta o no. La casella va riempita rispondendo SI o NO.
- **Interpretazione corretta**: qualora la valutazione precedentemente effettuata fosse ritenuta non corretta, nella casella viene specificato il Part-of-Speech corretto.
- **Eventuali note**: indicano informazioni aggiuntive che con il semplice tag non è possibile fornire.

Questo lavoro di analisi è stato ripetuto manualmente per tutti i 250 commenti. Nella tabella 3 sono riportati i tag utilizzati per l'etichettatura del **che**. Nel caso del pronome relativo (PRnn), interrogativo (PQnn) ed esclamativo (DEnn) e nel caso dell'aggettivo interrogativo (DQnn) le lettere "nn" indicano che il programma non specifica il numero e il genere.

PRnn	Pronome relativo
CS	Congiunzione subordinativa
CC	Congiunzione coordinativa
PQnn	Pronome interrogativo
E	Preposizione
DQnn	Aggettivo interrogativo

A	Aggettivo
DEnn	Pronome/aggettivo esclamativo
SP	Nome proprio
SFP	Nome comune

Tabella 3. Legenda tag utilizzati

5.3 Analisi dettagliata degli errori

Osserviamo innanzitutto quante volte occorre ciascuna categoria grammaticale, riassumendo tale dato nella tabella delle occorrenze totali che segue. A questi dati devono essere aggiunti i casi di D.A. (3) e quelli particolari (4) discussi in seguito.

Tag	Totale occorrenze
PRnn	123
CS	91
CC	9
DQnn	10
E	6
PQnn	2
SP	1
SFP	1

Tabella 4. Occorrenze totali per ciascuna categoria grammaticale

Percentuale di errore sul totale della categoria:

Tag	Occorrenze ✗	%
PRnn	29	23,57
CS	19	20,87
CC	9	100
DQnn	10	100
E	6	100
PQnn	2	50
SP	1	100
SFP	1	100

Tabella 5. Percentuali di errore per ciascuna categoria grammaticale

Sono stati sottoposti al vaglio di Tanl tutti i commenti precedentemente scelti, con i seguenti risultati: 167 interpretazioni risultano essere corrette mentre sono stati commessi 79 errori di etichettatura (32,11%). La tabella sottostante evidenzia le tipologie di errore e il numero di occorrenze per ciascuno di essi, specificando nel dettaglio il tag scorretto assegnato dalla pipeline e quello corretto successivamente inserito.

Tanl ✗	Corretto ✓	Occorrenze	% occorrenza
PRnn	CS	22	27,84
CS	PRnn	16	20,25
CC	PRnn	5	6,32
DQnn	A	5	6,32
PRnn	PQnn	4	5,06

E	CS	3	3,79
DQnn	CS	3	3,79
D.A.	D.A.	3	3,79
E	PRnn	2	2,53
PRnn	DE	2	2,53
PQnn	A	1	1,26
CS	PQnn	1	1,26
DQnn	DE	1	1,26
E	CS	1	1,26
CC	DQnn	1	1,26
PRnn	A	1	1,26
CC	Congiunzione comparativa	1	1,26
SP	PRnn	1	1,26
SFP	CS	1	1,26
CC	CS	1	1,26
CC	A	1	1,26
CS	Locuzione congiuntiva	1	1,26
DQnn	Locuzione congiuntiva	1	1,26
CS	Locuzione congiuntiva	1	1,26

Tabella 6. Analisi degli errori

Come possiamo osservare l'errore di etichettatura più frequente prodotto da Tanl, con il 27,84 % dei casi, è rappresentato dal caso in cui la pipeline ha assegnato al *che* la funzione di pronome relativo, quando l'interpretazione corretta risultava essere congiunzione subordinativa. Il secondo errore più frequente, non molto distante con il 20,25 % dei casi è la situazione contraria: Tanl ha etichettato il *che* come congiunzione subordinativa invece che come pronome relativo. Questi due casi costituiscono assieme circa il 50% degli errori di etichettatura totali.

È stato deciso di racchiudere nel tag denominato D.A. i casi di difficile assegnazione, quelli a cui è davvero complicato riuscire ad attribuire un valore poiché collocati all'interno di una frase formulata in modo grammaticalmente scorretto.

Un'altra importante considerazione da fare è che Tanl è case sensitive, cioè è in grado di distinguere due parole uguali in base all'uso di lettere maiuscole o minuscole. L'interpretazione del **che** assegnata dalla pipeline ad una stessa frase scritta con le minuscole e le maiuscole è differente, come nei seguenti esempi:

" **Che** Dio vi protegga " → DQnn

"CHE DIO VI PROTEGGA" → CS

Nella prima frase il *che* viene etichettato come pronome determinativo mentre nella seconda come congiunzione subordinativa.

L'intero commento scritto in maiuscolo, come nel caso illustrato, cambia completamente l'interpretazione data dallo strumento. Sono stati rintracciati altri due esempi particolari:

[...] NIENTE DI QUESTE MULTINAZIONALI **CHE** VOGLIONO DIRCI COSA MANGIARE.

SPERIAMO **CHE** IL GOVERNO E L'EUROPA SI PRONUNCI.

Le etichette assegnate dalla pipeline al *che* sono rispettivamente di nome proprio (SP) invece di pronome relativo e di nome comune (SFP) invece di congiunzione subordinativa. Se le stesse frasi vengono scritte in minuscolo Tanl invece attribuisce al *che* la funzione corretta.
Un altro esempio che merita la nostra attenzione si può considerare osservando la forma singolare e plurale del pronome dimostrativo quello:

*[…] ma noi siamo **quello che** facciamo e non **quello che** diciamo di essere → CS* ✗

*Populisti sono **quelli che** hanno ragione di lamentarsi → Prnn* ✓

Nel primo caso, si rileva un errore di etichettatura poiché Tanl assegna al pronome la funzione di congiunzione subordinativa, quando è pronome relativo. Nel secondo caso invece, dove il pronome è al plurale, viene fornita l'interpretazione corretta.
Infine si riscontra che la pipeline Tanl non ha etichette per il *che* polivalente, per il *che* locuzione congiuntiva e per il *che* congiunzione comparativa.

5.4 Discussione dei casi particolari

Nell'analisi dei commenti ai post di Maurizio Martina sono emersi alcuni casi, particolarmente interessanti ai fini di questa ricerca, in cui interpretare correttamente la funzione del *che* è stato complicato.

*"[…] sono la conseguenza del disastro economico culturale che avete combinato in ormai anni **che** state al governo"*

Alla prima analisi di questa frase il *che* appare un pronome male usato al posto di *in cui*; in verità questo uso del *che* è attestato nella lingua orale e scritta, come afferma Luca Serianni, per collegare una dipendente a una subordinata e "si parla di *che* subordinante generico, o *che* polivalente" (Serianni 2016, p.569). In questo caso si può definire un *che* polivalente con valore temporale.

"[...] non ha niente a ***che*** *spartire con tutti gli indagati del tuo partito [...]"*

Il modulo *avere a che fare*, presente anche nell'edizione Ventisettana dei Promessi Sposi, è frequentemente utilizzato nell'italiano moderno, scritto e parlato, in particolare nelle frasi di segno negativo. Come osserva Ornella Castellani Pollidori (2004, pp.425-450) ciò che colpisce nel modulo *avere a che fare* è la presenza contemporanea della preposizione e del pronome: uno dei due elementi pare essere di troppo, dato il senso transitivo di *fare*. Per analogia il processo è arrivato a coinvolgere anche il costrutto *non aver nulla a che spartire*. Entrambi i moduli si inseriscono all'interno della diffusa tendenza ad estendere l'uso del *che*, con significato generico, "anche per introdurre subordinate che nell'italiano standard avrebbero più spesso congiunzioni subordinanti semanticamente più precise" (Fiorentino, *Enciclopedia dell'Italiano*, 2010).

"E certo ***che*** *voto i 5 stelle [...]"*

Per inquadrare questo esempio possiamo fare riferimento ai tipi di frasi nucleari la cui testa è costituita apparentemente da un aggettivo o un nome; in realtà si tratta di "assertori" (Marchello – Nizia 1999, p.68), un gruppo di strutture che servono a introdurre frasi nucleari come nel seguente esempio:

certo che vengo.

Nel caso preso in esame il che svolge la funzione di congiunzione che introduce una soggettiva in dipendenza da un assertore.

*"Mamma mia **che** teste disabitate **che** avete"*

I due *che* presenti nella frase rientrano nella categoria del che esclamativo (aggettivo esclamativo con variante intensiva). Questo costrutto si adopera frequentemente davanti a un aggettivo e, nonostante sia criticato da molti grammatici, è oggi frequentemente utilizzato nel parlato e nello scritto (Serianni 2016, p.324).

5.5 *Che* polivalente

Nell'italiano standard la congiunzione *che* introduce alcune tipologie di subordinate: le oggettive, le soggettive, le dichiarative e le relative oltre che in altri di tipi di subordinate avverbiali. Oggi nel parlato colloquiale e spontaneo si tende, con sempre maggiore frequenza, ad estendere l'uso del *che* per introdurre subordinate che nell'italiano standard sono introdotte da congiunzioni diverse e semanticamente più idonee. Come osserva Berruto (1998, p.68) questo fenomeno si inserisce all'interno della più generale tendenza alla ristandardizzazione[10] della lingua. In particolare si parla di *che* polivalente laddove si adopera la congiunzione per introdurre subordinate esplicative-consecutive, causali, consecutivo-presentative, relative temporali, finali, enfatizzanti-esclamative e pseudorelative. Si annovera tra i casi del *che* polivalente anche l'estensione del che in luogo delle forme relative *preposizione + cui* e *preposizione + art. + quale* (Palermo 2015, p.208). In alcuni casi il *che* polivalente si accompagna a un pronome clitico che sostituisce l'elemento

10 La ristandardizzazione è il fenomeno che determina la progressiva accettazione nelle grammatiche di fenomeni già presenti nell'uso da secoli ma tenuti ai margini dell'italiano normato.

relativizzato (Treccani, voce *che* polivalente), come nell'esempio che segue:

*"è una cosa **che** l'ha detta il ministro"*

Nel nostro corpus questa particolare forma compare una sola volta, come commento a un post di Maurizio Martina:

*"[…] la proposta **che** vuoi proporla agli altri."*

Le attestazioni del *che* polivalente sono di tipo relativo-temporale e causale in cui *che* viene spesso usato come aferesi[11] di perché. Ne abbiamo un esempio nel nostro corpus di commenti:

"In *bocca al lupo ma datevi da fare **che** il paese è in ginocchio"*.

In alcuni casi il *che* polivalente, violando in modo particolarmente evidente le regole di formazione della frase relativa, viene escluso dallo standard e qualifica l'enunciato come popolare (italiano popolare). Tale fenomeno è efficacemente rappresentato dal seguente esempio:

"*tu mi hai fatto convincere a me di una cosa **che** io prima non ero convinto.*"

Appartengono alla sfera dell'italiano popolare anche l'uso del che polivalente per rafforzare un'altra congiunzione ("quando *che* ci vediamo?") ed infine i casi di *che* polivalente il cui valore sintattico non può essere stabilito ("prestami la penna *che* te la do subito?").

11 Fenomeno fonetico che consiste nella caduta di uno o più foni all'inizio della parola

Strumenti di analisi linguistica

6.1 Il progetto Universal Dependencies

Questa prima analisi effettuata sulla funzione assegnata al *che* contiene alcuni tag (ad esempio DQnn e A) la cui denominazione è stata successivamente unificata secondo lo standard UD (Universal Dependencies), in modo tale da renderla universale per tutte le lingue. Infatti, quello delle Universal Dependencies[12] è un progetto che nasce nel 2014, finalizzato alla definizione di uno schema di annotazione sintattica a dipendenze applicabile a tutte le lingue. Questo framework è composto da una open community con oltre 200 contributori che producono più di 100 treebanks in oltre 70 lingue. L'obiettivo fondamentale del progetto è quello di sostituire tutti gli schemi finora esistenti e, per farlo, si propone di fornire un tagset che permetta di enfatizzare le similarità fra lingue senza però appiattirne le differenze quando necessario (Nivre et. al., 2016, pp.1659-1666). L'annotazione UD segue un principio lessicalista della sintassi, per cui le parole sono le unità minime dell'analisi grammaticale. Le caratteristiche morfologiche sono codificate come proprietà delle parole e non vi è alcun tentativo di segmentare le parole in morfemi. La segmentazione è un compito non banale in molte lingue poiché varia a seconda del sistema di scrittura e della lingua stessa.

Le risorse linguistiche messe a disposizione da UD sono costruite attraverso il progressivo susseguirsi di certe operazioni. Queste sono, in ordine, tokenizzazione, annotazione morfologica e annotazione sintattica. Nello schema UD la descrizione morfologica di una parola sintattica consiste di tre livelli di rappresentazione:

- un *lemma*, corrispondente al contenuto semantico della parola e determinato da dizionari e lessici specifici.

12 http://universaldependencies.org/

- un *part-of-speech tag*, corrispondente alla categoria lessicale astratta della parola. Nel nostro lavoro di analisi abbiamo preso in considerazione i due campi di CPOSTAG e di POSTAG (descritti in dettaglio nella sezione 7.1).
- un insieme di tratti linguistici (*features*) che riguardano le proprietà lessicali e grammaticali associate al lemma o alla specifica forma della parola.

Generalmente i lemmi sono determinati da dizionari o lessici specifici per ogni lingua, mentre le *part-of-speech tag* e le proprietà grammaticali vengono estratti da un inventario universale predefinito. Per quanto riguarda le *feature* vengono fornite informazioni aggiuntive sulla parola, sulla sua categoria grammaticale e sulle proprietà morfosintattiche. Ciascuna *feature* si presenta nella forma Nome=Valore e non vi sono restrizioni sul numero di tratti linguistici che possono essere applicati ad una parola. Di seguito si riportano le tabelle con la lista di POS e *feature* disponibili nella versione 2.0.

Classe aperta di parole	**Classe chiusa di parole**	**Altro**
ADJ: aggettivo	ADP: apposizione	PUNCT: punteggiatura
ADV: avverbio	AUX: verbo ausiliare	SYM: simbolo
INTJ: interiezione	CCONJ: congiunzione coord.	X: altro
NOUN: nome	DET: determinante	
PROPN: nome proprio	NUM: numerale	
VERB: verbo	PART: particella	
	PRON: pronome	

	SCONJ: congiunzione subordinante	

Tabella 7. Part-of-speech tag

Tratti lessicali	**Tratti flessivi**	
PronType: tipo pronominale	***Nominali***	***Verbali***
NumType: tipo numerale	Gender: genere	VerbForm: forma verbale
Poss: possessivo	Animacy: simile al genere	Mood: modo
Reflex: riflessivo	NounClass: categoria lessicale di nomi	Tense: tempo
Foreign: parola straniera	Number: numero	Aspect: aspetto
	Case: caso	Voice: diatesi
Abbr: abbreviazione	Definite: stato	Evident: funzionalità specifica lingua turca
	Degree: grado di comparazione	Polarity: val. Pos/Neg
		Person: persona
		Polite
		Clusivity

Tabella 8. Tratti linguistici (features)

6.2 UDPipe

È una pipeline addestrata per la tokenizzazione, l'etichettatura, la lemmatizzazione e l'analisi delle dipendenze dei file CoNLL-U. UDPipe rappresenta un'eccellente fonte di dati e può essere addestrato con dati annotati nel formato CoNLL-U. Probabilmente l'uso più comune di UDPipe[13] è quello di tokenizzare, taggare e analizzare l'input presumendo che questo sia nella codifica UTF-8 già tokenizzato e segmentato. Oppure può trattarsi anche di un semplice testo che verrà tokenizzato e segmentato automaticamente, esattamente com'è stato fatto per l'analisi dei commenti a post di Maurizio Martina.

Come si può osservare dalla figura 2 c'è la possibilità di scegliere tra le 71 lingue presenti e di usufruire di alcune opzioni avanzate in cui si può specificare il formato di input e se viene spuntata l'opzione tokenizer l'input viene assunto come testo semplice e viene tokenizzato utilizzando il tokenizzatore del modello.

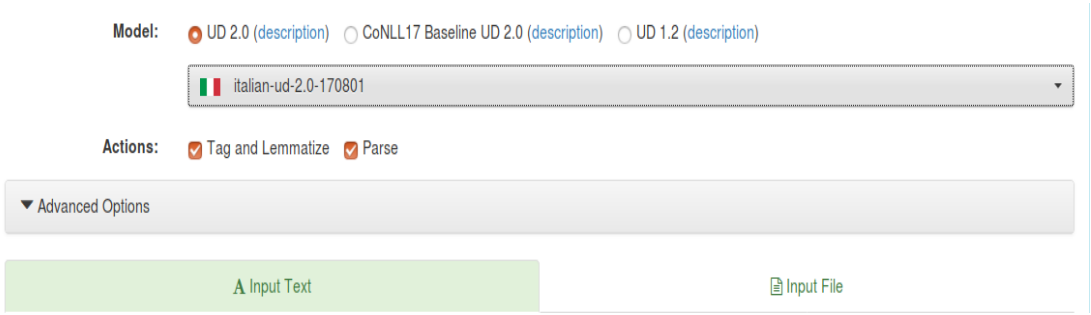

Figura 2. Interfaccia UDPipe

Inoltre il testo può essere analizzato optando per i due seguenti metodi:

- **Input Text**, inserendo direttamente il testo nell'area predisposta.
- **Input File**, caricando il file dal proprio computer.

13 http://lindat.mff.cuni.cz/services/udpipe/run.php

I modelli con il quale è stato effettuato il lavoro sono basati esclusivamente su treebanks di Universal Dependencies 2.0. UDPipe è un software gratuito distribuito con Mozilla Public License 2.0 e i modelli linguistici sono gratuiti per uso non commerciale e distribuiti sotto licenza CC BY-NC-SA.

Risorse di apprendimento

6.2.1 ISDT

La ISDT[14] (*Italian Stanford Dependency Treebank*) nacque per essere la prima risorsa italiana annotata secondo il formalismo *Stanford Dependencies* e fu rilasciata in occasione del *dependency parsing shared task di Evalita 2014*. Venne poi utilizzata come punto di partenza per la definizione, tramite conversione, di IUDT, il corpus annotato secondo il modello delle Universal Dependencies, pubblicato per la prima volta nel gennaio 2015 (come dichiarato sul sito ufficiale UD).

Il training set della risorsa è costituito da 13,121 frasi, che si suddividono in 257,616 tokens, estratte da testi appartenenti a diversi generi testuali: articoli giornalistici, testi di ambito giuridico e articoli di giornale scritti in italiano semplificato. Ai fini della nostra ricerca questa risorsa è stata utilizzata con Tanl e con UDPipe.

6.2.2 PoSTWITA

La crescente dipendenza da Internet e dai social media nella vita di tutti i giorni hanno portato alla rapida espansione dei contenuti generati dagli utenti. L'interesse per la valutazione automatica dei testi dei social media, come i tweet sta crescendo considerevolmente: Twitter è tra i principali fornitori di questo tipo di contenuti. PoSTWITA[15] è

14 https://github.com/UniversalDependencies/UD_Italian-ISDT/blob/master/README.md

15 http://corpora.ficlit.unibo.it/PoSTWITA/

una raccolta di tweet italiani, annotati secondo lo standard Universal Dependencies, che possono essere sfruttati per l'addestramento dei sistemi di PNL per migliorare le loro prestazioni sui testi dei social media. È stato creato arricchendo il dataset per il lavoro di EVALITA 2016[16] delle Part-of-Speech tagging dei social media (Bosco 2016) e rappresenta la lingua italiana di questi ultimi. Il lavoro sul Part-of-Speech tagging reso possibile da Fabio Tamburini (Università di Bologna), Oronzo Antonelli (Università di Bologna), Alberto Lavelli (FBK, Trento) e Alessandro Mazzei (Università di Torino) si è principalmente concentrato su testi standardizzati per molti anni. Il corpus originale consiste di 6,438 tweet nel set di sviluppo e 300 tweet nel test set (4.759 token) annotati dal punto di vista morfologico. Il training set è costituito da 5,368 tweet (99,441 parole). Il processo di conversione e di annotazione sintattica è stato eseguito alternando fasi di scripting automatico e di revisione manuale. Anche i risultati di analisi sono stati sottoposti a una revisione manuale da parte di due annotatori indipendenti, fondamentale per valutare l'affidabilità dell'annotazione. Il motivo per cui è stata considerata la risorsa PoSTWITA è che il genere testuale dei tweet si avvicina molto al lavoro che è stato effettuato in questa tesi avendo caratteristiche molto simili ai commenti di post su Facebook.

16 EVALITA 2016 è un'iniziativa di AILC (Associazione Italiana di Linguistica Computazionale).

Risultati

7.1 Gold Standard (GS)

Per confrontare il corpus di commenti provenienti da Facebook con diversi strumenti di analisi linguistica è stata necessaria la creazione di un Gold Standard test set, ovvero la porzione del test corpus annotato a mano che rappresenta l'output "corretto" di riferimento, idealmente privo di errori. Il nostro Gold Standard di riferimento è composto da 504 *che,* selezionati dai 3772 commenti a post di Maurizio Martina. Il progetto UD mette a disposizione le risorse linguistiche annotate in un formato chiamato CONLL-U, una versione rivisitata del formato CONLL-X, codificate in file *plain text* (UTF-8). Le righe di commento sono introdotte dal simbolo #, mentre le righe di parola contengono l'annotazione di un token, al quale sono associate diverse tipologie di informazione ciascuna contenuta in un campo separato dagli altri attraverso una singola tabulazione. I campi rappresentano le seguenti informazioni:

1. **ID**: indice di parola, numero intero pari a 1 all'inizio di ogni frase
2. **FORM**: forma della parola o simbolo della punteggiatura
3. **LEMMA**: lemma o radice della parola
4. **CPOSTAG**: part-of-speech tag universale
5. **POSTAG**: part-of-speech tag specifico della lingua in analisi
6. **FEATS**: lista di tratti morfologici
7. **HEAD**: testa del token corrente, che può essere il valore dell'ID o lo zero (0)
8. **DEPREL**: relazione di dipendenza del token corrente con l'HEAD (root se e solo se la Head è uguale a 0) o un particolare sottotipo specifico della lingua in analisi

9. **DEPS**: lista di dipendenze secondarie
10. **MISC**: qualsiasi annotazione eventuale

La creazione di un Gold standard ha reso possibile il confronto con queste risorse addestrate su ISDT e PoSTWITA. In figura 3 possiamo osservare un esempio in formato CONLL-U del nostro Gold di riferimento.

SENT	ID	FORM	LEMMA	CPOS	POS	FEATS: TRATTI MORFOLOGICI
1.	1.	Le	il	DET	RD	Gender=Fem|Number=Plur|PronType=Art
	2.	generazioni	generazione	NOUN	S	Gender=Fem| Number=Plur
	3.	che	che	PRON	PR	PronType=Rel
	4.	non	non	ADV	BN	PronType=Neg
	5.	vi	vi	PRON	PC	Clitic=Yes|Number=Plur|Person=2|PronType=Prs
	6.	votano	votare	VERB	V	Mood=Ind|Number=Plur|Person=3|Tense=Pres

Figura 3. Gold standard

7.2 Risultati a confronto

Per svolgere al meglio il lavoro è necessario avere a disposizione strumenti di annotazione linguistica efficienti.

Come affermato da Nivre (2015, pp.3-16) per avere uno strumento di annotazione linguistica utile ed efficace è fondamentale che sia riconoscibile dal soddisfacimento di questi quattro requisiti:

1. *Robustezza*, sistema in grado di analizzare qualsiasi frase in input;
2. *Disambiguazione*, riuscire a selezione l'analisi corretta tra quelle possibili;
3. *Accuratezza*, quantificare l'accuratezza dell'annotazione, quindi la precisione assegnata alla frase di un testo dal parser;
4. *Efficienza*, capacità di analisi linguistica con il minimo impiego di risorse e di tempo.

Come metrica di valutazione per riconoscere la correttezza delle risposte del sistema e misurare la copertura dello stesso si utilizzano rispettivamente i valori di *precision* e di *recall*. La *precision* è calcolata come il rapporto fra il numero di elementi etichettati correttamente e il numero totale (output corretti + ouput errati) di elementi a cui è stata assegnata quella data etichetta. La *recall* invece è il rapporto fra il numero di elementi etichettati correttamente e il numero totale di elementi del test set che effettivamente presentano quella relazione di dipendenza (indipendentemente dalla valutazione del parser). Generalmente si fa riferimento anche al valore di *F-score*, ovvero la media armonica fra *precision* e *recall*.

I risultati dei file in formato CONLL-U sono stati processati tramite comandi del sistema Unix quali grep e cat. In particolare è stata utilizzata un'espressione regolare (v. fig.4) per estrarre le righe contenenti i risultati degli strumenti di analisi linguistica, per la successiva elaborazione attraverso uno script di valutazione scritto in Python.

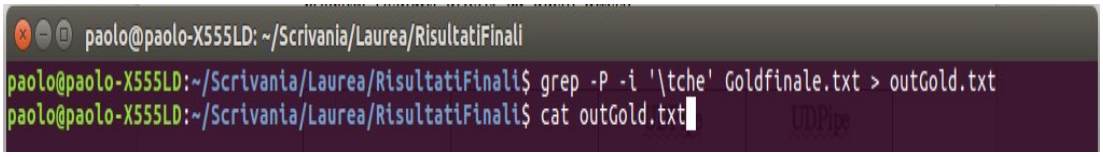

Figura 4. Comandi shell

È stato possibile mettere a confronto il Gold con i risultati ottenuti dalle analisi condotte con:

- Tanl addestrato su ISDT;
- UDPipe addestrato su PoSTWITA;
- UDPipe addestrato su ISDT.

I primi dati riguardano l'accuratezza del PoS tagging. È necessario specificare però che questo dato espresso in percentuale riguarda esclusivamente le occorrenze di *che*. Si tratta quindi di una valutazione ristretta al part-of-speech del *che* e non a quello globale. L'ipotesi di ricerca iniziale prevedeva che il risultato dell'analisi avrebbe messo in luce la maggiore affidabilità di UDPipe addestrato su PoSTWITA per la sue similitudini con il linguaggio dei commenti a post di Facebook. Tuttavia, come si può osservare nella tabella seguente l'analisi ha prodotto risultati diversi da quelli attesi.

	Tanl addestrato su ISDT	**UDPipe addestrato su PoSTWITA**	**UDPipe addestrato su ISDT**
Accuracy %	80,36%	78,17%	81,15%
Interpr. Corrette	405/504	394/504	409/504

Tabella 9. Accuratezza degli strumenti sulle occorrenze di che

È interessante notare che sia Tanl e sia UDPipe addestrati sulla risorsa ISDT possiedono un'accuratezza maggiore rispetto a UDPipe addestrato su PoSTWITA. Un motivo da non trascurare a sostegno di questo dato potrebbe essere la differenza numerica tra i due training set di Tanl addestrato su ISDT e UDPipe addestrato su PoSTWITA: il primo, con 257.616 tokens, è più grande del secondo, con 99.441 parole.

7.3 Analisi dettagliata degli strumenti di analisi linguistica

Tanl addestrato sulla risorsa ISDT:

Tag	Interpr. Corrette	Precision	Recall	F-score
PR	227	82,85	87,64	85,18
PQ	2	66,67	33,33	44,44
E	0	0	0	0
CC	1	12,50	14,29	13,33
CS	168	84	79,62	81,75
DQ	7	41,18	63,64	50

UDPipe addestrato su PoSTWITA:

Tag	Interpr. Corrette	Precision	Recall	F-score
CS	174	75	82,46	78,56
PR	207	84,49	79,92	82,14
DE	7	43,75	70	53,85
E	0	0	0	0
DQ	6	60	54,55	57,14

UDPipe addestrato sulla risorsa ISDT:

Tag	Interpr. Corrette	Precision	Recall	F-score

PR	241	81,69	93,05	87
PQ	1	100	16,67	28,57
E	0	0	0	0
SP	0	0	0	0
DE	0	0	0	0
CS	160	87,91	75,83	81,42
DQ	7	36,84	63,64	46,87

Le tre tabelle mostrano chiaramente che una caratteristica accomuna questi strumenti di analisi linguistica: la confusione nell'interpretazione corretta fra pronome relativo e congiunzione subordinativa. Questo costituisce l'errore di interpretazione principale perché per gli altri casi diventa difficile giudicare per la mancanza di frasi sufficienti per effettuare un'analisi attendibile.

Conclusioni

L'analisi è stata svolta adottando un approccio linguistico-computazionale grazie all'utilizzo delle risorse linguistiche annotate manualmente o semi-automaticamente del progetto Universal Dependencies. Attraverso la creazione di un Gold standard e il successivo confronto tra quest'ultimo e i risultati ottenuti dagli strumenti di analisi linguistica, si è resa possibile la valutazione dell'accuratezza riguardante le occorrenze di *che*. Abbiamo visto l'utilizzo del *che polivalente*, elemento con una vasta polisemia di impieghi, che come osserva Berruto si inserisce all'interno della più generale tendenza della lingua.

Gli strumenti che hanno permesso di ottenere i dati oggetto della discussione sono la pipeline Tanl addestrata sulla risorsa IUDT,

convertita successivamente in ISDT, la catena di annotazione linguistica UDPipe addestrata anch'essa su ISDT e su PoSTWITA. L'analisi effettuata avrebbe avuto un ulteriore giovamento confrontando un corpus simile, per esempio composto da commenti a post di personaggi pubblici o pagine di attualità. Come già detto, tra le tante tipologie di produzione linguistica che offre Facebook, i commenti a post di personaggi pubblici necessitano di studi più specifici. Potrebbe dimostrarsi ancora più interessante estendere il Gold con il Part-of-speech tagging per ogni parola in modo da ottenere un grado di accuratezza degli strumenti di analisi linguistica a livello globale. Sarebbe utile ed interessante estrarre dei pattern di utilizzo dell'italiano neo-standard come il *che polivalente* per procedere al riconoscimento semi-automatico di tali fenomeni, fornendo un tag più specifico per indicarli, dato che attualmente i diversi strumenti non li possiedono. In conclusione si vuol far riflettere sul fatto che i risultati ottenuti dimostrano chiaramente che le potenzialità degli strumenti sono enormi, sebbene informazioni altrettanto ricche siano ancora nascoste all'interno delle risorse linguistiche. Per questo motivo può ancora essere fatto tanto nello sviluppo degli strumenti per realizzare studi di questo genere.

Bibliografia

Antonelli, Giuseppe (2017), *L'italiano nella società della comunicazione 2.0*, Bologna, il Mulino

Barbera, Manuel (2013), *Linguistica dei corpora e linguistica dei corpora italiana. Un'introduzione.*, Milano, Qu.A.S.A.R s.r.l

Bentivegna, Sara (2012), *Parlamento 2.0*, Milano, Franco Angeli.

Cristina Bosco, Fabio Tamburini, Andrea Bolioli, Alessandro Mazzei. 2016. *Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian task*. In: Proceedings of Evalita 2016

Desideri, Paola. 1993. *L'italiano della Lega/1,* in: "Italiano e oltre", VIII, pp. 281-285.

Epifani, S., Jacona, A., Lippi, R., e Paolillo, M. (2011), *Manuale di comunicazione politica in Rete. Costruire il consenso nell'era del Web2.0*, Roma, Apes.

Lasswell, Harold e Fox, Merritt B (1979), *The Signature of Power: Buildings, Communication, and Policy*, New Brunswick, N.J.: Transaction Book.

Montemagni, Simonetta (2013) *Tecnologie linguistico-computazionali e monitoraggio della lingua italiana*. In "Studi Italiani di Linguistica Teorica ed Applicata (SILTA)", Anno XLII, 1.

Nivre Joakim. *Towards a universal grammar for natural language processing*. In International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2015.

Nivre Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and others. *Universal dependencies v1: A multilingual treebank collection.* In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).

Ornella Castellani Pollidori (2004), *In riva al fiume della lingua. Studi di linguistica e filologia (1961-2002)*, Roma, Salerno editrice.

Palermo, Massimo (2015), *Linguistica italiana*, Bologna, Il Mulino.

Serianni, Luca (2016), *Grammatica italiana. Italiano comune e lingua letteraria*. Novara, De Agostini Scuola SpA.

Simone, Raffaele (2010), *Enciclopedia dell'italiano*, Roma, Istituto della Enciclopedia italiana.

Spina, Stefania (2012), *Openpolitica. Il discorso dei politici italiani nell'era di Twitter*, Milano, Franco Angeli Editore

Tavosanis, Mirko (2011), *L'italiano del web*, Roma, Carocci

Tavosanis, Mirko (2016), *Il linguaggio della comunicazione politica su Facebook, in: L'italiano della politica e la politica per l'italiano, Atti del XI Convegno ASLI Associazione per la Storia della Lingua Italiana (Napoli, 20-22 novembre 2014)*, a cura di Rita Librandi e Rosa Piro, Firenze, Cesati Editore.

Tortorelli, Maria Cristina, *Analisi linguistica di commenti ai post delle pagine Facebook dei Ministri della Repubblica Italiana, elaborato di laurea magistrale*, corso di laurea in Informatica umanistica, Università di Pisa, anno accademico 2016-2017.

Sitografia

Facebook for developers: http://www.developers.facebook.com

ISDT:https://github.com/UniversalDependencies/UD_ItalianISDT/blob/master/README.md

PoSTWITA: http://corpora.ficlit.unibo.it/PoSTWITA/

Python, sito della comunità ufficiale italiana: https://www.python.it/

Scraper dei commenti: https://github.com/minimaxir/facebook-page-post-scraper

Treccani, voce *che* polivalente

http://www.treccani.it/enciclopedia/che-polivalente_(Enciclopedia-dell'Italiano)/ (visitato il 30 ottobre 2018)

UDPipe: http://lindat.mff.cuni.cz/services/udpipe/run.php

Universal Dependencies: http://universaldependencies.org/