UNIVERSITÀ DEGLI STUDI DI PISA

FACOLTÀ DI LETTERE E FILOSOFIA

FACOLTÀ DI SCIENZE MM. FF. NN.

CORSO DI LAUREA IN INFORMATICA UMANISTICA

Il linguaggio dei social network. Costituzione e analisi di un corpus.

Relatore

Prof. Mirko Tavosanis

Correlatore

Dott.ssa Maria Simi

Candidato

Matteo Natali

Anno Accademico

2009/2010

INDICE

1 - Introduzione

Il presente lavoro di tesi intende condurre un'analisi linguistica dei tipi di comunicazione che caratterizzano i *social network*. Con tale termine si fa correntemente riferimento, nell'ambito del web, a tutti quei siti che offrono la possibilità di instaurare una rete sociale "virtuale". Mettendo a disposizione degli utenti vari servizi, i social network offrono l'opportunità di mantenere i legami con le proprie amicizie e farne nascere di nuove. La tesi analizza quantitativamente un corpus di testi raccolti su Facebook e suddivisi in categorie. A sostegno dell'analisi è stato realizzato uno strumento informatico costituito da una sezione che permette la raccolta dei dati e da una sezione dedicata all'analisi quantitativa. Il sistema di catalogazione e raccolta è basato su un database MySQL, volto a raccogliere i testi e i dati degli utenti, e su un sistema di gestione dello stesso con maschere scritte in PHP. Attraverso le maschere è possibile inserire testi e utenti del campione, specificando anche dati relativi al sesso e all'età, e visualizzare gli elementi raccolti sia su schermo che tramite un foglio excel. Nelle sezioni dedicate all'analisi quantitativa è possibile consultare i dati relativi alla media delle parole, dei caratteri e delle emoticon nelle categorie dei testi raccolti. Tramite la realizzazione di filtri di ricerca viene offerta la possibilità di consultare i risultati relativi all'analisi sia in base alle categorie di testi presi in esame sia in base a dati personali degli utenti del campione quali età e sesso.

2 - Funzionamento e descrizione dei social network e di Facebook

Entrare a far parte di un social network è molto semplice: infatti è solamente necessario registrare il proprio profilo specificando alcune informazioni aggiuntive, gli amici, gli hobby, gli interessi così come le esperienze di studio e lavoro.

Una volta che il profilo è stato creato è possibile espandere la propria rete invitando gli amici a partecipare, oppure aderendo a "gruppi" di persone che condividono passioni e interessi mettendo in comune molte informazioni. Tutte queste possibilità permettono di aggregare gli utenti del social network, di stringere nuove amicizie e reperire contatti.

Il *social network* a cui ho deciso di rivolgere l'attenzione è Facebook (http://www.facebook.com/), oggi il più frequentato e utilizzato. Secondo i dati forniti dal sito stesso, Facebook ha festeggiato il suo sesto compleanno il 5 febbraio 2010 e ad avere circa 400 milioni di utenti in tutto il mondo.

La storia viene riassunta sul blog ufficiale del sito (http://blog.facebook.com/blog.php): Facebook è nato nel 2004 da un progetto di uno studente dell'Università di Harvard, Mark Zuckerberg, e si è esteso poi ad altri atenei, fino a raggiungere l'utenza mondiale. Oggi secondo Alexa (http://www.alexa.com/), azienda statunitense sussidiaria di Amazon.com che si occupa di statistiche sul traffico di internet, è uno dei dieci siti più visitati del web.

2.1 - Alcuni dati statistici sull' utenza Facebook in Italia.

Facebook mette a disposizione alcuni dati sull'utenza per chi vuole creare banner pubblicitari rivolti a specifiche fasce di pubblico (l'indicazione mi è stata fornita da Giorgio Taverniti, fondatore di uno dei più attivi forum italiani dedicato al SEO e allo sviluppo web: http://www.giorgiotaverniti.it/). Accedendo a questo servizio, che troviamo al link http://www.facebook.com/advertising/ (Fig. 1), e simulando la creazione di un banner possiamo dunque accedere ai dati, che sono soltanto una stima ma che ci possono dare un' idea del fenomeno Facebook in Italia.

Figura 1 - Pagina principale di Facebook Avertising

Iniziando quindi la nostra simulazione, nella form che ci viene presentata, possiamo scegliere alcuni criteri per individuare il target dei destinatari delle inserzioni (Fig. 2).

I principali sono:

- Posizione geografica
- Età
- Livello di istruzione
- Sesso

Inoltre sono disponibili ulteriori filtri di ricerca quali situazione sentimentale, posto di lavoro etc.

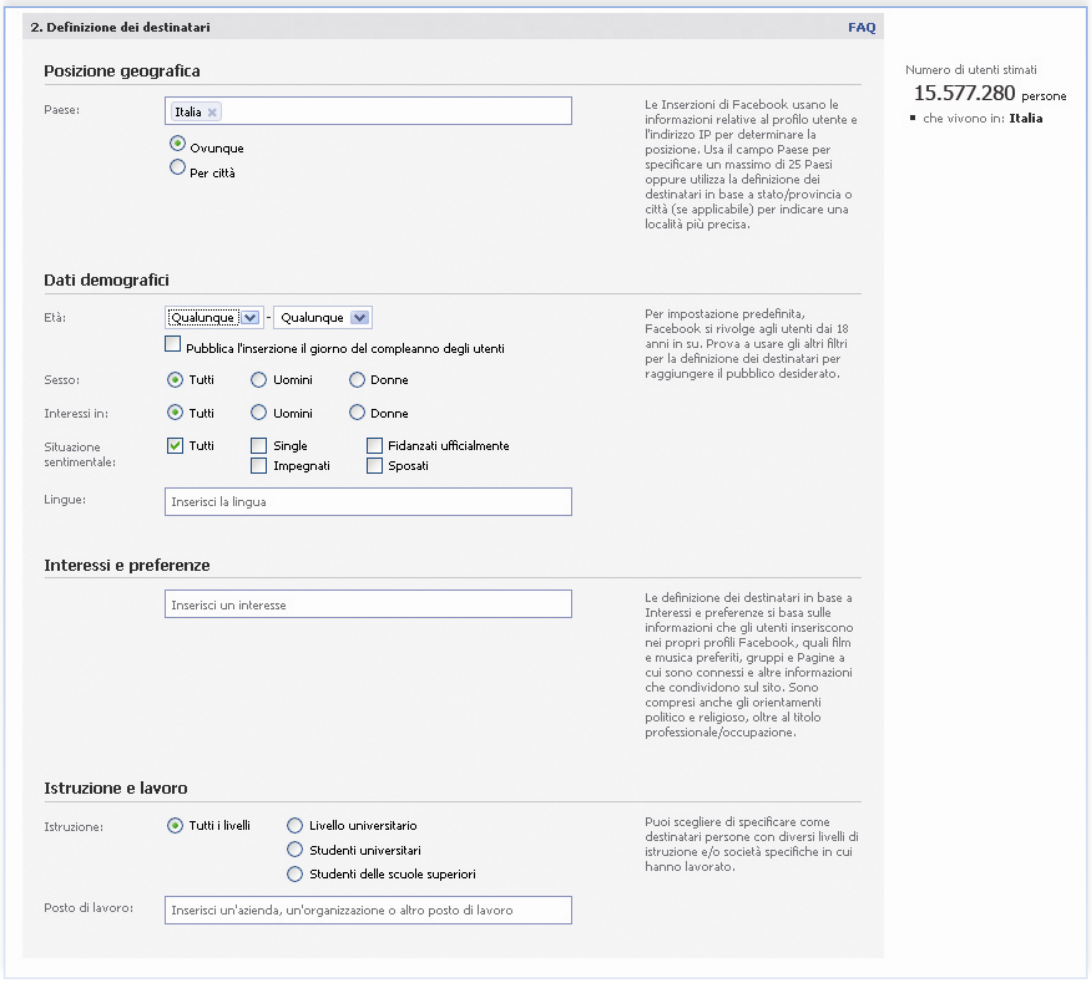

Figura 2 - Filtri di definizione destinatari di Facebook advertising

Per includere nel target il maggior numero di utenti, selezionando come posizione geografica l'Italia, impostando i filtri relativi all'età su "qualunque" e quelli relativi al sesso, agli interessi e alla situazione sentimentale su "tutti", si ottiene una stima di 15.577.280 utenti al 18 marzo 2010.

Attraverso la medesima funzione possiamo ricavare ulteriori dati utili a capire meglio il fenomeno.

Suddivisione utenza di Facebook in base al sesso e all'età

Nella tabella sottostante (Tab. 1) vengono esposti i valori relativi alla suddivisione in base al sesso e alle fasce di età.

ETA’	UOMINI	DONNE	TOTALE	SATISTICHE ISTAT
0-18	**1.481.580**	**1.456.080**	**2.937.660**	**10.804.889**
19-24	**1.681.080**	**1.589.600**	**3.270.680**	**3.720.187**
25-29	**1.179.960**	**1.087.560**	**2.267.520**	**3.555.430**
30-35	**1.240.880**	**1.036.160**	**2.277.040**	**5.309.530**
36-45	**1.559.760**	**1.116.960**	**2.676.720**	**9.794.423**
46-55	**696.540**	**478.260**	**1.174.800**	**8.239.702**
56-OLTRE	**360.420**	**172.360**	**532.780**	**18.620.907**
TOTALE	**8.200.220 (54%)**	**6.936.980 (46%)**	**15.137.200**	**60.045.068**

Tabella 1 - Suddivisione utenza di Facebook in base al sesso e all'età. Confronto con dati ISTAT.

Dai dati esposti possiamo osservare che in generale l’utenza italiana è formata in maggioranza da individui di sesso maschile (il 54% contro il 46%) e che la fascia di età più rappresentata è quella compresa tra i 19 e i 24 anni, seguita da quella degli adolescenti fino ai 18 anni. Interessante notare che anche la fascia di utenti con età compresa tra i 36 e i 45 anni è abbastanza consistente, leggermente superiore alle due fasce che la precedono.

È, però, importante ricordare che, non essendo possibile accedere a dati precisi e ufficiali sul numero di utenti registrati su Facebook in Italia, quella qui rappresentata è soltanto una stima approssimativa e che, essendo realizzata per fini commerciali (la realizzazione di banner pubblicitari a target mirato), i dati possono non essere affidabili. Infatti, incrociando i dati ottenuti con le statistiche demografiche ISTAT (reperibili sul sito http://demo.istat.it/index.html e relative alla popolazione residente al 1 Gennaio 2009 per età, sesso e stato civile), il numero di utenti di Facebook risulta essere sovrastimato, soprattutto per quanto riguarda la fascia di età compresa tra i 19 e i 24 anni, in cui i dati hanno una differenza minima.

2.2 - Struttura della pagina di Facebook

L'interfaccia del sito rispecchia pienamente lo stile Web 2.0 (Fig. 3), ed è molto semplice e intuitiva. Dopo essersi registrato, l'utente può accedere alla propria area personale e quindi gestire i diversi servizi che la piattaforma offre.

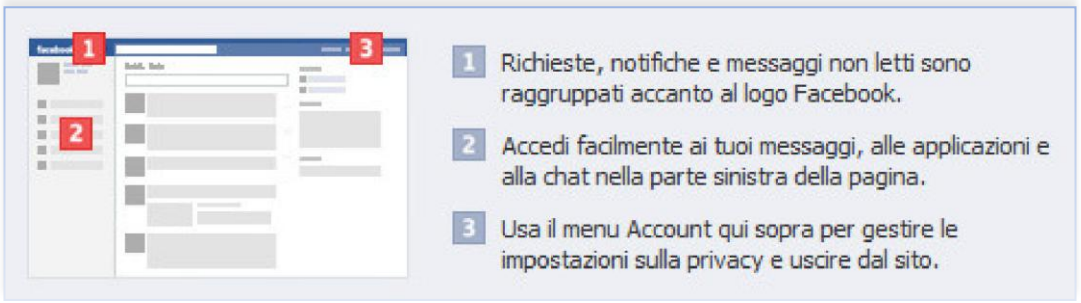

Figura 3 – Struttura della pagina di Facebook

Come possiamo osservare dalla figura che Facebook stessa ci mette a disposizione per spiegare l'interfaccia sul blog ufficiale (http://blog.facebook.com/), nella barra orizzontale sulla parte superiore della pagina ci sono icone che, selezionate, aprono menù a tendina che presentano le ultime notifiche, le richieste e i messaggi. Ogni volta che, per esempio, viene ricevuta una notifica, compare una piccola bolla rossa nell'angolo a sinistra vicino alla barra di ricerca.

Nella parte sinistra, al di sotto della foto del profilo, troviamo invece due gruppi di menù laterali. Con il primo gruppo di menù possiamo navigare attraverso i contenuti dei nostri contatti. Nel secondo gruppo invece possiamo accedere alle applicazioni e ad eventuali giochi.

Nella parte destra della finestra di navigazione infine c'è la possibilità di gestire alcune richieste, avere "suggerimenti" di amicizia ed alcuni link utili (Fig. 4).

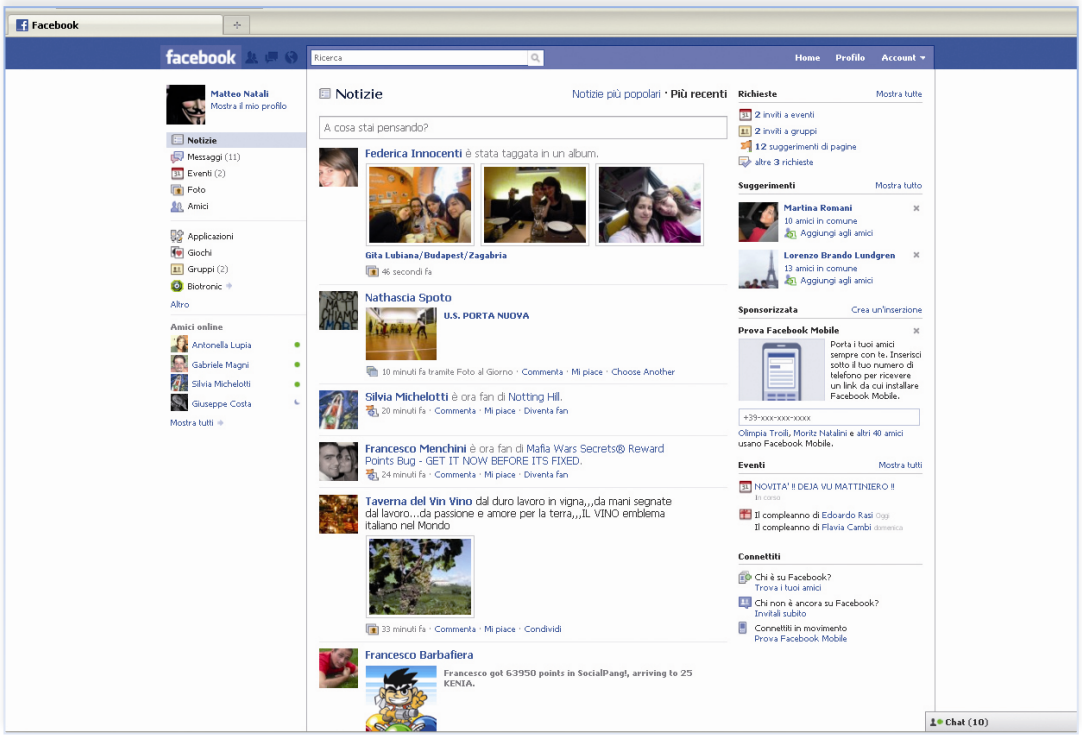

Figura 4 – Pagina principale di Facebook

Registrandosi, accedendo al sito ed esplorandolo un po' salta subito all'occhio come Facebook venga utilizzato e come mantenga la promessa, che spesso ricorre nelle lettere aperte agli utenti del sito, del suo stesso fondatore ovvero quella di "*rendere il mondo più aperto e connesso*". Infatti la piattaforma viene utilizzata principalmente come strumento utile per rimanere in contatto con gli amici e le persone che interessano ma anche per pubblicare e condividere file multimediali come foto o video (creando dei veri e propri album) e condividere informazioni sugli eventi grandi e piccoli.

Ogni utente di Facebook ha la possibilità di pubblicare e condividere testi in varie modalità utilizzando un proprio spazio personale del profilo (bacheca), più o meno accessibile agli altri utenti a seconda del grado di privacy scelto, oppure commentando ed esprimendo giudizi nei confronti di testi scritti da altre persone o di pubblicazioni di link od "eventi", chiamando così in causa un'ampia dimensione di ipertestualità.

2.3 - La bacheca

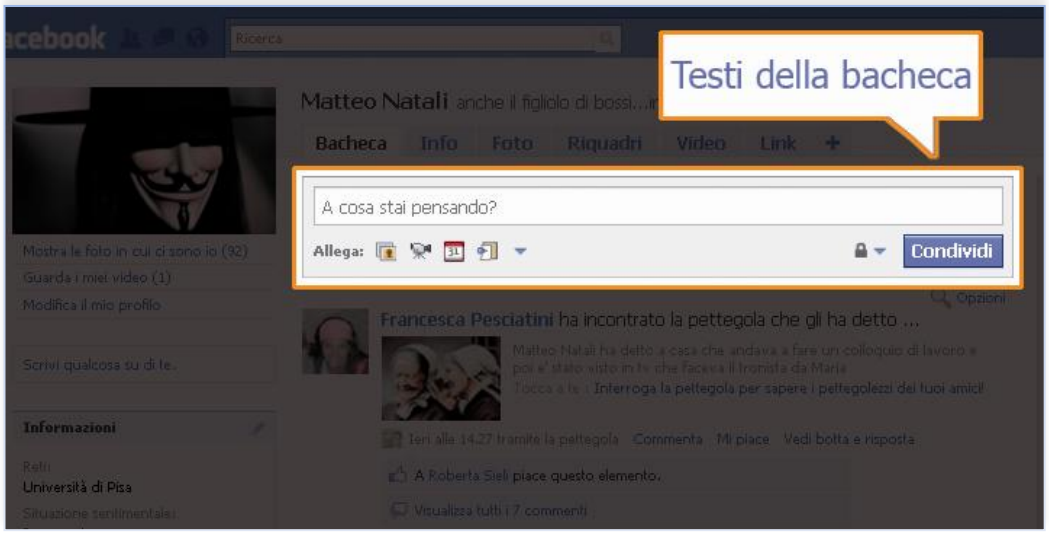

Figura 5 – La bacheca di Facebook

La bacheca di Facebook, evidenziata all'interno della figura soprastante (Fig. 5), rappresenta un po' il "cuore pulsante" di ogni profilo. Infatti permette di aggiornare il proprio stato e condividere contenuti. È possibile inoltre aggiungere contenuti sulle bacheche dei propri amici usando la casella apposita che appare in alto sui loro profili e che nella versione inglese del sito viene chiamata "wall".

All'interno della casella di testo si legge un testo breve che invita gli utenti ad aggiornare il proprio stato. Interessante osservare la frase "a cosa stai pensando?" che spinge l'utente a scrivere e condividere in tempo reale pensieri e contenuti. A prima vista, i testi che sono presenti nella bacheca del profilo di un utente sono per lo più testi brevi, raramente costituiti da più di due frasi; vedremo in seguito, al capitolo 4.3.1, se ciò corrisponde al vero.

Naturalmente quello che possiamo leggere sulla bacheca di un utente è regolato dal grado di privacy che l'utente stesso ha impostato.

2.4 - I commenti

Figura 6 – I commenti di Facebook

I commenti agli status degli utenti, invece, sembrano essere costituiti da testi più ampi rispetto a quelli della bacheca. Anche in questo caso vedremo al capitolo 4.3.2 se ciò corrisponde al vero. Ogni utente, in funzione del grado di privacy, può commentare sia il semplice stato che foto, link e video condivisi. Ogni volta che un utente ha commentato una pubblicazione, al di sotto del contenuto oggetto del commento compare un campo che contiene il testo scritto dall'utente, evidenziato ancora una volta in figura (Fig. 6). Tutto ciò è strutturato come una vera e propria conversazione e non sono presenti limitazioni di numero ai commenti che possono essere inseriti.

Oltre a ciò esiste anche la possibilità di comunicare attraverso messaggi privati di posta e attraverso una chat, limitando l'accessibilità del testo solo a determinate persone.

2.5 – Gli studi su Facebook: una panoramica

Naomi S. Baron, docente di linguistica alla American University a Washington DC, è una delle principali autorità nello studio del linguaggio nell'era del computer. Al centro dei suoi studi troviamo il linguaggio degli studenti dei college americani nei programmi di messaggeria istantanea, nella posta elettronica, nei telefoni cellulari, nei blog e social network e nel mondo "mobile" in generale. Naomi Baron, in un capitolo del suo libro *Always on* (Naomi S. Baron 2008), affronta la questione dell'uso del linguaggio all'interno di Facebook, da lei considerato il più importante tra i social network, conducendo un'analisi sulle principali caratteristiche della piattaforma e concentrandosi sui dati di utilizzo degli studenti del college. Dopo averne ripercorso la storia dalla nascita, passa a descriverne, in generale, le funzioni basilari, la struttura dei profili e l'utenza. Nella descrizione del profilo, la cui interfaccia è cambiata nel corso degli anni, l'autrice elenca le parti principali: la foto di riconoscimento, le informazioni di contatto e gli interessi. Una volta iscritti il modo principale per stabilire relazioni online è entrare in contatto con qualcun altro oppure entrare a far parte di un "gruppo", persone che condividono interessi. L'autrice passa poi in rassegna i metodi di interazione tra gli utenti disponibili nel 2006, anno dell'analisi: i messaggi privati e la bacheca. La parte principale del suo studio però si concentra sul modo in cui i giovani usano la piattaforma per costruire rapporti sociali e interagire con essi.

Nella primavera del 2006 sottopose, infatti, un questionario costituito da 8 principali categorie di domande (demografiche, modi di usare il sito, informazioni personali, amici, interazione online etc..) a 60 studenti, divisi equamente tra maschi e femmine e con una età media di circa 20 anni. Dal questionario emersero alcuni dati interessanti. Il primo dato indica che i maschi, rispetto alle femmine, utilizzano molto più frequentemente Facebook, connettendosi spesso ma per brevi lassi di tempo, mentre le femmine vi trascorrono molto più tempo (circa 45 minuti) e visitano in media circa 7 profili nell'arco di 24 ore. Tra gli studenti del campione il 55% di loro si connette al sito giornalmente. Altro dato che è emerso dal questionario riguarda le informazioni personali e il numero di amici. È risultato infatti che solo il 5% degli utenti cambia l'immagine personale sul proprio profilo più di una volta a settimana e

solamente il 7% modifica le informazioni personali con la stessa frequenza. Per quanto riguarda la cerchia di amici, gli utenti del campione intervistato hanno dichiarato che, sebbene la media di amicizie si aggiri intorno a 229 persone, le persone con cui hanno rapporti regolari e che possono essere considerati "veri" amici sono solo 65 per le femmine e 78 per i maschi. Naomi sostiene che è abbastanza comune da parte dei giovani avere un gran numero di amici ("*the more Friends the merrier*" Baron 2008, 88), come se ciò fosse uno sport o un passatempo; anzi alcuni studenti intervistati si sono sentiti in imbarazzo nel dichiarare un esiguo numero di amici. Questo discorso vale anche per la registrazione ai "gruppi": la media di registrazioni per ogni utente si aggira intorno a 15 gruppi ma gli studenti ne frequenterebbero attivamente soltanto 2.

Passando ad analizzare l'interazione online degli studenti del campione, Naomi Baron ha raccolto dati sull'uso dei messaggi privati e della bacheca. Per quanto riguarda i messaggi privati solo il 7% del campione ha dichiarato di usarli più di una volta al giorno o giornalmente, mentre il 10% non li ha mai utilizzati. In media meno di un messaggio al giorno per utente è stato spedito nell'arco di una settimana. La bacheca viene utilizzata in modo simile: circa l' 8% degli utenti scrive sulla bacheca di un altro utente, mentre il 60 % dichiara di non averne mai fatto uso. Inoltre i testi che vengono pubblicati sono per lo più messaggi di saluto e solo un quarto degli utenti scrive messaggi articolati volti a scambiare informazioni. Curioso osservare che molti degli utenti intervistati hanno dichiarato di scrivere sulle bacheche altrui solo per essere letti dalle altre persone ("*so other people can also see what i'm saying*" Baron 2008, 90); solamente 4 studenti su 60 ha utilizzato la bacheca del proprio profilo.

Nella parte del questionario riservata alla privacy e all'accesso libero sul proprio profilo da parte di altri utenti, i 2/3 del campione ha risposto di non aver niente in contrario se i visitatori sono studenti coetanei ma che non sarebbero d'accordo se a visitarlo fossero datori di lavoro o laureati perché il profilo non corrisponderebbe a come vorrebbero presentarsi: Facebook in sostanza è per gli amici ("*this is for my friend*" Baron 2008, 91).

Nell'ultima parte della sua analisi Naomi Baron ha messo a confronto i programmi di messaggeria istantanea con Facebook. Dal paragone è emerso che gli studenti spendono molto più tempo utilizzando i programmi di messaggeria (circa 2,18 ore contro 40 minuti di Facebook) anche se va sottolineato che, alla fine, questi programmi spesso vengono solo aperti e non utilizzati attivamente per ore.

Confrontando infine il ruolo "sociale" dei due mezzi di comunicazione è emerso che gli studenti assegnano loro diverse funzioni: Facebook sta diventando il network in cui presentarsi agli altri "ufficialmente" mentre i programmi di messaggeria mantengono il loro ruolo base di comunicazione online tra individui.

Nel panorama dell'italiano in generale, quello che viene usato all'interno di Facebook rientra nei canoni dell'italiano elettronico, ovvero di quell'italiano utilizzato nella comunicazione dell'era informatica. Giuseppe Antonelli, docente di Linguistica italiana all'Università di Cassino, nel libro *L' italiano nella società della comunicazione* (Giuseppe Antonelli 2007) ne tratteggia le caratteristiche. Secondo l'autore nella odierna società della comunicazione i testi sono diventati brevi, rapidi e facili da leggere. Questo può essere riscontrato in alcuni esempi di testi prelevati dal corpus:

- Un'altra perla di Mai dire Tv...il predicatore!!!Cioè un genio! :-)

- Canzone brutta...

- uahuhauah!

- ahahahaha... ben ti stà!!! :p

Al centro dei dialoghi a più voci, spesso, c'è la "chiacchiera spicciola", come sostiene l'autore, e la comunicazione ha uno scopo ludico, caratterizzata da un parlare veloce e da uno scambio di informazioni altrettanto veloce (Antonelli 2007, 144-145). Troviamo inoltre la tendenza a una scrittura colorita e a una espressività molto accentuata; a ciò contribuiscono, all'interno delle frasi, molti segnali discorsivi e interiezioni che danno al testo un andamento parlato e un tono espressivo. Anche l'uso delle emoticon risponde a questa esigenza di espressività, andando a marcare e a sottolineare il senso della frase. Per quanto riguarda la sintassi, si devono alla scrittura veloce e poco meditata alcuni aspetti caratteristici dell'italiano digitato quali

la scarsità di proposizioni subordinate e un raro uso di elementi connettivi. Il tono è quasi sempre informale favorendo un lessico colloquiale, che non disdegna parole e modi regionali o dialettali. Altra caratteristica che va nella direzione dell'espressività è rappresentata dall'uso di iterazione vocalica e punteggiatura enfatica (Antonelli 2007, 153-159). come vediamo in alcuni esempi tratti dal corpus:

- ... E CHE TE LO DICO A FARE ?!?!?!?....

- nooooooooooooooooooooooooooooooooooooo...x le undici...nooooooooooooooooooooooooooooooo...ahuahuahuhauhuahuhauh

- Ehmmmmmmm.....cosa significa????Booohhhhh

3. - Il corpus di riferimento

L'analisi sarà condotta prendendo in esame testi prodotti da utenti e raccolti in un corpus avente le seguenti caratteristiche:

1) estensione del corpus: testi scritti da un campione di 408 persone nell'arco di circa un mese (novembre 2009), per una dimensione totale di 20.905 tokens.

2) tipologia dei testi: data l'impossibilità di accedere ai dati relativi alla posta privata e ai messaggi di tutti i soggetti del campione, sono stati presi in considerazione solo i testi della bacheca, i commenti ai testi della bacheca e i commenti alle varie pubblicazioni.

Non sono state invece prese in considerazioni le "citazioni" di testi di brani musicali o poesie e di tutti i testi non prodotti dagli utenti del campione.

3.1 - Gestione informatica dei dati

Il procedimento che ho adottato per la raccolta del materiale consiste nel "copia e incolla" delle categorie di testi sopra elencate per ogni profilo facente parte del campione all'interno di un database MySQL, uno dei database open source più utilizzati al mondo. Si è scelto questo metodo di raccolta dei testi, che può risultare lento e macchinoso di fronte a un metodo più automatico, in primo luogo per la necessità di classificare e verificare ogni testo al momento dell'inserimento e, infine, perché la struttura di Facebook, e quindi la disposizione dei tag HTML nel codice sorgente, si aggiorna frequentemente rendendo non facile l'ideazione di un metodo automatizzato.

L'ambiente di sviluppo è rappresentato dal programma EasyPHP, scaricabile gratuitamente dal sito http://www.easyphp.org/: un pacchetto software da installare in locale e che comprende un server Apache, un database MySQL, PHP, così come strumenti di sviluppo e di amministrazione.

Il database permette di immagazzinare, gestire e recuperare dati in modo intuitivo e funzionale. Infatti è possibile intervenire sui record in modo molto rapido e sfruttare tutte le possibilità che un database professionale offre rispetto a una semplice trascrizione su file: per esempio, la manipolazione delle stringhe. L'inserimento e l'interrogazione dei dati vengono gestiti attraverso maschere realizzate in PHP che permettono un facile utilizzo del database.

La codifica che ho deciso di utilizzare, sia per il database sia per le pagine di inserimento e consultazione, è UTF-8, una codifica dei caratteri Unicode in sequenze di lunghezza variabile di byte che permette di trattare senza problemi di compatibilità molti simboli o caratteri particolari che compaiono all'interno dei testi.

All'interno del database MySQL ho utilizzato tre tabelle per raccogliere i dati necessari all'analisi linguistica (Fig. 7):

1. **Tabella "testi"**
2. **Tabella "categorie_testi"**

3. Tabella “utenti”

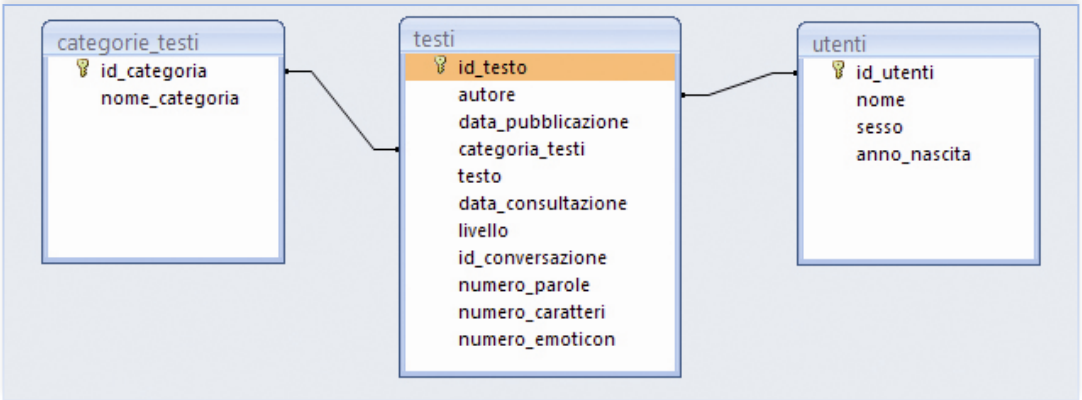

Figura 7 – Schema del database

La tabella “testi” è costituita dai seguenti campi:

- **id_testo** (chiave primaria): un campo ID univoco della tabella;
- **autore**: campo in cui viene specificato il riferimento alla tabella utenti e che contiene il nome dell’utente del campione;
- **data_pubblicazione**: campo in cui viene inserita la data di pubblicazione dei testi prelevati;
- **categoria_testi**: campo in cui viene definito il riferimento alla tabella categoria_testi e che contiene il nome della categoria dei testi che abbiamo deciso di prendere in esame
- **testo**: campo che contiene il testo scritto dagli utenti del campione
- **data_consultazione**: campo in cui viene specificata la data in cui il testo del campione è stato prelevato.
- **livello**: campo in cui viene specificato il livello del testo per stabilirne la priorità.
- **id_conversazione**: campo ID che identifica ogni conversazione.

- **numero_parole:** campo in cui è inserito il numero di parole che il testo contiene.
- **numero_caratteri:** campo in cui è inserito il numero di caratteri che il testo contiene.
- **numero_emoticon:** campo in cui è inserito il numero di emoticon che il testo contiene.

La tabella "categorie_testi" è costituita da:

- **Id_categoria** (chiave primaria): un campo ID univoco della tabella;
- **nome_categoria**: campo in cui viene specificato il nome della categoria dei testi

La tabella "utenti" è costituita da:

- **id_utenti** (chiave primaria): un campo ID univoco della tabella;
- **nome**: campo in cui viene specificato il nome degli utenti che fanno parte del campione preso in esame.
- **sesso**: campo in cui viene specificato il sesso dell' utente del campione.
- **anno_nascita**: campo in cui viene specificato l'anno di nascita dell' utente del campione.

3.2 - Raccolta e catalogazione testi – premessa

Per effettuare la mia raccolta ho seguito un criterio molto semplice. Data l'impossibilità, dovuta al sistema di privacy del sito, di accedere a profili casuali degli utenti di Facebook, ho rivolto l'attenzione alla mia rete personale e al gruppo di persone con cui ho stretto legami di amicizia. Ho perciò visitato i profili con cui ho

rapporti e di cui posso consultare la pagina principale, scelto il periodo di tempo che mi interessava esaminare (come già detto, il mese di novembre), e infine ho inserito e catalogato i testi all'interno del database attraverso la funzione copia e incolla. Vediamo in dettaglio il procedimento.

3.3 - Raccolta e catalogazione testi – popolamento tabella utenti.

I lavoro di raccolta testi e popolamento del database si è articolato in vari passaggi. Per prima cosa ho ritenuto necessario catalogare e inserire alcune informazioni relative agli autori dei messaggi presi in considerazione. Attraverso una maschera realizzata in PHP ho quindi popolato la tabella utenti del database MySQL e specificato il nome dell'utente del campione (che rimane comunque anonimo sia nel foglio excel scaricabile che a video e identificabile soltanto attraverso la numerazione progressiva del campo id_utenti della relativa tabella), il sesso, che può essere utile per approfondire alcune analisi, e l'anno di nascita in modo da eseguire ricerche anche in base alle fasce di età del campione (Fig. 8). Purtroppo per alcune persone (188 utenti) non è stato possibile accedere all' anno di nascita e quindi è stato assegnato loro un anno fittizio "0000".

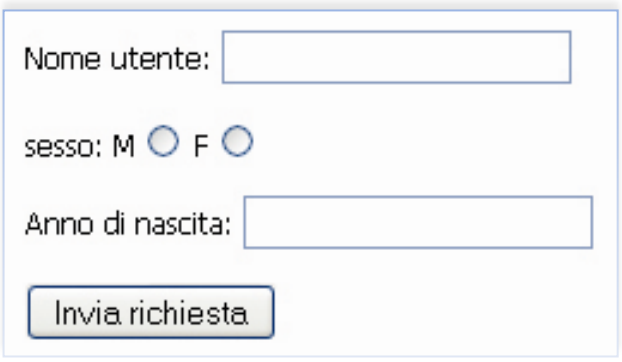

Figura 8 – Maschera inserimento utenti

3.4 - Raccolta e catalogazione testi – definizione delle categorie.

Le categorie dei testi presi in esame sono raccolte all'interno della tabella "categorie_testi". La struttura della tabella è molto semplice e al suo interno sono già inserite le tre categorie di testi che ho scelto di collezionare:

- **Testo bacheca**: sotto questa categoria inserisco tutti i testi scritti in bacheca.
- **Commento bacheca**: categoria che raccoglie tutti i testi che hanno funzione di commento ad un testo scritto in bacheca
- **Commento pubblicazione**: visto che Facebook offre la possibilità di condividere vari elementi (video, foto o link) sotto questa categoria ho raccolto i testi che sono stati scritti al momento della pubblicazione o come commento di questi elementi.

Quindi al momento dell' inserimento del testo ogni stringa del campione viene catalogata secondo queste categorie.

3.5 - Raccolta e catalogazione testi – inserimento dei testi nel database e nella tabella "testi".

Questa è la fase principale del lavoro di raccolta. Una volta copiato il testo dalla pagina dell'utente del campione e dopo aver inserito i dati relativi all' utente stesso nella tabella "utenti" sono passato alla catalogazione attraverso una nuova maschera di inserimento in PHP (Fig. 9).

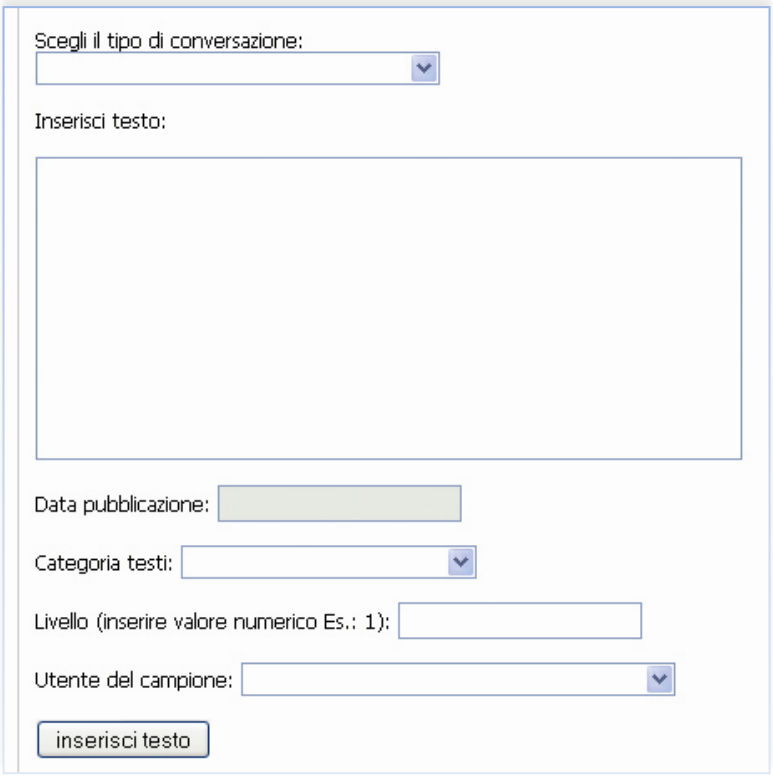

Figura 9 – Maschera inserimento testi

All'interno della maschera, per cercare di ricostruire le conversazioni che si vengono a formare all'interno di Facebook tra gli utenti, per prima cosa è possibile scegliere se il testo raccolto fa parte di una conversazione precedente oppure no. Agendo sul campo SELECT verrà assegnato un ID al testo che identificherà o meno l'appartenenza a una conversazione (Fig. 10).

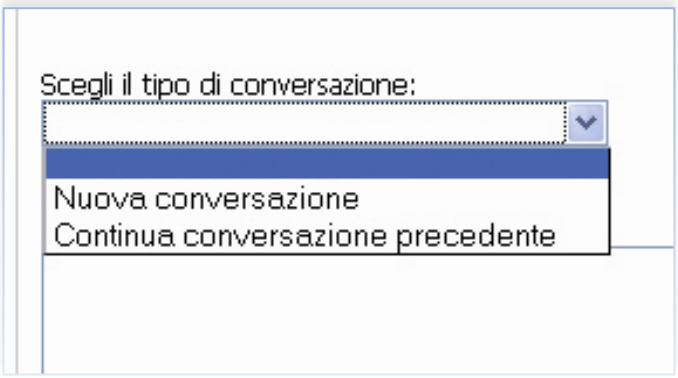

Figura 10 – Scelta del tipo di conversazione

Successivamente nel campo TEXT sottostante possiamo inserire il testo copiato preso in esame e assegnare alcuni parametri quali data di composizione e, nel campo SELECT successivo, scegliere a quale categoria della tabella "categorie_testi" appartiene (Fig 11).

Figura 11 – Scelta della categoria dei testi

Fatto ciò incontriamo un campo molto importante per la ricostruzione di una conversazione, denominato Livello, che stabilisce il turno della conversazione. Definendo un valore numerico in ordine crescente all'interno a partire da 1, possiamo ricostruire l'ordine dei messaggi di una conversazione: i testi che risulteranno contrassegnati con il numero di livello 1 saranno quelli che danno inizio alla conversazione, 2 i secondi e così via… L'ultimo campo da specificare è rappresentato da una SELECT in cui scegliere l'utente del campione che ha scritto la frase, in modo da collegare la frase ad un suo autore (Fig. 12).

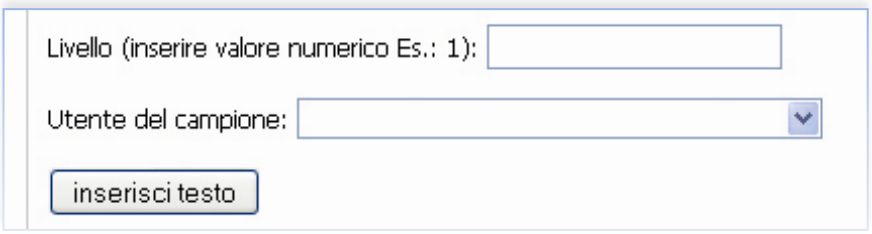

Figura 12 – Scelta del livello e dell'utente del campione

Con questo ultimo passaggio il testo verrà inserito nel corpus, opportunamente catalogato. Come possiamo osservare dall'immagine sottostante (Fig. 13), il risultato definitivo raccoglie le informazioni inserite, scaricabili anche in un foglio excel oltre che visibili nel browser.

	A	B	C	D	E	F	G	H	I
1	Conversazione	Data Pubblicazione	Categoria Testi	Testo	Data Consultazione	Livello	Numero Parole	Numero caratteri	Numero Emoticon
2	1	30/11/2009	commento pubblicazione	30-nov	21/01/2010	1	2	10	0
3	2	25/11/2009	testo bacheca	Terminato il laboratorio alla Coop. Inizio prossimo laboratorio 20 gennaio	21/01/2010	1	10	64	0
4	3	23/11/2009	testo bacheca	corri, corri, corri... quando si è preso l'aire, come si fa a fermarsi?	21/01/2010	1	14	52	0
5	4	22/11/2009	commento pubblicazione	e ora... alla romana: "Mo, me fermo"	21/01/2010	1	7	23	0
6	4	22/11/2009	commento pubblicazione	le tue note le ho tra i preferiti ma poi le raccogliamo tutte	27/01/2010	2	13	49	0
7	4	22/11/2009	commento pubblicazione	ciao Wanda se vuoi vederle tutte insieme vai a questo indirizzo httpwwwadaascarinamecitazioniindexhtml e li rimarranno. Le note proseguono fino al 31 dicembre...	27/01/2010	3	22	137	0
8	4	23/11/2009	commento pubblicazione	wow ada sei troppo forte! che bella collezione.	27/01/2010	4	8	38	0
9	4	23/11/2009	commento pubblicazione	sono meravigliose, è un momento di benessere per la mia vista e il mio cervello leggere questi contenuti...grazie ancora	27/01/2010	5	20	99	0
10	5	21/11/2009	commento pubblicazione	bellissima, Ada! E' un altro dei tuoi prodigi?	27/01/2010	1	8	35	0
11	5	21/11/2009	commento pubblicazione	come sempre	27/01/2010	2	2	10	0
12	5	22/11/2009	commento pubblicazione	Il prodigio è di chi l'ha scritta. Io ho solo sparso qualche semino.	27/01/2010	3	14	54	0
13	5	23/11/2009	commento pubblicazione	bellissimo mi hai fatto ricordare cosa era emerso del laboratorio che ho fatto ad un corso di formazione ecm con il Prof. Duccio Demetrio, ho ancora il librettino che nè è uscito	27/01/2010	4	32	147	0
14	6	20/11/2009	testo bacheca	ad Anghiari c'è il sole	27/01/2010	1	6	19	0
15	6	20/11/2009	commento bacheca	sei di nuovo qui?	27/01/2010	2	4	13	0
16	6	20/11/2009	commento bacheca	ho provato a chiamarti a casa, ma non ho trovato nessuno. sono a palazzo testi con Tramma.	27/01/2010	3	17	71	0
17	7	19/11/2009	testo bacheca	Lunedì è passato e anche il martedì e il mercoledì. Il giovedì è di nuovo partenza.	27/01/2010	1	16	72	0
18	8	19/11/2009	commento pubblicazione	con qualcuno non funziona	27/01/2010	1	4	22	0
19	8	19/11/2009	commento pubblicazione	no no io sono per risolvere subito il problema!!!!!!!!!	27/01/2010	2	9	38	0
20	9	17/11/2009	commento pubblicazione	Sorridere fa bene	27/01/2010	1	3	15	0
21	9	18/11/2009	commento pubblicazione	:-))	27/01/2010	2	0	0	1
22	10	18/11/2009	commento pubblicazione	l'incontro tra l'infinito e il finito...	27/01/2010	1	8	30	0
23	11	16/11/2009	testo bacheca	riparte il lunedì	27/01/2010	1	3	16	0

Figura 13 – Foglio Excel scaricabile

Facendo sempre riferimento all'immagine, nella colonna A denominata "Conversazione" troviamo un numero progressivo. Qualora il numero sia lo stesso per più righe del foglio, esse rappresentano più battute di una stessa conversazione. A

questo punto sarà la colonna denominata “Livello” a indicarci l’ordine delle battute. Nella colonna B troviamo la data di pubblicazione, nella C la categoria di appartenenza dei testi e nella D i testi raccolti. Le ultime colonne contengono la data in cui testi sono stati raccolti, il numero di parole, il numero di caratteri e il numero di emoticon per ogni testo.

4 - Analisi del corpus – Gli utenti

Interrogando la tabella “utenti” possiamo analizzare i dati degli utenti del campione sulla base dei parametri selezionati.

Suddivisione utenza in base al sesso

Partiamo con l’esaminare attraverso una tabella la ripartizione degli utenti in base al sesso e alle fasce di età (Tab.2).

ETA’	UOMINI	DONNE	TOTALE
0-18	**3**	**9**	**12**
19-24	**23**	**44**	**67**
25-29	**40**	**35**	**75**
30-35	**26**	**8**	**34**
36-45	**15**	**7**	**22**
46-55	**4**	**3**	**7**
56-OLTRE	**0**	**3**	**3**
NON DICHIARATA	**84**	**104**	**188**
TOTALE	**195 (48%)**	**213 (52%)**	**408**

Tabella 2 – Suddivisione utenti del corpus in base al sesso

Da una prima analisi, quindi, risulta che il corpus è composto da 408 utenti di cui 213 donne (52%) e 195 uomini (48%).

La fascia di età con il numero maggiore di utenti è quella compresa tra 25 e 29 anni (75 utenti), seguita dalla fascia immediatamente precedente che comprende persone tra i 19 e i 24 anni (67 utenti). Molti però sono coloro che hanno deciso di non specificare l'età: 188 persone di cui 84 uomini e 104 donne. Possiamo comunque affermare che l'utenza presa in considerazione è verosimilmente composta soprattutto da giovani.

Confrontando la tabella che riporta i dati specifici degli utenti del campione con quella che ho realizzato in precedenza per riassumere l'utenza italiana di Facebook (Tab. 2), possiamo osservare come il mio campione non è rappresentativo e non rispecchia le stime che abbiamo visto.

Infatti, all'interno del campione, la percentuale di donne è superiore rispetto a quella degli uomini, mentre nelle stime italiane è maggiore la percentuale degli uomini. Anche nelle distinzioni degli utenti per fasce di età troviamo delle differenze. Infatti all'interno della suddivisione in base all'età degli utenti italiani che abbiamo visto in precedenza, quella più consistente è rappresentata dalla fascia di età compresa tra i 19 e i 24 anni, seguita da quella degli adolescenti fino a 18 anni; nel nostro caso, invece, la più consistente è quella tra i 25 e i 29 anni seguita da quella tra i 19 e i 24 anni.

4.1 - Analisi quantitativa del corpus – premessa

Per effettuare analisi quantitative, relative alla media delle parole e dei caratteri all'interno del corpus, ho deciso di utilizzare alcune funzioni che PHP mette a disposizione per lavorare sulle stringhe di testo.

Per contare le parole all'interno delle stringhe ho utilizzato la funzione EXPLODE(): questa funzione ha il compito di suddividere una stringa sulla base di un dato separatore e rilasciare il risultato in un array. Per far svolgere correttamente il proprio dovere alla funzione ho scelto come separatore un semplice spazio vuoto. Inoltre, essendo mia intenzione svolgere alcune analisi sull'utilizzo delle emoticon

all'interno dei record, occorre che il testo preso in considerazione sia strutturato in modo da non avere punteggiatura né le combinazioni di caratteri che formano, appunto, le emoticon. Una volta quindi che il testo sarà "standardizzato", soddisfacendo i requisiti della funzione, interrogheremo il database con una query SQL e, attraverso un ciclo WHILE(), applicheremo la funzione a ogni stringa immagazzinata.

4.2 – Standardizzazione dei testi

Il processo di standardizzazione consiste in alcuni passaggi. I testi raccolti all'interno del database non hanno tutti la stessa struttura: infatti ogni utente ha un proprio modo di scrivere e di utilizzare "emoticon", punteggiatura e spazi tra i costituenti della frase. Per fare in modo che tutti i testi abbiano la stessa struttura quindi ho proceduto in questo modo:

1. Ideazione e realizzazione di una espressione regolare volta a trovare e sostituire le emoticon all'interno delle frasi e ad eliminarle (verranno conteggiate a parte nell'analisi quantitativa).
2. Eliminazione della punteggiatura
3. Eliminazione degli spazi in eccesso
4. Esecuzione query di analisi
5. Inserimento nei campi numero_parole, numero_caratteri, numero_emoticon della tabella testi del numero di parole, caratteri ed emoticon

Vediamo in dettaglio i diversi passi.

4.2.1 - Isolamento delle emoticon

Durante la raccolta dei testi ho potuto rendermi conto di quanto le emoticon siano largamente utilizzate. Nella mia analisi ho deciso di trattarle a parte e quindi di non considerarle come parole. Le emoticon utilizzate sono molto varie, dalle più comuni (come ad esempio ":)" o ":(" a quelle più particolari (Es.: "è_é", "=0))") . Da quanto ho potuto vedere molto frequente è anche l'uso (24 volte all'interno del corpus) di un simbolo, il ♥, che in genere viene utilizzato per esprimere affetto o partecipazione. Il simbolo viene inserito digitando la combinazione dei due caratteri <3 e trasformato in ♥ dalla piattaforma Facebook.

Visti i presupposti ho deciso di togliere dai testi le varie combinazioni di caratteri che formano le emoticon utilizzando una funzione PHP per isolare le emoticon. La funzione che ho utilizzato è preg_replace() la quale mi permette di fare una ricerca nelle stringhe di testo secondo le regole di una espressione regolare e sostituire tutti i risultati della ricerca.

Una espressione regolare (o regexp, regex, RE, tutte abbreviazioni di "regular expression") è un'espressione costruita secondo una sintassi predefinita che permette di descrivere un insieme di stringhe. Le RE sono spesso utilizzate da editor di testo per la ricerca e la sostituzione di porzioni del testo, ma trovano anche ampi utilizzi nella programmazione. All'interno di una espressione regolare, la maggior parte dei caratteri corrisponde semplicemente a se stessa, quindi *a* corrisponderà alla stringa "*a*"; stessa cosa per una stringa costituita da caratteri ordinari, come ad esempio *spam*. Alcuni caratteri assumono invece significati particolari e vengono chiamati metacaratteri. Questi sono:

```
[ \ { | ( )  ^ $ ? + * .
```

e hanno diversi significati: possono essere quantificatori o ancoraggi o identificare gruppi e intervalli. La funzione che io ho utilizzato è questa:

```
<?php

$string2 = preg_replace ('#((♥+)|((:|=|;)\s?-
?_?o?0?\)+)|((=|:|\|)\'?\s?-?\(+)|([^A-Za-z0-9]:\'\s?-
?\<+)|([^A-Za-z0-9]:\s?-?\>+)|(\^\s?_*-*\.?\s?\^)|([^A-Za-z0-
```

```
9](=|:|;)\s?-?D+)|((X|:)\s?-?P+)|([^A-Za-z0-9]:\s?-?p+)|(:\s?-
?\|)|([^A-Za-z0-9](:|=)\s?-?\]+)|([^A-Za-z0-9]:\'?\s?-
?\[+)|([Xx]\s?-?D)|([^A-Za-z0-9]:\s?-?\*+)|(\*\s?_*.?\*)|([-
=]\s?_*\.?\s?[-=])|([^A-Za-z0-9]:\s?-?/)|([^A-Za-z0-9][oO]\.?-
?_*[oO][^A-Za-z0-9])|(\>_*\.?\<)|([^A-Za-z0-9]è\.?_*é)|([^A-
Za-z0-9]:\.?-?O+)|(:°D+)|([^A-Za-z0-9];\s?-?\)+)|([^A-Za-z0-
9];\s?-?\]+)|([^A-Za-z0-9];\s?-?P+)|([^A-Za-z0-9];\s?-
?p+)|(\>-?\))|(X\s?-?\)+)|(x-?_?x)|(X\s?-
?\(+)|((u|U)_(u|U))|(:-S+))#', ' ', $string);

?>
```

La funzione è costituita da tre parti:

- La prima definisce che cosa cercare nelle stringhe tramite espressione regolare.
- La seconda definisce con che cosa sostituire gli eventuali caratteri o parole che "matchano" con l'espressione regolare.
- La terza definisce su quale stringa, associata ad una variabile, eseguire le operazioni precedenti.

Nella prima parte, in pratica, ho elencato le varie combinazioni di caratteri che volevo evidenziare all'interno delle stringhe; ogni combinazione è separata dal metacarattere | che ha il significato di "oppure". Dato che ho l'opportunità anche di definire quante volte un carattere può comparire ho utilizzato anche i quantificatori seguenti:

- **?** che ha significato "0 oppure 1";
- ***** che ha significato "0 o più";
- **+** che ha significato "1 o più" .

Per fare un piccolo esempio, la funzione che ho realizzato andrà a ricercare tutti i caratteri "♥" o " :) " oppure " :P " o altri all'interno delle stringhe di testo. Una volta trovati li sostituirà con uno spazio vuoto.

In seguito, per verificare la funzionalità e la correttezza della funzione, ho eseguito un controllo a campione su circa 200 record, interrogando il database e confrontando le stringhe ottenute dopo l'applicazione della funzione con le originali e verificando che tutte le combinazioni indicate dall'espressione regolare fossero trattate nel modo voluto e quindi eliminate.

4.2.2 – Eliminazione della punteggiatura e degli spazi in eccesso.

Un altro step per la standardizzazione del testo è rappresentato dall'eliminazione di punteggiatura e spazi in eccesso all'interno della stringa di testo. Per effettuare questo passaggio ho utilizzato una nuova funzione PHP, str_replace(), e la funzione preg_replace() precedente.

La funzione str_replace() effettua la sostituzione di tutte le occorrenze di una stringa all'interno di un'altra stringa. Ho quindi inizializzato un array, che non è altro che un contenitore per stipare i dati, contenente tutti i caratteri di punteggiatura da togliere. In modo simile alla funzione vista in precedenza ho sostituito i caratteri di punteggiatura con uno spazio bianco all'interno delle stringhe di testo.

```
<?php

$caratteri= array(".", ",", "-",
"!",";",":","_","'","?","=","(",")","'","\"","<",">","°","§","
*","^","/","\\","`");

$string3= str_replace($caratteri, " ", $string2);

?>
```

L'ultimo passo nel procedimento di standardizzazione è costituito dalla rimozione degli spazi bianchi doppi o in eccesso all'interno della stringa, che si sono creati con le varie sostituzioni, attraverso la funzione preg_replace() e una nuova espressione regolare:

```
<?php

$string4= preg_replace('/  +/', ' ', $string3);

?>
```

Con questa espressione abbiamo creato e ordinato il testo all'interno delle stringhe e non resta che analizzarle con apposite funzioni che vedremo in seguito.

4.3 – Analisi: media delle parole e dei caratteri all'interno delle stringhe di testo.

Dopo la standardizzazione delle stringhe ho eseguito il calcolo della media delle parole e dei caratteri all'interno di esse. Anche in questo caso ho utilizzato una funzione PHP che prende il nome di EXPLODE(), la quale ha il compito di suddividere una stringa sulla base di un dato separatore e rilascia il risultato in un array. Per interrogare il database ho realizzato una maschera che mi permette di effettuare varie analisi con la possibilità di interagire con alcuni filtri (Fig. 14).

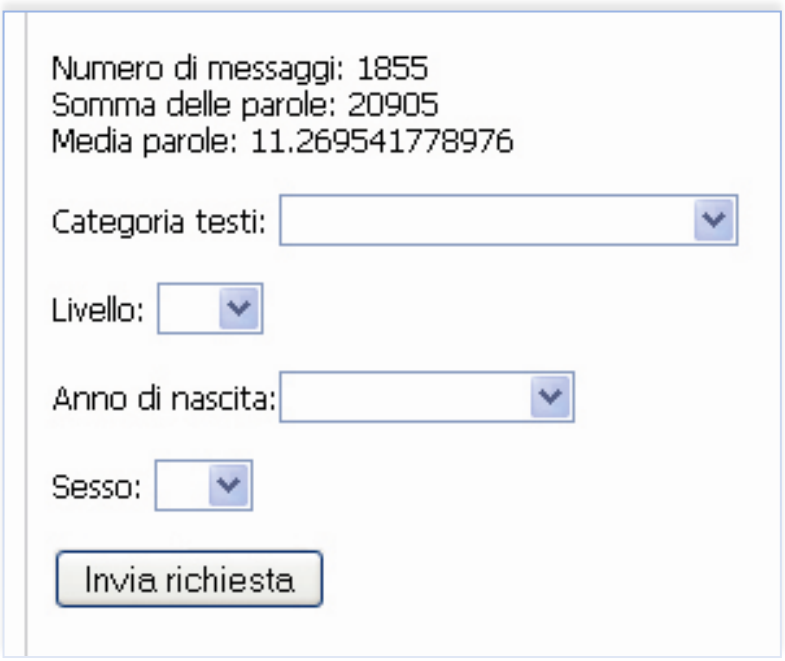

Figura 14 – Maschera di interrogazione analisi delle parole

Dalla figura possiamo ricavare un primo dato relativo al corpus in generale. Il numero dei messaggi (senza distinzione tra le categorie di testo) è 1855 e la somma delle parole è di 20905. La somma delle parole è stata calcolata tramite la funzione di cui parlavo poco prima applicata ad ogni stringa di testo in questo modo:

```
<?php

$par=explode(' ',$row['testo']);

$num=count($par);

?>
```

Nella prima parte della funzione, ogni stringa di testo viene smembrata ogni volta che il compilatore incontra uno spazio. In questo modo ogni sequenza di caratteri delimitata da spazi viene inserita all'interno di un array. Nella seconda parte del procedimento di conteggio viene utilizzata la funzione COUNT() che ritorna un numero intero che indica la dimensione dell'array, ovvero il numero dei suoi elementi (le parole nel nostro caso). Dopo aver effettuato il calcolo per ogni stringa presente nel database ho inserito il risultato corrispondente nel campo numero_parole della tabella testi con una semplice query di aggiornamento.

```
<?php

$query2="update testi set numero_parole = '$num' where
id_testo = '$row[id_testo]'";

$result2=mysql_query($query2, $db) or die();

?>
```

Quindi con semplici calcoli possiamo arrivare alla media generale che è di circa 11,3 parole per messaggio (Tab. 3). Nel calcolo della media delle parole ho trovato anche segni aritmetici, come ad esempio il + o il –, utilizzati come parole vere e proprie. Nei confronti di questi segni ho utilizzato il seguente criterio: se l'utente lo può leggere ad alta voce come una parola, è una parola. La stessa cosa è valida anche per i numeri scritti in cifre.

Utilizzando la funzione STRLEN(), che restituisce il numero di caratteri di una stringa, sulle stesse stringhe ma togliendo del tutto gli spazi tra le parole, possiamo ottenere anche il numero dei caratteri (Tab. 3) che ho inserito nel campo numero_caratteri della tabella testi.

NUMERO MESSAGGI IN TUTTO IL CORPUS	**1.855**
SOMMA DELLE PAROLE DI TUTTO IL CORPUS	**20.905**
MEDIA DELLE PAROLE DI TUTTO IL CORPUS	**11,26**
MEDIA DEI CARATTERI DI TUTTO IL CORPUS	**51,92**
RAPPORTO CARATTERI / PAROLE DI TUTTO IL CORPUS	**4,60**
PERCENTUALE DI MESSAGGI IN BACHECA COMMENTATI	**48,16%**

NUMERO DI TESTI DI LIVELLO 1	770
NUMERO DI TESTI DI LIVELLO 2	**373**
NUMERO DI TESTI DI LIVELLO 3	**232**
NUMERO DI TESTI DI LIVELLO 4	**156**
NUMERO DI TESTI DI LIVELLO 5	**103**
NUMERO DI TESTI DI LIVELLO 6	**70**
NUMERO DI TESTI DI LIVELLO 7	**45**
NUMERO DI TESTI DI LIVELLO 8	**29**
NUMERO DI TESTI DI LIVELLO 9	**22**
NUMERO DI TESTI DI LIVELLO 10	**15**

Tabella 3 – Visione generale del corpus. Media parole e caratteri.

Nella tabella soprastante ho effettuato anche analisi sul rapporto tra il totale dei caratteri e quello delle parole, sulla media dei commenti ai testi della bacheca e il numero di testi per livello. Il risultato che ho ottenuto per quanto riguarda la lunghezza media, in termini di caratteri, delle parole all'interno del corpus è di 4,60. La percentuale dei messaggi della bacheca commentati è del 48,16%: questo significa che i commenti ai messaggi scritti in bacheca sono abbastanza frequenti. Dal numero di testi per livello possiamo osservare che all'aumentare di quest'ultimo il numero di testi diminuisce progressivamente (Grafico 1).

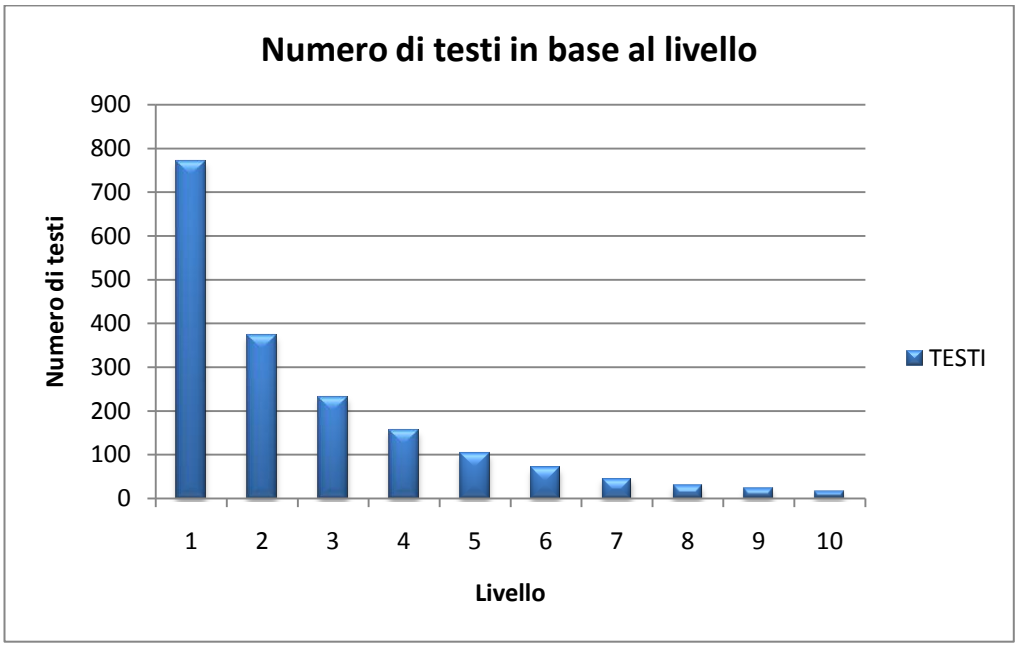

Grafico 1 – Testi per livello

Tramite i filtri è possibile approfondire l'analisi incrociando alcuni dati relativi ai testi e agli utenti:

- Categoria testi: attraverso la SELECT è possibile specificare su quali tipi di testi effettuare l'analisi (testo bacheca, commento bacheca, commento pubblicazione).

- Livello: all'interno della SELECT sono indicati i vari gradi di livello che possiamo usare per l'analisi. Il numero massimo selezionabile è il massimo rilevato al momento della catalogazione dei testi.

- Anno di nascita: con questo filtro possiamo scegliere la fascia di età degli utenti su cui effettuare l'analisi.
- Sesso: con questo filtro è possibile selezionare il sesso degli utenti.

Adesso sfruttando il filtro relativo alla categoria dei testi effettuerò alcune analisi sulla media delle parole e dei caratteri nelle stringhe di testo.

4.3.1 – Calcolo della media delle parole e dei caratteri nei testi della bacheca.

Il calcolo della media delle parole e dei caratteri all'interno del testo della bacheca si base sui seguenti dati (Tab. 4).

	TUTTI GLI UTENTI	19-24 ANNI	25-29 ANNI
NUMERO MESSAGGI DELLA BACHECA	**463**	**131**	**81**
SOMMA DELLE PAROLE DELLA BACHECA	**5.921**	**1.754**	**877**
MEDIA DELLE PAROLE DELLA BACHECA	**12,78**	**13,38**	**10,82**
MEDIA DEI CARATTERI DELLA BACHECA	**59,20**	**59,83**	**50,14**
RAPPORTO CARATTERI / PAROLE DELLA BACHECA	**4,62**	**4,46**	**4,63**

Tabella 4 – Media caratteri e parole nei testi della bacheca

Quindi possiamo dire che all'interno dei testi della bacheca, che necessariamente hanno livello 1, la lunghezza media si attesta sulle 13 parole circa e circa 60 caratteri. La lunghezza media delle parole all'interno di tutti i testi della bacheca raccolti è di 4,62 caratteri. Dato che il campione è costituito da due fasce di età principali, quella compresa tra 19 e 24 anni e quella compresa tra 25 e 29 anni, possiamo effettuare una ulteriore analisi specificando anche l'età degli utenti del campione. Come

possiamo notare la media delle parole utilizzata nei testi della bacheca è maggiore nella fascia di età compresa tra 19 e 24 anni. La media dei caratteri è di 60 circa nella fascia più numerosa (19-24 anni) e di 50 circa in quella compresa tra 25 e 29 anni. Interessante osservare, anche in questo caso, il rapporto tra caratteri e parole: 4,46 nella prima fascia contro 4,63 nella seconda.

4.3.2 – Calcolo della media delle parole nei commenti della bacheca.

Per effettuare una analisi generale relativa ai commenti della bacheca ci basiamo sui dati in tabella (Tab. 5).

	TUTTI GLI UTENTI	19-24 ANNI	25-29 ANNI
NUMERO MESSAGGI DEI COMMENTI DELLA BACHECA	**655**	**178**	**133**
SOMMA DELLE PAROLE DEI COMMENTI DELLA BACHECA	**7.469**	**1.935**	**1.630**
MEDIA DELLE PAROLE DEI COMMENTI DELLA BACHECA	**11,40**	**10,87**	**12,25**
MEDIA DEI CARATTERI DEI COMMENTI DELLA BACHECA	**59,20**	**50,16**	**55,73**
RAPPORTO CARATTERI / PAROLE DEI COMMENTI DELLA BACHECA	**3,67**	**4,61**	**4,54**

Tabella 5 – Media delle parole nei commenti della bacheca

Dunque la media all'interno dei commenti alla bacheca è di circa 11,4 parole, 52,20 caratteri e un rapporto di 3,67. Possiamo utilizzare gli stessi criteri dell'analisi precedente sui testi della bacheca relativi alle fasce di età per confrontare i dati. La media delle parole è maggiore nella fascia di età compresa tra 25 e 29 anni con un rapporto tra caratteri e parole del testo di 4,54.

Data la possibilità per questa categoria di testi di interagire anche con il filtro livello è interessante calcolare la media man mano che il livello aumenta. Faremo una analisi di esempio sui commenti della bacheca dal livello 2 al livello 6. I dati estrapolati sono riassunti in tabella sottostante (Tab. 6).

	LIVELLO 2	LIVELLO 3	LIVELLO 4	LIVELLO 5	LIVELLO 6
NUMERO MESSAGGI	**223**	**134**	**91**	**61**	**46**
SOMMA DELLE PAROLE	**2.044**	**1.457**	**1.066**	**818**	**665**
MEDIA DELLE PAROLE	**9,16**	**10,87**	**11,71**	**13,40**	**14,45**
MEDIA DEI CARATTERI	**42,02**	**50,23**	**51,78**	**60,85**	**65,58**
RAPPORTO CARATTERI / PAROLE	**4,58**	**4,61**	**4,42**	**4,53**	**4,53**

Tabella 6 – Media delle parole in base al livello

È interessante notare come al salire del livello, quindi man mano che la conversazione attraverso i commenti si allunga, aumenta la media delle parole utilizzate e quindi anche quella dei caratteri. Possiamo quindi affermare che le battute della conversazione sembrano farsi più lunghe e articolate all'aumento del livello. Vediamo qui di seguito alcuni esempi di conversazioni raccolte che possono aiutarci a capire meglio.

Nella tabella sottostante (Tab. 7) è presentata una conversazione estratta dal database. Nelle prime due colonne sono specificate le iniziali dell'autore e la categoria del testo. Nella colonna "testo" abbiamo il testo raccolto e nella colonna "livello di conversazione" ho inserito un numero che indica il turno della conversazione. Possiamo vedere da questi esempi di conversazione come, ad una iniziale frase della bacheca, seguono frasi più articolate.

AUTORE	CATEGORIA TESTO	TESTO	LIVELLO DI CONVERSAZIONE	NUMERO PAROLE
M. G.	testo bacheca	ANNA ANNA ANNAAAA! Ho vinto con Marcoooooo! =D ora sono io la preferita veroo???	1	13
M. G.	commento bacheca	...non ci provare nemmeno Mari, tanto non ce la fai...	2	10
A. F.	commento bacheca	Ma che devo fare con voi 2? Ahhh	3	8
M. G.	commento bacheca	niente, solo apprezzarmi di più...	4	5
A. F.	commento bacheca	Comunque Marco prende un bonus perché è andato a Montecarlo e ha pure fatto finale	5	15
M. G.	commento bacheca	UFFAAAAAAA! perchè lui ha sempre scusanti??!?!? NON è GIUSTO! io cambio maestra e pure palestraaa! Ma che maestra sei..come puoi far soffrire così la tua allieva??? Per chi poi....	6	30

A. I. C.	testo bacheca	Le PISANE in trasferta a FERRARA spopolano ANCORA! COMPLIMENTI CARA COLLEGA DOTTORANDA xxxxx xxxxxxx!	1	14
L. T.	commento bacheca	Grande!! Dai, lo sapevamo tutti ma non osavamo dire niente per non illuderla...qualcuno si ritita sempre!	2	17
L. T.	commento bacheca	ritira*	3	1
A. I.C.	commento bacheca	Sapevate tutti?? ke cattivi!!!! potevate dirlo almeno a me!!! non glielo avrei detto!	4	13
L. T.	commento bacheca	ma non che ERA successo, che SAREBBE successo! La notizia l'ho appresa ora da te :)	5	16
L. F.	commento bacheca	grazie ragazze!!!ancora non ci credo! l'ho saputo ieri sera.....sono ancora sotto shock!!!!! grazie cara collega dottoranda xxxx xxxxxx xxxxxx!!!! faremo furore!!!	6	24

A. I.	testo bacheca	..ma tu guarda te se per colpa di alcune persone che NON svolgono bene il loro lavoro di infermiere, mamma deve patire come un cane!?!?!?!? andate a lavorare in fabbrica!!!	1	30

D. T.	commento bacheca	eh guarda, non sai come ti capisco, è successo anche a mia mamma... sarebbero da ingabbiare alcuni!!	2	17
A. I.	commento bacheca	Davvero..trattano lei e altri pazienti in un modo osceno!	3	10
L. F.	commento bacheca	ehi cosa è successo????	4	4
A. I. C.	commento bacheca	Ci sono 3 infermiere che trattano i pazienti come fossero sacchi di patate! tra cui mamma..che, x colpa loro, le hanno dovuto cambiare tipologia di sacchetto! l'hanno rovinata! le hanno procurato un distaccamento della stomia..'ste deficienti! tu sentissi come urlano alle vecchiette lì ricoverate! da denuncia! alla signora in camera con mamma, stamani non hanno somministrato le pasticche x il cuore!!! questa donna è cardiopatica! il figlio, quando è arrivato, ha fatto casino, dicendo ke se ne sarebbero andati via!!! c'ha ragione..	5	86
L. F.	commento bacheca	ma che gente disumana c'è al mondo!!! ma fate casino anche voi! non si trattano così le persone malate! andate a lamentarvi dalla capo-infermiera o dal primario!	6	29
L.	commento bacheca	purtroppo ci sono passata anche io e capisco la tua rabbia. ci vuole pazienza (perchè il sistema non cambia) e determinazione (per agire laddove risulta necessario). Stai attenta a tutto e se devi rompi le palle a tutti. Un abbraccio.	7	40
S. V.	commento bacheca	Quando c'era mia nonna abbiamo portato tutte le medicine da casa e dovevamo anche stare attenti che gliele dessero davvero! Ma dov'è tua madre, a Pescia? Postaccio. E poi quando si agitava, giù di sedativi. Sembrava matta. Non è che fosse tutta rifinita, ma a casa non aveva le allucinazioni almeno! E così alla fine io, il babbo e una badante ci scambiavamo all'ospedale. Ho studiato analisi II accanto al suo letto..	8	75

A. I. C.	commento bacheca	No, Sara, non è a Pescia..è all'ospedale Versilia a Camaiore e finora non avevamo avuto problemi, fino a quando abbiamo scoperto che queste 3 infermiere sono delle arpie! e grazie a loro, è probabile che mamma debba tornare in sala operatoria per sistemare il guaio che LORO hanno combinato!	9	51

Tabella 7 – Esempi di commenti della bacheca

4.3.3 – Calcolo della media delle parole e dei caratteri nei commenti delle pubblicazioni.

Passiamo infine ad analizzare i commenti delle pubblicazioni procedendo come nell'analisi precedente. I dati su cui ci basiamo sono elencati in tabella (Tab. 8).

	TUTTI GLI UTENTI	19-24 ANNI	25-29 ANNI
NUMERO MESSAGGI DEI COMMENTI DELLE PUBBLICAZIONI	**736**	**126**	**258**
SOMMA DELLE PAROLE DEI COMMENTI DELLE PUBBLICAZIONI	**7.499**	**1.020**	**2.814**
MEDIA DELLE PAROLE DEI COMMENTI DELLE PUBBLICAZIONI	**10,18**	**8,09**	**10,90**
MEDIA DEI CARATTERI DEI COMMENTI DELLE PUBBLICAZIONI	**47,29**	**36,39**	**51,04**
RAPPORTO CARATTERI / PAROLE DEI COMMENTI DELLE PUBBLICAZIONI	**4,64**	**4,49**	**4,68**

Tabella 8 – Media delle parole e dei caratteri nei commenti delle pubblicazioni

Possiamo notare dunque come una media generale delle parole dei commenti delle pubblicazioni sia di circa 10 parole per testo raccolto. Curioso osservare come gli utenti compresi tra 25 e 29 anni sono molto più attivi nel commentare le pubblicazioni rispetto agli utenti della fascia precedente. La media delle parole è di

circa 8,10 per questi ultimi e di 10,90 per i primi. Passiamo quindi ad analizzare e riassumere i dati secondo i livelli (Tab. 9).

	LIVELLO 1	LIVELLO 2	LIVELLO 3	LIVELLO 4	LIVELLO 5
NUMERO MESSAGGI	**306**	**150**	**98**	**65**	**42**
SOMMA DELLE PAROLE	**2.423**	**1.445**	**1.012**	**929**	**553**
MEDIA DELLE PAROLE	**7,91**	**9,63**	**10,32**	**14,29**	**13,16**
MEDIA DEI CARATTERI	**36,85**	**44,44**	**48,23**	**65,01**	**63,76**
RAPPORTO CARATTERI / PAROLE	**4,65**	**4,61**	**4,67**	**4,54**	**4,84**

Tabella 9 - Media delle parole e dei caratteri in base ai livelli

Anche in questo caso, come si evince dalla tabella 9, all'avanzare del livello dei commenti aumenta anche la media delle parole e quindi la conversazione sembra farsi ancora una volta più articolata. Il rapporto della media dei caratteri con quella delle parole è di circa 4,65. Eventuali emoticon presenti all'interno dei testi non sono state calcolate nella media delle parole. Analizziamo adesso alcuni esempi come fatto in precedenza (Tab. 10).

AUTORE	CATEGORIA TESTO	TESTO	LIVELLO DI CONVERSAZIONE	NUMERO PAROLE
B. P.	commento pubblicazione	ahah...con un governo così vedo bene soltanto il ministero dell'ambiente XD	1	12
G. S.	commento pubblicazione	voglio minimo minimo quattro portaborse	2	5
C.P.	commento pubblicazione	wow ministro dei beni culturali...ahah XD	3	6
L.B.L.	commento pubblicazione	ahhh ora ho capito hehe cribbio, grazie a tutti e buon lavoro!!!	4	12

B.P.	commento pubblicazione	cadino:beni culturali=biankiz:economia la costante è:una materia che ci fa schifo!! XD	5	14
C.P.	commento pubblicazione	muahahah è vero...	6	3
B. P.	commento pubblicazione	perché come dice il benve: "gli economisti sono dei matematici frustrati, che si divertono a fare gli apprendisti stregoni inventandosi iperboli e parabole!!!" XD grande!!	7	24

B. P.	commento pubblicazione	oh oh... forse ci becca...	1	5
B. L.	commento pubblicazione	hehehehe, cm festaiolo ci siamo o meglio festinaiolo, inteligente, dibende dai punti di vista, cribbio XD	2	15
Bi.P.	commento pubblicazione	ma silvio tu non sei intelligente però..."sei tutto maschio!!" XD "less bad that silvio there is!!" eheh...	3	17
B.L.	commento pubblicazione	non parlare in quella sporca lingua comunista transalpina, cribbio!!! si dice "meno male che silvio c'è", e basta, cribbio!!! c c c c cr cr cr cr CRIBBIOOOOOOOOO!!!!!!!!!!!!!!	4	29
B. P.	commento pubblicazione	oh scusa...volevo tradurla in tutte le lingue, come l'internazionale...silvio è un fenomeno che dev'essere universale, perché noi italiani dovremmo continuare a tenerlo egoisticamene per noi ?	5	29

L. P.	testo bacheca	Bianchy, basta tristezza! ci sono anch' io per te! ♥	1	9
S. M.	commento bacheca	grande spoon river!!!!!!!!!!	2	3
L. P.	commento bacheca	Ehilà, chi si vede. E' sempre un piacere per me!	3	10
L. P.	commento	It' s a pleasure, sir Ilvio.	4	6

	bacheca			
S. M.	commento bacheca	piacere di che????ma va là....	5	6
L. P.	commento bacheca	Se mi vedessi anche solo una volta cambieresti idea...hihi ♥	6	10
B. P.	commento bacheca	Già grande Spoon River...lo sto rileggendo per l'ennesima volta...e ogni volta scopro qualcosa che non avevo notato o non avevo capito..una frase che avevo interpretato in un altro modo... E poi ho scoperto che Thomas Rhodes era il Berlusconi di Spoon River e il giudice, quello di DeAndré, il nano, era il suo Ghedini XD	7	58
S. M.	commento bacheca	m vattela pija n'saccoccia...lorena di chicazzo sei....	8	9
L. P.	commento bacheca	Il mio nome è Lorena, xxxxxxx xxxxxx, tu piuttosto? caro Ilvietto...sono sicura che ti stravolgerei l' esistenza se mi vedessi anche solo una volta ♥	9	25

Tabella 10 – Esempi di commenti della pubblicazione

4.4 – Analisi dell'utilizzo delle emoticon all'interno del testo.

L'analisi adesso si concentrerà sull'utilizzo delle emoticon all'interno del corpus. Per emoticon intendiamo le riproduzioni delle principali espressioni facciali e che indicano la presenza di uno stato d'animo. Questo stato d'animo viene reso graficamente attraverso l'utilizzo di combinazioni di caratteri. Abbiamo visto come nelle precedenti analisi le emoticon siano state escluse nel calcolo della media delle parole e dei caratteri. Questo perché l'uso delle emoticon mi è sembrato particolarmente rilevante e quindi interessante da trattare singolarmente.

Per procedere nella analisi ancora una volta mi sono servito della standardizzazione dei testi attraverso la procedura trattata poco prima. L'unica differenza è stata quella di marcare tutte le emoticon, anziché eliminarle come avevamo fatto prima,

attraverso l'uso della precedente espressione regolare e della funzione preg_replace() in PHP. Vediamola in dettaglio:

```
<?php

$string2 = preg_replace ('#((♥+)|((:|=|;)\s?-
?_?o?0?\)+)|((=|:|\|)\'?\s?-?\(+)|([^A-Za-z0-9]:\'\s?-
?\<+)|([^A-Za-z0-9]:\s?-?\>+)|(\^\s?_*-*\.?\s?\^)|([^A-Za-z0-
9](=|:|;)\s?-?D+)|((X|:)\s?-?P+)|([^A-Za-z0-9]:\s?-?p+)|(:\s?-
?\|)|([^A-Za-z0-9](:|=)\s?-?\]+)|([^A-Za-z0-9]:\'?\s?-
?\[+)|([Xx]\s?-?D)|([^A-Za-z0-9]:\s?-?\*+)|(\*\s?_*.?\*)|([-
=]\s?_*\.?\s?[-=])|([^A-Za-z0-9]:\s?-?/)|([^A-Za-z0-9][oO]\.?-
?_*[oO][^A-Za-z0-9])|(\>_*\.?\<)|([^A-Za-z0-9]è\.?_*é)|([^A-
Za-z0-9]:\.?-?O+)|(:°D+)|([^A-Za-z0-9];\s?-?\)+)|([^A-Za-z0-
9];\s?-?\]+)|([^A-Za-z0-9];\s?-?P+)|([^A-Za-z0-9];\s?-
?p+)|(\>-?\))|(X\s?-?\)+)|(x-?_?x)|(X\s?-
?\(+)|((u|U)_(u|U))|(:-S+))#', '#emoticon#', $string);

?>
```

Come si può osservare, nella seconda parte della funzione abbiamo indicato che ogni sequenza di caratteri all'interno del testo che corrisponde ad una emoticon venga sostituita con la parola "#emoticon#". A questo punto basta ricercare all'interno delle stringhe di testo la parola che abbiamo deciso sostituire la sequenza di caratteri. A livello di codice il tutto si traduce con l'utilizzo della funzione substr_count() che restituisce il numero di volte in cui una parola è contenuta all'interno di una stringa.

Ecco il codice:

```
<?php

$conta_emoticon=substr_count($row['testo'], "#emoticon#");?>
```

Dopo aver ricavato il numero delle emoticon in ogni testo, inseriamo il dato all'interno del campo numero_emoticon della tabella testi.

Per eseguire alcune analisi ho utilizzato poi la maschera realizzata in precedenza per applicare alcuni filtri di ricerca. Ecco quindi riassunti in tabella (Tab. 11) alcuni dati estrapolati dalle stringhe di testo:

	TESTO BACHECA	COMMENTO BACHECA	COMMENTO PUBBLICAZIONE
MEDIA GENERALE EMOTICON PER MESSAGGIO	**0,23**	**0,38**	**0,40**
MEDIA EMOTICON PER MESSAGGIO FASCIA DI ETA' 19-24	**0,26**	**0,34**	**0,36**
MEDIA EMOTICON PER MESSAGGIO FASCIA DI ETA' 25-29	**0,17**	**0,54**	**0,52**
MEDIA EMOTICON PER MESSAGGIO UOMINI	**0,10**	**0,25**	**0,27**
MEDIA EMOTICON PER MESSAGGIO DONNE	**0,31**	**0,46**	**0,47**

	TESTO BACHECA	COMMENTO BACHECA	COMMENTO PUBBLICAZIONE
NUMERO EMOTICON COMPLESSIVO	**110**	**254**	**295**
NUMERO EMOTICON FASCIA DI ETA' 19-24	**35**	**62**	**46**
NUMERO EMOTICON FASCIA DI ETA' 25-29	**14**	**73**	**136**
NUMERO EMOTICON MESSAGGI UOMINI	**18**	**58**	**74**
NUMERO EMOTICON MESSAGGI DONNE	**92**	**196**	**221**

Tabella 11 – Media e numero di emoticon

All'interno delle celle della tabella possiamo leggere la media per le relative categorie di testi e con alcuni filtri di ricerca applicati. Inoltre viene indicata anche la somma effettiva delle emoticon che soddisfano i criteri di ricerca.

Molto interessante notare come l'uso delle emoticon sia maggiore nei commenti ai testi della bacheca e soprattutto nei commenti alle pubblicazioni. In effetti l'espressione frequente di uno stato d'animo attraverso l'emoticon in risposta a una pubblicazione di un utente era stata notata nel momento di raccolta dei testi e forse favorita dal fatto che molti utenti esternano, nelle bacheche e nelle pubblicazioni, i propri stati d'animo. Questi possono essere facilmente commentabili attraverso l'uso delle cosiddette "faccine": un modo veloce e a volte più intuitivo. Come possiamo osservare nel corpus l'uso delle emoticon è maggiore nella fascia di età compresa tra i 25 e i 29 anni e soprattutto nei commenti alle pubblicazioni.

Altro dato curioso è rappresentato dall'uso in base al sesso. Possiamo osservare infatti che le emoticon vengono utilizzate in numero maggiore dagli utenti di sesso femminile.

5 – Analisi contrastiva con il linguaggio nei web forum.

Avendo completato le analisi sul corpus possiamo confrontare i risultati con quelli emersi dallo studio del linguaggio dei forum e presenti nell'articolo *Building a corpus of Italian Web forums: standard encoding issues and linguistic features* di Mirko Tavosanis e Silvia Petri (Tavosanis Petri 2009).

Come dice l'articolo (Tavosanis Petri 2009, 115), i web forum sono considerati i più popolari strumenti di interazione testuale tanto che una statistica Eurostat ci mostra che circa il 50% degli europei ha pubblicato almeno un testo nei forum nell'anno precedente lo studio. L' analisi è stata condotta tramite la formazione di un corpus di testi provenienti da 4 forum italiani con caratteristiche e strumenti di amministrazione diversi. I forum sono:

- **Accademia della Crusca:** contiene discussioni formali sul linguaggio italiano;
- **ADI – Associazione dei dottorandi e dottori di ricerca italiani:** forum che contiene discussioni di studenti e dottori di ricerca;

- **HTML.it:** forum tecnico sul linguaggio HTML e altri linguaggi di programmazione;
- **NGI :** social forum che ha come utenti i partecipanti ad un gioco di ruolo online.

I forum sono stati scelti in modo da avere differenti esempi di forma e stili di comunicazione. La codifica del corpus è stata effettuata sulla base di una DTD Xml e ogni elemento all'interno dei testi raccolti è stato opportunamente classificato. Le principali caratteristiche del corpus sono riassunte in questi punti:

- Numero di posts: 1.186
- Numero totale di parole: 150.115
- Media delle parole per messaggio: 125

Da questi primi dati possiamo effettuare un confronto tra i due corpus che riassumo in tabella (Tab. 12).

	CORPUS TESTI FACEBOOK	CORPUS WEB FORUM
NUMERO MESSAGGI IN TUTTO IL CORPUS	**1.855**	**1.186**
SOMMA DELLE PAROLE DI TUTTO IL CORPUS	**20.905**	**150.115**
MEDIA DELLE PAROLE DI TUTTO IL CORPUS	**11,26**	**125**

Tabella 12 – Media delle parole nel corpus

Dai dati esposti risulta evidente che la media delle parole per messaggio è di gran lunga maggiore all'interno dei post dei forum che nei testi di Facebook. Interessante quindi notare che la lunghezza dei messaggi rappresenta una differenza importante tra le due piattaforme di comunicazione; il forum sembra essere più adatto a

raccogliere frasi articolate e discorsive mentre in Facebook i messaggi sono più brevi e probabilmente più "istintivi".

Altro dato che possiamo confrontare è quello relativo all'uso delle emoticon. Nella tabella sottostante (Tab. 13) vengono riassunti i risultati dello studio sui due corpus:

	CORPUS TESTI FACEBOOK	CORPUS WEB FORUM
PERCENTUALE EMOTICON PER MESSAGGIO	**35,5%**	**44%**
PERCENTUALE EMOTICON IN RAPPORTO ALLE PAROLE	**3,15%**	**0,34%**

Tabella 13 – Percentuale di emoticon

Come possiamo osservare la percentuale di emoticon è maggiore all'interno dei web forum. Tuttavia anche in Facebook vi è un largo uso di emoticon nonostante non sia presente una interfaccia che permette di inserirle in modo visuale all'interno dei messaggi come in molti forum e che ne semplifica l'inserimento (Fig. 15).

Figura 15 – Maschera inserimento testo nei forum

Altro dato interessante è rappresentato dalla percentuale delle emoticon in rapporto alle parole. In questo caso la percentuale è di 3,15% nel corpus di testi di Facebook mentre nel corpus web forum è dello 0,34%; quindi rispetto al numero delle parole la percentuale è maggiore in Facebook.

Bibliografia e Sitografia

Antonelli, G. (2007), *L'italiano nella società della comunicazione*, Il Mulino, Bologna

Baron, (2008), *Always On. Language in an Online and Mobile World*, Oxford University Press

Petri, S. & Tavosanis, M. (2009), *Building a corpus of Italian Web forums: standard encoding issues and linguistic features*. "Journal for language technology and computational linguistics" (JLCL), 25, 1, 2009, pp. 121-133.

Wikipedia. (2010). (http://www.wikipedia.org).

Facebook (2010) (http://www.facebook.com/)

Forum per webmaster Giorgio Taverniti (2010) (http://www.giorgiotave.it/forum/)

Forum HTML.it (2010) (http://www.html.it/)

Alexa the web information company (2010) (http://www.alexa.com/)

Blog ufficiale di Facebook (2010) (http://blog.facebook.com/)

Mappe, Popolazione, Statistiche Demografiche dell'ISTAT (2010) (http://demo.istat.it/index.html)