Facebook and the Real World: Correlations between Online and Offline Conversations

Fabio Celli
University of Trento
fabio.celli@unitn.it

Luca Polonio
University of Trento
luca.polonio@unitn.it

Abstract

English. Are there correlations between language usage in conversations on Facebook and face to face meetings? To answer this question, we collected transcriptions from face to face multi-party conversations between 11 participants, and retrieved their Facebook threads. We automatically annotated the psycholinguistic dimensions in the two domains by means of the LIWC dictionary, and we performed correlation analysis. Results show that some Facebook dimensions, such as "likes" and shares, have a counterpart in face to face communication, in particular the number of questions and the length of statements. The corpus we collected has been anonymized and is available for research purposes.

Italiano. *Ci sono correlazioni tra l'uso del linguaggio nelle conversazioni su Facebook e faccia a faccia? Per rispondere a questa domanda, abbiamo raccolto delle trascrizioni di conversazioni di gruppo tra 11 partecipanti e campionato i loro dati Facebook. Abbiamo annotato automaticamente le dimensioni psicolinguistiche per mezzo del dizionario LIWC e abbiamo estratto le correlazioni tra le due diverse tipologie testuali. I risultati mostrano che alcune dimensioni linguistiche di Facebook, come i "mi piace" e il numero di condivisioni, correlano con dimensioni linguistiche dell'interazione faccia a faccia, come il numero di domande e la lunghezza delle frasi. Il corpus e' stato anonimizzato ed e' disponibile per scopi di ricerca.*

1 Introduction and Background

In recent years we had great advancements in the analysis of communication, in face to face meetings as well as in Online Social Networks (OSN) (Boyd and Ellison, 2007). For example, resources for computational psycholinguistics like the Linguistic Enquiry Word Count (LIWC) (Tausczik and Pennebaker, 2010), have been applied to OSN like Facebook and Twitter for personality recognition tasks (Golbeck et al., 2011) (Schwartz et al., 2013) (Celli and Polonio, 2013) (Quercia et al., 2011). Interesting psychological research analyzed the motivations behind OSN usage (Gosling et al., 2011) (Seidman, 2013) and whether user profiles in OSN reflect acual personaliy or a self-idealization (Back et al., 2010).

Also Conversation Analysis (CA) of face to face meetings, that has a long history dating back to the '70s (Sacks et al., 1974), has taken advantage of computational techniques, addressing detection of consensus in business meetings (Pianesi et al., 2007), multimodal personality recognition (Pianesi et al., 2008) and dectection of conflicts from speech (Kim et al., 2012).

In this paper we make a comparison of the linguistic behaviour of OSN users both online and in face to face meetings. To do so, we collected Facebook data from 11 volunteer users, who participated to an experimental setting where we recorded face to face multiparty conversations of their meetings. Our goal is to discover relationships between a rich set of psycholinguistic dimensions (Tausczik and Pennebaker, 2010) extracted from Facebook metadata and meeting transcriptions. Our contributions to the research in the fields on Conversation Analysis and Social Network Analysis are: the release of a corpus of speech transcriptions aligned to Facebook data in Italian and the analysis of correlations between psycholinguistic dimensions in the two settings.

The paper is structured as follows: in section 2 we describe the corpora and the data collection, in section 3 we explain the method adopted and report the results, in section 4 we draw some conclusions.

2 Data and Method

We collected 11 volunteer Italian native speakers, who provided the consent to use their Facebook metadata, and organized meeting sessions with them to collect spoken linguistic data. The meetings consist in sessions of one hour, where participants, 6 in the first session and 5 in the second one, performed free multi-party conversations. Groups were balanced by gender and aged between 18 and 50 years. There were no restrictions, predefined task or topic to elicitate speech. In order to prevent biases in the interactions we put in the groups persons who do not know each other.

We recorded and manually transcribed a corpus of spoken conversations from the meeting sessions, splitting utterances by turns where a speaker ends its speech or is interrupted by another speaker. Then we annotated each utterance with dialogue act (DA) labels. To select DA labels we referred to Novielli & Strapparava (Novielli and Strapparava, 2010), who performed a dialogue act annotation on meetings transcriptions in Italian. We just added the label "laugh" to their label set. The final dialogue act label set we used is reported in Table 1. The agreement on the annotation of

label	description	example
Req	Questions	what's your name?
St	Statements	Today is sunny
Op	Opinions	I think that..
Agr	Acceptance	ok for me!
Rej	Rejection	no, thanks
In	Opening	hello!
End	Closing	goodbye!
Ans	Answers	My name is ..
Lau	Laughs	haha

Table 1: Dialogue act label set.

dialogue act labels between 2 non-expert labelers is $k = 0.595$ (Fleiss et al., 1981). This moderate agreement score, and the feedback from the annotators, indicate that the task is hard due to the presence of long and complex utterances.

We aligned the data from spoken conversations with public data from the participants' Facebook profiles. Using Facebook APIs, we collected data from 6 months before the meeting session to 1 year later. We collected public status updates, including text messages, links, pictures, and multimedia files posted and received on the participants' walls. We distinguished between statuses posted

metadata	description
fb-friends	number of friends
fb-pics	number of photos
fb-comm	avg number of comments received
fb-likes	avg number of likes received
fb-p-tot	count of all P's posts
fb-p-usr	posts by P on his/her wall
fb-p-oth	posts by others on P's wall
fb-shared	posts of the P shared by others
fb-text	count of textual posts
fb-media	count of non-textual posts
fb-chars	average characters in posts
fb-words	average words of posts

Table 2: Description of Facebook metadata collected.

by the users and statuses posted on the users' wall by others. Eventually we computed the numerical metadata reported in table 2 and we analyzed the textual pots.

We anonymized both the transcription and the Facebook data. The final corpus contains 2 audio files (one hour each) with transcriptions (about 21000 tokens and about 1600 utterances in total; 1750 words and 133 utterances on average per participant), and Facebook data of the participants (about 80000 tokens, about 5800 posts including multimedia status updates). We automatically annotated the textual data in the corpus with the Italian version of LIWC (Alparone et al., 2004). Doing so, we annotated words with 85 psychological dimensions, such as linguistic categories (verbs, prepositions, future tense, past tense, swears, etc.), psychological processes (anxiety, anger, feeling, cognitive mechanisms, etc.), and personal concerns (money, religion, leisure, TV, achievement, home, sleep. etc.). In the next section we report the results of the analysis of the data collected.

3 Experiments and Results

Scope From a communication analysis perspective, face to face meetings and Facebook are two very different settings: in Facebook the communication is written, asynchronous, mediated and with an audience that is a mix of friends and unknown people. On the contrary in face to face meetings the communication is oral, synchronous, not mediated, and the audience is unknown people. In a theory of communication (Shannon and Weaver, 1949), illustrated in Figure 1, all those levels are variables related to the sender, receiver and medium. Here we restrict the scope of this

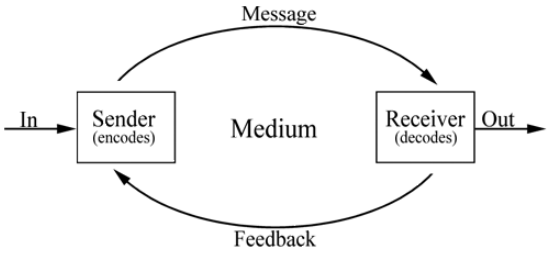

Figure 1: Schema of communication as transmission. We limit the scope of this work to the message level.

work to the analysis of message level, leaving to future work the possibility to extend this analysis to the characteristics of the media or the participants.

Experiments First of all we analyzed the topics in Facebook and meeting transcriptions. We removed the stopwords and we generated two word clouds with the 70 most frequent words in each dataset with 5 as minimum term frequency. We report the word clouds in Figure 2. The comparison of the two clouds reveal that participants

Figure 2: Word clouds of the 70 most frequent words in meeting transcriptions and Facebook data

to the experiments in Facebook discussed and planned actions ("*dormire*", "*andare*") places ("*rimini*", "*copenhagen*") and times ("*sera*", *stasera*", "*domani*") while in meetings they told and discussed mainly about places ("*bologna*", "*rimini*") and people ("*tipo*", "*gente*").

In order to discover relationships between psycholinguistic dimensions in Facebook and face to face meetings, we labelled the texts with LIWC, and we computed how much the psycholinguistic dimensions correlate in the two settings. We observed few, but strong, significative correlations (for significative we mean correlations with p-value smaller than 0.05 and correlation greater than 0.5), reported in table 3.

Word type (LIWC-it)	corr. to both settings
Anxiety	0,510***
Anger	0,580***
Feel	0,571***
Future	-0,532**
Home	-0,715*
TV	0,711*
sleep	0,537***
swears	0,696**

Table 3: Pearson's Correlations between LIWC dimensions in texts from Facebook profiles of the participants and face to face meeting. Only dimensions significantly correlating are reported. Significance is ***=p-value smaller than 0.001; **=p-value smaller than 0.01; *=p-value smaller than 0.05.

The dimensions with strong correlation are related to powerful emotions, difficult to control, like anxiety and anger, but also to the tendency to express feelings and emotions with words. Swears, that is the dimension with the highest combination of *correlation coefficient* and significance, is related as well to a dimension difficult to control. Maybe less interesting for our purposes are other dimensions with high correlations related to the content of discourse, like "home", "TV", "future" and "sleep". We ran automatic topic modeling with a Hierarchical Latent Dirichlet Allocation (Teh et al., 2006) (Blei et al., 2003) to reveal that participants spoke about "TV" and "sleep" in both settings, but about "home" and "future" only in Facebook and not in face to face meetings. This is why these values are negative.

We also compared behavioral data from Facebook and meetings. In particular we computed the correlations between Facebook metadata and dialogue acts annotated in meeting transcriptions, plus metadata from face to face meetings, namely the average length of utterances in words and characters. Results, reported in Table 4, show that

	f2f-req	f2f-st	f2f-op	f2f-agr	f2f-rej	f2f-in	f2f-end	f2f-ans	f2f-lau	f2f-words
fb-friends	0,243	0,130	-0,047	-0,298	-0,080	0,166	-0,475	-0,206	-0,063	-0,156
fb-pics	0,167	-0,157	0,281	-0,198	-0,410	-0,078	-0,253	0,163	-0,185	-0,084
fb-comm	0,439	-0,295	-0,003	0,464	-0,036	-0,287	0,297	-0,525	0,173	-0,064
fb-likes	**0,698***	-0,379	0,308	-0,276	-0,033	0,064	0,383	-0,230	-0,143	0,079
fb-p-tot	0,533	-0,078	-0,020	0,286	-0,117	-0,147	-0,240	-0,553	0,107	-0,135
fb-p-usr	0,140	-0,176	-0,297	0,230	0,174	0,311	-0,475	0,094	0,066	-0,157
fb-p-oth	-0,140	0,176	0,297	-0,230	-0,174	-0,311	0,475	-0,094	-0,066	0,157
fb-shared	-0,204	**0,698***	0,384	-0,352	-0,060	-0,206	-0,292	-0,155	-0,272	**0,619***
fb-text	-0,043	-0,096	-0,142	0,417	0,123	-0,336	0,427	-0,427	0,420	-0,100
fb-media	0,043	0,096	0,142	-0,417	-0,123	0,336	-0,427	0,427	-0,420	0,100
fb-chars	0,305	0,193	0,276	-0,042	-0,209	-0,475	0,269	-0,442	-0,161	0,309
fb-words	0,247	0,215	0,217	-0,005	-0,166	-0,453	0,275	-0,426	-0,124	0,283

Table 4: Pearson's correlations between metadata from Facebook and dialogue act labels from face to face meetings. *=p-value smaller than 0.05.

there are few, but very interesting, significative correlations. The number of likes received by the participants on Facebook correlate positively with a tendency to ask questions in meetings. This is quite surprising and perhaps reveals a will to engage the audience asking questions. Crucially, other significative correlations are related to shares generated in Facebook by the participants. In particular this is correlated with long statements in face to face meetings. In practice, people posting contents that are reshared online, in face to face meetings tend to produce long statements and talk more than the others.

4 Discussion and Conclusions

In this paper, we attempted to analyse the correlations between psycholinguistic dimensions observed in Facebook and face to face meetings. We found that the type of words significantly correlated to both settings are related to strong emotions (anger and anxiety), We suggest that these are linguistic dimensions difficult to control and tend to be constant in different settings. Crucially, we also found that likes received on Facebook are correlated to the tendency to ask questions in meetings. Literature on impression formation/management report that people with high self-esteem in meetings will elicit self-esteem enhancing reactions from others (Hass, 1981). This could explain the link between the tendency to ask questions in meetings with unknown people and the tendency to post contents that elicit likes in Facebook. Moreover, the tendency to ask questions in spoken conversations is correlated to observed emotional stability (Mairesse et al., 2007) and that emotionally stable users in Twitter tend to have more replies in conversations than neurotic users (Celli and Rossi, 2012). We suggest that the correlation we found can be partially explained by these two privious findings.

Another very interesting finding is that the tendency to be reshared on Facebook correlates to the tendency to speak a lot in face to face meetings. Again, literature about impression formation/management can explain this, because people with high self-esteem tend to engage people and to speak a lot, while people adopting defensive strategies tend to be assertive less argumentative. In linguistics it is an open debate whether virality depends from the influence of the source (Zaman et al., 2010) or the content of message being shared (Guerini et al., 2011) (Suh et al., 2010). In particular, the content that evokes high-arousal positive (amusement) or negative (anger or anxiety) emotions is more viral, while content that evokes low arousal emotions (sadness) is less viral (Berger and Milkman, 2012). Given that the tendency to express both positive and negative feelings and emotions in spoken conversations is a feature of extraversion (Mairesse et al., 2007), and that literature in psychology links the tendency to speak a lot to extraversion (Gill and Oberlander, 2002), observed neuroticism (Mairesse et al., 2007) and dominance (Bee et al., 2010). we suggest that the correlation between long turns in meetings and highly shared contents in Facebook may be due to extraversion, dominance and high self-esteem.

We are going to release the dataset we collected on demand.

Aknowledgements

We wish to thank the artist Valentina Perazzini for the contribution in the collection of data and Luca Rossi (University of Copenhagen) for the discussions.

References

Francesca R Alparone, S. Caso, A. Agosti, and A Rellini. 2004. The italian liwc2001 dictionary. Austin, TX: LIWC.net.

Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*.

Nikolaus Bee, Colin Pollock, Elisabeth André, and Marilyn Walker. 2010. Bossy or wimpy: expressing social dominance by combining gaze and linguistic behaviors. In *Intelligent Virtual Agents*, pages 265–271. Springer.

Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research*, 49(2):192–205.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Danah Boyd and Nicole Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Fabio Celli and Luca Polonio. 2013. Relationships between personality and interactions in facebook. In *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, pages 41–54. Nova Science Publishers, Inc.

Fabio Celli and Luca Rossi. 2012. The role of emotional stability in twitter conversations. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 10–17. Association for Computational Linguistics.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236.

Alastair Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 253–262. ACM.

Samuel D Gosling, Adam A Augustine, Simine Vazire, Nicholas Holtzman, and Sam Gaddis. 2011. Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, 14(9):483–488.

Marco Guerini, Carlo Strapparava, and Gözde Özbal. 2011. Exploring text virality in social networks. In *Proceedings of ICWSM*, pages 1–5.

Glen R Hass. 1981. Presentational strategies and the social expression of attitudes: Impression management within limits. *Impression management theory and social psychological research*, pages 127–146.

Samuel Kim, Fabio Valente, and Alessandro Vinciarelli. 2012. Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5089–5092. IEEE.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500.

Nicole Novielli and Carlo Strapparava. 2010. Exploring the lexical semantics of dialogue acts. *J Comput Linguist Appl*, 1(1-2):9–26.

Fabio Pianesi, Massimo Zancanaro, Bruno Lepri, and Alessandro Cappelletti. 2007. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3-4):409–429.

Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. 2008. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 180–185. IEEE.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Andrew H Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):773–791.

Gwendolyn Seidman. 2013. Self-presentation and belonging on facebook: How personality influences social media use and motivations. *Personality and Individual Differences*, 54(3):402–407.

Claude E Shannon and Warren Weaver. 1949. *The mathematical theory of communication*. University of Illinois press.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, pages 177–184. IEEE.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. 2010. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, pages 599–601.