ROBERTA FACCHINETTI
Department of Foreign Languages and Literatures
University of Verona
Verona, Italy
roberta.facchinetti@univr.it

English in Social Media: A Linguistic Analysis of Tweets

1. Aim and scope of the study

A number of recent studies have addressed issues concerning Computer Mediated Communication (henceforth CMC), either providing classification schemes for the features of computer-mediated discourse and of the grammar electronic language in particular (Herring 2007, 2012) or focussing on the stylistic diversity of Internet language (Crystal 2006) – also from a sociolinguistic perspective (Androutsopoulos 2011) – or again attempting a methodological reflection on the use of data in language-focused research on CMC (Androutsopoulos&Beißwenger 2008, Facchinetti 2013). However, the linguistic specificities of social networks have not been tackled in detail so far; this is particularly the case of the micro-blogging platform Twitter.

Twitter was launched in 2006 with the purpose for users to answer the question "What are you doing?". The original prompt was soon changed into "What's happening?", since its users exploited it more to report and comment on what was happening around them than to update on their status. Since then, it has been largely used as an information platform, and indeed it has been the main communication channel among Iranian protesters in the aftermath of the 2009 elections and one of the most exploited communicative tools for news-sharing in the 2011 Arab spring. As is well known, messages posted on Twitter are called *tweets* and are limited to at most 140 characters. Messages generally appear in reverse chronological order on the public timeline on Twitter.com, as well as on the individual user's Twitter page. As for the largest majority of social networks, the posts concern a wide range of topics and cover different modalities, from personal information to news, from pure verbiage to links to images, videos and sound.

Structurally, a tweet features two-parts: the first one provides information on the author's identity – namely the username, usually accompanied by a picture – and the message; the second one contains elements automatically produced by the software, technical data about the tweet – for example the time it was posted, its Internet origin, and response options, like *Reply*, *Retweet*, *Favorite*, and *More*, as in Figure 1:

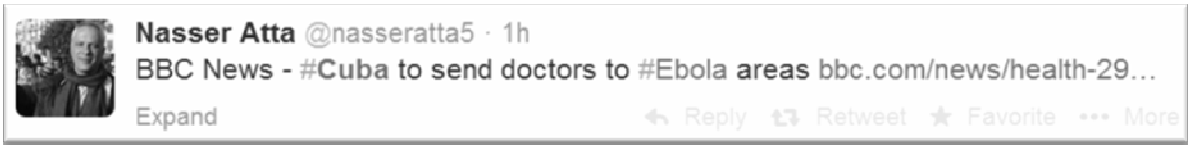

Fig.1: A typical tweet

The mark-up language of tweets as a means of conversation threading has been investigated largely with reference to the functions and uses of the @sign (Honeycutt & Herring 2009, to the practice of re-tweeting (boyd, Golder&Lotan 2010), and to the function of the hashtagZappavigna (2011); in contrast, the linguistic specificities of the message itself, has been the object of very little research so far. Zappavigna (2011) has explored how language is used to build community; Danescu-Niculescu-Mizil, Gamon&Dumais (2011) have studied linguistic style accommodation on the basis of a Twitter conversational dataset, while Hu, Talamadupula&Kambhampati (2013) have addressed the debate about the position of Twitter's language in the language spectrum of various well established CMC channels (like SMS, chats and blogs). However, Twitter language still remains largely unexplored, particularly with reference toits distinguishing linguistic features.

The present paper intends to bridge this gap and to investigate some linguistic traits that are potentially peculiar to the messages produced on the Twitter platform, bearing in mind that (a) the character limitation imposed by Twitter makes it a particularly interesting context of language use for researchers to delve into, and (b) since tweets are publicly accessible; this facilitates their circulation and consequently their language is likely to have an impact on linguistic practices, both online and offline. To carry out the analysis, a corpus has been specifically compiled, which will be illustrated in Section 2.[1]

2. The corpus

A corpus has been generated by selecting a number of *tweets* from *Twitter Search*, a search engine allowing access to tweets posted in real time and containing a search term. The engine also includes the facility 'Worldwide Trends', corresponding to the ten most trendy subjects of discussion on Twitter at the moment of the inquiry; they mostly consist of two- or three-word phrases or hashtags[2].

To compile the corpus, sample tweets were collected from the 'Worldwide Trends', ensuring the possibility of analysing naturally occurring public exchange, presumably among people from all over the world, using English

[1] The present paper will provide an overview of some prominent features detected in the corpus, while further detailed results are dealt with in Facchinetti&Caleffi (forth.); the corpus itself has been developed by PhD student Paola Maria Caleffi.
[2] The hashtag is the # symbol followed by one or more words, one after the other with no space in-between.

as their means of communication, irrespective of their L1. The samples were collected over a period of four months (March-June 2012) by selecting 'Worldwide Trends' at different times during the day and storing all the tweets that appeared in the stream over the following 5-10 minutes. Eleven trends were chosen:

Happy Birthday Twitter
Good Monday Morning
#TheySay
Gareth Barry
Elizabeth Tower
#WhatUniversityHasTaughtMe
Bieber OurBoyfriend
#BelieveItOrNot
Tony Blair
Tomorrow is June
Nap or Food

For each trend, 100 tweets were analysed, making a total of 1,100 tweets. All the samples of the parts automatically generated by the software were cleaned, as the interest was only on analysing only the message; the resulting data were then transformed into both *.docx* and *.txt* files. The final corpus is made up of 98,366 characters and 16,250 running words.

The linguistic analysis of the corpus was focussed on typography, orthography, morphology, syntax, and lexicon, with special attention to the following patterns:

- Abbreviations
- Capitalisation
- Emoticons
- Fragmented syntax
- Message length
- Neologisms
- New word formatives
- Non-alphabetic keyboard symbols
- Numbers and letters to substitute words
- Predications
- Punctuation
- Spellings imitating causal, dialectal pronunciation, prosody and non-linguistic sounds

3. Findings

Overall, the average length of a tweet in the corpus is 89.43 characters, just a bit more than half of the characters allowed by the system; this indicates that the limitation in the number of characters is not an issue for twitterers to convey their messages.

With specific reference to TYPOGRAPHICAL features, the most prominent aspect yielded by the corpus is the heterogeneity of elements appearing in tweets. Indeed, they are characterised by a mingling of non-alphabetic keyboard symbols, emoticons and sequences of keyboard characters imitating facial expressions, URLs, usernames (often containing numbers and letters), hashtags, unorthodox alternating of lower- and upper-case, and finally repeated punctuation marks and letters. A tweet may therefore look like examples (1-3):

(1) Bieber Our Boyfriend :"""""""""""">

(2) "@Mikhail010: #They Say love between best friends is wrong, but no one knows and cares about you like that person, so..why not?" our story baby

(3) Gareth Barry is OUT of the EURO'S... hhhmmmm not bothered

This mingling forces the linguist to manual analysis of every single tweet, since it is virtually impossible to exploit traditional software for corpus analysis. However, this heterogeneity of elements is not peculiar to Twitter; rather, it seems to be a consequence of the shift from "page to screen" (Snyder 1998) that has accompanied the rise of CMC. The brevity of tweets only makes this peculiarity more noticeable.

Emoticons in particular substitute elements of face-to-face interaction and express the writer's attitude towards both the topic and the addressee. Hence, they carry a pragmatic 'weight' by conveying convey personal opinions, feelings, and attitudes of the interlocutors, without the writer having to use too many characters. In the corpus, only a total of 323 emoticons were recorded, far fewer than expected; hence, emoticons do not appear to be a distinctive feature of the language used by twitterers, nor to play a major role in conveying the message.

With reference to ORTHOGRAPHY, although the findings provide examples of inconsistent or deviant use of capitalisation, on the whole data testify to the fact that the tweets in the corpus largely conform to the standards. Indeed, only 127 tweets (11.5%) show deviations from standard capitalisation rules, with only 3.6% of tweets beginning with the lower case. In contrast, with reference to punctuation, deviation from standard rules seems to be more noticeable. Indeed, up to 43.6%of tweets show some kind of punctuation un orthodoxies:(a) tweets written without any punctuation, (b) tweets where punctuation is used improperly, inconsistently or only at the end of the message, (c) erratic ellipsis dots, as in (4):

(4) damp monday morning! hoping for a good week ahead! Wish the same with you guise! ????

There are also cases of total omission of punctuation, which does not seem to be ascribable to the need to save characters, since, as previously mentioned, the average length of the tweets in the corpus is lower than the maximum number of characters allowed by the system. Moreover, the findings show that the shorter the tweet,

the higher the lack of punctuation marks. Once again, however, random or inconsistent punctuation is not unique to Twitter, being typical of other CMC settings too.

Moreover, spellings imitating casual or dialectal pronunciation are very limited – most of them being hapaxlegomena in the corpus. Abbreviations/misspellings of grammar words/phrases are also very limited (157 in total), the most frequent being the following:

- *u/ya*(= *you*)
- *ur* (= *your*)
- *gonna* (= *going to*)
- *your* (= *you are*)
- *-in* (instead of *-ing*)
- *cos/coz* (= *because*)
- omission of the standard apostrophe (as for *dont* instead of *don't*)

Again, abbreviations of content words/phrases and initialisms(like *bf* instead of *boyfriend*) are even more reduced in number and, with the exception of *lol*(= *loughing out loud*) and *omg* (= *oh my god*), they occur from a minimum of one to a maximum of five times each, as in (5):

(5) Bieber Our Boyfriend in just #5DAYS cant wait to be his boy friend causei know he wouldnt let me go. lol

These results related to orthography seem to indicate that character limitation does not affect the writer's linguistic choices in this respect. Once again, word shortening is by far more noticeable in text-messaging and, most notably, it occurs in other digital settings as well (Crystal 2006).

As for MORPHOLOGY, with the exception of the use of abbreviations, the corpus does not yield any distinctive new pattern of word formation, nor does it provide a significant number of neologisms. Only one example has been recorded of verb reduplication (*nudge nudge wink wink*). Furthermore, ten neologisms were recorded, one of which being the phrase *born day* for 'birthday' and the others obtained through blending (*jeliebers*,*beliebers*, *bieberconda*, *twitthearts*, *twitterverse*, *tweoples*, *tweet♥s*,*tweetiepies, fap*) as in (6) and (7):

(6) Good morning twitterverse happy monday!

(7) Bieber Our Boyfriend, well..even the bieberconda? ;D okay then

Moreover, the word *muchly* occurred once as instance of misuse of an inflectional suffix, along with one noun-to-verb conversion (*@justinbieber cannot TT,* TT normally being used as the abbreviation for 'trending topic'). Also with reference to morphology, then, in the corpus under scrutiny no prominent distinctive feature has been recorded that may identify the language used in Twitter messages as peculiar to this modality.

With reference to SYNTAX, tweets pose a few issues, first of all the heterogeneity of elements appearing within a tweet, along with the presence of non-clausal material (e.g. *Wow, yes, this*) and sentence fragments, as in (8):

(8) Wow, loving the positive vibes this Monday! Good Morning

Secondly, occurrences of unorthodox and inconsistent use of capitalisation and punctuation do not always allow for a clear identification of the syntactic nature of the writer's 'utterances'. In particular, omission or inconsistent use of such indicators as punctuation marks can blur sentence boundaries and hinder disambiguation. This proved to be a major problem when trying to collect data on the type and number of sentences a tweet is on average made up of, since tweets frequently displayed punctuation (and capitalisation) unorthodoxies, as in (9):

(9) *Tomorrow is June all ready for summer* or*idont understand people sometimes keeping it in is better then speaking out loud #BelieveItOrNot)*

Following Crystal's classification (Cystal 1988: 16-18), in the classification of sentences, the following were taken into consideration:

- minor sentences formulae
- words/phrases used as exclamations
- emotional noises
- initialisms
- elliptical constructions/verbless clauses
- non-clausal fragments in general.

The analysis yielded a total of 2,297 sentences, respectively 1,486 major (65%) and 811 minor (35%), meaning that the tweets in the corpus are composed of 2 sentences on average, mostly simple sentences and largely of the declarative type (73%).

A final element of syntax that was analysed is given by predications, namely words through which the writer wants to communicate an action (e.g.**shrugs**) or a state (e.g. **mind blown**) rather than say something, the corpus has yielded only seven examples (**shrugs**, **tumbletweed**, **exasperated voice, gentle glottal stop**, ***throws glitters & sunshine***, *~fangirling~*, **facepalm** and **mind=blown**). Once again, the findings do not provide enough evidence to claim that predications may be a typical feature of the language of *Twitter*.

LEXICALLY, occurrences of slang are limited throughout the corpus, each item appearing only once or up to six times (like *yay/yey*). Similarly, interjections (e.g. h*uzzah*; w*ow*) and written representations of sounds (e.g. *blah blahblah; huh?!*) are limited as well, most patterns also appearing only once, as in (10):

(10) huh?? #hotslap!!! RT@damolaade At least I know one in YOU RT"@moncherrygaga: **#**TheySay: most Nigerians are razz, true???

Finally, the tendency to use colloquial words seems to decrease when the topic or the subject being dealt with becomes more important or serious. However, overall, the extremely limited number of such colloquialisms does not allow any conclusions with reference to the degree of informality of language of Twitter.

4. Conclusions and further food for thought

The data testify to the fact that none of the features one might expect to be distinctive of Twitter due to the character limitation imposed by the medium seem to be prominent. In particular, abbreviations and other types of shortening techniques (like the use of emoticons, non-alphabetic symbols or numbers and letters to substitute words) are only occasionally employed. Moreover, twitterers do not make use of all the characters allowed by the system; indeed. Other features – poor legibility, inconsistent and erratic use of punctuation, phonetic spelling and spelling replacing prosody and non-linguistic sounds, as well as fragmented syntax – are listed in Herring's inventory as "the set of features that characterize the grammar of electronic language" (Herring 2012: 1) and therefore not Twitter-specific. Therefore, the corpus data suggest that the linguistic peculiarities of Twitter are one of the many alternatives bridging the gap between the two discrete poles of speech and writing, and a further confirmation of the hybrid nature of the language used in CMC.

The element that seems particularly distinctive of Twitter is brevity and conciseness; indeed, as has been observed in the corpus, twitterers use only a bit more than half of the 140 characters allowed by the system; though brevity is exhibited also by texts produced via text messaging and instant messaging, in those digital contexts it does not seem to be accompanied by the same degree of expressiveness.

From the methodological point of view, the data have cast light on the inadequacy of conventional tools for the linguistic analysis of the hybrid nature of CMC language, and on the challenge of handling the often heterogeneous and non-canonical linguistic forms emerging from social media, like the use of hash tags on Twitter.Specifically, in the first place, the present study highlights the difficulties of carrying out linguistic analysis of Twitter language and, possibly, of some types of CMC; indeed, in the first place, an investigation into the linguistic features of Twitter messages can hardly be carried out by means of traditional tools for linguistic investigation. Rather, many of the aspects selected require time-consuming manual analysis. Features of this kind include, among others, occurrences of types of emoticons and logograms, capitalisation, frequency of upper case, frequency of repeated punctuation marks, non-standard spellings and abbreviations, up to syntactic features like sentence type.

Secondly, it is difficult to accurately count the number of words constituting a corpus like the one exploited for the present analysis, since the 'word counter' application of Microsoft Office includes in the count all the 'disturbing' elements that appear in a tweet, such as emoticons, hashtags, @characters followed by a username, as well as URLs. In particular, hashtags are counted as one unit, irrespective of the number of words in the unit, and the same is for @usernames and URLs, although these are of no relevance for linguistic analysis.

Both content wise and methodologically, all the above-mentioned aspects need to be kept in consideration for further studies on CMC; indeed, far from being conclusive, the present results call for further and larger-scale research, in order to highlight not only the specificities of Twitter, but also its distinctiveness in relation to other social media, bearing in mind key aspects/variables like the writer, the addressee, and the aim of the message.

References

Androutsopoulos, J. (2011). "Language Change and Digital Media: A Review of Conceptions and Evidence". In Kristiansen T. &Coupland, N. (eds.) *Standard Languages and Language Standards in a Changing Europe.* Oslo: Novus Vorlag, pp. 145-161.

Androutsopoulos, J. &Beißwenger, M. (2008). "Introduction: Data and Methods in Computer-Mediated Discourse Analysis". *Language@Internet*, 5/2. <www.languageatinternet.org/articles/2008/1609>

Boyd, d., Golder, S. &Lotan, G. (2010). "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter". *Proceedings HICSS-43. IEEE: Kauai, HI, January 6,* pp. 1-10.

Crystal, D. (1988). *Rediscover Grammar*. London: Longman.

Crystal, D. (2006). *Language and the Internet* (2nd ed.). Cambridge: Cambridge University Press.

Danescu-Niculescu-Mizil, C.; Gamon, M. &Dumais, S. (2011). "Mark My Words! Linguistic style accommodation in Social Media". *Proceedings of the 20th International Conference on World Wide Web*, ACM, New York, USA, pp. 745-754.

FacchinettiR. (2013), Modal verbs in news-related blogs: When the blogger counts. In: Marin-Arrese, Juana I. / Carretero, Marta / ArúsHita, Jorge / van der Auwera, Johan (eds.), *English Modality: Core, Periphery and Evidentiality*, Berlin: Mouton de Gruyter, pp. 359-377;

Facchinetti R. and Paola Maria Caleffi (forth.) From English to Twenglish: A new language variety?, Proceedings of the International Conference on Language Variation and Change in Postcolonial Contexts, Fisciano, Salerno, Italy, 18-19 April 2013.

Herring, S.C. (2007). "A Faceted Classification Scheme for Computer-Mediated Discourse". Language@Internet, 4/1. <http://www.languageatinternet.org/articles/2007/761>

Herring, S.C. (2012) "Grammar and Electronic Communication". In Chapelle, C. (ed.) *Encyclopedia of Applied Linguistics*. Hoboken, NJ: Wiley-Blackwell Publishers. <http://ella.slis.indiana.edu/~herring/e-grammar.pdf>

Honeycutt, C. & Herring, S.C. (2009) "Beyond Microblogging: Conversation and Collaboration viaTwitter". *Proceedings of the Forty-Second Hawai'i International Conference on SystemSciences* (HICSS-42). Los Alamitos, CA: IEEE Press. <http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf>

Hu, Y.; Talamadupula, K. &Kambhampati, S. (2013). "*Dude. Srly*?: The Surprinsingly Formal Nature of Twitter's Language". *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*, Boston, USA.http://www.public.asu.edu/~yuhenghu/paper/icwsm13.pdf

Snyder, I. (ed.) (1998). *Page to Screen: Taking Literacy into the Electronic Era*. London; New York: Routledge.

Zappavigna, M. (2011). "Ambient Affiliation: A linguistic Perspective on Twitter". *New Media&Society,* 13/5: 788-806. <nms.sagepub.com/content/13/5/788.full.pdf+html>