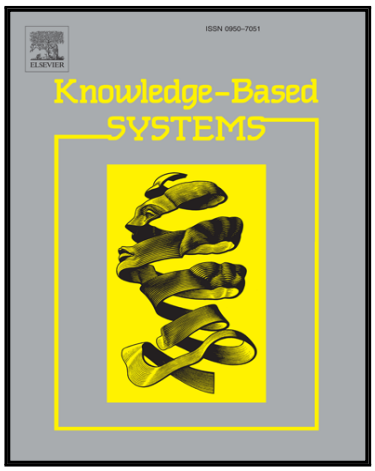
An automated system for grammatical analysis of Twitter messages. A learning task application

Oussalah, Mourad; Escallier, B.; Daher, D.

DOI:
10.1016/j.knosys.2016.02.015



An Automated System For Grammatical Analysis of Twitter Messages. A Learning Task Application

M. Oussalah , B. Escallier , D. Daher

PII: S0950-7051(16)00101-5
DOI: 10.1016/j.knosys.2016.02.015
Reference: KNOSYS 3435

To appear in: *Knowledge-Based Systems*

Received date: 13 April 2015
Revised date: 10 November 2015
Accepted date: 20 February 2016

Please cite this article as: M. Oussalah , B. Escallier , D. Daher , An Automated System For Grammatical Analysis of Twitter Messages. A Learning Task Application, *Knowledge-Based Systems* (2016), doi: 10.1016/j.knosys.2016.02.015

An Automated System For Grammatical Analysis of Twitter Messages. A Learning Task Application

M. Oussalah[1], B. Escallier, D. Daher

University of Birmingham, School of Electronics, Electrical and Computer Engineering
Edgbaston, B15 2TT, Birmingham, UK

Abstract:

This paper describes an educational study involving the use of Twitter as a way to enhance High School students' interaction while improving the linguistic quality of their messages. For this purpose, an interactive system has been developed for Twitter collection and analysis from grammatical perspective. The automated system involves a comprehensive data normalization phase, which allows us to identify any unknown token, and a grammatical analysis system. The latter makes use of a logical reasoning on bi-gram token representation as well as a simple rule-based reasoning in case of named-entity detection. The developed system allows the user to perform spatial, topic-based or identity-based search functionalities. Besides, the system generates interrupt to moderator (s) together with some statistical parameters related to user activity as soon as a linguistic inconsistency has been detected in order to take relevant course of actions. The automated system allows us to identify both the text normalization issues and the grammatical inconsistencies. The latter makes use of logical reasoning using bi-gram Wikipedia matching. A statistical analysis of tweet messages gathered from students that took part to this study has been carried out. Besides, the contribution of the peers to the improvement of the linguistic quality of users' messages has been quantified and investigated. The study demonstrates the interest of the participants to this new learning experience and evaluates the influence of the peers on their writing skills. Especially, the visibility and noticeability of Twitter messages to a large audience have been found to contribute widely to raise students' awareness about the linguistic quality of their messages. The study has also revealed the predominance of the slang language in their daily Twitter writings. Such abbreviations have shown to pose the greatest challenge for any automatic text analysis. Similarly named-entity identification and handling have also been shown to be very challenging, especially, given the nature of Twitter messages where capitalizing is often employed for emphasize as well.

Keywords: Data mining, Twitter, Social network, Learning.

1. Introduction

The boom in social networking sites in the recent decade has provided unprecedented amount of information to various users, especially, teenagers and student community. This, in turn, has influenced and challenged the standards approaches to learning. Among these social network tools, one shall focus on Twitter [6, 17], which, since its introduction in 2007 has gained more than half billion active users and more than two billions tweet searches made every day! Typically, each tweet is made of up to 140 characters, which can also be embedded with precise geo-location information to carry short conversations [25]. This microblogging service has been found particularly useful in disaster monitoring, opinion influence as demonstrated in events like the 2009 US Airways Flight 1549 incident, the 2009 Iranian presidential election, Japan's earthquake/Tsunami, recent Arab uprising [27] and happiness quantification [30]. In all previous examples, the efficiency of Twitter as a way to convey real time information about the status of the underlying scenes, and, possibly, influence the

[1] Corresponding author: M.Oussalah@bham.ac.uk ; Tel:+44-121-4143128; Fax: +44-121-414-4291

public opinion, was clearly demonstrated. A key feature of Twitter is its open access, such that whenever a user enables unrestricted access, his/her tweets can appear on the public timeline, a running stream of tweets by registered users observable by anyone [18]. Especially, this offers the users the possibility to follow other (more popular) users, e.g., stars, corporates, which, in turn, enable the users to get the last update information regarding key events or news that are of interest to the user. Nevertheless, the tweet size restriction of 140 characters, which becomes even smaller when markup syntax and URLs were used together with high frequency of tweet messages sent by active users, some of which were only spam, renders the usage of correct English very difficult and very challenging. On the other hand, the size restriction substantially increases the usage of abbreviations and slang words in tweet messages, some of which are ill-known by the community. This triggered an open debate whether Twitter can be of any benefit to learning community from linguistic perspective. Independent journalist Sirucek [33] claimed that Twitter is where "grammar comes to die". Indeed, many argued that the size restriction caused users to sacrifice normative grammar to communicate tweets as a sign of language degradation. In the same context, Borau et al. [8] found that the character limit and dictionary usage limited the use of communication strategies, thus, concluding Twitter does not help in building strategic competence. Typically, shortening message is achieved through intensive use of abbreviations, phonetic substitutions, deletion of selected words and characters, which obviously can seriously hamper the understandability of the content of the message. Studies such as [32] concluded that the high noise level in the content of Twitter messages renders the use of any standard natural language processing technique for content mining pretty challenging and almost impossible as they contain highly non-standard orthography that would make commonly employed automatic text processing tools nearly void. However such opinion is also counterbalanced by many other findings, which expressed strong positive influence of Twitter on learning task. Arguments for such trend are well founded and seem very intuitive. First, a general scepticism about new technology was often raised by some linguistic scholars who initially claimed, for instance, that telegraph, next phone messages would kill correct English, but none of these did happen [12]. Second, as a communication tool which can generate growth and value, traditional business writers have already accommodated the inherent restriction to send clear and well-understood messages to (potential) customers. Third, with the advances in e-learning technology and related academic programs, the role of (online) recommended system is highly stressed [29]. Therefore interaction tools like Twitter would potentially be beneficial. Fourth, since the importance of peers-interaction in learning cannot be ignored and given the role of social network, including Twitter, in peers' formation and interaction, the impact of Twitter in students' academic performance is straightforward [11]. Indeed, peer networks have been demonstrated to promote prosocial behaviors, such as extracurricular participation and leadership [10]. Fifth, although it has been acknowledged that the research into the applicability of microblogging like Twitter in the language classroom is currently in its embryonic phase, its role in improving English for non-English speakers has been reported by many studies, see for instance [1] and references therein. Especially, it has been reported that given the short encoded messages, learners can follow daily conversations of native speakers as well as their peers, and engage in the underlying learning process. Antenos-Conforti [4] and Godwin-Jones [19] reported that social networking systems create new opportunities for teaching and learning given their ability to instantly engage the whole learning community. In this respect, one shall also mention the popular experience of Brazilian Red Balloon English School where pupils learned Grammar through identifying and correcting mistakes of their stars' (followers') tweet messages [9]. Another study performed by Grosseck and Holotescu [20] claimed that Twitter is the best place to practice a wide variety of expressions and fixed phrases. In the same spirit, Antenos-Conforti [4] argued that Twitter is effective because of two main reasons. First, a single tweet can trigger communication between socially connected users. Second, Twitter lowers affective filters so that any intentional thought can be transmitted. Ulrich et al. [36] analyzed student interaction using a microblogging network designed for English language learning in a Chinese university. The authors found that students tended to interact with those of the same gender, to self-initiate replies to tweets, and to favor public communication. In his book on the use of Twitter for Japanese learners, Homma [24] identified four reasons why Twitter is a good platform to study English. First, Twitters systems appears to be intuitive because of its simplicity and easy classification capability through hash-tags for instance, which provides a sound overall picture, as well as the possibility of initiating new communication based on previous state. Second, it increases users' motivation for participation to convey their daily desires and needs. Third, the size restriction makes it easy to follow up and replying accordingly as it takes only few seconds to do so. Fourth, it presents a visual medium analogy in the sense that after using Twitter for a while, people can always look back to what they have written, and see the extent of their improvements, if any.

Strictly speaking, the previous review indicates that the large extent of the studies involved in the use of Twitter for English studies have been primarily performed with non-native English speakers, although *bad* writing is found to be occurring with native English speakers as well. This partly motivates our current study to design and implement an educational case study involving Twitter with reasonably good English speakers. Besides, given the public nature of Twitter, our goal is to enlarge the domain of interaction of the learners beyond the school or classroom boundaries. For this purpose, an automated system has been designed to capture the geolocated tweets

in the West Midland area, including the tweets submitted by the users not involved in this study, and automatically query the tweet messages for incorrect grammatical constructions and word inconsistencies. More specifically, in the light of our previous work [31], an open architecture allowing us to collect geolocated tweets together with all student participants is put forward. Next, standard natural language processing is extended through the use of a comprehensive normalization stage, while a bigram-representation model in conjunction with corpus matching were employed for grammatical analysis. Finally, the contribution of the peers' interaction in the learning process is quantified and discussed. Typically, the main research questions that we aim to investigate throughout this study are:

i) Will students perceive Twitter as beneficial for practice, proceduralization and memorization of new grammatical constructions?
ii) Does Twitter encourage new forms of linguistic constructions and language distortion?
iii) How does interaction among learners take place?
iv) How to design the software architecture to deal with grammatical analysis of the tweet messages.

Section 2 of this paper describes the experimental setup highlighting the participant population and the design experiment. Section 3 emphasizes the Twitter collection software. Next, basis of textual analysis system and text normalization are described in Section 4. Section 5 highlights the grammatical analysis system, while results in terms of data statistics, users' interaction and related discussions are reported in Section 6.

2. Experimental Setup

2.1 Participants

The sample for this study consists of high school student population in West Midland area. We initially targeted schools with good achievement records in order to maximise our chance of recruiting good students. We meant by *good* students those who already master basic rules of English (normative) grammar and willing to positively collaborate in this study. The participants are selected on voluntary basis through a general questionnaire sent via Twitter, school internal notice board announcements, and link through Birmingham City Council. The students were asked to communicate the detail of their twitter accounts (usernames, backup email address), and requiring them to activate the public status and the geolocation information in their Twitter profile. On other hand, other potential participants have also been identified by our automated system through querying users' profile and picking up those which contain high school information in their profiles, so that a (personal) tweet message will be sent. No specific guidelines have been provided to participants except encouraging their active involvement and make more than 10 tweets per (week) day on average, either through replying to existing requests or trends (discussions) or initiating a new topic of interest. The students are also encouraged to identify and follow key organizations that deliver messages related to their core interests. This includes mainly leisure activities, local sport news and school activities. The list of the main followers employed by the students is also shared among all participants so that anyone can pick up the same followers if he/she wishes to. Besides, pointers to online resources about the Twitter manipulations have also been provided to students. Two teachers have also been volunteered to interact when needed.

In total, we quoted 133 students who took part to this study. About 70 % of these students were considered as positively involved in the sense that they generate more than 10 tweets per day on average for the period of April-June 2012 (around 10 weeks), which includes Easter break where student activity is pretty low. Students came from different ethnic backgrounds including Indian, Chinese and Pakistanis, although no assumptions were made regarding the performance of various ethnic backgrounds.

One shall also mention that many of the students, about 54, have only decided to participate in response to our query sent through Birmingham City Council, and no face to face meeting has taken place to meet up the participants. A short questionnaire has also been communicated to the students to describe their feeling and appreciations after this experience. In short, most students expressed their enthusiasms into discovering the various features of Twitter and enjoying the online interaction with their peers as well as making new friends as will be detailed later on.

Although the participants have been encouraged, if using their smartphones, to activate the geolocation tag in order to allow our system to capture the tweets using bounding box queries that will be explained later on, not all students were endowed with active GPS enabled devices as several users do use their laptop or PC to send their tweets. Consequently, in such cases, the tweets are only captured using the known user identity of the participants. A total of 73256 tweets were generated by the participants for this study.

2.2 Design experiment

Aim of the study

The study aims to design an automated system that will analyse the linguistic quality of the student tweet messages and to reflect on the learning process of the student through enforcing their interactions with their peers and the moderators. Especially, the users were asked to reply to individual queries as well as comment on a specific discussion trend or initiate a new stream of discussion according to his/her interests in subjects related to leisure, sport or education.

Intervention

When needed, the volunteered instructors and/or moderators (authors) intervene to guide or motivate the users by sending them stimulating tweets. We also tried deliberately to avoid commenting on issues related to alcohol, smoking, dating in order to avoid hurting personal feelings or opinions. Most of the users do use their smartphones for writing up their tweets.

Instances of tweets used for stimulating discussion include:

(i) @ Villa match today was one of the worst I've seen
(ii) @ Do u like yesterday Allison Janney tweet about her future?
(iii) @ Anyone fancy online gaming this afternoon?
(iv) @ Anyone seen big brother last night? Missed it.
(v) @ Can u rephrase ur msg pls?
(vi) @ Can you guess how many tweets followed ur discussion last nigt? Etc

The above examples are motivated as follows. The first instance (i) is an attempt to match users' interest in football. The second one encourages the users to go into celebrity webpages and check the tweet he/she sent and, possibly, influence the decision to add it as part of his/her followers. Tweet (iii) enhances one-to-one interaction through gamming. Tweet (iv) encourages the students to write descriptive sentences to highlight the key events. The example (v) attempts to persuade the student to check the grammar and construction of the sentence (s). Finally, the last tweet motivates the user to use Twitter platform to learn how to count number of users following discussion and so on. Figure 1 provides an overview of the types of interactions endorsed by this experiment. In essence, the students interact with each others by commenting on generic trend/news from public tweets by tracking relevant topics (Hashtags), replying to individual request of their peers or moderator, or initiating a new discussion trend. The outcome in terms of student tweets generated through such interaction is inputted to the automated textual analysis system that will be described in Section 4 in order to assess the linguistic quality and, at some extent, reflect on the learning process.

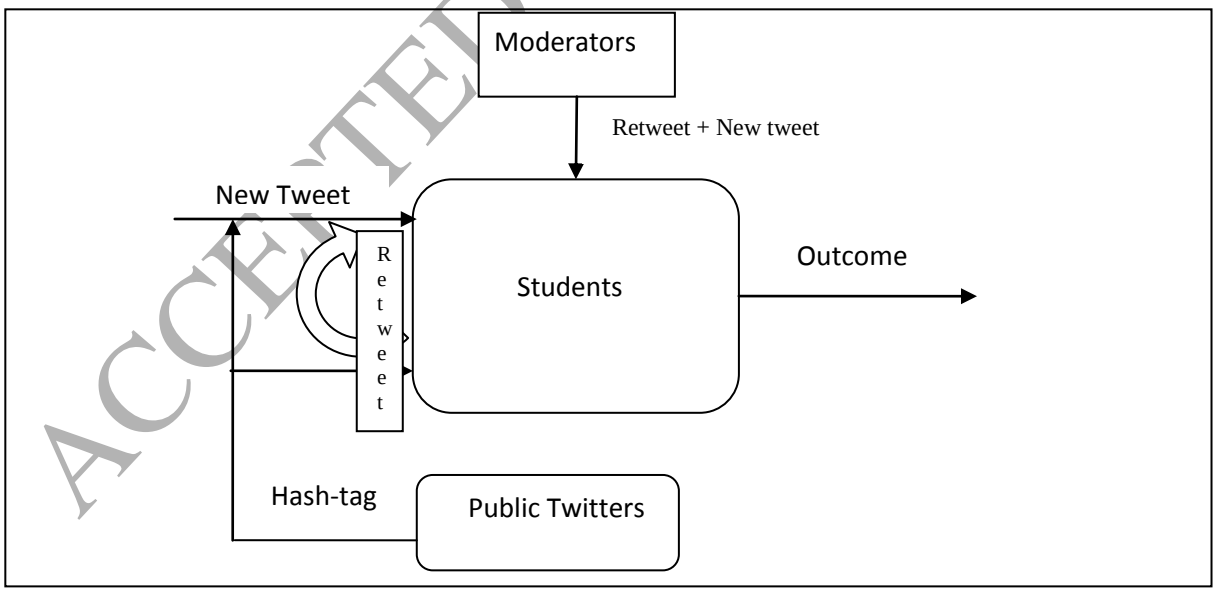


Figure 1. Overall of Students' interaction for Twitter generation

Method

The generic method to achieve the aforementioned aim of this study involves both design of automated system for tweet message analysis from linguistic perspective that could trigger action from moderator to reinforce interaction and statistical analysis of the inconsistencies, interaction and peers' influence. Figure 2 summarizes the key-stages of such analysis that will be detailed in Section 4 and Section 5.

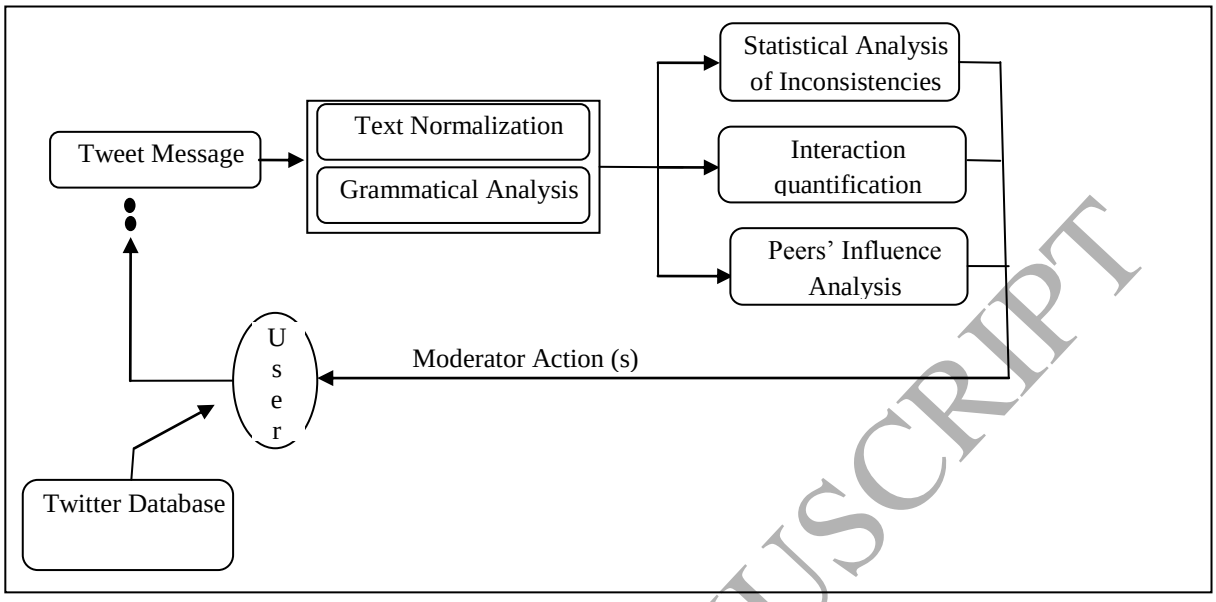


Figure 2. Overview of the Twitter-based analysis method

3. Data Collection Software

The developed software is built on our previous work in designing and implementing software architectures for Twitter collection and automatic analysis [31], with the difference that the current architecture also tracks a set of users constituted of the students (participants) based on their Twitter Ids regardless whether they were issuing geolocated tweets or not. More specifically, the data collection software is based on the following principles:

- Use of Django environment as a web platform connecting with Twitter Streaming API to retrieve Twitter data. The system makes use of the open-source JDBC driver to store tweets received from the Streaming API in a structured SQLite database. This enables us to capture all tweets originated from west midland area, provided the geolocation feature is activated by the users.
- Use of bounding box technique where the geometrical area over which tweets are collected is described using the latitude/longitude coordinates of the two adjacent points delimiting west midland area. In addition, the set of Twitter Identities for students participating to the study were manually entered into the system to enable the software to track their tweets regardless their actual locations.
- Integration of spatial queries into the developed software using UK Ordinance Survey Code-Point Open and 1:50k scale gazetteer libraries [15], which allows the system to query using postcodes and named entities.
- WordNet lexical database [16] was employed to derive semantically equivalent wordings, which are then integrated into the search mechanism of the query-system.
- Structuring the database such that it contains most of Twitter attributes including tweet message, user identity, screen-name, date, in-reply-to, accountancy information on number of followers and friends, number of tweets generated by user, location information, etc.

Especially, it is possible to collect tweets using either keywords based search, bounding box description of the area, city name, postcode, user Id, or any combination of the above. Users can specify whether the retrieval results should be displayed on map (using the maps.html template discussed later) or exported to a CSV format.

The running software has been tested during the period of April 2012 to June 2012, which includes the Easter break holiday where only a tiny students' activity has been recorded as well as beginning of exam period. In total, more than 300,000 tweets have been collected among which around 42,000 were originated from students. It should be noted that the ratio of total number of tweets (in west midlands area) to the student population does not provide a full picture of the total number of generated tweets in the region. Indeed, this is widely explained

by the fact that the system only captures geolocated tweets, which in only a tiny fraction (less than 4%) of total number of tweets generated by Twitter users.

4. Textual Analysis and Normalization

4.1 Difficulty and challenges

From a grammatical perspective, it should be noted that most of smart phones as well as web applications that can be used to send tweet messages do have an automated spellchecker to suggest spelling corrections to the user. This suggests that the orthographical and/or grammatical errors in tweets are rather intentional errors. Nevertheless, one of the most important sources of errors consists in the shortening of the words so that the new word cannot be found in any standard English dictionary, but, often, such shortening form becomes universally known in Twitter community. This includes for instance *u* for *you*, *2moro* for *tomorrow*, *y* for *why*, *pls* for *please*, *omg* for *oh my god*, *msg* for *message*, *ur* for *your*, etc. Although, one notices the wide spread of many commonly employed shortening words, there is also emerging tendency from the students to create their own abbreviations through cutting words or using smiling, which makes the message sometimes confusing. It is therefore worth investigating the extent to which the shortening words generated by the students deviate from the commonly employed abbreviations.

4.2 Principle of textual analysis and normalization

The textual analysis involves natural language processing of the messages in order to identify language distortion. First, one shall mention that special attention has been focused on word shortening language. Strictly speaking, the word shortening is not typical to Twitter but goes back to SMS (short message services) employed since the introduction of mobile phone for texting, but later on, this has also been expanded to many other internet based communication such as email, chat rooms and instant messaging. This has led to development of language for SMS as pointed out by Bodomo [7]. Many dictionaries/glossaries for SMS text messages have emerged. This includes Vodacom [37], NetSpeak [12], Lingo2Word[2], webopedia[3], among others, which offer a wide range of list of acronyms with their associated meaning in Standard English.

In this course, our textual analysis takes into account this new SMS language, which is then accommodated into the developed system. More specifically, one uses the Netlingo[4], which contains one of the largest lists of acronyms. Nevertheless, an important part of the textual analysis is dedicated to text normalization and data cleaning in order to bring the text into an acceptable level of semantic meaning (see, e.g., [26] for overview of challenges in Twitter normalization). For this purpose, several commonly occurring miss-conceptions were handled. This includes, for instance:

- -Ill placed quotations, which often induce error when attempting to identify a named entity, composed term or phrase. For instance, “today is payday!! :\, 2099485, “DavidC” where :\, would need to be searched and replaced with :\” (missing closing or opening quotation marks), in order to bring the text file to a cleaner state.
- - An escape character placed in front of quotation marks wrongly leads to nulling the real closing quotation mark. Consequently cautious is required when using automated tokenizers to deal with quotation marks as this may yield substantial loose of information conveyed by the text message.
- - As the URLs tend to get split up and transformed by some Twitter related applications, a cautious attitude is to remove them by the application, which would avoid further processing.
- - Some characters, e.g., “@” and “#” convey special semantic meaning in tweet messages, therefore special care should be given to the handling process of messages containing such characters.

From an implementation perspective, given its high indexing capability as well as advanced search functionalities, the open source Apache Lucene [23] has been used as a basis for the normalization task, where several integrated tokenizers have been employed and accommodated in order to meet the constraints Twitter. More formally, the following summarizes the key amendments:

- - UAX29URLEmailTokenizer already implemented in Lucene system was used to identify and remove the URL terms, which includes all http, ftp, https, ::// schemes, hostnames with registered top level domain, e.g., “.com”, IPv4 and IPv6 address format as well as email addresses. This used word boundary rules from Unicode standard annex UAX#29. The reason why this preprocessing should come first is that any

[2] http://www.lingo2word.com
[3] http://www.webopedia.com/quick_ref/textmessageabbreviations.asp
[4] http://www.netlingo.com/acronyms.php

standard application of word tokenizer prior to URL extraction would lead to URL being split down, which, in turn, renders its identification almost impossible.

- LetterTokenizer was next employed to extract all individual words, referred to as tokens, from the text message where all non-letter characters were discarded. More specifically, the letter tokenizer defined tokens as a maximal string of adjacent letters where a token is deemed to be separated by a space on its left and its right hand sides.
- StopFilter was employed to discard all common words by comparing each to the stop word list. The latter differed from the Lucene default list in order to avoid possible conflict with the (potential) abbreviation list. The use of such filter ultimately reduced the computational complexity as smaller number of tokens will be looked at in dictionary. This also allowed us to handle the excessive use of punctuation marks in tweet messages often employed to attract user attention.
- KStem stemmer was employed in order to recover the elemental root of the word. For instance, *animals* is converted into *animal*, *connected*, *connecting* are converted into *connect*. This was motivated by the fact that in standard dictionary, often only elemental words are present.
- ASCII Folding Filter was used in order to convert alphabetic, numeric, and symbolic Unicode characters which are not in the first 127 ASCII characters (the "Basic Latin" Unicode block) into their ASCII equivalents, if one exists. However, only those characters with reasonable ASCII alternatives were actually converted in order to enable other subsequent filters to work better with the few non English tweets that managed to get into the database and for any unconventional characters used in tweets.
- A database containing both the English corpus and WordNet set of synsets was created and linked to Lucene system. The outcomes of the KStem are then matched to entries of the database. This operation is instantaneous because of the structure of MySQL database and efficiency of the search operations in Lucene. If a match was found, the search operation terminates.
- If in the above operation, no match were found neither in English corpus, nor in WordNet database, the token was then matched against the Netlingo list of abbreviations. If a match was found, then the abbreviation was substituted by its corresponding interpretation.
- If Netlingo matching did not yield any positive outcome, then further examinations will be looked at. First the excessive repetition of letters, usually three or more identical characters, in a single token (word) was identified and removed. Indeed, in tweet messages, emphasize is often performed using repetition of letters as in “Goshhh”. To correct for such occurrences, misspelled words that contained repeated sequential letters were substituted by corresponding word containing only one single occurrence of the repeated character. Then the new word was matched again to the above databases and abbreviation list. Second, the occurrence of character “@” or “#” initiated new reasoning because of its special interpretation in Twitter.
- If “@ username” was found, this typically corresponds to a retweet to a Twitter user defined by his/her username. For instance, “@Joe, I’ll do it later” indicates a message “I’ll do it later” sent to user “Joe”. Besides, its removal would not alter the syntactic meaning of the message, which entails removal of expression “@Joe”. However, when the syntactic meaning is compromised by the removal, as in the tweet message “I will go out with @Joe”, the underlying expression is kept unchanged. Kauffmann (2010) pointed out that if the @username is at the beginning of the tweet, then only the subsequent terms can be used in this analysis. If the @username is followed by a word that is a coordinating conjunction, subordinating conjunction, preposition, or a verb, it is necessary to keep the @username in the tweet. If it is not the first word, then the part of speech of words on both sides can be used to help disambiguate @username. Similar reasoning applies to “# tag” (hash-tag) where the tag usually corresponds to the topic to which the tweet pertains to. However, sometimes Twitter users employ the symbol “#” for emphasize as well, so making the decision to delete the expression systematically in the tweet message is not always rational since this may alter the syntactic meaning of the message. Indeed common observations indicate that the use of hashtag terms in the middle of the sentence is usually meant to stress on their importance to the syntax of the sentence, and, therefore, should not be deleted. While hashtags at the beginning or at the end of tweet messages are usually much more difficult to disambiguate.
- If a given token still cannot be matched despite the previous reasoning, then we look at presence of capital letter at the start of the word. Its presence would be interpreted as a named entity, so the token was ascribed to the list of nouns. Otherwise, an anomaly was trigged and a message will be generated to the moderators to take any relevant actions.

The use of previous cleaning and normalization strategies allowed us to generate semantically clean phrases. Still the syntax and grammar need to be checked upon. This is the task of the automatic grammar checker that will be described in the next section. A generic diagram summarizing the main steps in tweet normalization with key tools employed is highlighted in Figure 3.

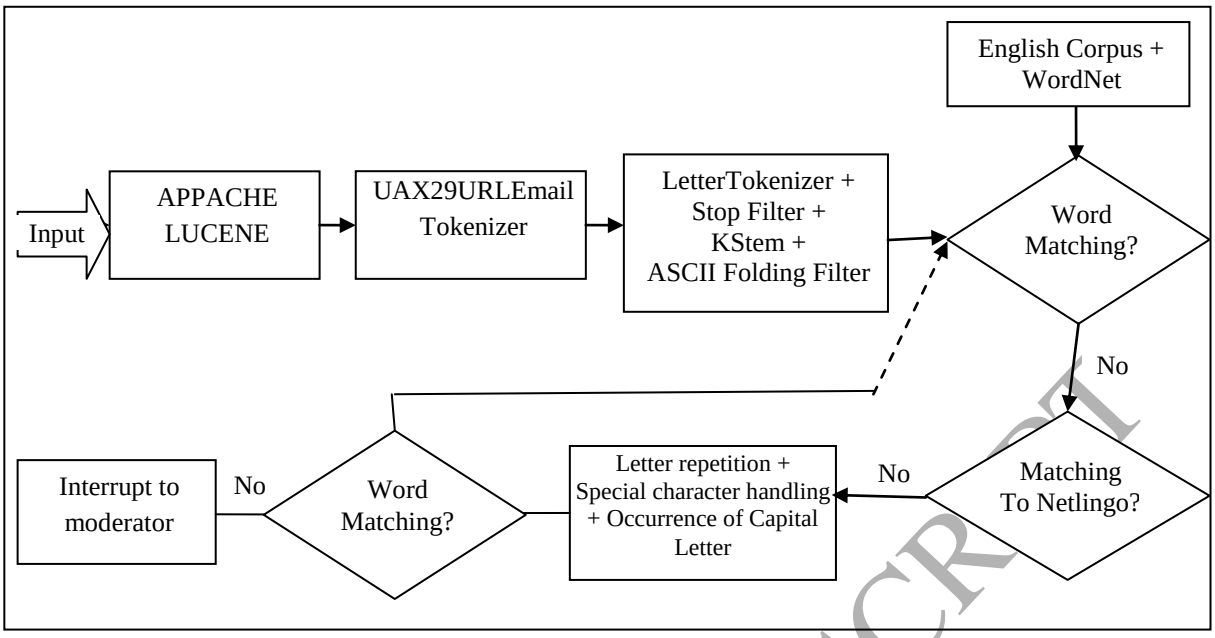


Figure 3. Generic diagram of text normalization

5. Grammatical analysis

5.1. Introduction

Grammar checking is the task of verification of the syntax and morphology of a given sentence according to the used language. Although such task is constantly performed by users or students when writing an essay or any message, designing an automated system to do so is not straightforward given the complexity of linguistic structures that may occur in natural language. Several research directions have been investigated to address the issue of automated grammar checker. This includes rule based approach that checks, for instance, the form of the verbs that follow nouns, see [28], for a state of the art related to studies focused on language learning. This approach was also the essence of the popular grammar checker of Microsoft word document as pointed out by Dolan et al. [14]. However such approach has also been proven to be quite limited as it fails to identify meaningless sentences as in "I went glass". Pattern recognition like approach has also been suggested for the task of automatic grammar checking [28] where a database is built of commonly employed mistakes, and then each sentence, or part of it, is matched to each entry of such database. Again, the approach may work well if the sentences of the document can be correctly and uniquely matched to an entry of the database. However, in the vast majority of cases, exact matching is not held. A third class of approaches that emerged in the last decade or so is related to the use of corpus statistics where the text is matched against a large database of correct wording (e.g., Wikepedia, English corpus, Google books n-gram corpus). Studies, such as those by Han et al. [21], for instance, proved the superiority of such approach over standard rule-based systems in the task of detecting article-related errors. Yin et al., [39], Whitelaw et al. [38] used the web as a corpus in conjunction to various degrees of processing of the input text, such as lemmatisation, POS-tagging and chunking. In this context, there is a growing interest in use of n-gram tagging decomposition where a set of n words are checked in the corpus. For instance given a sentence "He is ill today", which contains four-gram, then if a trigram grammar checker was used, the system will check for all consecutive trigram in the above sentence; namely, the expressions "He is ill" and "is ill today". The above two expressions are then matched to the corpus in the statistical sense as will be detailed later on. The relative success of statistical approach motivated us to adopt such method in the current study.

5.2 Methodology

The use of n-gram based methodology involves at least three main phases:

- Part of Speech Tagging. This consists in assigning appropriate word class and morphosyntactic features to each token in text, e.g., identifying nouns, verbs, adverbs, adjectives, pronoun, etc. Several open sources taggers are nowadays available in natural language processing community. In our system, we employed

Standford Tagger [35], due to its proven efficiency and portability to other languages. This is especially relevant to handle situations where a token may have multiple facets.

- Choice of corpus. Wikipedia (English) was used as the main source of corpus for our analysis. This contains more than four millions of structured articles. Although the size of such database is pretty huge, the easy indexing and availability of utilizing tools makes the search on such database pretty straightforward.
- Mathematical analysis. Statistical approach employing n-gram feature representation based on the assumption that the correctness of a sentence can be derived just from probabilities of all n-grams in the sentence was used as the main approach for our methodology. Nevertheless, the presence of unknown named-entity triggers (grammatical) rule-based approach. The letter is restricted to the form of the two tokens that precede and follow the underlying named entity.

In order to exemplify the above reasoning, let us consider a simple sentence "Anyone fancy going out tonight?" and let us consider a bigram feature. In this respect, the correctness of the above sentence is evaluated as

P("Anyone fancy going out tonight?") = P(Anyone|<start>).P(fancy|Anyone).P(going|fancy).
P(out|going).P(tonight|out).P(?|tonight)

To estimate the first single probability, e.g., probability that "Anyone" is a start of a sentence, one requires to look at corpus and find out the frequency of occurrence of sentences beginning with "Anyone". Similarly, the second probability P(fancy| Anyone) requires us to search in the corpus for those sentences containing the (ordered) pair (Anyone, fancy), e.g., the token "anyone" is immediately followed by "fancy", occurring in the corpus. Therefore, the occurrence of each pair involved in the above probabilities enables us the calculus of individual probabilities.

Strictly speaking, the calculus of individual probabilities from the corpus can be conducted in two distinct ways:

- Logical quantification. In this case, the probability takes 0-1 value where, for instance, P(Anyone|<start>) will be assigned a value 1 as soon as there is at least one sentence in the corpus which starts with token "Anyone". Similarly, P(fancy|Anyone) will be set to one as soon as there is one sentence in the corpus which contains the composed expression "Anyone fancy". However, whenever one of the corresponding pairs of words cannot be found in the corpus, the underlying probability will be equal to zero. Consequently, the overall probability of the whole sentence is set to 1 if all pairs can be found in the corpus, otherwise, the overall probability is set to zero. This agrees with the logical viewpoint where the interest is to find out whether the underlying sentence is true or not in view of information available from corpus.
- Frequency based quantification. In contrast to the first evaluation, the frequency of the occurrence of the pair in the corpus will be taken into account in the probability eliciting process. For instance, P(Anyone|<start>) is given by the ratio of the number of sentences in the corpus starting with "Anyone" over the total number of sentences in the corpus. This approach has been implemented by most researchers employing n-grams related method. So, the overall probability is compared against a fixed threshold beyond which the sentence is deemed to be grammatically correct. Nevertheless, it is clear that given the size of the corpus in terms of number of sentences, this often induces relatively small probability values, which, in turn, substantially affects the overall probability. This makes the choice of the above threshold very debatable. Although, it is always intuitively valid to stipulate that the overall sentence is deemed to be grammatically correct in the sense of existence of related constructions from the corpus as soon as the overall probability is strictly greater than zero regardless the value of the threshold. On the other hand, the choice of the threshold can also be employed to enquiry the familiarity of the sentence construction in the sense that commonly employed expressions trivially gain higher probabilities. A possible remedy to low probability values attached to bigram probability evaluations consists in the use of a ratio of number of sentences containing the underlying pair of tokens over all sentences containing the two words of the tokens in whatever disposition. For instance, P(Anyone| <start>) will be evaluated as the ratio of the number of sentences in corpus that start with "Anyone" over the total number of sentences in the corpus that contain the word "Anyone" regardless whether it is at the beginning, end or somewhere in the middle of the sentence.

Given the nature of the requests of our developed system which attempts to determine whether a given phrase /sentence is correct from both syntax and grammatical viewpoints, without generating a possible correction as

this is left to moderator to decide on appropriate course of action, the logical representation sounds more appropriate.

From an implementation perspective, one should notice the following. First, all bigrams generated by Wikipedia (English) are stored in MySQL database because of its proven efficiency and reliability. Besides, it allows us to use multiple clusters where parts of database are stored in distinct machines, which, in turn, improves the speed considerably. Second, the original sentences of normalized text carried out in the first phase (normalization stage) are restored and constitute part of the input to the grammar checker system. Strictly speaking, the sentence restorer module discards those phrases which are deemed to be incomplete sentences. This includes, for instance, those sentences, which after the normalization stage yield one single token. This would excludes many text messages, which are only formed of very short messages, e.g., "Hello", "OK", "love", "sleep well", "look at www.bham.ac.uk", "It is September 11th ", "It connects new connection" from passing to grammatical module as after normalization (URL and stop-word removals, stemming) stage, such phrases get reduced to a single token or none. Third, Standford Tagger was employed to identify the part of speech of each token. This allowed us to overcome situations in which the same word, although correct, can have multiple facets in terms of part of speech (e.g., the same word is used both as noun and verb). Notice that the case of an unknown token, except named-entity, is completely discarded as this should have already been detected in the normalization stage as pointed out in previous section, which would prevent the sentence to be passed to grammar checker system. If the underlying named entity is found in Wikipedia corpus, the system carries on to next stage, otherwise alternative reasoning will be used as will be explained later on. Fourth, if the sentence does not contain one of the newly introduced named-entity, the corpus is used for the purpose of grammar checking task by matching each consecutive pair of token to the corpus. The determination of the grammatical status of the sentence is given in the following axiomatic description.

Let a sentence S be given as a sequence of its N tokens T_1, $T_2,\ldots,T_N$. Let $G(.)$ be the binary variable associated to the entity in $(.)$ and indicating its grammatical status (1 for valid grammar construction and 0 otherwise)

Proposition 1
$G(T_i,T_{i+1}) = 1$ as soon as the bigram (T_i,T_{i+1}) occurs in Wikipedia database, i=1 to N-1

Proposition 2

$$G(S) = \min_{i=1,N-1} G(T_i, T_{i+1})$$

Especially Proposition 2 materializes the fact that as soon as one bigram cannot be matched in the corpus, the whole sentence is deemed to be grammatically incorrect.

In case of presence of (unknown) named-entity, the reasoning will be different for the above. First, consecutive bigram matching with that of Wikipedia will still be employed even if the pair does contain the named-entity. If no matching is found, a simple rule-based system will be employed. The latter makes use of the outcome of the Tagger and check for basic grammar match of the underlying pair. For instance, if the following token is a verb, then the form of the verb will be checked. If it is a noun then the rule checks for the presence of possessive form. To axiomatize such reasoning, let us consider again a sentence S with tokens T_1, T_2,..,T_N and assume that T_p, $p < N$ stands for such named-entity. Let $R(.)$ be the binary variable associated to the rule-based grammar status of (.).

Proposition 3

$$G(S) = \min\left(\min_{\substack{i=1,N-1 \\ i\neq p,\ i+1\neq p}} G(T_i, T_{i+1}), R(T_p, T_{p+1}), R(T_{p-1}, T_p) \right)$$

Proposition 3 again materializes that if a single grammatical rule involving the named-entity fails, then the overall grammatical status of the whole sentence vanishes. Similarly, the sentence is deemed to be grammatically incorrect if one of the bi-gram fails.

Next, simple statistical information is generated from each user profile. This includes the extent to which semantically and grammatically correct messages have been sent as well as the percentage of incorrect constructions. The latter are also classified into two classes: wording and grammar construction. The first class is generated as soon as the token cannot be found in dictionary, corpus or abbreviation list. While the grammar class refers to those sentences whose individual words were valid while the overall sentence scores zero according to the constructed grammatical system. More specifically, let U_w and U_G be the statistics information regarding the wording and grammatical inconsistencies, respectively, generated by user U, then:

$$U_w = \sum_i M(S_i)$$
$$U_G = \sum_i \left(1 - G(S_i)\right)$$

Where $M(S_i)$ stands for the total number of mismatches with respect to database normalization (English corpus, WorldNet, abbreviation list, named entity) occurred in sentence S_i generated by user U. The sum operation in the above expressions is with respect to all sentences generated by user U. Normalization of the above expressions is ensured by dividing the outcomes by the total number of sentences generated by the user. Other performance indices include the percentage of stop-word, numbers of links in the text. An overview of the grammar analysis system is highlighted in Figure 4.

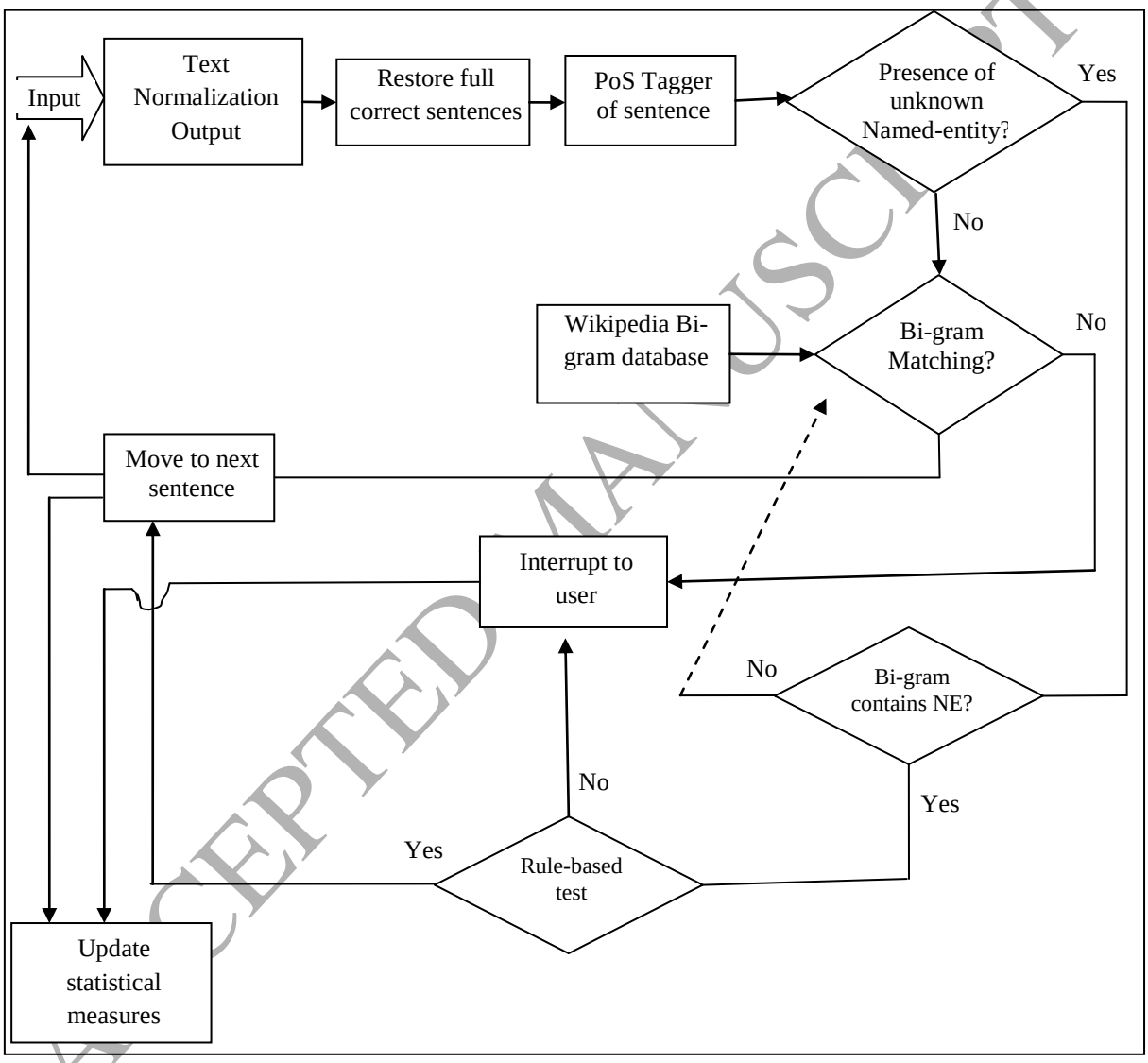


Figure 4. Overview of grammatical checker system

It should be noted that the above architecture has also its own limitations, which can be summarized in the following:

- The restriction to bi-gram analysis is often insufficient to identify complex grammar structures as many studies have rather employed larger number of tokens (3,4 and/or 5-grams or even sometimes higher). Nevertheless it is rationally acceptable that Twitters often employ simple sentences where bigram are often enough for regular checking. On the other hand, the use of successive pairwise matching to Wikipedia database allows us to check the occurrence of whole sentence in database.

- The use of Wikipedia database, although, technically speaking, sounds rational due to occurrence of most daily conversation types in its documents; still the diverse nature of Twitter messages renders the exact matching sometimes negative. Indeed, it often occurs that many events cannot be exactly matched to their counterparts in Wikipedia documents. For instance football match result, tennis results, number of beers that he/she bought, etc., which are by nature context dependent and very diverse, are rarely matched to Wikipedia documents.
- The later observation puts strong emphasize on the limitation of the logical based reasoning in n-gram related approach. Indeed, the use of stochastic reasoning with adequately chosen threshold would trigger positive matching despite occurrence of few mismatches. However, in our case, the moderator makes use of statistical information as an aid for the decision-making process. For instance the percentage of occurrence of numbers, stop-words in the text or named-entity provides a rough idea to the moderator whether the mismatch is genuine or only due to presence of such attributes.
- The simplified postulate that all unknown word with capital letter at start are considered as named-entity and tagged as noun is a very approximate way to handle the challenging task of named-entity recognition. Indeed, it occurs in Twitter that the capital was used to strengthen emphasize on specific issue in the message, therefore, coinciding such token with named-entity maybe far to be realistic. On the other hand, it also occurs that although being genuine named-entity, the token does not start with a capital letter as in “ibm”, “hp”, which degrades the information content of the message, as an interrupt of unknown token will be sent to moderator without any further analysis.
- The use of grammatical rules on bigram associated to named-entity is also very simplistic as it ignores scenarios in which the agreement between named-entity and its predecessor or successor is not needed as such agreement could take place with later successors or later predecessors as in “IBM today is gaining”, “many companies, say IBM, increase their shares”. Consequently, restricting the grammatical rule to the two adjacent tokens to the named-entity may lead to compromising results. Although, one also acknowledges that the occurrence of such sentences in Twitter messages is rather not common due to dominance of simple and short messages.
- The normalization stage carries out also a lot of inherent limitations. For instance, the appropriate choice of stop-word list is always problematic as this would lead to generate erroneous messages of incomplete sentence. For instance if the pronouns were discarded then the sentence “I see” becomes useless. That is why cautious should be considered when selecting the list of stop-words. Similar remark applies when using Porter stemming as well. This motivates us to allow more flexibility in this operation. For instance, if the normalization of the sentence yields one or zero token then limited stop word list will be invoked and stemming stage will be removed if the normalization still yield a single or zero token.
- The automated grammatical checker system is not meant to suggest alternative possible corrections to user unlike some commonly employed systems, e.g., Microsoft word grammar checker. This was forced on purpose to capture as much users’ interactions as possible and, on the other hand, leave more freedom for moderator to suggest appropriate actions.
- The use of a single abbreviation list as that used from Netlingo appears to be insufficient to capture the wide range of slogan words employed in Twitter. Nevertheless the intensive and parallel use of several aggregation lists simultaneously is very challenging due to presence of many conflicting interpretations of same slogan words according to distinct acronym lists. Consequently such parallelism cannot be achieved without intensive manual check and conflict resolution stage, which suggests the use of only one single, although possibly debatable, list.

6. Results and discussions

6.1 Statistical analysis of textual inconsistencies

The interest focused on the overall statistics related to tweets generated by all students. For this purpose, Table 1 summarizes the statistics in terms of U_W and U_G values averaged over all users. Especially the results highlight the percentage of inconsistencies due to token mismatch and grammatical constructions.

Table 1. Comparison of token mismatch and grammatical errors in Twitter database

	Ratio	Total number
Proportion of total sentences with token mismatch	73.45 %	107600
Proportion of sentences with grammatical inconsistencies (after passing token matching)	18.18 %	7071

The results pointed out in Table 1 show the dominance of word-mismatch in Twitter messages, which account for more than 73% of the total sentences generated by users, despite the use of comprehensive lexical databases as well as acronyms interpretations. Table 1 also points out that sentences which are free of token mismatch are less likely to fail in grammatical inconsistencies as only around 18% of those sentences were found grammatically incorrect. This would suggest that users who employ correct wording are likely to convey grammatically correct sentences as well. On the other hand, Table 1 also discards those sentences, which passed the normalization test, but did not go through grammatical checker because of the number of tokens left after the filtering phase is less than one. Such sentences are still considered, by abuse, to be grammatically correct as in "Hello", "See http://www.bham.ac.uk". In other words, this again confirms the predominance of the typographical mistakes in tweet messages, which sometimes renders the grammatical test void.
Next, instead of taking the overall sentences generated by all users, one focuses on individual user's score; namely, the overall inconsistency ratio in terms of U_W and U_G generated by each user (student). In this course, Figure 5 and Table 2 summarize the statistics of the underlying variable in terms histogram representation and first/second order statistics, respectively.

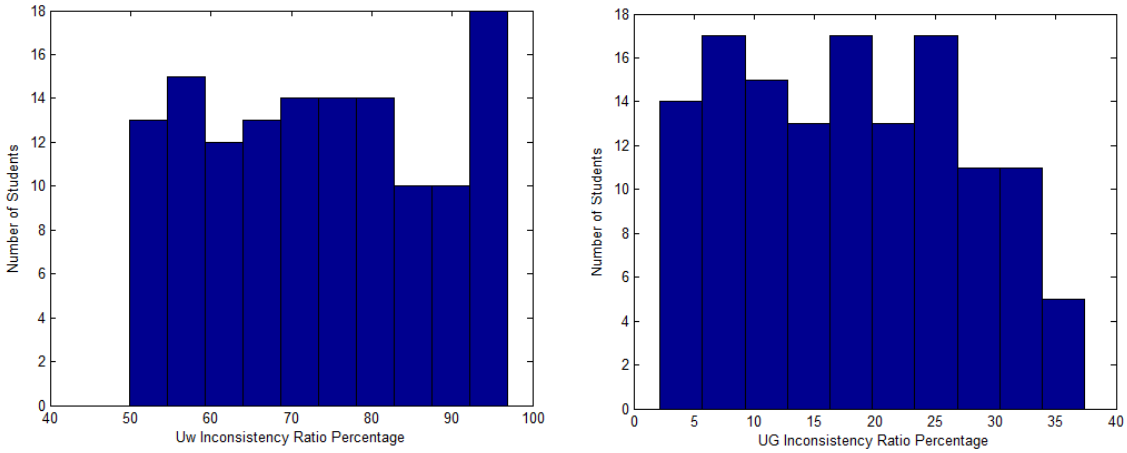


Figure 5. Histogram of U_W and U_G with respect to number of students

Table 2. Comparison of token mismatch and grammatical errors in Twitter database

Variable	Mean	Standard deviation	Max	Min
U_W	73.45 %	13.82%	96.9%	49.9%
U_G	18.18 %	9.4%	37.4%	2.2%

Figure 5 indicates that even the best students over the long compilation of messages only achieve 50% success rate in terms of tokens matching but can achieve high success rate in grammatical constructs as shown by the minimum value of U_G.
An interesting question is to see whether the effect of retweeting and possibly moderator intervention influences the linguistic quality of the users' messages. For this purpose, we considered the number of inconsistencies generated by each user in the first 5 weeks of the experiment and that generated by the same user in the last 5 weeks. The rationale behind this dichotomy is to hypothesize that the students would have learned from their peers and moderators' comments to improve the linguistic quality of their messages. To this end, we employed a statistical paired difference test using Wilcoxon signed-rank test [18]. The latter allowed us to circumvent the assumption of the Gaussian distribution of the sample, which does not hold as can be noticed from the histogram representation in Figure 5. Besides, we focused on one-side tail test to test whether the strict improvement of the performance. More formally, the null and alternative hypotheses are:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 < \mu_2$

Where μ_1 (resp. μ_2) is the population mean rank for scores (inconsistency percentage in case of U_W or U_G) for the first 5 weeks (resp. the last 5 weeks) of the experiment. Table 3 summarizes the key findings including the result of this statistical test.

Table 3. Comparison of token mismatch and grammatical errors in Twitter database

Random Variable		Mean	Standard deviation	Max	Min
U_W	First 5 weeks	78.7 %	23.13%	100%	49.6%
	Last 5 weeks	68.58%	14.33%	94.4%	43.4%
U_G	First 5 weeks	18.18 %	9.4%	37.4%	2.2%
	Last 5 weeks				

Wilcoxon signed-rank test: H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 < \mu_2$, significance 5%

	Degree of freedom	Hypothesis supported	p
U_W	224	H_1	$< 10^{-10}$
U_G	224	H_1	$< 10^{-10}$

Statistical results summarized in Table 3 confirmed that the interaction with peers and the moderators positively influenced the linguistic quality of the students.

We also focused into the type of grammar mistakes which are often faced by the students. For this purpose, we have performed a manual checking on a randomly selection (around 10%) of total grammatical incoherent sentences. We borrowed the classification employed by Ahangari and Barghi [2], and manually evaluated the proportion of each category in the randomly selected sample of incoherent sentences. The summary of this analysis is highlighted in Table 4.

Table 4. Category of grammatical mistakes

Category	*Frequency*
Adverbial Clauses of Contrast	22.6 %
Inversion	14.1 %
Definite Article	41.7 %
Conditional Sentence	11.8 %
Coordination	31.5 %
Relative Clauses	23.4 %
Adjectives	16.6 %
Agreement	19.7 %
Verbs	44.8 %

In order to exemplify the most frequent inconsistencies encountered in our analysis of tweet messages, Figure 6 highlights the most common (top 10) word mismatches encountered through monitoring all tweets generated by students regardless whether they were sent to their peers or not (typically the total number of tweets generated ranges from one to 3 times those sent to the peers as will be seen later on). The results showed that 8 out 10 of the most mismatches are due to new abbreviations that cannot be mapped to Netlingo's list. While only two are due to other incorrect wording, e.g., space character missing in case of "nowplaying" and unknown expressions in case of "u wind.", which is converted into "you wind." that cannot be mapped into Wikipedia database, so assumed to be grammatical inconsistent.

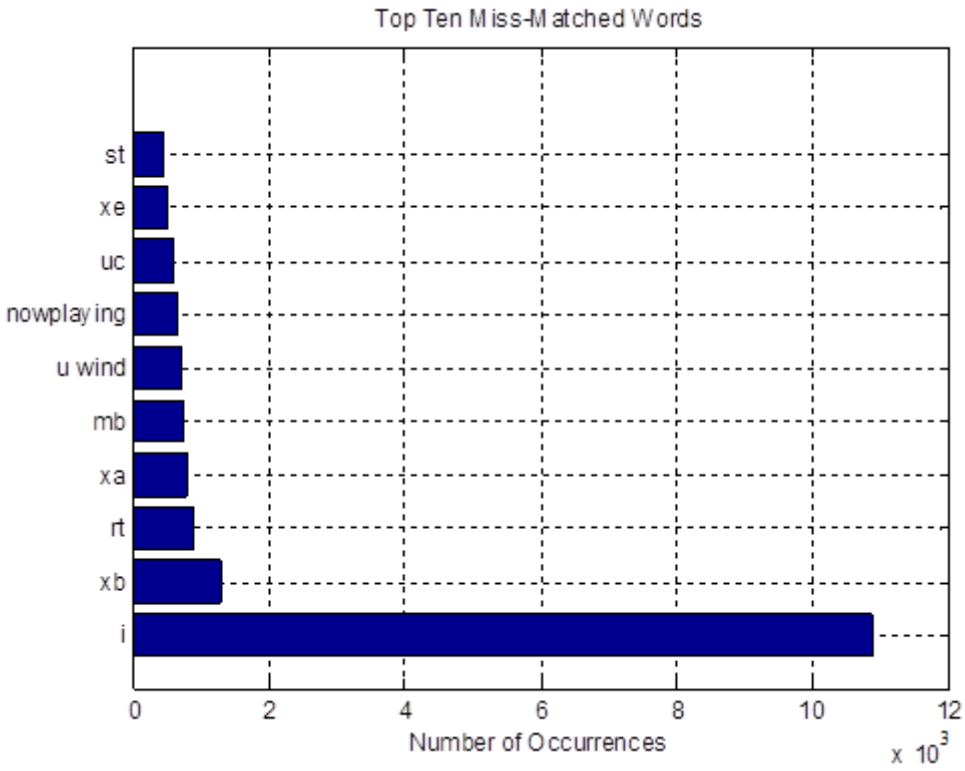


Figure 6. Top Ten Most Frequent Word Anomalies

Similarly, Figure 7 highlights the 10 most common grammatical errors recorded by our system.

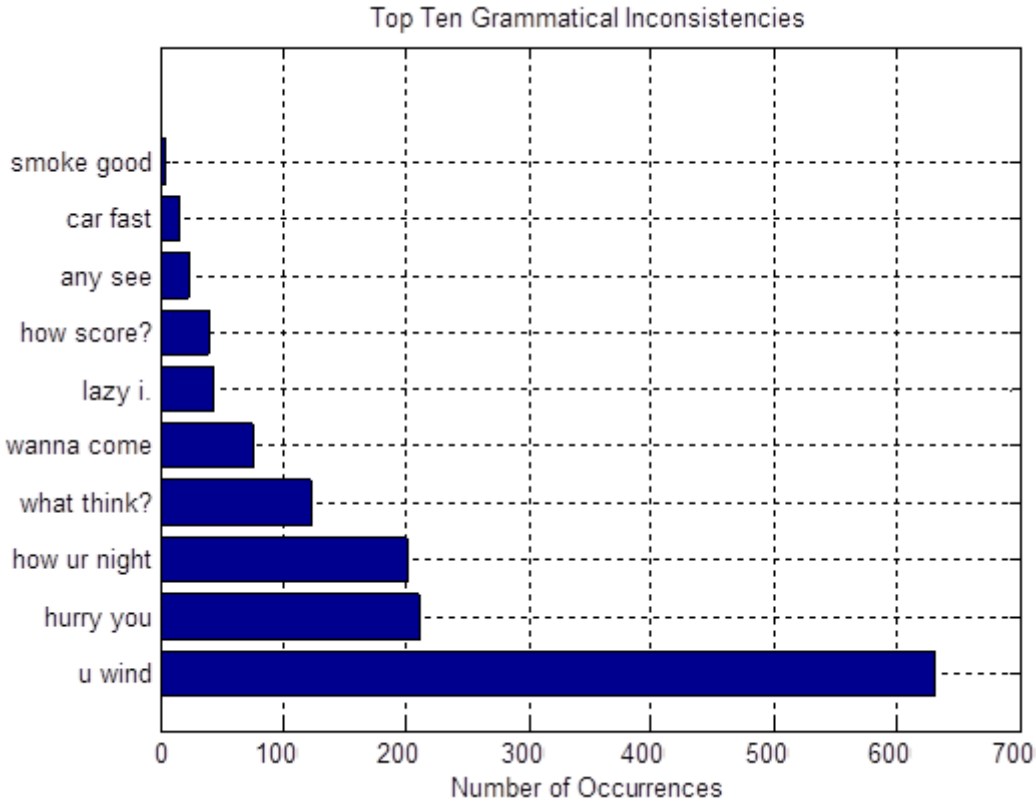


Figure 7. Main common grammatical inconsistencies in Twitter corpus

The results in Figure 7 show that the main source of grammatical incoherence arises from deliberate behaviour of the Twitter user to shorten the sentences by omitting some words, mainly verbs in the phrases as pointed out in the above graph. Indeed, for instance, "how is your night" is shortened to "how ur night", "what do u think?" is reduced to "what think?" "lazy I am" into "lazy I", "hurry up you" into "hurry you", "you wind me" or "you wind up" into "you wind", "car is fast" into "car fast", possibly "anyone see it" into "any see", etc.
Strictly speaking the use of abbreviation together with the tendency to omit some words increase the likelihood of grammatical inconsistencies. For instance, one can speculate that a user did forget to add the symbol "1" in "any1" to stand for anyone. On the other hand, it is also worth pointing out that the use of bi-grams which includes punctuation points increases the chance of grammatical consistencies. Indeed, one may expect for example that if one omits the end point in "you wind." will turn the grammatical inconsistency into a fully grammatical correct phrase. Therefore, omitting the punctuation points in the sentence, although reduces the semantic meaning of the underlying text, would obviously decrease the extent of the grammatical inconsistencies in the sense of bi-gram Wikipedia matching. The correction of a given grammatical fault is not always uniquely performed as it maybe subject to multiple interpretations. For instance, "car fast" maybe "car is fast" (omitting verb "is") and maybe "fast car" –by interchanging the order of the two words. In other words, a grammatical mismatch can also be rooted to sequential order of the tokens in the sentence.

6.2 Student interaction quantification

An interesting question is the contribution of students' interaction to the overall learning process as reflected by their Twitter messages or direct re-tweets. From this perspective, one shall mention the following:

- All tweets are made readable for all students provided they wish to open them and read the content but it is a common practice that Twitter users read only those tweets they are interested in. Especially, when the name or Id of the person is explicitly mentioned in the tweet, there is a high likelihood that the underlying user will open such message.
- Although it is recommended that students put their peers as followers in their Twitter account to follow up their activities, conversations and share opinions of their mates, it turned out that many students select only a subset of their mates as followers or friends. This can be explained by both the desire to keep the friendship list small and the non-interesting discussion topics initiated by some mates.
- Students are more tempted to reply to their mates when a direct re-tweet was used.
- The connection to the local server, which supports our extended search capabilities, although is made available to all students if they want to perform extensive search on Twitter database collected so far, has not attracted too much interest from the students as only a tiny proportion (less than 15%) of them have reported to be using it on regular basis. This can be partly explained by the previous argumentation where some students are only interested into activities related to their close peers or friends.
- The students have also been made aware of existence of other social media monitoring tools like Hootsuite where the students can query Facebook, Twitter, Linkedln, among others, and link to relevant tutorials has been provided. Only few students have reported testing Hootsuite tool as will be reported later on. Nevertheless, although, we can physically monitor the number of accesses to the developed server application related to Twitter search, one cannot do so for other social media monitoring tools so that we can only rely on student feedback for such purpose.

To quantify the interaction among the students in terms of number of followers, Figure 8 illustrates the proportion of the student population among the total number of followers for the ten most active users.

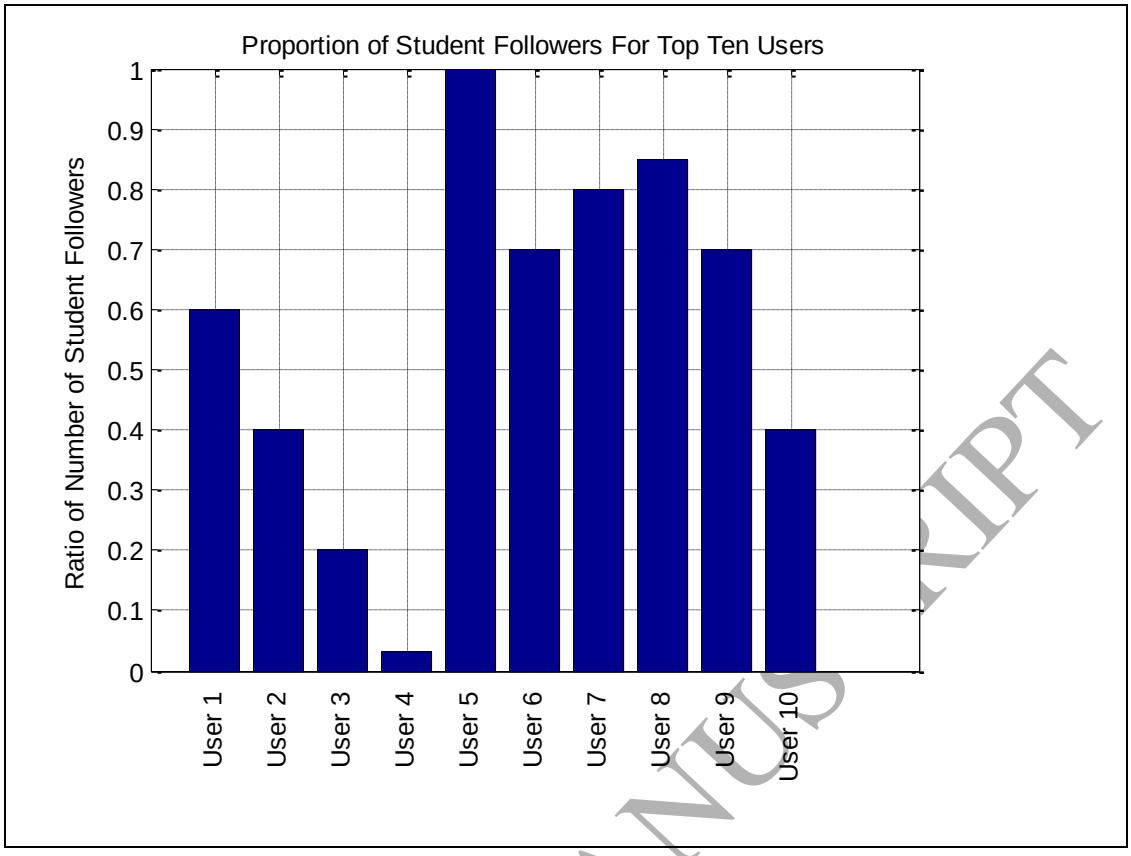


Figure 8. Proportion of student followers with respect to total number of followers for each user

Results in Figure 8 demonstrate that, for most active users, the proportion of students in the user's follower list is not negligible and varies from 3% till 100% for User 5 where all his/her followers are constituted of student population. The graph also highlights the trivial fact that most Twitter users do have their own list of followers depending on their own interests, which may not overlap with that of their peers. On the other hand, we shall also point out that the number of student followers over time varied from more than hundred to less than 20 at the end of the study. This is mainly due, as already pointed out, to either the desire to keep the follower list to close circle or the non-interest of topics developed by other peers.

Strictly speaking, although the proportion of student followers is a rough measure to quantify the interaction among students, it may not provide an overall picture about the real scenarios that may have been taken place during the interaction process, especially with respect to the contribution of the group to the learning task. That is why highlighting instances of tweet messages might be more relevant for such purpose. Another metric of interest is the activity of the students users in terms of tweets generated to their peers throughout this study. In this respect, Figure 9 highlights the activity of the most active users in the first 5 weeks.

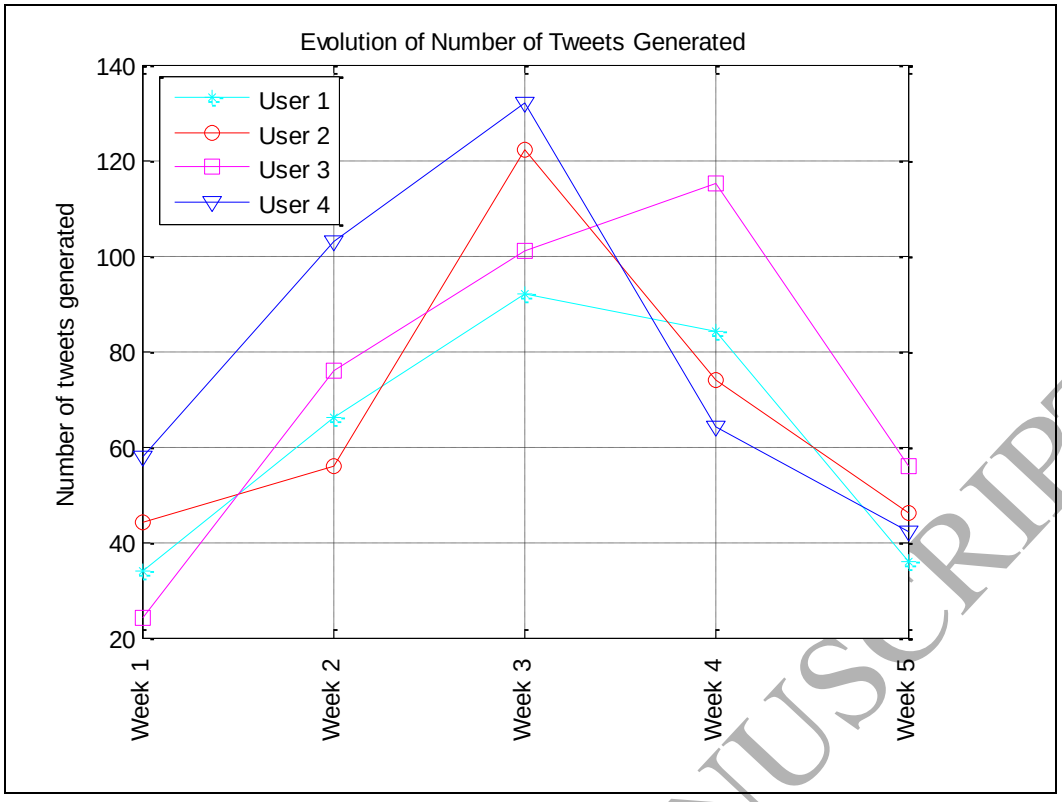


Figure 9. Evolution of number of tweets generated by the four most active users.

Figure 9 highlights the observation that students make a slow start at the beginning then get familiarized while enjoying hot discussion topics and possibly making new connections and/or new friends yielding larger number of tweets before the motivation decreases due to possible *boredom* or less interest, which, in turn, is translated into a reduced number of initiated tweets. For the sake of the consistency, one shall also mention that there were other important external factors that forced the student activity down in the last weeks. This was related to the exam period and its preparation. On the other hand, the tweets recorded in Figure 8 were only related to peers' discussion topics as indicated by Hashtag or the occurrence of one of the student ID in the body of the tweet message. This also suggests that the active users maybe issuing other tweets not necessary related to this study. To quantify the proportion of such behaviour, Figure 10 provides the ratio of the tweets related to this study over the total number of tweets generated by the ten most active users. The results show that active Twitter users do have a range of topics that do not want to share with the rest of the group. This includes mainly family related messages as well as intimate and love discussion that most users decided not to share.

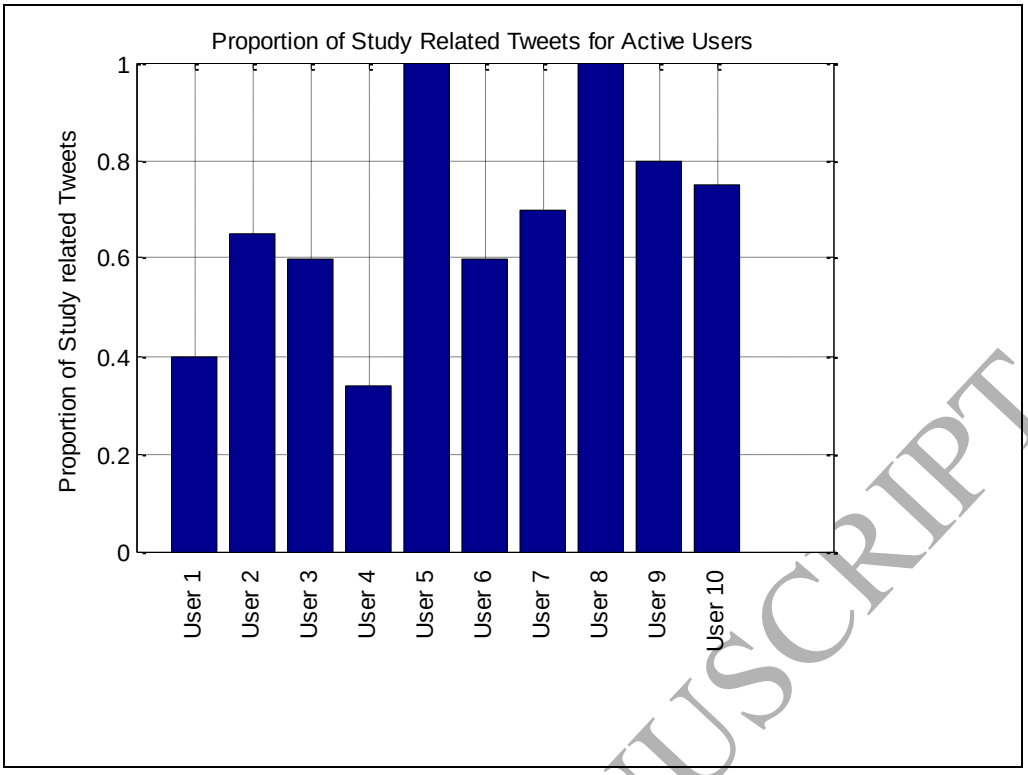


Figure 10. Proportion of study related tweets for active users

Peer's influence in Grammatical spelling correction

Instead of looking to overall (linguistic) quality of tweets generated by individual user as performed in previous subsection, one shall focus here to individual performance in response to re-tweet relationship. More specifically, given that *User " i"* retweet to *User "j"*, the question is whether an improvement of linguistic quality of *User "i"* message entails an improvement of that of *User "j"* or not. In this respect, three cases can be distinguished. This concerns situations in which the message of *User "j"* can be present either better or at least similar (resp. worse) linguistic quality as that of *User "i"*, which corresponds to positive influence (resp. negative), otherwise, the influence is deemed to be neutral if such conclusions cannot be inferred.

Indeed, it often occurs that a message even if it contains inconsistent wording, still can be understood by their mates who may reply using the same level of inconsistencies or worse. On the other hand, we have also noticed situations in which the user does correct his spelling or wording in response to his peer's tweet. Examples of such behaviour are provided in Fig 11 and 12.

-Good morning all. Hope all's well in ur world. Had a bit of a fall last night and feeling sorry for myself. Hurt my knee and it kills!
-@greeneyes_uk hope you are feeling ok, you old folk need to watch out for falls x
- You Need to revise your spelling baby
- @greeneyes_uk sorry, meant your old folkneeded // grammatical correction
-@Ozzrahog cheeky! Does ur mobile no end 008? I know random question, but good reason lol

-Thx @Nayfz for sharing ur personal views on my account...
-@mistry267 Thanks your**
-@mazhancock quite you. I'll have you know its rather comfy out here
-@bengrindle_ really though? It's not cold? Great spelling btw ;) // comment to enforce good spelling
-GOOO Hurry!!! // emphasize through capitalization
-@mazhancock it's not actual cold haha and I cant spell for shit!
-@bengrindle_ haha ahh fair play! Naaaah you really can't ;)

-Mmmm Caramel Macchiato and breakfast in bed :-) love u @absoluteben
-@jasonbetts @absoluteben wow. Go u. How is the costa caramel coffee?
-@JamesMcCafferty @absoluteben it's not Costa.. I'd rather drink shit // Correction of wording
- Thanks for pointing this out. Well done! // Moderator comment to encourage self correction!
-My accountent has stopped working to order a curry #skiever
-@Ste_CAFC Don't just need someone to help you with numbers, you need someone to help you with words too.
-@CB_YNWA what did I spell wrong?
-@Ste_CAFC Accountant and Skiver.
-@CB_YNWA everyone will know what I'm on about shh. Snoggrass is a cunt.
-@Ste_CAFC Haha why is he?
- @CB_YNWA plays for Leeds lol

Figure 11. Examples of Tweet messages influenced by peers

// PATTERN : User corrected his mistake himself
@JoanneHannam stupid touch screen n predictive text ... My spelling looks awful sorry

@KrisFoster92 I might waie one now haga in bored // one of worst spelling ever

// PATTERN : Both friends make same mistake :

-Can't believe u r still so chipper with so much on your plate! Wish I could help u with your move! And tks 4 ur support babe. It means LOT!
*-@Briarwitch hehe I know I even surprise myself! Haha ugh the move! *groan* LOL! Ur welcome, anytime :) Xx*

// PATTERN : User is correcting by friend's retweet but he corrects it by irony :

-@Destructor no no. Happy new yeat to you China.
-@sevitz you made it a whole 21 minutes before making your 1st spelling mistake of 2011. Is that a record? Happy new one guv'nor.
-@mattverso that's a good 15 mins longer than I expected. Happy gnu ear china.

Figure 12. Examples of Tweet messages with specific patterns

Next, in order to quantify the extent of the three influence types (positive, negative and neutral), we carried out a statistical analysis of all pairwise retweet operations generated by this study (a total of 2382 tweets containing retweet operations were examined). Table 5 summarizes the extent of the neutral, negative and positive influence in the examined retweet messages.

Table 5. Peers influence on spelling and grammatical errors

	Percentage	Total number of retweets
Proportion of Positive Influence	4.5%	107
Proportion of Negative Influence	15.8 %	377
Proportion of Neutral Influence	79.7 %	19000

Table 5 indicates that when chatters do understand each other, they do not bother much asking for clarification or rewording original messages, which partly explain the highly dominant neutral influence in our study. Another interesting result corresponds to the dominance of negative influence over positive influence, which is justified by the fact that many users when retweeting on a tweet containing inconsistent wording or grammar fault are not bothered much by correcting such faults.

6.3 Discussions

- At the end of this study, the students have also been asked to report their responses regarding the usefulness of this Twitter experience. They have been asked to provide their answers in the scale (Not Useful, Moderately Useful, Useful, Very Useful) as well as any written comment if any. The responses in terms of usefulness of the study are summarized in Table 6 below.

Table 6. Overview of participant evaluation of Twitter learning experience.

Not Useful	Moderately Useful	Useful	Very Useful	No opinion
1.5 %	8 %	33%	56 %	1.5%

Results of Table 6 show the large satisfaction of participants with the new learning experience. Especially, it turns out that the novelty of such experience in the education setting likely played great impact to ensure student satisfaction.

On the other hand, many of students' comments paid tribute to their interesting experience with this study, which allowed them to discover various features of Twitter system and advanced social learning interactions. At the same time, at least for some of them, knowing that their messages can be monitored and seen by their peers and close friends provides further motivation to pay more attention to their spelling to avoid possible disappointing comments from their mates or moderators.

- Educational and psychological studies acknowledged the importance of social influences and peer interactions as an important factor in learning, besides no assumptions were made regarding the positivity of the outcomes of peer influence [22]. This fully agrees with the outcome of our study where, although the importance of peer interaction through Twitter has been unanimously recognition by the participants as demonstrated by feedback analysis in Table 6, the outcome of the interaction in terms of enhancing quality writing is not always positive as testified by results in Table 5.
- Freedom of discussion topics initiated by students in their interaction with their peers also played an important role in enhancing their engagement as it is trivially difficult to motivate students if no interest has been shown. Indeed, Alton-Lee et al. [5], noticed that informal talks plays a large role in fostering students' learning of concepts and ideas of lessons, no matter whether it occurs during teacher related lessons or in independent activity time. That is because informal talk promotes cognitive restructuring and co-construction of ideas between peers.
- The novelty of Twitter like study has also played key role in enhancing students' interest. Indeed, although students were very familiar with various social networking tools including online chat, Facebook, messenger, among others, many of them discovered the advanced features of Twitter for the first time. Therefore, the curiosity in reading the peers' and mates' messages and comments is worth reporting. Their feelings of enjoyment were consistent with previous findings of students' perceptions of microblogging [4], and also agreed with the dominant perception among researchers that the use of microblogging, which includes Twitter, in education research is still in its infancy.
- Unlike in standard class-room related grammar constructions and learning where the students have little audience for their productions, Twitter provided a huge audience to students where all mates and even all Twitter community can monitor their outcomes. This obviously has intuitively both positive and negative effects. Positive effects include increased student attention in their writing being known that he/she is watched by several people and yielding more normative sentences. But this can also affect negatively student's performance in the sense that it may put more pressure causing the student to use more automated filters, which would cancel the benefit of peers' interaction that may help student to focus on the forms of his/her tweets. Indeed DeKeyser [13] reported that having declarative knowledge in working memory during practice is an essential part of skills acquisition.
- From the student feedback perspective, this study revealed that the majority of students felt Twitter helped memory consolidation. Indeed, construction of short sentences, trying to be brief and concise in the tweet messages, always make students remembering their basic grammar and correct wording schemas. Even

sometimes some students argued that they checked the spelling of the sentence in Google before putting it up on Twitter message! Besides students often do remember their mistakes especially when highlighted to all audience, learning new constructions and common miss-conceptions, which overlaps with other related studies, see, for instance [4].

- The analysis of the common construction errors generated by the developed grammar checker system showed that a large number of such errors are rather due to unknown abbreviations. Indeed, the limitations of the acronym list employed by our system are clear in the sense that it only conveys a small number of abbreviations contained in tweet messages. This also partly explain the dominance of neutral and negative effects in Table 6 because, trivially, whenever an abbreviation is mutually understood by users, it is no longer a source of concern, while the automated system fails to detect so. Loosely speaking the dynamic aspect of slang vocabulary, which evolves both in time and space dimension, has already been reported by other studies, see, for instance [12]. Consequently, restricting the acronyms to a single list developed at a given time and space cannot be ambitious to accommodate all slang words that may occur in this study. Many slang words are indeed well understood in a given area but not elsewhere. For instance "bham" is well understood to stand for "Birmingham" but not identified as so by Netlingo and many other slang dictionaries.
- The issue of named entity pointed out in Section 4 has shown to be of little influence in overall given the low number of occurrences of such cases in student tweet corpus. Indeed, the students use of capitalizing mostly occurred either with genuine named-entities or verbal expressions to highlight some emphasize as in "GOOO hurry!!" message in Figure 11.
- This study also revealed that the use of punctuation, although plays an important role in academic writing, maybe problematic when it comes to automatic message analysis. This is because the use of punctuation in Twitter messages is often not consistent, and many users make use of interrogation and exclamation points for emphasize or surprize highlighting, which, in turn, may reduce the likelihood of positive matching to Wikipedia corpus. A possible solution to this effect consists of avoiding the use of punctuation points in bigram representation. This would ultimately increase the proportion of grammatically correct sentence by a non-negligible proportion. An evaluation of such cases indicated that a reduction of percentage of sentences with grammatical inconsistencies from 18.8% to 16%.
- The study pointed out that the interaction was rather more important with a (relatively) small circle of close friends as demonstrated by the (relatively) high number of messages arising from this interaction. Although, some students have reported using advanced visualization tools like Hootsuite, which is rather a social media aggregator allowing the user to add various social media accounts, then create side-by side streams of different functions of the social media tool instead of several clicks, required when Twitter was used alone. In this respect, an investigation of cases where student interacted with users outside such circle revealed that this was primary related to either the desire to pick up interesting topics through hash-tag and next sharing it with peers or the need to update with last news related information concerning some specific events, e.g., latest football match score, player or artist news, among others. Nevertheless, the study revealed that only a tiny proportion of the student population actually employed Hootsuite tool and queried the whole tweet database for some search output.
- Throughout this study most statistical analysis has been performed over the overall number of tweets generated by the student (s), not necessary sent to his/her peers. This allowed us to quantify the influence of our educational study on the writing skills of the students regardless whether this is destined for the class purpose or not. Indeed, needless to say that such influence can better be impacted if it can be demonstrated both within and outside the class room context, which motivated our current approach for such analysis.
- The role of the moderator (s) in this study is kept very minimal. It was mainly focused on enhancing student motivation if there is a low activity, stimulating debate through possible direct questions, charming and encouraging users of well written messages, issuing retweets to seek clarification or force student to rewrite his/her message using better wording and highlighting links to illustrate better use of Twitter. However, the moderator provided no technical (academic) grammatical lesson or forced students explicitly to do so. But, obviously, implicitly, students were motivated to refresh their grammar and spelling background to avoid, possibly, unwanted comments from moderator (s) or their peers. This contrasts with previous Twitter educational studies where full academic backgrounds were provided. Nevertheless such approach has been conducted as so, because in contrast to the aforementioned related educational Twitter studies, the participants were reasonably good English speakers and already achieved good grade at their GCSE English. Besides, it was not part of the study to allow concepts be learnt through explicit teacher intervention but rather espouses frequency approach of learning [28] where students learn from frequency of constructions in the input and repeated exposure to exemplars especially from their peers.

7. Conclusion

This paper reported on the use of Twitter in an educational study involving reflection on linguistic aspect of tweet messages. Especially, an interactive system has been developed for Twitter collection and analysis from grammatical perspective. The latter, after a comprehensive normalization stage that involved URLs suppression, tokenization, abbreviation matching to known acronym lists and matching to corpus constituted of WordNet and English corpus database, made use of logical approach on bigram representation in conjunction with Wikipedia database. The system offered increased interaction capabilities as well as possibility to generate worry signal to moderators. Although the implementation of the automated system was very challenging due to inherent imprecision pervading natural language processing and the nature of tweet messages, which were dominated by intensive use of (sometimes random) abbreviations and incomplete sentences, the outcome reinforced the results from other educational studies employing microblogging. Indeed, the study demonstrated the interest of the vast majority of the students to the new learning experience as testified by the outcome of the students' feedback. Besides, the novelty aspect of Twitter feature played probably a key part in enhancing student motivation. On the other hand, the visibility and noticeability of Twitter messages enforced the students to pay more attention to the linguistic quality of the messages, which led to increase in linguistic quality and reduced grammatical mistakes. The study has also revealed the predominance of the slang language. Such abbreviations have shown to pose the greatest challenge for automatic text analysis [3,6, 34] especially given its dynamic aspect both in time and space in the sense that new abbreviations are still being generated by Twitter users, which quickly got adopted by other users. On the other hand, abbreviations employed in one region may not be applied in another region. This makes any effort to create a universal slang list very difficult, if not impossible. The use of bigram representation together with Wikipedia matching in the framework of logical reasoning, although can be justified when simple sentences were used, revealed that the limitations of such approach that cannot be ignored when more complex sentences were employed. Another challenge that rose from this study related to the handling of named-entity. Indeed, our approach considered a very simplistic assumption where an unknown entity starting with a capital letter is assumed to be a new named entity. However such reasoning is intuitively very debatable as well since Twitter users tend to use capitalization for emphasize. In conclusion, the outcome showed that Twitter were consistent with skills-acquisition theory and increased the likelihood of form-function meanings becoming proceduralized and entrenched in long-term memory [13].

Acknowledgment

This work is partly supported by EPSRC Bridging the Gap Project 2009-2011. The authors are thankful for Birmingham City Council who contacted school willing to participate on voluntary basis into this study.

References

[1] A. Acar and N. Kimura, Twitter as a Tool for Language Learning: The Case of Japanese Learners of English, *The Eighth International Conference on eLearning for Knowledge-Based Society, 23-24 February 2012, Thailand, pp.1-14.*

[2] S. Ahangari and A. H. Barghi, Consistency of measured accuracy in grammar knowledge tests and writings: Toefl Pbt, Language Testing in Asia, Vol. 2. No.2, 2012, pp. 5-21.

[3] E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011.

[4] E. Antenos-Conforti, Microblogging on Twitter: Social networking in Intermediate Italian classes. In L. Lomicka, & G. Lord (Eds.), *The next generation: Social networking and online collaboration in foreign language learning*, (pp. 59), 2009

[5] A. Alton-Lee, G. A.,Nuthall, and J. Patrick. Reframing classroom research: A lesson from the private world of children. Harvard Educational Review, 63(1),1993, 50–84.

[6] A. Bifet, A., and R. Kirkby *Data Stream Mining. A Practical Approach*, MOA, The University of Waikato Press, 2009.

[7] A. B. Bodomo. Computer-mediated communication for linguistics and literacy: Technology and natural language education. Hershey, PA, USA: IGI Global, 2009.

[8] K. Borau, J. Feng, R. Shen, C. Ullrich. Microblogging for language learning: Using twitter to train communicative and cultural competence. *Lecture Notes and Computer Science, 2009. Vol 5686.* DOI: 10.1007/978-3-642-03426-8_10

[9] R. Broderick, Brazilian School Kids Have Been Learning English By Correcting Celebrities' Grammar On Twitter, *BuzzFeed,* 2013, http://www.buzzfeed.com/ryanhatesthis/brazilian-school-kids-have-been-learning-english-by-correcti Accessed in September 2013
[10] B. B. Brown, Peer groups and peer culture. In Feldman, S. S. and Elliott, G. R. (eds.) At the Threshold: The Developing Adolescent, pp 171–196. Cambridge, MA: Harvard University Press, 1990
[11] R. Crosnoe, S. Cavanagh, G. H. Elder, Jr Adolescent Friendships as Academic Resources: The Intersection of Friendship, Race, and School Disadvantage. Sociological Perspectives. 2003;46:331–52.
[12] D. Crystal, Language and the Internet. Cambridge, U.K.: Cambridge University Press, 2001
[13] R. M. DeKeyser, Beyond focus on form: Cognitive perspective on learning and practising second language grammar. In C. Doughty and J. Williams (eds.). *Focus on form in classroom second language acquisition* (pp. 42-63). New York: Cambridge University Press, 1999.
[14] W. B. Dolan, L. Vanderwende, S. D. Richardson. Automatically Deriving Structured Knowledge Base from On-Line Dictionaries. Proceedings of the Pacific ACL. Vancouver, BC, 1993.
[15] Dracos., Open Data Gazetteer - GitHub. Available at: https://github.com/dracos/opendata-gazetteer/, Access May 2011.
[16] C. Falbium, WordNet, *An Electronic Lexical Database* (Language Speech and Communication), MIT Press, 1998.
[17] D. Foster, A Typical Twitter Use, Gigaom, June, 2009, http://gigaom.com/collaboration/a-typical-twitter-user/ [Date Accessed: 02/04/2012]
[18] J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*, 5th Ed., Boca Raton, FL: Chapman & Hall/CRC Press, Taylor & Francis Group, 2011.
[19] B. Godwin-Jones, Emerging technologies: Web-writing 2.0: Enabling, documenting, and assessing writing online. *Language Learning & Technology, 12* (2), 2008, 7–13.
[20] G. Grosseck and C. Holotescu. Can we use Twitter for educational activities? The 4th international conference eLearning and software for education, Bucharest, 2008.
[21] N-R. Han, M. Chodorow, C. Leacock. Detecting Errors in English Article Usage by non-Native Speakers. Natural Language Engineering, 12(2), 2006, pp. 115–129.
[22] W. W. Hartup. Constraints on peer socialization: Let me count the ways. *Merrill-Palmer Quarterly Journal of Developmental Psychology*, 45(1), 1999, 172–183.
[23] E. Hatcher, O. Gospodnetic, and M. McCandless, *Lucene in Action*, 2nd edition, Manning Publication, Stamford, USA, 2010.
[24] M. Homma. Twitter Eigogakushuuhou- Way to study English using Twitter. Tokyo: Discover Twenty-one, 2010.
[25] C. Honeycutt and S. C. Herring.. Beyond microblogging: Conversation and collaboration via Twitter. *Proceedings of the Forty-Second Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Press, 2009
[26] J. Kaufmann and J. Kalita. Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing*, Kharagpur, India, 2010
[27] V. Lampos and N. Cristianini, Tracking the flu pandemic by monitoring the Social Web. *2nd International Workshop on Cognitive Information Processing*. Elba Island, Italy. 2010.
[28] C. Leacock, M. Chodorow, M. Gamon, J. Tetreault. Automated Grammatical Error Detection for Language Learners. USA: Morgal and Claypool, 2010
[29] O S. Martínez, C. P. G. Bustelo, R. G. Crespo, E. T. Franco, Using recommendation system for E-learning environments at degree level, International Journal of Artificial Intelligence and Interactive Multimedia, 1(2), p. 67-70.
[30] F. Mochón , O. Sanjuán, A First approach to the implicit measurement of happiness in Latin America through the use of social networks, Special Issue on AI Techniques to Evaluate Economics and Happiness, DOI: 10.9781/ijimai.2014.252
[31] M. Oussalah, F. Bhat, K. Challis, and T. Schnier, A Software Architecture for Twitter Collection, Search and Geolocation Services, Knowledge Based Systems, 37, 2013, 105-120.
[32] K. Puniyani, J. Eisenstein, S. Cohen, and E. P. Xing. Social links from latent topics in microblogs. In Conference on Social Media, page 31, June 2010.
[33] S. Sirucek, Twitter, where grammar comes to die. *Huffington Post, 2010. See* http://www.huffingtonpost.com/stefan-sirucek/twitter-where-grammar-com_b_379191.html
[34] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005.
[35] K. Toutanova and C. D. Manning, Enriching the Knowledge Sources Used in a Maximum Entropy Part of Speech Tagger. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000). Hong Kong, China, 2000, pp. 63-70.

[36] C. Ullrich, K. Borau, and K. Stepanyan, (2010). A Social Network Analysis Perspective on Student Interaction within the Twitter Microblogging Environment". Los Alamitos, CA, USA., pp. 70-72. IEEE Computer Society.
[37] Vodacom, https://www.vodacommessaging.co.za/dictionary.asp? (accessed in October 2013)
[38] C. Whitelaw, B. Hutchinson, G. Y. Chung, G. Ellis. Using the Web for Language Independent Spell Checking and Autocorrection. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 890–899, 2009.
[39] X. Yin, J. Gao, W. B. Dolan. A Web-based English Proofing System for English as a Second Language Users. Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India, 2008, pp. 619–624.