Key identifiers and spelling conventions in MXit-lingo as found in conversations with Dr Math

L BUTGEREIT, RA BOTHA AND M VAN DEN HEEVER[1]

Abstract: Different human languages look different from other human languages. To use a term from the computer industry, each human language has its own "look and feel". European English speakers can easily recognise a phrase such as "*Comment allez-vous?*" as being written in French while the phrase "*¿Habla usted español?*" is written in Spanish. Each language has its own letter frequencies, word frequencies and other identifiers. This paper describes key identifiers in MXit lingo as found in Dr Math conversations. MXit is a mobile instant messaging system which originated in South Africa and is expanding to other countries. Dr Math is a mobile tutoring system which uses MXit as a communication protocol. Primary and secondary school pupils can receive help with the mathematics homework using the Dr Math tutoring system. The pupils use MXit on their cell phones and the tutors use traditional Internet workstations. After exploring *how* MXit lingo is written, this paper will briefly explore *why* MXit lingo is written the way it is. By identifying and describing the orthographic conventions visible in the spelling of MXit lingo, although with some theoretical support, insight into the purposeful and functional nature of written, mobile communication will be revealed. In highlighting spelling that is influenced by Black South African English, an attempt will be made to contribute to the empirical development of a field of study that explores the construction of words used in South African mobile communication.

Keywords: MXit, Math, letters, writing, orthography.

Disciplines: Linguistics, mathematics, information technology

Background

The written word has existed for thousands of years (Schmandt-Besserat, 1996). The development of civilisation and the development of writing are so intertwined that it is hard to imagine one without the other (Coulmas, 1991). Although systems of writing have existed for these thousands of years, they did not develop uniformly around the world. There are

[1]. Laurie Butgereit (lbutgereit@csir.co.za) is a PhD candidate in the School of ICT at the Nelson Mandela Metropolitan University, Port Elizabeth. She is also a principal technologist at Meraka Institute, CSIR, Pretoria. Reinhardt A Botha is a professor in the School of ICT and the Institute for ICT Advancement at the Nelson Mandela Metropolitan University, Port Elizabeth. Michelle van den Heever holds a teaching degree and an Honours degree in Applied Language Studies with an interest in multilingual mother-tongue-based education. Butgereit's research was partially supported by the Rupert Gesinstigting Award provided by the Rupert Family Trust, while Botha is partially supported by the NRF.

different systems of writing in different parts of the world. In addition, the systems of writing are not static. They change.

Short electronic messages are a rather specific form of writing. Short electronic messages are perceived by the originating author to be extremely temporary. One only has to sit in divorce court for a few days, however, to realize that the perceived temporary nature of short electronic messages is a fallacy. Short electronic messages are as permanent as any other digital content. However, since authors *perceive* them to be temporary, they do not put as much care and forethought into the composition of short electronic messages as they do into longer messages such as letters, reports, or blogs.

One example of short electronic messages are messages sent via cell phones or mobile phones including SMS messages (also known as text messages). When cell phones first became available to the general public, they consisted of only a numeric keypad with a handful of supporting keys to cater for symbols. Letters are obtained by pressing a specific numeric key more than once within a short period of time. Some researchers even considered the experience to be unpleasant and hard to use (Boothe, 2010). This led to many short abbreviations being used. Although modern cell phones now often have predictive text dictionaries and QWERTY keyboards, the SMS abbreviations have persisted.

Another example of short electronic messages which are perceived to be extremely temporary in nature is Twitter. Twitter is a microblogging facility where participants post 140 character tweets (or status updates) about themselves (Kwak, Lee, Park, & Moon, 2010). Because of the size limitation, Twitter abounds in abbreviations and acronyms. Recent research has also shown that Twitter users have even developed various regional dialects (Eisenstein, O'Connor, Smith, & Xing, 2010).

MXit is a communication system which uses Internet technologies over cell phones (Chigona, Chigona, Ngqokelela, & Mpofu, 2009). MXit has become South Africa's most popular social network service working on common feature phones (only requiring an Internet connection and an execution environment such as Java). MXit presents low barriers to adoption (Donner, 2010). Children and teenagers are usually introduced to MXit via other children and teenagers.

This paper will specifically deal with key identifiers found in conversations using MXit as a medium. The paper will also identify some new spelling conventions which are present in such conversations and present a linguistic basis for these conventions.

MXit-Lingo

The English term *lingo* is often considered to be a slang term for a dialect of a language or for a vocabulary of a specific industry or body of knowledge. However, published researchers have already used the term *SMS lingo* when referring to the short abbreviations used in SMS messages (Aw, Zhang, Xiao, & Su, 2006; Fung, 2005; Ng'ambi & Knaggs, 2008; Raghunathan & Krawczyk, 2009). The term *SMS lingo* has also been used in applications for patents in the United States (Dantzler Jr, Wyatt, Johnson, & Bufmack, 2009). The term *Net lingo* has been used when discussing the influence CMC (computer mediated communication) has on offline writing tasks (Wei, 2007). The term *IM lingo* has been used by published researchers when referring to the short abbreviations common to IM (Instant Messaging) conversations (Dorfmeister, 2007).

The term *MXit lingo* can, therefore, be used to describe the specialised vocabulary, spelling, syntax, and grammar which are used when communicating using MXit as a medium.

Synchronous vs Asynchronous

The terms *synchronous* and *asynchronous* can be confused when people from different disciplines use the terms.

When computer scientists speak of *synchronous communication,* they refer to the fact that when a message is sent from one computer, the receiving computer must immediately reply with some sort of acknowledgement message. To a computer scientist, the common HTTP (HyperText Transfer Protocol) is a synchronous protocol because when a user makes an HTTP request via a browser, the receiving server must reply immediately with the requested webpage. To a computer scientist, MXit is an *asynchronous* protocol. When a message is sent via MXit, the server which receives is not required to reply.

When linguists speak of the *synchronous communication,* they refer to the fact that when a human being says something in a conversation, the other person immediately replies. With reference to Sali Tagliamonte's (2008) description of online chatting, the MXit protocol is *synchronous* because it supports two human beings communicating with each other in a reciprocal (turn-taking) manner. To a linguist, a bulletin board system running over HTTP would be considered to be *asynchronous* communication. When a message is posted on a bulletin board, there is no guarantee that another person will reply.

Dr Math

Dr Math is an on-going project hosted at Meraka Institute which allows primary and secondary school pupils to converse with tutors in mathematics about their mathematics homework problem (Butgereit, 2011). The pupils use MXit on their cell phones and the tutors use traditional Internet based workstations.

The pupils usually ask their questions in this MXit lingo. Some of the questions are straight forward and simple:

> *i wnt 2 knw hw 2 fnd th nth term tht they always ask abt*

Letter Frequencies Others can be quite involved:

> *nadeem and jeny keep fit by skiping. nadeem cn skp 90 tyms pe min, when he starts training. each week he increase dis by 5 per min. jeny starts wt 60per min n increase dis by 10 per week. after hw many weeks wl deir number of skipz b the same*

The pupils can be quite energetic in stating their questions using MXit lingo:

> *hai, cn u help me simplfy((cubc root of 2000)/cubic root of 2003)*cubc root 3003 /cubc root 3000*

The Dr Math project was started in January 2007 at Meraka Institute. At that point in time, the initial research was focused on whether or not school pupils, especially high school pupils, would even want to chat about mathematics on their personal cell phones. At that point in time, there were concerns that perhaps high school pupils might not want to *contaminate*

their personal cell phones with conversations about mathematics homework. Since 2007, however, over thirty thousand pupils have used the Dr Math service.

The Dr Math research project involves minor children without parental permission. For that reason, a procedure was followed to ensure the safety and security of the minor children were protected. All tutors sign codes of conduct defining the types of conversations they can have with the pupils. All tutors also sign consent forms that their conversations can be recorded for research purposes. All pupils receive daily messages that their conversations are recorded for research purposes. An ethics clearance was issued by the Tshwane University of Technology for the Dr Math project.

The corpus of conversations which form the basis of this paper was taken from these recorded textual conversations between Dr Math tutors and pupils.

Research Methodology

In a way this research is ethnographic as the researchers describe the language of a culture-sharing group (Creswell, 2007). However, unlike a typical ethnographic study the researchers did not immerse themselves in the day-to-day lives of the studied group. The studied group, students using the Dr Math tutoring platform over MXit, were only observed in the sense that their textual interactions were recorded; no further contact with the observed group took place. As such it may share some common ground with the linguistic ethnographers that take language rather than culture as its principle point of analytic entry (Creese, 2008).

The researchers set out to identify the key identifiers found in conversations using MXit lingo. To this end the researchers employed basic descriptive statistics to identify interesting phenomena in the selected corpus. Thereafter the phenomena is qualitatively analysed and explanations for their existence in this cultural grouping offered by drawing on prior linguistic research.

The Corpus

The key identifiers and spelling conventions described in this paper are based on conversations about mathematics between pupils using MXit and tutors using conventionally sized computer workstations. These conversations occurred during the period of January 2010 through July 2011. In formulating these constructs, only the text typed by the pupil of the conversations was considered. Any statistics or numerical values presented do not include text typed by the tutor on a conventional computer workstation. Having said that, however, it is important to note that there are occasional examples of pupils retyping questions when tutors cannot understand MXit lingo.

> *wht r exponentz*
> *could you rephrase the question?*

These types of requests from tutors influence the pupils into spelling words correctly and in full. In addition, there are situations where the pupil voluntarily starts using traditional English spelling rules to help the tutor better understand. These types of situations affect the vocabulary that the pupils use. These examples have been retained in the corpus. The result

of this is that some of the statistics provided by this MXit language construct may not hold true for conversations where both parties are conversing in MXit lingo.

The corpus was manually cleaned in the following manner:

1. The Dr Math tutoring platform has a CAPTCHA (Completely Automated Public Turing Test to tell Computers and Humans Apart) challenge which a pupil receives as his or her first message when contacting a tutor in a given day. The CAPTCHA challenge and the response that the pupil types in have been removed from the corpus.
2. The Dr Math tutoring platform has a number of *bots* or automated facilities which are available to the pupils. Occasionally pupils attempt to access these *bots* but their requests are badly formed and are sent to tutors instead. These failed attempts to access the *bots* or automated replies have been removed from the corpus.
3. Messages of zero length have been removed from the corpus.
4. Any messages that may have information which may identify any of the minor children (such as telephone numbers or names) have been removed.
5. The statistical analysis was conducted in a case insensitive manner.

After manually cleaning, the corpus contained 48,858 messages which had been sent from pupils to tutors using MXit.

In analysing this corpus it was important to clarify the definition of a *word*. The definition that was used for this step of the research was a sequence of letters and/or digits and/or the handful of specific symbols including @ or $. The remaining symbols including a space were considered to be word delimiters.

Letter Frequencies

The frequencies of letters in the MXit based corpus differ from the frequency of letters in traditionally spelled English documents. Data columns 1 and 3 (English Rank and English Frequency) in Table 1 show the rankings and frequencies of letters in traditionally spelled English documents as reported by Solso and King (1976). Those frequencies are based on a corpus of approximately one million words. All hyphenated words, words containing apostrophes and numbers were excluded.

Data columns 2 and 4 (MXit Rank and MXit Frequency) in Table 1 show the letter rankings and frequencies found in the Dr Math MXit based corpus. Words containing embedded numerals such as *n0t* and *h3llo* were retained but only the letters were counted.

The differences in the frequencies and the percent change in the frequencies are displayed in data columns 5 and 6. Letters where the absolute value of the percent change is greater than 10% are highlighted in **bold** print.

As can be seen from Table 1, the frequency of sixteen letters varied by more than 10% (positive or negative) between English and the MXit corpus for Dr Math. The sixteen differences are the letters *A, C, E, F, G, J, K, M, O, Q, R, U, W, Y, X,* and *Z*. Each of these letters will be discussed in detail.

Letter	English Rank	MXit Rank	English Frequency (as %)	MXit Frequency (as %)	Frequency Difference	% Change in Frequency
A	**4**	**3**	**7.61**	**8.43**	**0.82**	**10.78**
B	20	20	1.54	1.56	0.02	1.30
C	**12**	**15**	**3.11**	**2.61**	**-0.50**	**-16.08**
D	11	11	3.95	3.91	-0.04	-1.01
E	**1**	**1**	**12.62**	**9.63**	**-2.99**	**-23.69**
F	**15**	**19**	**2.34**	**1.75**	**-0.59**	**-25.21**
G	**17**	**17**	**1.95**	**2.15**	**0.20**	**10.26**
H	9	8	5.51	5.07	-0.44	-7.99
I	5	4	7.34	7.48	0.14	1.91
J	**24**	**26**	**0.15**	**0.25**	**0.10**	**66.67**
K	**22**	**21**	**0.65**	**1.46**	**0.81**	**124.62**
L	10	10	4.11	4.21	0.10	2.43
M	**14**	**13**	**2.54**	**3.38**	**0.84**	**33.07**
N	6	5	7.11	7.10	-0.01	-0.14
O	**3**	**6**	**7.65**	**6.83**	**-0.82**	**-10.72**
P	16	18	2.03	2.06	0.03	1.48
Q	**25**	**24**	**0.10**	**0.45**	**0.35**	**350.00**
R	**8**	**9**	**6.15**	**4.73**	**-1.42**	**-23.09**
S	7	7	6.50	6.02	-0.48	-7.38
T	2	2	9.33	9.25	-0.08	-0.86
U	**13**	**12**	**2.72**	**3.67**	**0.95**	**34.93**
V	21	23	0.99	1.04	0.05	5.05
W	**18**	**14**	**1.89**	**2.84**	**0.95**	**50.26**
X	**23**	**22**	**0.19**	**1.18**	**0.99**	**521.05**
Y	**19**	**16**	**1.72**	**2.55**	**0.83**	**48.26**
Z	**26**	**25**	**0.09**	**0.43**	**0.34**	**377.78**

Table 1: Comparison Letter Ranks and Frequencies English vs. MXit

The Letter A

The higher percentage frequency of the letter <A> in the MXit corpus when compared to English is due to two reasons. One reason is that the letter <A> is used often in mathematics to represent areas, angles, and other measurements in analytical geometry. The following examples illustrate the high use of the letter <A> in mathematics messages:

Can i relate that to cos (a + b)
Wats the first sTep to do when we fynd such a problem: a^2 -(b+c)^2
find the products 5a-[3a-4{a-2(a+5)} -(2a+3)]

Another reason is the high usage of the word <da> in place of the word <the> and the high use of the suffix <-a> in place of the suffix <-er>. For example:

but da question is even on one dice and odd on da oda dice
and da form for da nth trm of geomtr seq?
And wat abt da uda 3 rules?

These two phenomena also affect the statistics for the letter <R> and the letter <E>.

The Letter C

The letter <C> has a lower percentage in the MXit research corpus than in the traditionally spelled English corpus. The reason for this is that the hard <C> sound is often spelled with a <K> in MXit lingo. For example:

at skul m strtin 2 do it so whtz it al abt
nah itz kwl so hw du u solve simaltaniouz equationz
area of sirkle pls

This also affects the frequencies of the letter <K> when compared to Standard English.

The Letter E

Although <E> has the same top frequency and rank in both corpora, the lower percentage in the MXit based corpus can be explained by two MXit conventions. The first convention is that the numeral <3> is often used in place of the letter <E> in words such as <n3d> (need), <b3low> (below), <b3n> (been), <h3lo> (hello). The second convention is that the letter <E> is often omitted or exchanged for a different vowel. For example, although the word <the> occurred 7,248 in the corpus, the alternate spelling of <da> appeared 2,252 times contributing to a lower percentage of the letter <E> in the corpus. Another example is the word <need> which appeared 1055 times in the corpus and an alternate spelling of <nd> which appeared 1,100 times. Table 2 provides a number of similar examples of alternate spellings of common words where the letter <E> is omitted thereby contributing to the lower percentage of the letter <E> in the corpus.

Word	Count
have	988
hav	365
hv	515
the	7,248
da	2,252
time	254
tym	41
line	241
lyn	108
like	567
lyk	477
need	1,055
nd	1,100
help	2,459
hlp	454
please	582
pls	374
plz	717

Table 2: Common words with <E>s and their alternate spellings

The Letter F

The letter <F> has a lower percent frequency in the Dr Math MXit based corpus than in traditional English. This is due to a number of common English words which contain the letter <F> being spelled differently in MXit.

The wordfrequency.info website sells word list frequencies (Davies, 2012). However, it also releases free of charge one list of word frequencies which is based on the top 5 000 words in American English. According to the website, this list of 5 000 words is gleaned from a corpus with over 400 million words.

According to wordfrequency.info (Davies), the English word <of> is the fourth most common word used in the English language. In the Dr Math MXit corpus, however, the word <of> is only the sixth most common word. Often, pupils use just the letter <O> to represent the entire word <of>. For example:

a youn couple decided to take a bank loan o 1200000
in oder to purchase a hme
wat r laws o fractorization?
i meant o the quadratic

In addition, wordfrequency.info (Davies) states that the word <for> is the 13th most common word in traditional English and the word <first> is the 86st most common in traditional

English. In the Dr Math MXit corpus, the word <for> is 31st most common word and the word <first> is merely the 181st most common word. This is due to the words often being spelled as <4> and <1st> respectively. For example:

lets wrk wit d 1st 1 it luks mre easier 2 me
im struggle hw 2 find derivative from 1st principle
u luk 4 da lcd 1st
i 4got da exponent lawz

All of these alternate spellings contribute to the lower percentage frequency of the letter <F> in the Dr Math MXit research corpus.

The Letter G

The letter <G> has a higher percentage frequency in the Dr Math MXit corpus. There are two reasons for this. The first reason is the common MXit initialisms or acronyms such as <omg> (indicating Oh My God or Goodness), <g2g> and <gtg> (indicating Got To Go), and various methods of indicating a grin or smile such as <(g)>. For example:

um... explain hyperbola to me(g)
cant u check it 4 me(g)

The second reason is mathematical. The letter <G> is often used in algebra to indicate a second function in and expression. The letter <G> is also in common mathematical terms such as graph, trigonometry and angle. For example:

ja plz m given an equation wch says g(x)=sin(x+30)and
h(x)=tan(1/2x) and m supose 2 find da domain of g(x)

The Letter J

The letter <J> has a percentage occurrence which is substantially higher in the Dr Math MXit corpus. This is due to both the multilingual culture of South Africa and the fact that many Afrikaans speaking pupils use the English based Dr Math tutoring system. Words such as <ja>, <jah>, and <jammer> are common in the Dr Math MXit corpus. The simple letter <J> often represents the Afrikaans word <ja> (yes) or <jy> (you). Even though Dr Math is an English based tutoring system, these abbreviations abound. The top five words containing the letter <J> can be found in Table 3. In addition, Afrikaans words such as <jy>, <jou>, <jep>, <joh>, <jip>, <jap>, <julle>, and <jes> appear in the corpus.

Word	Count
Just	514
Jst	420
Ja	394
Jah	103
J	60

Table 3: Top five words containing <J>

The Letter K

The letter <K> has a substantially higher percentage frequency in the Dr Math MXit corpus than in traditional English. This is due to two reasons. One reason is that the letter <K> is

often used to spell a hard <C> sound in words such as <school> (often spelled <skewl>). Another reason is the high prevalence of the various spellings of <OK> and <kewl>. In addition, because of the nature of the Dr Math tutoring environment, the word <ask> plus all its various derivations had a much higher rank in the MXit corpus. Table 4 provides a handful of examples of words containing the letter <K>.

Word	Count
Ok	1,483
k	1,062
ohk	375
kwl	137
Owk	95

Table 4: Examples of words containing the letter <K>

The Letter M

The letter <M> has a higher percentage occurrence in the MXit research corpus than in the traditional English corpus. There are two reasons for this. The first reason is a characteristic of the Dr Math tutoring environment where terms such as <mathematics> occur often. In addition the letter <M> is often used in mathematics to indicate the slope of a line.

The second reason is the pupils indicate that they are thinking about their problems by using words such as <mmm>, <hmmmm> and <uhmmm> with varying quantities of the letter <M>. For example:

mmm..w0ts x1 plus 16=19
? uhmmmm... i'm lost now
hmmm, now i've lost you
ummm abit c0nfusd

The Letter O

The letter <O> has a lower percentage occurrence in the MXit research corpus than in the traditional English corpus. This is due to the numeral <0> (zero) often being used in place of the letter <O> (oh) in words such as <*n0t*> (not), <questi0n> (question), <h0w> (how), <kn0w> (know). Looking through the entire MXit corpus and only looking at unique words (and not their frequency), there was a total of 1,336 unique words containing the vowel <O> but there were 1,255 unique words containing the numeral <0> in the place of the vowel <O>. If the frequency of these words is also taken into account (in other words, taking into account that the word <n0t> is used often by pupils), the frequency of the letter <O> combined with the numeral <0> when it indicates a vowel increases to 7.08%. In addition, the letter <O> is often omitted in common words as can be seen in Table 5.

Word	Count
how	1,945
hw	1,224
dont	504
dnt	1,195
know	605
knw	844
now	511
nw	665

Table 5: Common words with <O>s and their alternate spellings

The Letter Q

The letter <Q> has a substantially higher percentage frequency in the Dr Math MXit corpus when compared to traditional English. This is due to the nature of the conversations between pupils and tutors about mathematics. The highest ranging words containing the letter <Q> can be found in Table 6.

Word	Count
Question	576
Equation	548
Square	298
Equations	288

Table 6: Highest Ranking words containing the letter Q

The Letter R

The letter <R> has a lower percentage frequency in the Dr Math MXit corpus. This is due to the suffixes <-er> and <-or> being replaced by a simple <-a>. For example:

its a numbaz wif powas
i dnt undastand
i jst knw dat sin ova cos is tan

The Letter U

The letter <U> has a much higher percent occurrence in the Dr Math MXit corpus than in traditional English. This is due to two reasons. The most important reason for the use of the letter <U> on its own is to indicate the word <you>. According to the wordfrequency.info website (Davies, 2012), the word <you> is the 14th most common English word. In the Dr Math MXit corpus, the word <u> indicating the word <you> is the 3rd most common word in the entire corpus.

Another reasons for the increased percent occurrence of the letter <U> is its presence in common mathematics and tutoring terms such as <question>, <equation>, <number>, <calculate>, <formula>, <sum>, and <square>.

The Letter W

The letter <W> has a substantially higher percentage occurrence in the MXit research corpus than in the traditional English corpus. This is partially due to the MXit corpus being taken from a tutoring environment where the pupils asked questions beginning with words such as <what>, <why>, <how>? This phenomenon may not hold true for other MXit based corpora. Another reason for the increase in the use of the letter <W> is the word <mwah> which is used often in MXit based conversations and represents the sound of a kiss.

The Letter X

The letter <X> has an extremely higher percentage occurrence in the Dr Math MXit research corpus than in traditional English. This is also due to two reasons. The first reason is the nature of the mathematical conversations between Dr Math and pupils. The letter <X> is often used in mathematical formulae and when discussing the Cartesian coordinate system. For example:

Y is equal a.x^2 added to q
What is x if: 2x=3-(-8/2x)
Solv simultaneously for x and y X+y=5 and x-y=3

The second reason is a characteristic of the MXit medium where people often express affection and thanks with virtual "kisses and hugs" which are represented as a string such as <xoxoxo>. For example:

Bye:)xxx love you :)
Bye :-) xoxo :-)

The Letter Y

The letter <Y> has a substantially higher percentage occurrence in the MXit research corpus than in the traditional English corpus. This is due to two reasons. The first reason is that the English word <why> is commonly abbreviated to simply <Y> in MXit lingo. As can be seen in Table 7, the word <Y> is the third most frequent word containing the letter <Y>.

Word	Count
You	2,220
My	1,590
Y	1,407
Yes	1,353

Table 7: Top ranking words containing <y>

The second reason is the nature of the mathematical discussions taking place with Dr Math. The letter <Y> is used often in mathematical formulae and is one of the axes in the Cartesian coordinate system. For example:

Y=x+4/2x-5 i want an inverse of y
Y+2x=2 y^2+2x^2=3yx
In y=mxPlus c

The Letter Z

The letter <Z> has an extremely higher percentage occurrence in the Dr Math MXit research corpus when compared to traditional English. This is because of the nature of the mathematical conversations where <Z> is a common variable name. For example:

You get z= 226?
It says; sum of n terms z=a('r^n'-1) ol dvde by r-1
3x+y-z+4x-2y+9z

Another reason is that the letter <Z> is often used in place of the word <is>. For example:

Wat z a radi ?
ok ths z confusn me
Wht type of a sequence z ds

Letter Frequencies Conclusion

This first portion of the paper presented statistics about letter frequencies which are present in MXit lingo as shown in conversations obtained from the Dr Math project. These statistics were compared to statistics which are present in traditionally spelled English. The next section of the paper will deal with the linguistic aspects of MXit lingo.

Vowel Reduction

Vowels are optional in longer words in MXit lingo as long as there are enough consonants to retain the recognisability of the word. A word such as <intercept> can be found in the MXit corpus and can be spelled in a number of different ways with varying numbers of vowels including <intercpt>, <intrcept>, and <intrcpt>.

A very short word such as <me>, however, will usually be written in full. For example:

cn u hlp me wth maths

This lack of vowels is not unique to MXit lingo. There are numerous human languages where vowels are optional in written form. These languages are known as abjad or consonantary languages (Golcher, 2007). Both Hebrew and Arabic are examples of abjad languages and have optional vowels in their written form (Abu-Rabia, 2001). The vowels in these languages can be omitted or they can be written as a type of diacritical mark. This is similar to English writings dotting the letter <i> or crossing the letter <t>.

Al-Sa'adi and Hamdan (2005) explain that vowels are often left out of words as long as the product (consisting mostly of consonants) is still able to express meaning of the full word. It is after consulting Al-Sa'adi and Hamdan that Shaw (2009) prefers to refer to this reduction of vowels as *abbreviation*. Therefore, the source word <equation> and the abbreviated word <equatn>, or even <eqtn>, are examples of abbreviations. The abbreviation of words is often

a space-saving mechanism and a character-saving mechanism. The more words which can be shorted, then the more words can be squeezed into a message that allows a limited, fixed number of characters available to the user. SMS (Short Message System) messages also inhibit the length of exchanges.

Letters Exchanged for Similar-sounding Letters

Letters are often exchanged for other letters that sound similar. This is true for both vowel sounds and consonant sounds.

hw do u workout tha smultinious equation

ther r no squres in ma fone

an irrashunal

i stil cnt fnd th ryt answr

no i dnt hav any bcoz de skulz r clozd

Number or letter homophones are numerals or letters, respectively, which are sounded the same as syllables or whole words. Al-Sa'adi and Hamdan (2005) distinguish between two forms of homophone engineering. By *homophone engineering*, the writers are referring to MXit users' use of similar-sounding letters and numerals in the place of whole words and syllables. One form of homophone engineering is when a letter or numeral that sounds the same as a word replaces the *whole word*. MXit users regularly replace the word <are> with the letter <r>, the word <why> with the letter <y>, and the word <you> with the letter <u>. This form of homophone engineering is called *letter homophone* usage by Shortis (2007), a student of Al-Sa'adi and Hamdan.

The other form of homophone engineering is when a letter or numeral that sounds the same as a syllable or part of a word replaces a syllable or part of a word. <Late> becoming <l8> is an example of the vowel sound /ei/, usually written as <-a-e> being replaced by the numeral <8>. <Great> being written as <gr8> is another example of what Shortis (2007) prefers to call a *number homophone*. <Before> being written as <b4> and <tomorrow> being written as <2morrow> are two more examples of number homophones.

It is important to note that these extracts are taken from conversations between pupils and tutors. These sentences and questions were created in a conversational manner. They were not written as a document. In computer-mediated spaces, like chat rooms and instant-messaging platforms, where users take turns to write and respond, *synchronous* (from a linguistic point of view) communication takes place (Al-Sa'adi & Hamdan, 2005; Shaw, 2009). They were written as communication between two people in a timely manner. The conversations are fast paced.

Th Written as D or F

Given the popularity in South Africa of African-American entertainment media, such as rap music tracks and music videos, it is not surprising that South African MXit users can imitate the unique pronunciation of what most linguists refer to as African-American Vernacular English (AAVE). It is with reference to 'counter-cultural' symbolism that has been

cultivated by popular African-American artists and entertainers that scholars like Deumert and Masinyana (2008) suggest that the imitation of the AAVE accent exists, and influences the spelling of CMC texts produced by South African youths. This may certainly be true, from an anthropological perspective. However, from a sociolinguistic perspective, it is reasonable to assume that South Africans (the majority of whom happen to be black) possess their own variety of spoken and written English, and do not only imitate selective elements of a variety of English that is owned by African Americans. There is, fortunately, well-established research that describes the pronunciation of Black South African English (BlSAfE). According to Van Rooy (2000), there is in fact an absence of the voiced [th] sound from BlSAfE. This implies that the [th] sound, as in [that] and [then], is not familiar to BlSAfE speakers. The closest approximation of the voiced [th] sound that speakers of BlSAfE can produce is the [d] sound, phonetically and orthographically (in writing). Therefore, the words <de> can appear for the word <the>, and the word <dose> can appear for the word <those>.

It is important to note at this point that the English spelling of <th> for the sound at the beginning of the word <the> and <this> is not a phonetic spelling. It has nothing to do with the sound of the letter <t> or the sound of the letter <h>. It is a spelling convention which has been adopted. As written English has developed over time, the [th] sound has previously been spelled in English with the letters <þ> and <ð> (Gregory, 2011).

New Suffix -a

With MXit lingo, the English suffixes of <-er> or <-or> are often replaced by <-a>. What follows is an explanation for the denoted vowel change. The schwa is the name of the vowel *sound* that is voiced with one's mouth open but without moving one's tongue or lips. The phonetic symbol for the schwa is /ə/. Although this sound is not represented in the 26-letter Roman alphabet employed by English, it is, according to McArthur (1998), the most common vowel sound in English. Instead of a single symbol to denote the /ə/ sound in written English, the suffix <-er> often denotes the /ə/ sound, as in <mother> and <number>. The <-or> suffix at the end of <doctor> and <motor> also denotes the /ə/ sound.

However, McArthur (1998) asserts that not all languages use the /ə/ sound, and he mentions African languages in particular. Therefore, if African-language speakers are not familiar with producing the /ə/ sound, the schwa would be absent from the African varieties of English they speak. That the /ə/ sound is not a feature of Black South African English (BlSAfE) is established by Van Rooy and Van Huyssteen (2000). They explain that the /ə/ sound from SAE is replaced by a different vowel sound by BlSAfE speakers. For example, the [a] sound in [wash], as pronounced by a speaker of SAE, is the closest approximation of the [a] sound BlSAfE speakers use to replace the /ə/ sound. This means that by using the special /a/ sound, speakers of BlSAfE would say [numba] instead of [numbə], or [ova] instead of [ovə]. According to Mesthrie (2005), among speakers of English as an additional language, there is an inclination to write words according their (non-standard) pronunciation. This means that the MXit users of BlSAfE write <numba> instead of <number>, and <ova> instead of <over>. Replacing the /ə/ sound that occurs in the middle of polysyllabic words with the different vowel sound is also a feature of BlSAfE. Thus the word <understand> can be written as <undastand>, and <another> written as <anada>. For example:

i dnt undastand
docta wat is sin ova cos
tl me wat a prime numba is

The top occurring words in this corpus using this suffix are <ansa>, <neva>, <ova>, <afta>, <anada>, <numba>, <eva>, <anoda>, <answa> and <otha>.

This <-a> suffix is not unique to MXit lingo. Linguists have categorised languages as being rhotic languages or non-rhotic languages by the way the speakers handle the letter <r> (Lindau, 1980). With rhotic languages, speakers pronounce a written <r> and with non-rhotic languages, speakers do not pronounce a written <r>. Within languages, however, there can be different accents or versions of the pronunciation of a sound. Such variations are called *allophones*. According to Van Rooy (2000), it is the occurrence of allophones for the pronunciation of the letter [r] that makes it difficult to define the true rhoticity of BlSAfE. In the interim however, Hartmann and Zerbian (2009) have managed to show that female speakers of BlSAfE speak it with more rhoticity than do their male counterparts. The ethical clearance of the Dr Math project prevents collecting demographic data about the users. Therefore a supporting and detailed explanation of the rhoticity of BISAfE cannot be supported by our data.

The MXit users spell under the influence of BlSAfE because of a universal, sociolinguistic element that encourages the use of non-standard forms of spelling. According to Sebba (2003), this sociolinguistic principle is the desire amongst youth not to conform to all conventions or accepted norms. Sebba (2003) goes on to explain that because written CMC tools of casual and social interaction (like MXit) are not regulated, some space for accent and spelling variation is created by and for users of such tools. When there is space to behave unconventionally, people, especially youths, will make a point of behaving unconventionally, according to Sebba (2003) and Shaw (2009). In intentionally deviating from (rebelling against) the norm, that is, the regularised spelling and writing that is demanded in scholarly contexts, MXit users are able to construct social identities for themselves.

Regularisation of Irregular Spelling

The regularisation of irregular spelling, as identified by Shaw (2009) includes:

<k> replacing hard <c> (crazy, krazy)

<s> replacing soft <c> (circle, sircle)

<z> replacing <s> (ladies, ladiez)

<ee> replacing <ea> (please, pleez)

<f> replacing <ph> (phone, fone)

<-shun> replacing <-tion> and <-sion> (addision, addishun)

<w> replacing <wh> (what, wat)

<-i-e> replacing <-igh-> (night, nite)

<w> replacing <-ou-> (out, owt)

Double letters need to be only written once to save space and typing effort. In Shaw's (2009) study of the regularisation of irregular spelling, he states that for the sake of transparent and

accommodating products of CMC, single consonants often replace sets of double consonants. Thus a word such as <small> can be written as <smal> or <borrow> can be written as <borow>. For example:

isnt range wen u subtract da smal value 4rm big 1
i have borowd 15000 n i hav 2 pay it aftr one year

Numerals and Symbols for Specific Sounds

Numbers and symbols can be placed in words to represent certain sounds. Linguists call these substituted sequences homophones (Shortis, 2007). For example, <l8r> can represent the word <later>, <d@> can represent the word <that>, <2morrow> can represent the word <tomorrow>

un4turn8ly

ok thx d@s gr8

w8 da 16 is multiplied by x

f9 thnks nd u

Numerals and Symbols for Specific Shapes

Specific numerals and symbols often replace specific letters. This is due to the shape of the numeral or symbol and not the sound of the number or symbol. Numerals which replace letters of a similar shape are usually called "homographs". Moran (2011) defines a "graph" as the most basic and single unit of a writing system. A *homograph* is thus a unit of writing (or graph) that has the same appearance as another unit with a different function. When writers of CMC texts use numeral graphs that resemble letter graphs to replace letter graphs, they are using "leetspeak'". Leetspeak is often typed out as, but not limited to, 'l337 5p34k'. Tavosanis (2007) explains that leetspeak, derived from "elite speak", is a term that was originally devised to describe the unique encoding used by online gamers.

In the MXit corpus the *zero* numeral <0> often occurs as a replacement for the letter <o>. That this occurs more often in this context, where Mathematics is the subject of the computer-mediated conversation, than it occurs in contexts where Mathematics is not the subject of conversation is an investigation that exceeds the scope of this paper.

i hv prblms wth the p0int 0f circle whch is n0t at 0rigin

iv g0t a prblm wth calculus

i n3d hlp on trigonometry

h3llo

Cheer$

Then ten most common words where numerals replace letters that are shaped the same are: <*not*>, <*l0l*>, <*questi0n*>, <*h0w*>, <*kn0w*>, <*n0u*>, <*n0w*>, <*g0t*>, <*d0nt*> *and* <*equati0n*>.

Acronyms, Initialisms, and Acrostics

MXit lingo is peppered with short acronyms, initialisms and acrostics which represent phrases.

LOL	Laughing Out Loud
G2G	Got to Go
OMG	Oh My God/Goodness
BRB	Be Right Back
WUD	What you doing?

This is not a new phenomenon in English. There are many acronyms, initialisms, and acrostics that have existed in the English language prior to chat protocols such as MXit.

Benelux	Belgium, Netherlands, and Luxemburg
Nato	North Atlantic Treaty Organisation
Scuba	Self contained underwater breathing apparatus
Laser	Light amplification by stimulated emission of radiation
Aids	Acquired Immune Deficiency Syndrome

Acronyms, initialisms and short forms have been used for thousands of years. Many Christians believe that the Greek word ΙΧΘΥΣ (fish) was an acrostic for **Ἰ**ησοῦς **Χ**ριστός, **Θ**εοῦ **Υ**ἱός, **Σ**ωτήρ which translates Jesus Christ, God's Son, Savior (Kidwell & Faiman-Silva, 2001; Tenšek, 2005).

The first documented written use of OMG can be found in the memoirs of Lord Fisher, Admiral of the Fleet. In 1917 he wrote to the Right Honourable Winston Churchill (as he was then known) "...I hear that a new order of Knighthood is on the tapis – O.M.G...." (Fisher, 1919).

Conclusion

MXit is a communication system which utilises Internet technologies over cell phones and is used primarily by young people (Chigona et al., 2009). MXit presents low barriers to adoption (Donner, 2010) and children and teenagers are usually introduced to MXit by other children and teenagers. Much has been written in the academic press and popular press arguing whether or not the use of MXit will enhance or harm traditional literary skills (Considine, 2004; Vosloo, 2009; Wei, 2007). As such, the study of MXit language or lingo is important to future research.

This paper presented a statistical analysis of letter frequencies obtained from a corpus of conversations held using MXit in a mobile mathematics tutoring environment. After presenting the statistical information, the paper then explored various linguistic and sociolinguist characteristics of communication using MXit. The paper argues that the language used in MXit is not just an arbitrary shortening of text, but follows specific rules and patterns. As such it can be argued that MXit language evolved (and still is evolving) in ways consistent with that found in the evolution of English language varieties.

The presence of rules and patterns in MXit language suggest several fields for further exploration. Researchers in the field of Natural Language Processing may be able to process the MXit language to translate it to 'normal' English. In tutoring environments, like Dr Math, this would enable tutors unfamiliar with the MXit language to also serve as tutors. Further analysis of the MXit text could be used to provide additional help to tutors, for example by providing similar conversations or links to topical areas. Specifically in the Dr Math environment being able to separate MXit text from mathematical formulas could assist tutors further by reformatting mathematical formulae in a more understandable way.

This paper represents a first step to understanding and explaining the MXit language and provides a solid foundation for further work in this language which is widely used among South African youth.

References

Abu-Rabia, S. (2001). The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew. *Reading and Writing, 14*(1), 39-59.

Al-Sa'adi, R. A., & Hamdan, J. M. (2005). "Synchronous online chat" English: Computer-mediated communication. *World Englishes, 24*(4), 409-424.

Aw, A. T., Zhang, M., Xiao, J., & Su, J. (2006). A phrase-based statistical model for SMS text normalization. *Proceedings COLING/ACL, July 17-21, 2006, Sydney, Australia,* 33-40.

Boothe, P. (2010). Using cell phone keyboards is (NP) hard. *Proceedings Fun with Algorithms: 5th International Conference, FUN 2010, Ischia, Italy, June 2-4, 2010,* 53-67.

Butgereit, L. (2011). C^{3}TO: a Scalable Architecture for Mobile Chat Based Tutoring. Unpublished Masters of Technology, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa.

Chigona, W., Chigona, A., Ngqokelela, B., & Mpofu, S. (2009). MXIT: Uses, perceptions and self-justifications. *Journal of Information, Information Technology, and Organizations, 4*, 1-16.

Considine, D. M. (2004). Linking the literacies: Teaching & learning in a media landscape. *Wisconsin State Reading Association Journal, 44*(5), 49-53.

Coulmas, F. (1991). *The writing systems of the world*. Massachusetts: Blackwell Publishers.

Creese, A. (2008). Linguistic ethnography. *Encyclopedia of Language and Education, 10*, 229–241.

Creswell, J. W. (2007). Qualitative inquiry and research design: Choosing among five approaches. *Thousand Oaks.*

Dantzler Jr, T. C., Wyatt, J. A., Johnson, R. C., & Bufmack, N. V. (2009). Hands free messaging, US patent application 2010/0298009 AI.

Davies, M. (2012). Top 5 000 lemmas. Word Frequency Lists and Dictionary from the Corpus of Contemporary American English. Retrieved from http://www.wordfrequency.info/5k_lemmas.asp

Deumert, A., & Oscar Masinyana, S. (2008). Mobile language choices the use of english and isiXhosa in text messages (SMS) evidence from a bilingual south african sample. *English World-Wide, 29*(2), 117-147.

Donner, J. (2010). Framing M4D: The utility of continuity and the dual heritage of" mobiles and development". *The Electronic Journal of Information Systems in Developing Countries, 44*(0).

Dorfmeister, J. (2007). Computer mediated communication. *Trends: A Collection of Literature Reviews Written by the MED Students from EDU 612, 1*(Spring), 26-33.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* , 1277-1287.

Fisher, B. J. A. F. (1919). *Memories.* London: Hodder and Stoughton.

Fung, L. M. (2005). SMS short form identification and codec. Unpublished Honours, National University of Singapore, Singapore.

Golcher, F. (2007). A stable statistical constant specific for human language texts. Presentation made at *International Conference RANLP – 2007, September 27-29, Borovets, Bulgaria*.

Gregory, G. (2011). Teaching and learning about language change (part one). *Changing English, 18*(1), 3-15.

Kidwell, B., & Faiman-Silva, S. (2001). Ukrainian egg art and customs. Unpublished course notes for AN 110-01 Introduction to Folklore, Bridgewater MA: Bridgewater State University.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web, April 26-30, 2010, Raleigh, North Carolina, USA,* 591-600.

Lindau, M. (1980). The story of /r/. *The Journal of the Acoustical Society of America, 67*(A1), S27.

McArthur, T. (Ed.). (1998). *Concise Oxford companion to the English language.* Oxford: Oxford University Press.

Mesthrie, R. (2005). Putting back the horse before the cart: The "spelling form" fallacy in Second Language Acquisition studies, with special reference to the treatment of unstressed vowels in Black South African English. *English World-Wide, 26*(2), 127-151.

Moran, S. (2011). An ontology for accessing transcription systems. *Language Sources and Evaluation*, *45*(September), 345-360.

Ng'ambi, D., & Knaggs, A. (2008). Using mobile phones for exam preparation. *Proceedings of the IADIS Mobile Learning Conference, April 11-13, 2008, Algarve, Portgual,* 35-42.

Raghunathan, K., & Krawczyk, S. (2009). CS224N: Investigating SMS text normalization using statistical machine translation.

Schmandt-Besserat, D. (1996). *How writing came about*. Texas: University of Texas Press.

Sebba, M. (2003). Spelling rebellion. *Discourse Constructions of Youth Identities, 110*, 151.

Shaw, P. (2009). L8r or l8a? rhoticity variation in computer-mediated communication. *Corpora and Discourse–and Stuff,* , 267-276.

Shortis, T. (2007). Gr8 txtpectations. *English Drama Media,* (June 2007), 21-26.

Solso, R. L., & King, J. F. (1976). Frequency and versatility of letters in the english language. *Behavior Research Methods, 8*(3), 283-283-286.

Tagliamonte, S. & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech, 83*(1), 3-34.

Tavosanis, M. (2007). A casual classification of orthography errors in web texts. *Proceedings of the International Joint Conference on Artificial Intelligence - Workshop on Analytics for Noisy Unstructured Text Data*, 99-106.

Tenšek, T. Z. (2005). The theology of images with a special emphasis on the patristic period. *Bogoslovska Smotra, 74*(4).

Van Rooy, B. (2000). The consonants of Black South African English: Current knowledge and future prospects. *South African Journal of Linguistics, Supplement 38, 35-54.*

Van Rooy, B. & Van Huyssteen, G. (2000). The vowels of Black South African English: Current knowledge and future prospects. *South African Journal of Linguistics, Supplement 38, 15-35.*

Vosloo, S. (2009). The effects of texting on literacy: Modern scourge or opportunity? *Shuttleworth Foundation,* 2-6.

Wei, K. C. (2007). The impact of using net lingo in computer mediated communication on off-line writing tasks.