

# Text Messaging: a tool in e-Health Services

Ademola O. Adesina  
inadesina@gmail.com

Kehinde K. Agbele  
agbelek@yahoo.com

Henry O. Nyongesa  
henrynyongesa@gmail.com

University of the Western Cape, Dept. of Computer Science,  
Natural Language Research Group,  
Cape Town, South Africa

**Abstract:** HIV/AIDS continues to be a menace to the global community, especially in sub-Saharan Africa and South Africa is not an exception. The infection rate is continues to grow, in particular, among the young adults. Cell phones have been identified as one of the tools that can be used to overcome the challenge of information dissemination regarding HIV/AIDS among young adults because of its acceptability within this age group. Access to appropriate information can be a powerful for prevention and management of many chronic illnesses, including hypertension, diabetes and HIV/AIDS. Within the young adults age group information access by use of Short Messaging Services (SMS) becomes particularly appealing. In this regard, it is proposed to provide access to carefully screened information on HIV/AIDS within the context of frequently asked questions (FAQ) system. However, automating SMS-based information search and retrieval poses significant challenges because of the inherent noise in SMS communications. In the paper, a special corpus of SMS messages was collected based on a standardised question-answer collection. The SMS messages were then analysed, transcribed and classified, with the aim of building a dictionary of SMS-speak to English translations, with reference to HIV/AIDS.

**Index Terms:** HIV/AIDS Management, SMS Communication, Mobile Information Access.

## I. INTRODUCTION

Text messaging has become ubiquitous, and it is used wherever mobile phone services are available. For many users, mobile phones function first as text messaging devices, and secondly as voice calling devices [1, 3]. Mobile phone users can maintain communication during situations in which a voice call is impractical, impossible or unacceptable. Text messaging has also provided avenue for participatory culture, as well as access to information on the move. It can also bring people together and create a sense of community through ‘Smart Mobs’ [4, 5]. This research has proposed application of text messaging as a means of access to health care information. Patients with chronic illnesses, such as, HIV/AIDS, cancer and diabetics can benefit from access to appropriate information for care and management of their conditions. This is more so in Africa, where levels of illiteracy and ignorance are relatively high.

On one hand, developing such an application inherently requires building of a specialized corpus of medical terminologies and information. On the other hand, accessing such information through SMS communication causes significant challenges because of the noisy nature of SMS text messages. SMS messages are associated with non-formal writing styles, including colloquialism, misspellings and homophonic abbreviations [12]. Such communication needs to be parsed into formal language, for example English, for it to be amenable to information search retrieval. In this paper, an exemplar corpus SMS messages was collected as an initial step to analysis of SMS communication, in order to build a SMS-English parser. We describe a simple experiment to collect sample SMS writing styles from a group of first year students at the University of the Western Cape. We also describe the proposed schematic for the SMS-English parser.

The rest of the paper is organised as follows. Section 2 elaborates different methodologies adopted in collecting SMS. Section 3 reveals the pre-processing techniques for the corpus and classification of various SMS collected. In section 4 the schematic model of the proposed SMS-query health care information system was discussed. Results and Discussion were centred on in Section 5. Finally, the conclusion was centred on in Section 6.

## II. METHODOLOGY

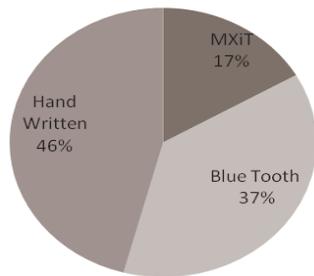
Short Messaging Service (SMS) messages were collected from a group of first year computer science students totalling 50 and a set of 25 questions were administered. The SMS communication appears in different forms because of flexibilities associated with this form of communication. A set of pre-formed questions related to HIV/AIDS were provided for all participants. The participants were then required to re-write the same questions, assuming they were personally sending the same question via an SMS message. Three different methods used in collecting question data sets: the use of the MXit platform, transfer of data via blue-tooth and hand-written on paper. MXit [7] is a free, instant messaging application developed in South Africa that runs on 3G General Packet Radio Service (GPRS/3G) mobile phones and on PCs. This technology allows the user to send and receive one-on-one text and multimedia messages to and from other users, as well as in general chat rooms. For all the methods used, laptop was configured to serve as database server. It

received all forms of text messages from the participants the way they all responded to the questionnaires.

Cell phones that have capacity for data transfer via blue tooth or infra red were used for the data transfer. The participants saved the SMS as a draft in their handsets, later transfer to a dedicated cell phone through Bluetooth or Infra red. From this the text messages are deposited into the database server.

The FAQs were written in hard copy and the participants were asked to write the SMS languages' equivalent. This was later transferred into the database server. This approach was widely acceptable because it was very fast to complete as no one struggled with the key-pad or the tiny screen of the cell phones.

The chart in Fig. 1 below represents the proportion of the data sets that were collected for the experiment. This shows that 9(17%) participants are on MXiT, 18(37%) participants have Bluetooth on their cell-phones and able to transfer the text messages through it while the remaining 18(46%) prefer to use hand writing participatory method.



**Fig1. Chart showing the relative frequency of the three datasets**

### III. PRE-PROCESSING THE CORPUS

After the end of the data collection there was a partial parsing by the use of the spelling checkers to confirm the regularity of the typographic errors, ungrammatical abbreviations, acronyms, colloquial words, duplicated words and so on. The in-depth study of the SMS collected showed the great variability of the words form. Getting a clear meaning of the words in the corpus may be difficult to the extent that one has to do some guess work in order to confirm the correct interpretations of the words that are under search. This perhaps makes the reading difficult or impossible- because of lack of vowels (e.g. 'nyt' for 'night') or much consonants joined together (e.g. 'trtmnt' for 'treatment'), even sometimes diphthong (e.g. 'ao' for 'how') appearing in the SMS formation.

Text messages readability is another major phenomenon that needs to be taken into consideration as some words for example, 'ask4trtmnt' meaning 'ask for treatment' are written without spaces, the upper and lower cases are sometime jam together e.g. 'disCHg' for 'discharge' and mixing letters and figure together of some words like '8s' for 'AIDS' and '4rm' for 'from' are languages that one needs to understand the

codes, usages and habits for perfect translation of the SMS language. Otherwise, this may have been taken as errors unless one is familiar with the usage to avoid misrepresentation.

It is imperative that the corpus be translated into structured language in order to facilitate the exploration of the messages. With this, the meaning can come forth, as the parsing of the unstructured SMS language to a more structured for instance, English language. Without the structured list it is difficult to get the utterance of a given grammar, but it becomes very easy to find the transaction of the corresponding SMS.

The table below shows some of the various ways in which the participants were able to text the corresponding questions.

**Table1. Examples of Some of the Questions and their corresponding SMS languages**

Question	SMS Representation
What is HIV?	<i>Wat is hiv, Wats hiv, Watz hiv, Wtz hiv Wht is hiv, Wat's hiv</i>
What is AIDS?	<i>Wat is aids, Wht is aids Wots aids, Wt s aids Wt's 8s</i>
How is HIV passed on?	<i>Auz hiv pasd on Hows hiv psd on Hw dos d hiv test wk Hw ds da hiv tst wrk</i>

The table below shows various ways in which the data sets were able to be classified.

**TABLE II: Listing of Identified Non-standard Orthographic Forms**

Form Type	SMS	Translation
Shortenings	<i>aft ar sexual wil</i>	after are sexual will
Contractions	<i>cn knw infctd shud hw abt ppl</i>	Can know infected should how about people
G-clippings	<i>anythin testd passd medicatn possbl thk</i>	anything tested passed medication possible think
Other clippings	<i>hav tel infctd wt's wot's wat's kno</i>	have tell infected what is what is what is know

Acronyms	<i>HIV</i> <i>AIDS</i> <i>STI</i>	Human Immunodeficiency Virus Acquired Immune Deficiency Syndrome Sexually Transmitted Infections
Letter/number homophones	<i>8s</i> <i>xx</i> <i>+ve</i> <i>4rm</i> <i>In4mation</i> <i>4</i> <i>odas</i> <i>abnoma</i> <i>possibu</i>	AIDS sex positive from information for others abnormal possible
'Misspellings' and typos	<i>rtlnshps</i> <i>dxcharge</i> <i>xx</i> <i>nite</i> <i>n</i>	relationship discharge sex night and
Non-conventional spellings	<i>shud</i> <i>odas</i> <i>no</i>	should others know
Accent stylisation	<i>cot</i> <i>genita</i> <i>invecton</i>	caught genital infection

#### IV. SMS INFORMATION ACCESS SYSTEM

Figure 2, below shows the schematic model of the proposed SMS-query healthcare information system.

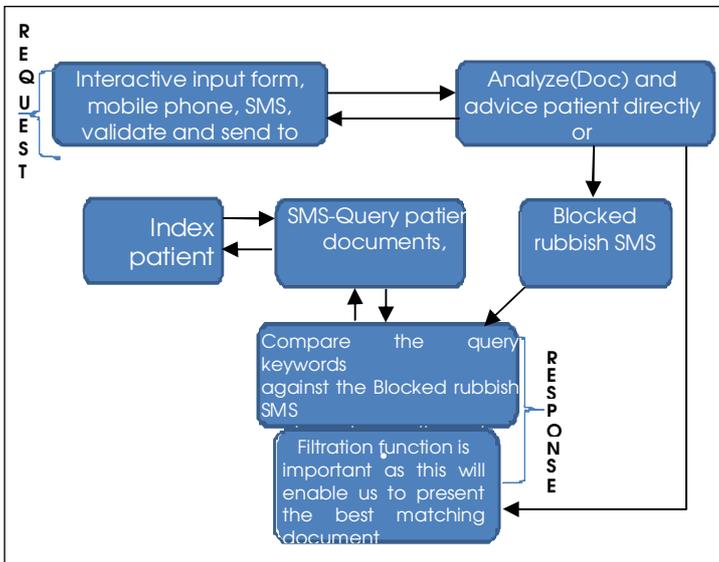


Fig 2. SMS-query Healthcare Information System

SMS text messaging using mobile phones has emerged as the dominant form of communication amongst certain healthcare providers of ICT users in sub-Saharan Africa especially in South Africa. Statistically SMS text messaging usage is more rampant than voice communication, especially among the age bracket 15-35years. Besides, there are services accessible through SMS by so-called wireless application service

providers (WASPs). Such services include, access to banking, job search and healthcare management. In healthcare, for example, access to SMS-query healthcare information system is a form of therapy. The model presented in this paper is based on these concepts of healthcare providers and mobile phone users. This is illustrated in figure 2, and is described in the following stages.

At the heart of the system are an information server and Intelligent Personal Information Agents designed to derive and present information appropriate for individual users. The major challenges that are addressed include the representation and modelling of individual user information needs. Personal information agents have been proposed to help information seekers to cope with the increasing amount of information available in the search engine. Personal information agents are intelligent assistants that perform several information-related tasks such as searching, filtering and learning information needs over time.

The service interface is the main forum for interaction with the system. It has two components: send and receive software installed on the mobile device. The send interface is used to login into the system in order to send queries, access and retrieve information.

The search agent is used to retrieve information in response to an incoming query, sent as an SMS. The Search Agent carries out processing function to compare the query keywords against the index keywords for each document, and return the best matching document for each SMS-query received.

The knowledge database is a repository of documents that are sent to patients in response to SMS-query. Each document held by the knowledge database has an associated index, which is a set of keywords (weights) that identifies the document. On request, document indices are sent to the Search Agent to be compared against incoming SMS-queries.

The Patient Agent contains basic and temporal information about each user. The basic information includes Names, Mobile No, and Password, ID etc. Queries, medications and appointment are temporal information logged during usage of the system.

The purpose of the information server on the service interface angle receives log-in request, free-form SMS, SMS-queries as input from the Mobile phone user; receives information from the knowledge database (e.g. FAQs) and free-SMS from human (e.g. physician) to be sent to users through Mobile phone user.

The challenge of information retrieval systems is to retrieve relevant documents in response to user queries [9]. A ranked retrieval system is a computer system managing a collection of documents, each containing text, and possibly other media. The collections are typically on the order of megabytes or gigabytes of text. The retrieval system does a number of

things, such as display document to the user and manages the database of documents. One of the most important functions of the retrieval system is to provide documents to users that satisfy their information needs. The process of representing an information need is referred to as a query formulation process [10].

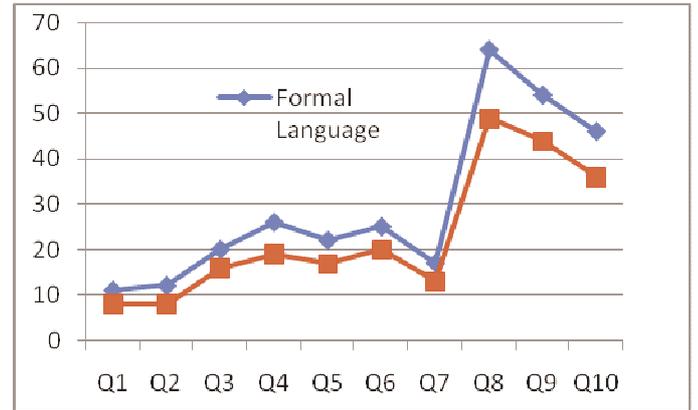
The workflow pattern is done in order to get the estimated relevance scores from the knowledge database within the healthcare context. A common method for query formulation is called the relevance feedback [11] and allows a user to interactively express an information requirement by modifying successive query inputs. The user is providing ‘feedback’ to the system that ‘relevant’ documents might look like the one indicated- thus relevance feedback. This in turn is used to improve the current search results for the user.

The system does this by estimating the degree of relevance of each document to the user’s statement of need (or query). These relevance estimates are used to rank the most relevant before the irrelevant. In theory, all the documents in the collection could be ordered by the retrieval system, so that the user can easily look through the ordered list and find the documents of interest. Evolutionary computation technique will be applied to a number of problems in text retrieval to validate their effectiveness.

## V. RESULTS/DISCUSSION

**TABLE III: Characterization Ration of the Formal and SMS Languages**

FAQ	No of Characters Represented (with spaces)	Average No. of Characters for SMS formulation
Q1	11	8
Q2	12	8
Q3	20	16
Q4	26	19
Q5	22	17
Q6	25	20
Q7	17	13
Q8	64	49
Q9	54	44
Q10	46	36



**Fig 2. Graphical Representation of characterisation of the Formal and SMS Languages**

## VI. CONCLUSION

In this paper, we have presented an early investigation into building an information access system based on SMS-based queries. The difficulties with SMS communication were explored, namely, the informal nature of this communication and the associated difficulty of searching and retrieving a standardised formal corpus of knowledge. We also presented a proposed information retrieval schematic based on SMS queries. This is a work in progress and we shall examine other approaches, in particular, phrase-based statistical method to normalise SMS messages, corresponds with manually normalised messages. The other stage of our research work will focus on how the machine translation can be adopted for parsing the non-standard texts into Standard English. The proposed system is allied to health care information access.

## VII. REFERENCES

1. Cedrick Fairo *et al.* (2009) *A Translated Corpus of 30,000 French SMS*, Centre de traitement automatique du langage, UCLouvain, Place Blaise Pascal 1, B-1348 Louvain-la-Neuve
2. How, Y., & Kan, M. (2005) *Optimizing Predictive Text Entry for Short Message Service on Mobile Phones*. In M. J. Smith & G. Salvendy (Eds.), *Proceedings of human computer interfaces international*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
3. Mohan S. Kankanhalli *et al.* (2002) *Watermarking of Electronic Text Documents*, Electronic Commerce Research, 2: 169–187 Kluwer Academic Publishers. Manufactured in the Netherlands.
4. [www.wikipedia.org](http://www.wikipedia.org) cited Nov 12, 2009
5. Crispin Thurlow, *Generation Txt? The sociolinguistics of young people’s text-messaging* Department of Communication, University of Washington, Box 353740,

Seattle, WA 98195, USA Thurlow@u.washington.edu  
<http://faculty.washington.edu/thurlow>

6. Catherine Quantin *et al.* *New Advanced Technologies to Provide Decentralised and Secure Access to Medical Records: Case Studies in Oncology*, Cancer Informatics 2009: 7 217 -229 available from <http://www.la-press.com>
7. As cited by [en.wikipedia.org/wiki/MXit](http://en.wikipedia.org/wiki/MXit) May2, 2010
8. Fairon, C., J. Klein, S. Paumier. (2006c). "*Faites don de vos SMS à la science*". *Étude du langage SMS au départ d'un corpus informatisé*.
9. C. N. Mooers (1976), Technology of information handling: pioneer's view, Bull AMIS, vol. 2, no8, p18-19.
10. R. Baeza-Yates and B. Riberio-Neto (1999), Modern Information Retrieval, Addison-Wesley, Essex.
11. G. Salton and C. Buckley (1997), Improving retrieval performance by relevance feedback, in Sparck-Jones K and Willett (eds), Readings in Information Retrieval, Morgan-Kaufman, ISBN:1-558-60454-5, p355-363.
12. Adesina A. O., Nyongesa H.O, Agbele K.K. (2010) *Digital Watermarking: The state-of-the-Art Review*" IST Africa 2010 Conference Proceedings Paul Cunningham and Miriam Cuningham (Eds) IIMC International Information Management Corporation ISBN: 978-1-905824-15-1

**Ademola O. Adesina**, received his Masters Degree in Computer Science from the University of Ibadan, Nigeria (2004). Presently is a doctoral student in the Natural Language Group, University of the Western Cape, South Africa. His research interests are in Digital Watermarking with special interest to Linguistic Steganography of the text messages, Machine learning and Mobile Communication.