A verbal autopsy corpus for machine learning of cause of death

Samuel Danso[1,3], Eric Atwell[1], Owen Johnson[1], Guus ten Asbroek[2], Seyi Soromekun[2], Karen Edmond[2], Chris Hurt[4], Lisa Hurt[2], Charles Zandoh[3], Charlotte Tawiah[3], Zelee Hill[2], Justin Fenty[2], Seeba Amenga Etego[3], Seth Owusu Agyei[2,3], and Betty R Kirkwood[2].
[1] University of Leeds ,[2] London School of Hygiene and Tropical Medicine , [3]Kintampo Health Research Centre, Ghana [4] University of Cardiff

Abstract

Unregistered deaths account for over 67 percent of global annual deaths. These deaths occur in poor resource countries where no or weak death registration system exists. Verbal Autopsy is a technique recommended by the World Health Organization (WHO) to determine the causes of deaths in countries with poor death registration systems. Traditional approaches to the analysis and prediction of causes of death from verbal autopsies have several limitations, and computational approaches are being developed to address. Ongoing research aims to extend the current computational approaches by employing corpus linguistics and natural language processing with machine learning approaches to predict causes of death from Verbal Autopsies. In this paper, we present the corpus to be used for this research. We demonstrate that the Verbal Autopsy corpus has properties of human language and similarities to other corpora. Apart from the primary objective of predicting causes of death, this corpus has potential, to be of interest for other linguistic research.

Introduction

The quest to achieve the Millennium Development Goals (MDGs – 4 & 5) set out by the United Nations in September, 2000 has led to increased attention to mortality reduction[1]. To achieve this goal, there is a crucial need for cause of death information at all levels[2]. This is to inform national and international health managers, policy makers and researchers about the trends in causes of death in order to develop strategies, design research interventions, and carry out sound budgetary allocations[3]. This implies the need for a robust death registration system to be able to achieve these targets[1]. However, the challenge is that over 67 per cent of the deaths that occur go undetermined, and these deaths occur mostly in countries that lack a robust death registration system, resulting in a huge information deficit[4]. In response, the World Health Organisation(WHO) recommended the use of the Verbal Autopsy technique for places where obtaining Cause of Death(CoD) information remains problematic[5].

The Verbal Autopsy (VA) technique involves interviewing people (such as relatives or caregivers) who were close to the deceased, or if possible, individuals who cared for the individual around the time of death, for events that led to the demise of the individual. This information usually contains signs and symptoms of possible illness that caused the death. Inferences are then made from the information gathered from relatives to establish the possible cause of death. VA is now an established technique used by health professionals in developing countries to estimate the distribution of diseases[6].

Traditionally, the standard approach employed in inferring from the VA information to ascertain the CoD has been the use of physicians. Even though Joshi et al.[7]argue that one physician review is enough, the mostly used method involves a minimum of two physicians, each taking about 20 – 30 minutes to go through the review process of a given VA and assign the possible CoD[8]. Where there is no agreement between the two physicians, a third is employed to go through the entire process to ascertain the CoD. The limitations of this approach such as high cost, inefficiencies , and inter-physician reliability are well documented[8]. The situation is compounded in regions where physicians are usually in short supply and this is where Verbal Autopsies are performed[9]. Further, even where physicians are present, their substantial workloads means the coding of Verbal Autopsies is placed very low in their lists of priorities making the entire process unsustainable[10].

The recent emergence of sophisticated computational algorithms and computing power[11], has given rise to attempts to use various computational approaches to classify causes of death[10, 12]. This has almost universally employed the use of closed-ended questions(referred to in this paper as closed part) where statistical inferences are made from the data[3, 8]. However, it has also been accepted that the closed part is invariably limited in capturing all potentially relevant information. It is also limited compared with physician review because the physicians base their coding of the CoD on both the closed part and open narrative (referred to in this paper as free text)[13]. This has led to the use of a combination of closed part and free text to maximize the yield of relevant information [14]. Most algorithms in use, however, do not allow for consideration of information yielded from the free text.

Ongoing research at University of Leeds in the United Kingdom aims to explore the application of Corpus Linguistics and Natural Language Processing (NLP) with Machine Learning in the analysis and the prediction of CoD from VA. In other words, this research is formulated as a Text Classification problem, which is related to other studies [15, 16], where each VA will be classified according to CoD. In this paper however, the focus is to present the corpus to be used in carrying out the research. We show that this corpus has properties similar to other referenced corpora and could be further explored for other corpus linguistics research.

The Verbal Autopsy Corpus

Source and acquisition process

The corpus for this research project is obtained from two large field trials carried out in Ghana, which led to the establishment of a surveillance system between December 2000 and July 2010[17, 18].The objective of the first trail by Kirkwood et al.[17] was to assess the effect of vitamin A supplementation in women in Ghana. The objective of the second trial also by Kirkwood et al. [18]aimed at the development of a feasible and sustainable community-based approach to improve newborn care practices in order to improve the survival of newborns. Data were collected during 4-weekly surveillance, which included all deaths in infants to 1 year and women of reproductive age. VAs were conducted for all of these deaths, and these form the content of the corpus for this research.

A total of over 10,000 individuals VA were done. This comprises of over 8,000 infants up to 1 year death(the infant corpus) and over 3,000 adults of women of reproductive age(the women corpus). The questionnaires on the VA interviews have been reviewed by physicians and the CoD for each individual has been ascertained. The corpus metadata also includes related information such as the instructions for the data collectors on how VA interviews are to be conducted and the instruction manual for clinicians on how to review and assign cause of death. Figure 1 depicts the process model of the collection.

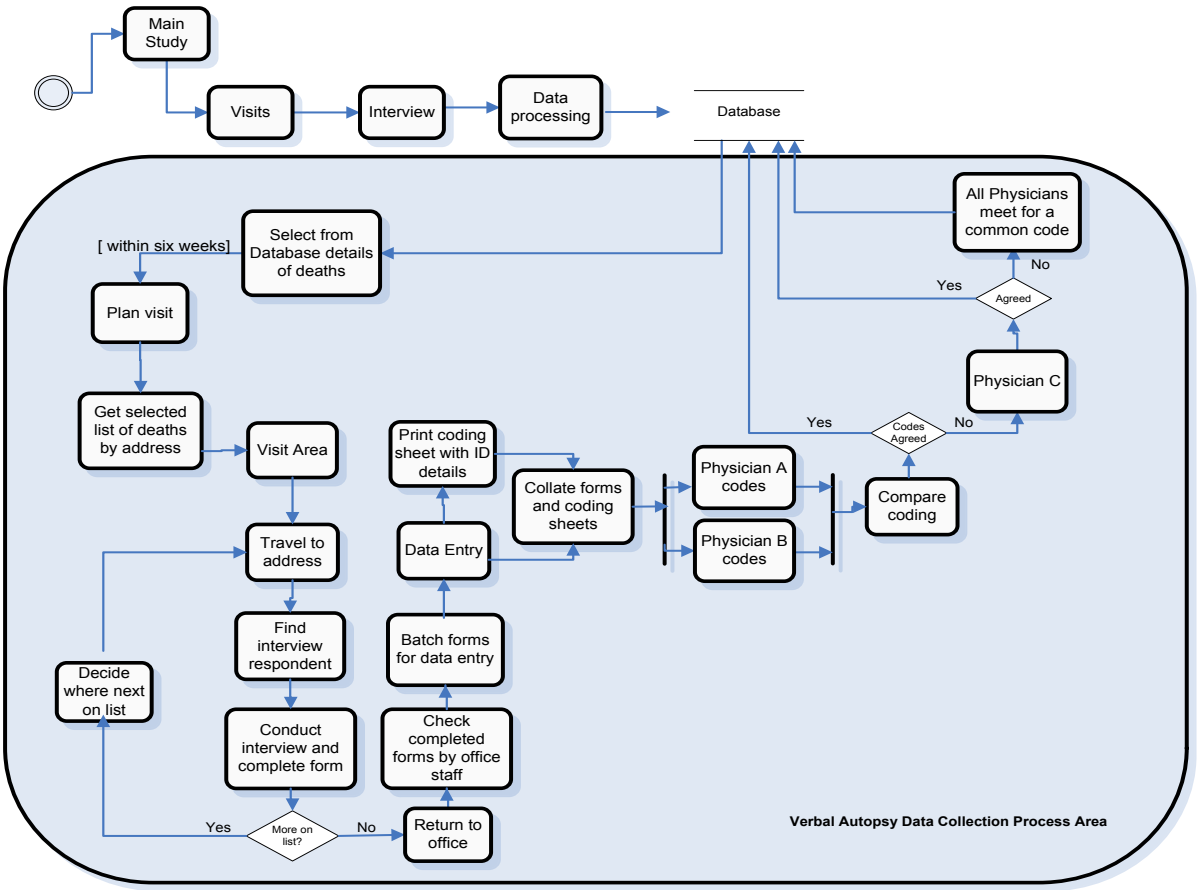

Figure 1 Process model of VA corpus collection

The collection of VA corpus as shown in figure 1 above was part of the studies designed and carried out according to well defined objectives as described earlier. Routine visits were made to participants of the study for data to be collected. The data, which included events that have occurred concerning the study participants (including deaths) at the time of visit

were processed and stored in a database. List of deaths was generated from the database, and this triggered the VA corpus collection process to begin. A study conducted suggests that a period between 1 and 12 months is enough to elicit reliable information from respondents[5]. However, it is probably better to conduct the interview as soon as possible after the death, but because researchers also have to respect the family's need to grieve, so a minimum of 6 weeks was set before conducting the interview. This is because it is believed that there will be recall challenges on the part of the respondent. Also, not allowing any time gap for mourning, could pose emotional challenges to the bereaved family.

Interviewers planed visits according to the list generated from the database indicating the address to visit. An Interviewer travelled to the address and identified an appropriate respondent according to the protocol, which suggests a close friend, relative or other care-giver of the deceased, deemed capable of providing accurate information. The interview was conducted and a questionnaire was filled in. The Interviewer then decides to move to another address for another VA interview or get back to the office. Basic quality control checks (eg no missing information) were performed on the questionnaires and then they are forwarded for data entry in batches.

Each questionnaire went through a transcription process where the entire content was entered by a data entry clerk. The content of the questionnaire is not to be changed at this stage, although inconsistencies in the data are flagged up by range and consistency checks during data cleaning of the coded part of the questionnaire. The general rule is to ensure that the information captured on the form is the exact copy that is entered by the data entry staff.

A 'coding sheet'[1] was generated with ID details of each dead person. The' coding sheets' were then collated with the Verbal Autopsy questionnaire. Each set of questionnaire with its corresponding coding sheet was given to two independent physicians for a review and CoD assigned based on their expert judgment. A comparison was done to establish the consistency between the two assigned CoDs. An agreed code was assigned when the two physicians agree. Otherwise, the forms were then passed on to a third physician for a review. The CoD assigned by the third physician was compared with the other two assigned causes of death, and when there was an agreement between any of the two already assigned CoD, then that CoD was assumed to be the final. On the other hand, when there was no agreement, a meeting is held between the physicians involved for an agreement to be reached. In an event where there is no agreement, the cause of death was declared 'undetermined'.

Background of interviewers

The project staff that perform the VA interview are not medically trained. They normally have completed high school and have attained fieldwork supervisory status. These people were identified and provided with basic training on how to administer the Verbal Autopsy questionnaire. For example a specific training may be given on how to probe in order to elicit the information relevant to enable physicians to arrive at the possible cause of death.

[1] This is used by the physicians during review of the questionnaire to record the cause of death

Characteristics of the corpus

The questionnaire can be grouped into identification, coded part and free text sections:
Identification: this part of the questionnaire contained basic contact details of the deceased. Information include unique ID number, compound number, name, and date of birth. This information is derived from the surveillance database and printed of the listing, which is transcribed onto the VA questionnaire by interviewer.
Coded part: the coded-part questions have in total over 200 variables. These variables are questions to elicit the presence of specific symptoms during the final illness. This information is often accompanied by a box for recording the length of time that the deceased experienced those symptoms.

Did she have fever?	1.Yes	2. No	8. NK	For how long? (in days) [99=NA]		

Figure 2: Question and response options provided to respondent during interview

Figure 2 for example is asking a mother of a dead child about whether the child had fever, and where the questions are not applicable, the double line is drawn through those questions and '99' is entered into the databases as the response value.
Free text: the purpose of the free text is to enrich the data collected in the coded part and both free text and the coded-part are used when making decisions about the CoD. The information collected included details of events surrounding death, socioeconomic characteristics, fertility, and obstetric history. The free-text part gives a verbatim account of the responses from the interviewee. The interview is conducted in the local language (Twi) and the interviewer translates into English as seen in figure 3 below.

It is worth noting here that the structure of the questionnaire used for the infants was different from the questionnaire used for the adult women. The free text part of the infant questionnaire had clearly defined sections that asked for specific information as shown in figure 3 below. The women questionnaire however did not have sections but had blank page that required the interviewer to write during the interview. It is also important to note that the VA questionnaires were designed based on the most standard methods available at the time of the design. The Infant questionnaire was based on the WHO standard, which was locally adapted by Edmond et al.[19] to collect the infant VA data. The adult women questionnaire on the other hand was designed based on the work carried out by Chandramohan et al. [20]and the questionnaire used in the study conducted by [21].

The free text part of the VA questionnaire is divided into sections. The infant questionnaire as shown in figure 3 for example have been categorised into three main themes; the pregnancy section elicits information about any experiences and sickness that were encountered during and into the end of the pregnancy. Other sections include incidents before delivery, during labour, and after delivery.

Can you tell me something about your pregnancy?
Movement of the baby in the womb started around the 6th month continuously till 9th month following the delivery. Although I did not encounter too many pregnancy complications, malaria persistently attacked me on the 7th month until I delivered. I suffered severely from anaemia which was diagnosed by a health worker when i visited hospital on the 8th month. Finally, I was not able to feed by self well when about a month to delivery due to lost of appetite. Sometime instead of feeding thrice a day, once daily becomes a problem for me.
Can you tell me something about your labour
the labour started around 1pm in the night following the flow of water approximately 4hours. All of a sudden I felt the baby coming therefore I decided to try my best as much as possible to deliver at home. To my surprise the baby came with her both legs which really made it difficult to deliver myself. Therefore the TBA in the village was called to assist yet it proved futile. thus my husband had to go and arrange for vehicle to take me to the nearest hospital facility remarked by the TBA. before the vehicle arrived i had finally delivered.
Can you tell me something about the baby?
the baby landed without breathing or crying, therefore I enquired from the TBA to know what has happened to my baby but the woman assured me that the child is weak so I should lie down for a while and feel comfortable for everything will be alright. after she had finished with me she confirmed the baby landed dead.
Can you tell me what happened after delivery?
the baby neither cried or nor breath after delivery
Any signs and symptoms before the death of the child ?
since the baby was very weak, he was put in an incubator but died after three hours of birth.

Figure 3 Sample of free-text questions and responses from VA questionnaire

Indicative word log-likelihood distribution in VA corpus

The transcription of the entire corpus to be used for the research is yet to be completed. However, to examine some characteristics in the corpus, we perform our analysis based on a sub-sample of 1,800 words out of an estimated 1.8million words in the entire corpus. This sub-sample is derived from only the infants VAs. We compare our sub-sample with the Brown corpus using frequency distribution with log-likelihood[22] as shown in table 1 below.

Word	Freq. in VA	Log-likelihood	Word	Freq. in VA	Log-likelihood
baby	55	534.806	incubator	4	50.636
was	107	209.167	gestation	4	50.636
pregnancy	12	133.976	till	7	47.138
delivered	15	127.619	Not	34	47.096
started	20	121.145	Given	12	46.599
hospital	16	117.894	immediately	8	42.246
delivery	11	101.828	cry	6	38.458
month	16	101.715	breathing	5	38.349
labour	9	100.469	she	21	38.200
died	14	96.608	Zakari	3	37.976
birth	12	86.229	Yefri	3	37.976
TBA	6	75.961	Nkoranza	3	37.976
very	20	69.379	CS	3	37.976
Akos	5	63.298	ANC	3	37.976
Pains	7	61.940	Womb	3	33.481
after	19	61.436	Drip	3	33.481
weak	8	61.370	normal	7	33.317
experiencing	6	58.041	suffered	5	31.592
Incubated	5	57.895	chills	3	31.252

Table 1 Most significant words in the VA corpus

The above table shows an indicative frequency of words with log-likelihood statistics generated in comparison with the Brown corpus. The table show the words that characterise the VA corpus as they appear at the top of the list with the highest log-likelihood scores. It is therefore not surprising to see words such as 'baby', 'pregnancy',' birth' and 'TBA' (Traditional Birth Attendant- a 'recognised ' person who assist women during delivery within a community) appearing at the top of the list, as these are typical of words to be found in narratives about circumstances surrounding pregnancy and birth. Similarly words that are common to both the VA and the Brown corpora are found at the bottom of the complete table, with least log-likelihood score, which we are unable to show due to limited space. Examples include 'possible', 'about', 'many', and, 'this'

Indicative distribution of causes of deaths

Again, this analysis is based on the current data we hold, and subject to change. These are to highlight some of the interesting characteristics about the corpus from a Text Classification perspective. Figure 4 and figure 5 depict the distribution of the causes of deaths among infants up to the age of 1 year and women respectively from the corpus as determined by physicians. The ICD -10 causes of deaths, which in this context, are classes or labels has a hierarchy of codes, in which diseases are assigned to differentiate specific types from general types of the same category. For example Maternal Disease 'Anaemia' and 'Eclampsia' as specific types of Maternal Disease. We show the breakdown of these classes in our corpus in the figures below.

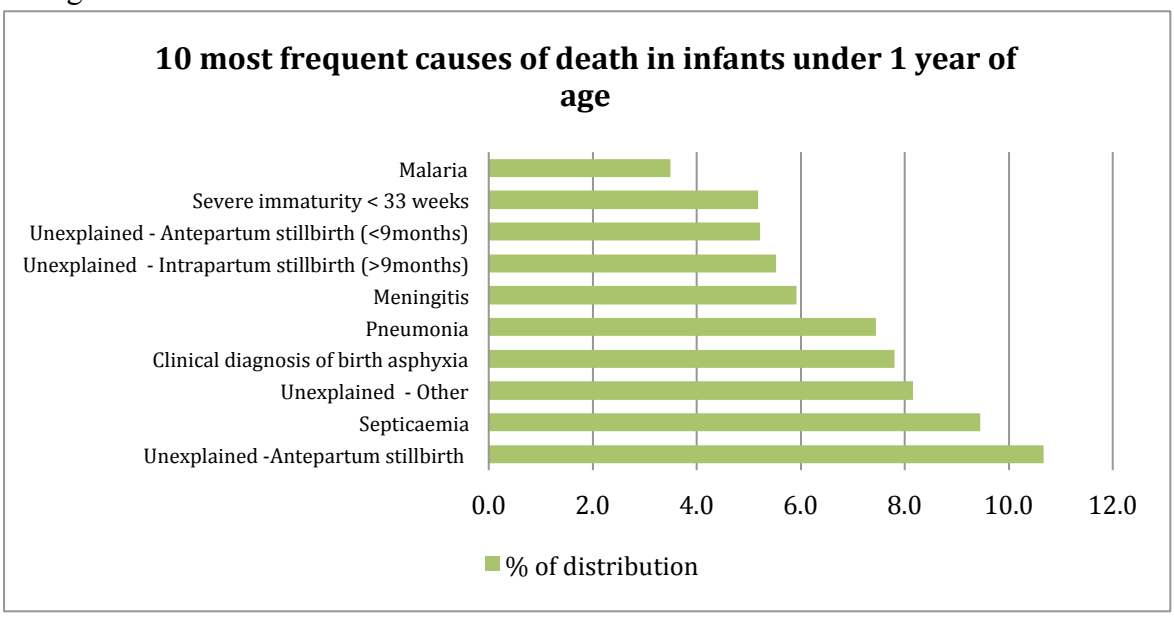

Figure 4: Most frequent causes of death in the infants' corpus found in a sub-sample of 2903

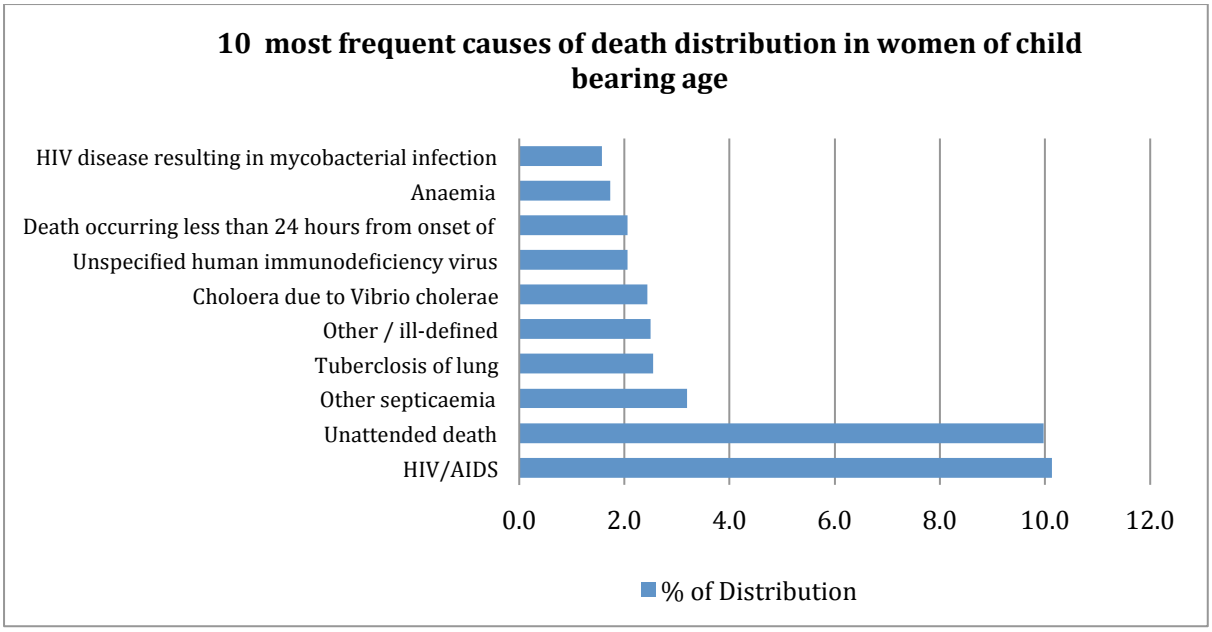

Figure 5: 10 most frequent causes of death in the women corpus found in sub-sample of 1,456

Figure 4 and 5 above show the 10 most frequent causes of death found according to the sub-sample of 2903 out of the over 8,000 expected from the entire infant corpus and sub-sample of 1,456 out of the over 3,000 expected from the adult women corpus respectively. Although these figures may change, they are indicative of the distributions expected to be found in the entire corpus.

Some characteristics about the corpus could be deduced from both figure 4 and 5 above. An obvious one is the difference that exists between the causes of death from both figures as one would not expect infants and adults to have the same causes of death categories. While in figure 4, the most frequent cause of death in babies is due to '*Unexplained - antepartum-stillbirth'*, '*HIV-AIDS*' is the most frequent cause of death in women in Figure 5. Another notable characteristic is the fact that there are significant overlaps in the classification from both figures. In figure 4 while there are some clear categories such as '*malaria*' and '*pneumonia*', there are also ill-defined and overlapping categories such as '*Unexplained Antepartum stillbirth < 9months*' and *'Unexplained stillbirth*'. Similarly, '*HIV disease resulting in mycobacterial infection*' and '*Unspecified human immunodeficiency virus'* as found in figure 5. Some of these overlapping categories are clearly similar in their reporting and content. Resulting overlap in keywords or features will pose difficulties for machine learning algorithms to differentiate between them. This is partly due to the classification system adapted in assigning the CoD, the International Classification of Disease (ICD-10), which is hierarchical in nature[23]. It has fine grained, which tend to share similarities and overlaps. The differences become much clearer at the top of the hierarchy, where there are no or less overlaps. Further analysis and re-classification of the current categories may be required to conform to the high level classification scheme.

Spelling and grammatical variations

Considering the context and the source in which this corpus was generated we anticipate various spelling and grammatical variations. Figure 6 below is an extract from an interview example of a mother who lost her child after birth, highlighting the variety of spellings and grammatical constructions that exist in the corpus.

"WHEN THE CHILD WAS SIXTEEN (16) DAYS OLD SHE FELL SICK WHICH LAUTED FOR THREE (3) DAYS BEFORE SHE DIED. THE CHILD HAVING DIFFICULT BREATING. ANY TIME, SHE BREATHS, YOU SEE A HOLE IN THE CHEST, AND ALSO MAKING NOISE IN THE CHEST. SHE HAD CONVULSION WHEN SHE WAS SEVERTEEN (17) DAYS OLD BEFORE SHE DIED THE FOLLOWING DAY. SHE ALSO HAD A BULGING FONTENED AND SEVERE HOT BODY WHICH LASTED FOR TWO (2) DAYS BEFORE SHE DIED. THE CHILD ALSO HAD A FIT WHICH SHE COULD NOT OPEN HER MOUTH."

Figure 6 extract of verbal autopsy interview with highlighted errors

Figure 6 above demonstrates the challenges that are inherited from the VA corpus. This corpus was collected by data collectors who are not native English speakers. Also, the data entry clerks may have difficulty in reading the written information or do not understand what is written during data entry. Thus, occasional grammatical errors are bound to occur. Additionally, there are misspellings and incorrect use of punctuation marks and use of local words which are not recognised in any lexicon resource. Some of these challenges are highlighted above, which show the level of complexity and the need for robust algorithms. From figure 6 the highlight in the first line is an example of misspellings. The word 'lasted' has been misspelled as '*lauted*'. Also, on the second line are two types of errors. The first is a grammatical error found in the phrase 'the *child having difficult breathing'* and the second is wrong use of punctuation as found in the text '*any time, she breaths, you see a…*'

Discussion

We base our discussion on the benchmarks set out to assess a modern corpus by McEnery and Wilson [24]. Their benchmarks include the format in which the corpus is stored, the sampling and its representativeness, and also status of reference standard. We discuss these attributes in turn. The issue of sparsity and imbalance as well as the quality of the text are also discussed.

Format of corpus

This criteria is concerned with whether the corpus is stored in a machine readable format. We have clearly demonstrated this by the fact that part of the corpus (the coded part) has been transcribed and we are in the process of entering the free-text collected onto a computer. This will result in the entire corpus being machine readable and conforming to the Text Encoding Initiative guidelines[25]

Finite corpus

This benchmark tends to establish the fact that a corpus must be finite in size. We estimate this from the sub-sample of 1,800 words (this is a total of 10 documents) used for illustration

in this paper. It can be extrapolated from this that an estimated finite size of 1.8million words can be obtained from the 10,000 documents, which are in the process of being digitised.

Referenced standard status

It is acknowledged that references can be found in the literature that point to other corpora and particularly within the biomedical domain , examples include, the GENIA corpus[26] and the i2b2 challenge corpus of medical narratives[27]. In other words these corpora have achieved the status of reference set out as a benchmark. The Verbal Autopsy corpus is the first to be used for corpus linguistic research and as such this publication is the standard reference.

Sample Representativeness

This benchmark tends to determine the representativeness of the corpus as a sample of the domain of interest. We can assert that the Verbal Autopsy corpus being presented here is a representative sample of the language used in Verbal Autopsies considering the source in which it was obtained. This sample was drawn from 7 contiguous districts within the Brong-Ahafo region of Ghana. The collection processes adopted were of the international standard employed globally. Based on the sub-sample of 1,800 words used for illustrations in this paper, further confirmation is carried out. We used Zipf's law to formally assess this. Zipf's law has been established as a useful test to observe the frequency distribution of words in a human language[28]. Using the word frequency distribution obtained from the sub-sample of 1,800 words, we plot a graph as shown below.

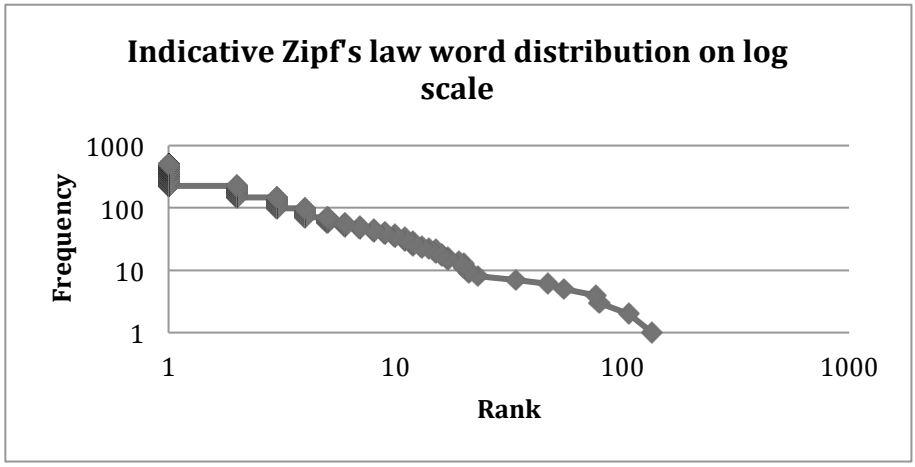

Figure 7 indicative word Zipf's law of distribution

Figure 7 shows a relationship between frequency of a each word and its position with respect to its rank as expressed by Zipf's law. It can be observed from the graph that it has a negative slope, which suggests the properties of representative sample as expected from Zipfs law[29]. Similar properties hold for distribution of categories of causes of death, where we tend to find many more causes of death as well as very few as observed in figure 4 and 5 above. It must be noted that the corpus does not cover adult males. However, we do not anticipate any difference between males and females in narratives and the cause of deaths.

Sparsity and imbalance

Another property that is obvious in any natural language text, which is also exhibited as one of the properties of Zipf's law is the issue of sparseness[30]. The distribution figures shown above suggest a high level of sparseness and imbalance from the text and also in the cause of death categories which is problematic for the NLP with machine learning standpoint [31]. The challenges of sparsity and imbalanced corpus have been recognised and documented, which have given rise to research into best approaches to handling datasets of this nature[32]. A recent solution propose by [33] is reported to have good tolerance and ability to handle skewed datasets.

Quality of text

[34] argues that misspellings and ill-formed grammar in corpora pose a great risk, and as such could introduce side effects in algorithm performance. However, the recent increase in the use of the web as corpus for carrying out NLP studies have led to the development of robust methods that are capable of dealing with ill-formed content in corpora[35]. This is particularly prominent in genre classification of web pages[36] and other forms of online classification[37]. This is because web documents are produced by people with varying levels of command of the English language resulting in text with variations in spelling and grammar. From this viewpoint, it can be argued that the Verbal Autopsy corpus has similarities with other corpora generated from the web that are used for corpus linguistics research.

Conclusion

The usefulness of causes of death information to the WHO and the international community has been well acknowledged as an important issue. Verbal Autopsy is a recommendation from WHO as a method of obtaining cause of death information. The limitations associated with the traditional methods of analysis have been discussed, and computational approaches are being developed to address these challenges. Ongoing research at Leeds University aims to extend these computational approaches by employing NLP with Machine Learning. We have presented in this paper a Verbal Autopsy corpus to be used in carrying out this research.

Although numerous corpora exist within the biomedical domain used for corpus linguistics research, this corpus is the first to be published generated from Verbal Autopsy. A considerable difference exists between this corpus and other published corpora within the biomedical domain. This is because whereas other corpora found in the literature are some form of discourse between experts within the biomedical domain, this corpus is generated from a discourse between interviewers and respondents who are both non medical experts. We have also demonstrated using Zipf's law to suggest that the language in this corpus has the properties of any human language and by extension any referenced corpus. This therefore suggests the representativeness of this corpus and it should be considered as a different genre within the biomedical domain.

Our current research is focused on predicting causes of death, which is formulated as a Text Classification task within NLP. We however think that this corpus presents an opportunity for exploration of other interesting medical language research questions that are outside the scope of our current work.

Acknowledgments

We thank the staff of Kintampo Health Research Centre for their hard work and dedication. Special thanks go to the computer centre staff, the field staff and the drivers. This corpus would not been collected without them. We also thank the doctors who took the pains to review all coding of verbal autopsies for us to have the causes of death. Their contributions a very much appreciated. We also thank the study participants who agreed to be part of the project and contributed to this corpus. Finally, we acknowledge the contributions of Angela Vega of London School of Hygiene and Tropical Medicine, who provided administrative support to the projects that, collected this corpus and Dr. Poorna Gunasekera who also played a key role in the co-ordinations of the Verbal Autopsy coding.

References

1. Murray CJL, Laakso T, Shibuya K, Hill K, Lopez AD: **Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015.** *The Lancet* 2007, **370:**1040-1054.
2. Filippi V, Ronsmans C, Campbell OMR, Graham WJ, Mills A, Borghi J, Koblinsky M, Osrin D: **Maternal health in poor countries: the broader context and a call for action.** *The Lancet* 2006, **368:**1535-1541.
3. Murray C, Lopez A, Feean D, Peter S, Yang G: **Validation of the symptom pattern method for analyzing verbal autopsy data.** *PLoS Medicine* 2007, **4:**1739 - 1753.
4. Fottrell E, Byass P, Ouedraogo T, Tamini C, Gbangou A, Sombie I, Hogberg U, Witten K, Bhattacharya S, Desta T, et al: **Revealing the burden of maternal mortality: a probabilistic model for determining pregnancy-related causes of death from verbal autopsies.** *Population Health Metrics* 2007, **5:**1.
5. Soleman N, Chandramohan D, Shibuya K: **Verbal autopsy: current practices and challenges.** *Bulletin of the World Health Organization* 2006, **84:**239-245.
6. Baiden F, Bawah A, Biai S, Binka F, Boerma T, Byass P, Chandramohan D, Chatterji S, Engmann C, Greet D, et al: **Setting international standards for verbal autopsy.** *Bulletin of the World Health Organization* 2007, **85:**570-571.
7. Joshi R, Lopez AD, MacMahon S, Reddy S, Dandona R, Dandona L, Neal B: **Verbal autopsy coding: are multiple coders better than one?** *Bulletin of the World Health Organization* 2009, **87:**51-57.
8. Byass P, Fottrell E, Dao Lan Huong, Berhane Y, Corrah T, Kahn K, Muhe L, Do Duc Van: **Refining a probabilistic model for interpreting verbal autopsy data.** *Scandinavian Journal of Public Health* 2006, **34:**26-31.
9. Usman N: **Developed world is robbing Africa of health staff: There are no easy answers.** *British Medical Journal (BMJ)* 2005, **331:**48-49.
10. Byass P, Kahn K, Fottrell E, Collinson MA, Tollman SM: **Moving from Data on Deaths to Public Health Policy in Agincourt, South Africa: Approaches to Analysing and Understanding Verbal Autopsy Findings.** *PLoS Medicine* 2010, **7**.
11. Cairncross F: *Death of distance : how the communications.* Boston: Harvard Business School; 1997.
12. King G, Lu Y, Shibuya K: **Designing verbal autopsy studies.** *Population Health Metrics* 2010, **8:**19.
13. Soleman N, Chandramohan D, Shibuya K: **WHO Technical Consultation on Verbal Autopsy Tools.** *Geneva* 2005.
14. Marsh DR, Sadruddin S, Fikree FF, Krishnan C, Darmstadt GL: **Validation of verbal autopsy to determine the cause of 137 neonatal deaths in Karachi, Pakistan.** *Paediatric and Perinatal Epidemiology* 2003, **17:**132-142.
15. Sebastiani F: **Text categorization.** *Text Mining and its Applications to Intelligence, CRM and Knowledge Management* 2005**:**109–129.
16. Sebastiani F: **Machine learning in automated text categorization.** *Association for Computing Machinery (ACM) Computing Surveys* 2002, **34:**1-47.
17. Kirkwood BR, Hurt L, Amenga-Etego S, Tawiah C, Zandoh C, Danso S, Hurt C, Edmond K, Hill Z, ten Asbroek G, et al: **Effect of vitamin A supplementation in women of reproductive age on maternal survival in Ghana (ObaapaVitA): a cluster-randomised, placebo-controlled trial.** *The Lancet* 2010, **375:**1640-1649.
18. Kirkwood B, Manu A, Tawiah-Agyemang C, ten Asbroek G, Gyan T, Weobong B, Lewandowski R, Soremekun S, Danso S, Pitt C, et al: **NEWHINTS cluster randomised trial to evaluate the impact on neonatal mortality in rural Ghana of**

routine home visits to provide a package of essential newborn care interventions in the third trimester of pregnancy and the first week of life: trial protocol. *Trials* 2010, **11:**1-11.

19. Edmond KM, Quigley MA, Zandoh C, Danso S, Hurt C, Agyei SO, Kirkwood BR: **Diagnostic accuracy of verbal autopsies in ascertaining the causes of stillbirths and neonatal deaths in rural Ghana.** *Paediatric and Perinatal Epidemiology* 2008, **22:**417-429.
20. Chandramhan D, Maude G.H, Rodrigues L.C, Hayes R.J: **Verbal Autopsies for Adult Deaths: Issues in their Development and Validation.** *International Journal of Epidemiology* 1994, **23:**213-222.
21. Kane TT, El-Kady AA, Saleh S, Hage M, Stanback J, Potter L: **Maternal Mortality in Giza, Egypt: Magnitude, Causes, and Prevention.** *Studies in Family Planning* 1992, **23:**45-57.
22. Rayson P, Garside R: **Comparing corpora using frequency profiling.** In. Association for Computational Linguistics; 2000: 1-6.
23. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA: **Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data.** *Medical Care* 2005**:**1130-1139.
24. McEnery T, Wilson A: *Corpus linguistics: an introduction.* Edinburgh University Press; 2001.
25. Consortium T, Sperberg-McQueen C, Burnard L, Bauman S: *Guidelines for Electronic Text Encoding and Interchange.* Humanities Computing Unit, University of Oxford; 2002.
26. Ohta T, Tateisi Y, Kim JD: **The GENIA corpus: An annotated research abstract corpus in molecular biology domain.** In. Morgan Kaufmann Publishers Inc.; 2002: 82-86.
27. Uzuner Ö, Sibanda TC, Luo Y, Szolovits P: **A de-identifier for medical discharge summaries.** *Artificial Intelligence in Medicine* 2008, **42:**13-35.
28. D bowski: **Zipf's Law: What and Why?** *URL: http://www ipipan waw pl/ldebowsk* 2000.
29. Li W: **Random texts exhibit Zipf's-law-like word frequency distribution.** *IEEE Transactions on Information Theory* 1992, **38:**1842-1845.
30. Goweder A, De Roeck A: **Assessment of a significant Arabic corpus.** In Association for Computational Linguistics (EACL). Citeseer; 2001
31. Lakeland C, Knott A: **Implementing a lexicalised statistical parser.**
32. Chawla NV, Japkowicz N, Kotcz A: **Editorial: special issue on learning from imbalanced data sets.** *SIGKDD Explor Newsl* 2004, **6:**1-6.
33. Li Y, Bontchenva K, Cunningham H: **Adapting SVM for data sparseness and imbalance: a case study in information extraction.** *Natural Language Engineering* 2009, **15:**241-271.
34. Ruch P, Baud R, Geissbühler A: **Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record.** *Artificial Intelligence in Medicine*, **29:**169-184.
35. Liu V, Curran JR: **Web text corpus for natural language processing.** In proceedings of the 11th meeting of the European Chapter of the Association for Computational Linguistics.; 2006: 233–240.
36. Meyer zu Eissen S, Stein B: **Genre classification of web pages.** *Advances in Artificial Intelligence* 2004**:**256-269.

37. Mishne G: **Experiments with mood classification in blog posts.** In proceedings of Association for Computing Machinery SIGRI 2005 Workshop on Stylistic of Text for Information Access. Citeseer; 2005: 19.