A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death

Samuel Danso, University of Leeds, and Kintampo Health Research Centre
Eric Atwell, University of Leeds
Owen Johnson, University of Leeds
Augustinus H. ten Asbroek, Academic Medical Centre, Amsterdam
Seyi Soromekun, London School of Hygiene and Tropical Medicine
Karen Edmond, University of Western Australia, Perth
Chris Hurt, Cardiff University
Lisa Hurt, London School of Hygiene and Tropical Medicine
Charles Zandoh, Kintampo Health Research Centre
Charlotte Tawiah, Kintampo Health Research Centre
Justin Fenty, London School of Hygiene and Tropical Medicine
Seeba Amenga Etego, Kintampo Health Research Centre
Seth Owusu Agyei, Kintampo Health Research Centre, and London School of Hygiene and Tropical Medicine
Betty R. Kirkwood, London School of Hygiene and Tropical Medicine

Abstract

This paper presents a method employed in building a semantically annotated corpus of 11,741 Verbal Autopsy documents, each annotated with Cause of Death, based on verbal records of deaths of mothers, stillbirths, and infants up to 1 year of age, captured for analysis in Ghana between December 2000 and July 2010. Verbal Autopsy is a technique which involves interviewing individuals (such as relatives or caregivers) who were close to the deceased, and if possible, who cared for the individual around the time of death, to document events that may have led to the individuals' death. The Verbal Autopsy technique is recommended by the World Health Organisation as a pragmatic substitute for a clinical autopsy to establish cause of death in regions such as sub-Saharan Africa where death may occur well away from clinical services. An evaluation is carried out based on established criteria to demonstrate that the Verbal Autopsy corpus possesses the qualities of many referenced corpora. The experiences drawn from the methods employed, with alternative approaches, may lead to a more efficient and cost effective corpus development framework.

1 Introduction

An annotated corpus is essential to the development and evaluation of automatic approaches in corpus linguistics research. The biomedical domain is one area that is witnessing a high growth of corpus based approaches to the development of automatic systems. This paper presents the method employed in creating a semantically annotated corpus of Verbal Autopsy documents. The content of this corpus is derived from two questionnaires for conducting Verbal Autopsy interviews to obtain cause of death information, the first for stillbirths and infants (less than 12 months of age) and the second for women of reproductive age.

1.1 Uses of Verbal Autopsy

Of the estimated 57 million deaths per year worldwide, 67 per cent are not medically certified with cause of death due to weak or negligible death registration systems (World Health Organisation 2004). During the Year 2000 Summit, the United Nations set out the Millennium Development Goals for improving world health and these included a determined focus on reducing child mortality and maternal deaths (Sachs and McArthur 2005). To achieve this goal, there is a crucial need for cause of death information from hospitals as well as deaths that occur outside hospitals. Meanwhile, globally, over 40 countries have employed Verbal Autopsy as a means to ascertain the likely cause of death (Fottrell *et al.* 2007). Cause of death information is vitally important for developing health interventions and disease treatment research (Kahn *et al.* 2000). Analysis of Verbal Autopsy data may be invaluable in revealing preventable illness. For example, a locally high death rate due to neonatal infection may help regional health managers respond with strategies such as targeted education on infection control. This information may similarly help to inform national and international health managers, policy makers and researchers about trends in causes of death, in order to develop strategies, design interventions, and carry out sound budgetary allocations (Murray *et al.* 2007a).

1.2 Problems associated with current automatic analysis of Verbal Autopsy

Typically, information gathered using Verbal Autopsies is captured on paper using standard questionnaires, which are then passed to physicians who review them to determine the most likely cause of death (Byass *et al.* 2006). The standard practice worldwide has been the use of a minimum of two physicians to give two independent assessments of each Verbal Autopsy, even though there is some evidence to suggest that one physician may be enough for this process (Joshi *et al.* 2009). The use of physicians for this process is characterised by several limitations: high cost, intra-physician reliability, repeatability, and time consumed (Byass *et al.* 2010). The cost of manual review and assignment of cause of death to Verbal Autopsy documents has not been formally evaluated. However, this exercise may be equated to assigning International Classification of Diseases (ICD) codes to clinical documents, the manual coding of which had an estimated cost of approximately $25 billion per year in the United States alone (Lang 2007). The problem is compounded where there is a shortage of medical personnel and generally this is the case in places where Verbal Autopsy is used.

Consequently, there is a growing interest in research in the use of computational approaches to classify causes of death (King *et al.* 2010; Byass *et al.* 2010), to address the limitations associated with time consuming and expensive physician reviews (Usman 2005). Verbal Autopsy questionnaires contain both structured data and open history narrative. Our literature review found that the computational approaches published so far have only made use of the structured data available (Byass *et al.* 2006; Murray *et al.* 2007b) while physicians have access to and make use of both the structured information and the open history narrative (Soleman *et al.* 2006). Our research is motivated by the belief that computational algorithms that can take into account information yielded from the free text in Verbal Autopsies may improve the contribution they can make to the UN's Millennium Development Goals.

1.3 Corpus based approach to analysis of Verbal Autopsy

The importance of annotated corpora as a resource for carrying out research into automatic analysis of biomedical language research has been well established (Pestian *et al.* 2007). There are numerous semantically annotated biomedical corpora built from biomedical text and made available publicly: GENIA, PennBioIE, Wincosine, Yapex, and GENTAG are examples. Clinical corpora with various levels and types of annotations are also being developed (Pestian *et al.* 2007; Kipper-Schuler *et al.* 2008). These annotations are determined by the intended application, which could be Information Extraction or Text Classification systems. However, there is no similar corpus available for Verbal Autopsy. The only referenced dataset was collected by the Population Health Metrics Consortium (Murray *et al.* 2011) to provide a gold standard for the evaluation of computational approaches to Verbal Autopsy analysis. However, its free-text part has been pre-processed and stored as part of the coded part database. The pre-processing involved for example the elimination of syntactic structures and linguistic information available in the textual data, removal of infrequent words, and consideration of only medically relevant concept-terms amounting to key information loss, and consequently rendering the data unsuitable for language research.

1.4 Verbal Autopsy as another genre of biomedical text

Corpora are generally grouped according to several characteristics: genre or document type, annotation type (syntactic or semantic), availability (either public or private), utility or popularity (how systems utilise them), and intended applications (Cohen *et al.* 2005). The Verbal Autopsy corpus, within this context, is considered a semantically annotated biomedical corpus of another type of genre. Unlike all other biomedical and clinical corpora, which are obtained from a discourse between experts, the Verbal Autopsy corpus was gathered by non-clinical staff interviewing relatives of the deceased and is therefore a discourse between people with limited knowledge or expertise in the biomedical field. Interviews are generally conducted in the local language of the interviewee; for our Ghana dataset, the responses were translated into English by the interviewer. The differences that exist between the language found in Verbal Autopsies and formal biomedical language mean that they must be considered another, rather unusual, sub type of biomedical text. Moreover, the inherent challenges and the complexities the Verbal Autopsy corpus exhibits are comparatively different to the other biomedical corpora. The language analysis systems developed by language researchers in the biomedical domain may not be immediately applicable to Verbal Autopsy and may require adaptation by language system developers and researchers.

1.5 Other potential use of Verbal Autopsy corpus

A window of opportunity exists for the development of prediction models for Verbal Autopsy which can be explored by language researchers. There is also an area of research which is concerned with exploring sociological issues within the biomedical domain particularly the interactions between patients and professionals. Even though the majority of this research has focused on doctor-patient interactions (Candlin and Candlin 2003), some research has considered the interactions between patients and other professionals, for example with nurses; pharmacists; physiotherapists; and occupational therapists. These studies have mainly employed traditional qualitative and quantitative approaches. There is however a paradigm shift in the approaches, and this is being driven by corpus linguistics due to the advantages it offers over established methods (Adolphs *et al*. 2004). For example, using corpus linguistics, Harvey *et al*. (2007) revealed how teenagers were able to discuss confidential health problems with a virtual doctor over the internet with less difficulty than discussions with a doctor in person. Similar analysis of a Verbal Autopsy corpus may lead to a better understanding of some of the sociological barriers that exist in the interactions between patients and the interviewers. This could reveal important findings that will inform Verbal Autopsy researchers and policy makers including the World Health Organisation and may help improve the quality of the information gathering process.

This paper reports on the work done to create an annotated corpus, which is being used in research into the development and evaluation of methods for automatic classification of diseases, at the University of Leeds in the United Kingdom. The research employs Corpus Linguistics, Natural Language Processing and Machine Learning approaches, and this requires an annotated corpus. Our review of the literature indicates that no effort has been made elsewhere to explore the development of a free text corpus of Verbal Autopsies. An earlier publication by Danso *et al*. (2011), demonstrated the potential use of this corpus using a small sub-set. This paper gives a detailed account of the development process and quantitative methods used to establish the full corpus that is now available. This detailed account will serve as reference guide to those who wish to use the corpus or undertake similar projects in the future.

2 Methods for building the Verbal Autopsy corpus

2.1 Source of the corpus and sampling

This corpus is obtained from two large field trials carried out in Ghana, which led to the establishment of a Verbal Autopsy surveillance system and ran between December 2000 and July 2010. The surveillance system covered seven contiguous, predominantly rural, districts within the Brong-Ahafo region of Ghana. The objective of the ObaapaVitA trial by Kirkwood *et al*. (2010a) was to assess the effect of weekly low-dose vitamin A supplementation in women of reproductive age in Ghana on pregnancy-related mortality, female mortality more generally, and peri-natal and infant mortality. The objective of the Newhints trial, also by Kirkwood *et al*. (2010b), was to develop a feasible and sustainable community-based approach to improve newborn care practices in order to improve the survival of newborns. Data was collected during four-weekly surveillance, which included recording all stillbirths, deaths in infants up to one year, and women of reproductive age using Verbal Autopsies conducted by field supervisors.

The corpus contains real Verbal Autopsy text as obtained from the interview and transcribed onto the Verbal Autopsy form. The sample contained all stillbirths and deaths in infants to the age of 12 months, which is referred to in this paper as the ‘infant subcorpus.’ Additionally, it also contains text about the causes of all deaths in adult women between the age of 15 and 45, which is referred to in this paper as the ‘women subcorpus’. The corpus contains a total of approximately 2.5 million words in 11,741 documents. Table 1 below shows the breakdown of the infant and women corpus with their respective number of documents and number of categories:

Table 1: Basic statistics of the Verbal Autopsy corpus

Subcorpus	Number of documents	Number of Cause of Death Categories	Number of words
Infant	8,212	23	1.5 million
Women	3,529	43	1 million

2.2 The interview questionnaire

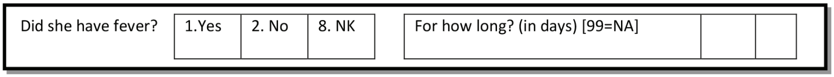
The questionnaire includes identification, coded part and free text sections.

Identification

This part of the questionnaire contained basic contact details of the deceased: identification number, compound number, name of subject, and date of birth. This information is derived from a surveillance database, which was set up by the ObaapaVitA and Newhints trials.

Coded part

The coded-part questions have in total over 200 variables for the infant and just over 270 for the women questionnaire. 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.



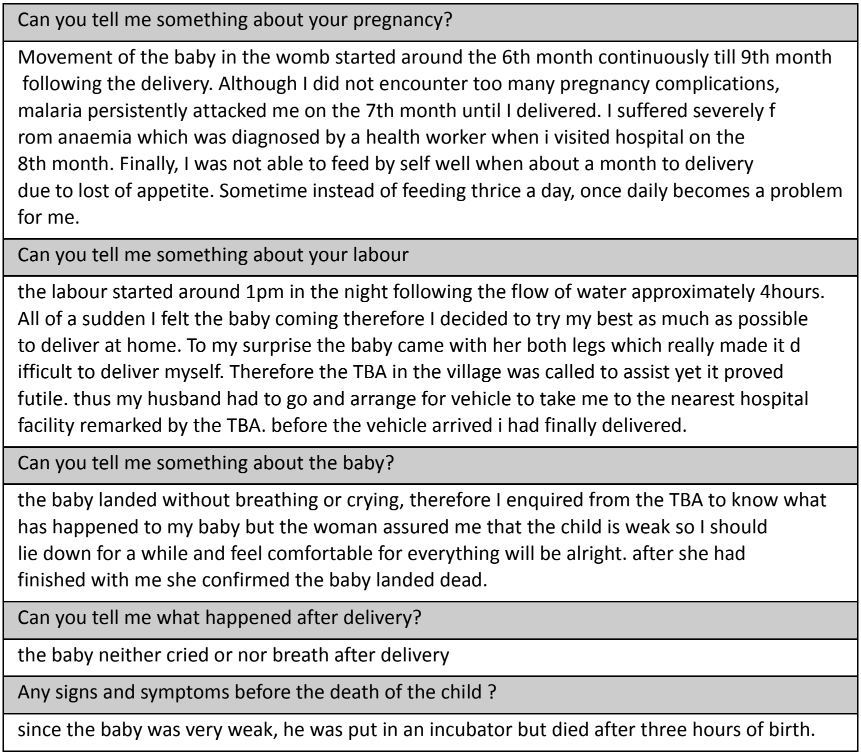
Did she have fever?	1.Yes	2. No	8. NK	For how long? (in days) [99=NA]		

Figure 1: Question and response options provided to respondent during interview

Figure 1 for example is asking a mother of a dead child about whether the child had fever, and where the questions are not applicable, a double line is drawn through those questions and *99* is entered into the databases as the response value to differentiate it from missing information.

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 by physicians when attempting to determine the cause of death. The free-text part gives a verbatim account of the responses from the interviewee. It is worth noting that the free text is a translation of the interview, which is conducted in a local language (Twi) and the interviewer in turn translates into English as seen in Figure 2:



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 f rom 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 d ifficult 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 2: A sample of free-text question and responses from Infant Verbal Autopsy questionnaire

Another notable difference between women and infant subcorpora worth pointing out here is that the structure of the questionnaire used for the infants was different from the questionnaire used for the women. The free text part of the infant questionnaire had clearly defined sections that asked for specific information as shown in Figure 3. The women questionnaire however did not have sections but had blank pages that required the interviewer to write during the interview. It is also important to note that the Verbal Autopsy questionnaires were designed based on the most standard methods available at the time of the design. The infant questionnaire was based on the World Health Organisation standard, which was locally adapted by Edmond *et al*. (2008). The adult women questionnaire was based on a validated instrument for adult deaths by Chandramohan *et al*. (1994) and an instrument for maternal deaths developed for a previous study by Campbell and Gipson (1993).

2.3 The selection and interview process

As indicated, the corpus was generated as part of field trials, where routine visits were made to participants of the study for data to be collected. The data, which included events that occurred concerning the study participants (including deaths) at the time of visit were processed and stored in a database. A list of infant and adult female deaths was generated from the database to indicate which interviews needed to be conducted. This list was shared among a group of interviewers who had been trained to conduct Verbal Autopsy interviews. An earlier study suggested that a period between one and 12 months after death is recommended to elicit reliable information from respondents (Soleman *et al*. 2006). Although it is probably better to conduct the interview as soon as possible after the death to avoid recall difficulties, the researchers also have to respect the family's need to grieve and a minimum of six weeks was therefore set before conducting the interview.

Background of the interviewers

The interviewers recruited for the Verbal Autopsy information collection were people who had completed high school, spoke the local language as well as English, and had attained fieldwork supervisory status with no medical training. These people were recruited and provided with basic training on how to administer the Verbal Autopsy questionnaire. For example, interviewers were given specific training on how to probe in order to elicit the information relevant to enable physicians to arrive at the possible cause of death. There was a mixture of both male and female interviewers and this did not seem to influence the data collection process.

2.4 The annotation process

For this research, annotation is defined as the process of reviewing the Verbal Autopsy document and assigning a cause of death code to it. We employed methods similar to the one described by Pestian *et al*. (2007) as a *democratic principle* approach to creating the final cause of death code for each Verbal Autopsy document.

A *coding sheet* was generated with ID details of each death for use by the physicians during their review. The coding sheets and Verbal Autopsy document were duplicated and a set passed to two different physicians. Each physician independently reviewed the Verbal Autopsy document and recorded their judgement as to the most likely cause of death on the coding sheet. The sets of documents were returned and the coding sheets input into a database where a comparison between the two assigned causes of death was made using custom made software. An agreed code was assigned when the two physicians agree. Otherwise, the process was repeated with a third physician. When there was an agreement between any two of the three assigned causes of death, then that was accepted; but when there was no agreement, a meeting was held between the physicians involved for an agreement to be reached. In the event where there was still no agreement, the cause of death was declared *unexplained*.

Alternative annotation process management models were explored to determine which would be least resource intensive without compromising on quality. The first model tried was that annotators were given basic training in the annotation task; then they were given the completed questionnaires to review and produce the annotations in their own time and preferred location. In a second model, employed later, annotators were assembled at a central location to carry out the annotation.

Background of annotators

Medically trained professionals were employed to carry out the annotation. The minimum qualification of an annotator was having completed a medical school and one year housemanship training in a hospital. To ensure high quality annotation and high level of agreement, all annotators were trained using standard annotation guidelines developed by the project management team, and led by paediatricians for the infant and maternal health experts for the women.

2.5 The transcription process

Tailor-made software was designed for this activity. The software had inbuilt spellchecking functionality that allowed typographical errors to be corrected. This however did not filter out all possible typographical and contextual errors. For example *died* and *dead* are both correctly spelt but *died* may be the inappropriate word contextually. Each questionnaire went through a transcription process where the entire content was entered by a data entry clerk. As a general rule, the content of the questionnaire was not to be changed at this stage. 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.

2.6 Processing and storage

The software used for the transcription stored the corpus in a Microsoft Access database. This format was not directly usable for corpus analysis although the corpus was now available in a digital format. Also, the causes of death codes assigned by the physician were stored in a separate database and there was the need to have all this information stored in a single file. The two separate databases were linked together using the unique identifier, which was assigned at the point of registering the study participant. The resulting database was further encoded into an XML format as this format allows all the individual files to be stored as a single file. An example of this is as shown in Figure 3:

```
<Cleaned_InfantVA_InfantVA_CoD>
    <infantid>KA017/1/08C1</infantid>
                <Pregnancy> I made eleven consecutive antenatal check–ups at birth xxx hospital
                and xxx maternity home respectively. I received two tetanus injections during the
                pregnancy and these, were the problems encountered : – painless bleeding at the
                initial stage of 2–3 months old (which I had drugs from xxx hospital for treatment),
                reported to have an anaemia at the age of 4 months old with headache/dizziness
                at the same time which, I had drugs and was advised to eat green vegetables.
                Besides these problems I had malaria (at the age of 3 months), severe abdominal
                pains (6 months old) restlessness after meals of fatal movement at the left rib side
                at my seventh month. My pregnancy ended, at 11 months old at prince of peace
                maternity home through normal birth.</Pregnancy>
                <Labour>Labour started around 6:00am and gave birth around 1:00pm but had
                water (green in colour with no bad scent) breathing during labour. Problems had
                during labour were: – lower abdominal pains and the umbilical cord which came
                out before the child followed.</Labour>
                <Baby>At birth, the child was an average in size with no malformation, but was
                having difficulty in breathing with blood flowing through the nose.</Baby>
                <AfterDelivery>At birth the blood flowing the nose was cleared, and artificial air
                was poured into the nose to assist him to breath well.</AfterDelivery>
                <IncidentToChildDeath>Child was having difficult breathing at birth and blood
                flowing through the nose as well. After clearing the blood from the nose an artificial
                air was pumped into the nose to assist him to breath well. The breathing rate was
                very fast and was abnormal. Even the breathing rate was very fast but at times, the
                breathing could stop for some time before it could come out again and the problems
                continued up to 12:00pm (mid–day) and finally died at the maternity home. Another
                symptoms detected on the child was the physical colour changed into green at
                birth.</IncidentToChildDeath>
    <CoD>Neonatal–other causes</CoD>
<Cleaned_InfantVA_InfantVA_CoD>
```

Figure 3: Various XML tags mapping to various sections of an Infant VA document with cause of death information merged

2.7 Anonymisation

As with other clinical data and text in particular, the identity of subjects in a Verbal Autopsy must be kept anonymous. To achieve this, basic details of all subjects were stored in a separate database and references were made to all documents using an ID which was allocated during the registration phase. The basic details database was obscured and was only accessible to the data collectors and trial management team. The challenge however was to find ways to deal with the named entity mentions and references made to individual details in the text. This challenge required careful handling since removing all the mentions and references in the text could lead to loss of potentially useful information. It was however found that sensitive information which seemed to need anonymisation in the free text could not mean much without the individual details. Even though there were occasional mentions of names of relatives, however, these named entities could not be traced without additional information such as location and address which had been removed. Pestian *et al.* (2007) made a similar observation in a corpus of clinical text, which was built for the development and evaluation of a multi-label classification shared task.

2.8 Ethical approval

Ethical approval was obtained from relevant ethical committees to carry out the two field studies. Consents were also sought from the participants of the field trials on the use of their data for research purposes.

3 Analysis of the corpus

McEnery and Wilson (1996) describe a modern corpus as a collection of text which has the following characteristics: machine readability; representative sample; corpus must have a finite size; and must have achieved status of standard reference. The attributes are expanded further: storage format of annotated corpus; and annotation levels. These attributes have been found to have effect on the use of corpora beyond their original purpose of creation (Cohen *et al*. 2005). The corpus building process described above suggests that the corpus is encoded in a format readable by machines as required. It has however not attained the status of reference, and this publication marks the starting point to achieving the referenced status. The following analysis shows the general characteristics of the Verbal Autopsy corpus.

3.1 Difference between infant and women subcorpora

Document length

Figures 4 and 5 show a wide spread of the length of the documents which make up the corpus:

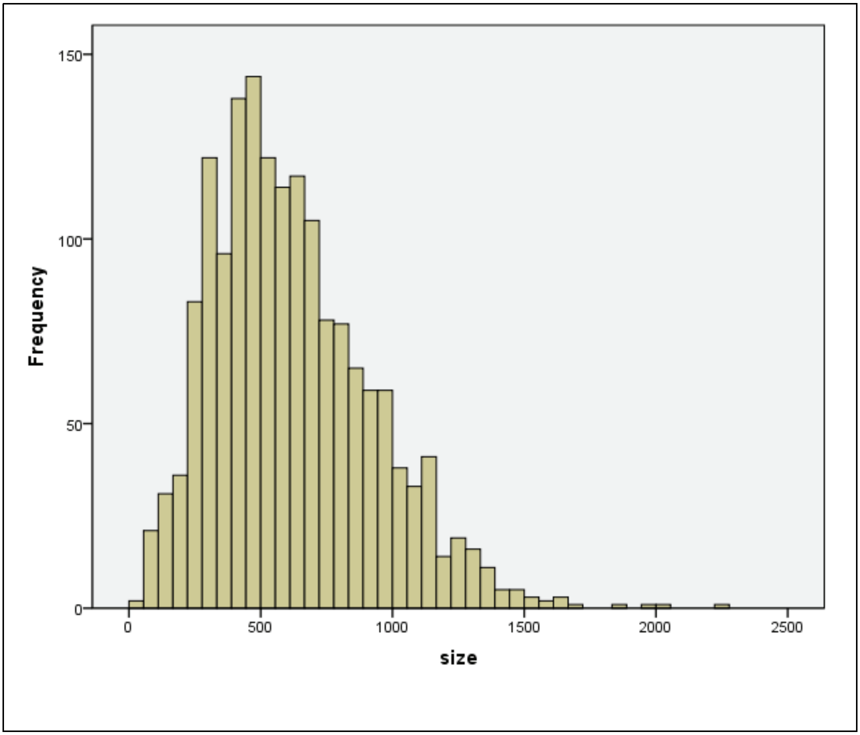


Figure 4: Distribution of document size of the infant subcorpus

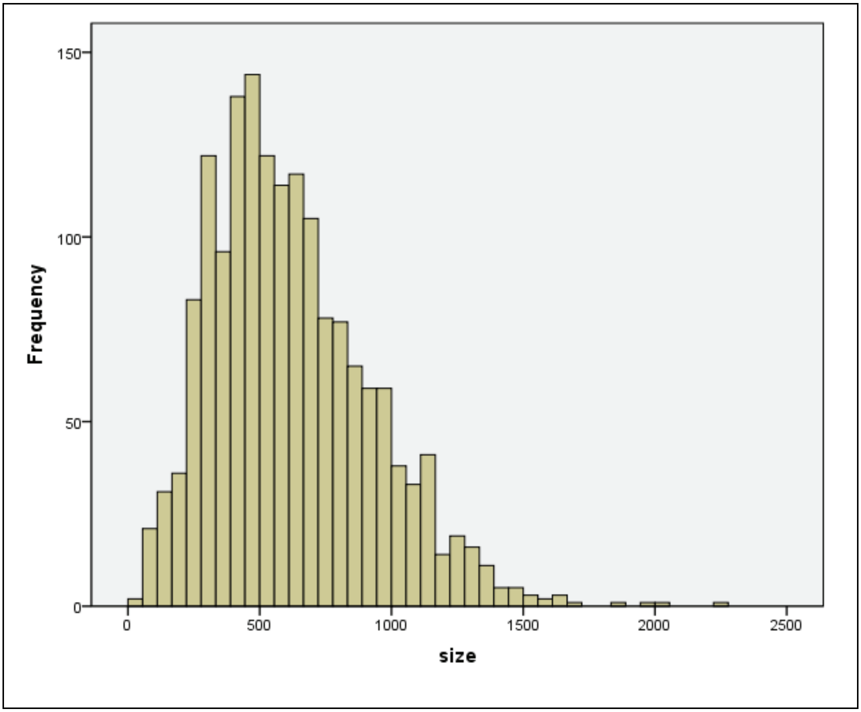


Figure 5: Distribution of document size of the women subcorpus

Figure 4 indicates the spread of the infant subcorpus, ranging between 15 and 550 words, with an average of 182. Similarly, Figure 5 also gives an indication of the spread of the women subcorpus length, ranging between 33 and 2200 words, with an average of 630 words. This suggests that the women subcorpus tends to contain much longer histories of events that led to the deaths compared to the infant subcorpus. Variation in document length plays a key role in computational methods and this characteristic is worth noting. For example, in document classification, it is important to consider normalisation during feature representation, which may take into account the length of the document (Leopold and Kindermann 2002).

Key-word comparison

The infant and women log-likelihood comparison: Table 2 is a log-likelihood key-word comparison between the infant and women subcorpora:

Table 2: Top 20 key-*words* in infant and women subcorpora

Infant		Women	
Log-likelihood	Word	Word	Log-likelihood
18457.747	I	She	19381.94
5003.486	baby	Her	4813.075
2824.311	my	Deceased	3800.457
2426.954	child	Complaining	3110.407
1946.707	birth	Complained	2451.803
1939.428	labour	Hospital	1264.405
1554.711	delivered	Admitted	1198.602
1508.423	pregnancy	Sent	1109.615
1475.825	delivery	Later	932.191
1288.805	he	Years	926.846
1173.189	me	For	883.649
906.209	during	Died	880.077
849.069	any	Complain	737.514
779.007	when	Became	728.608
738.985	normal	Admission	677.243
727.43	born	Heart	404.027
628.19	him	Typhoid	267.469
589.087	did	hypertension	120.225
548.812	immediately	Accident	118.613
538.709	average	Cancer	110.498

The table demonstrates some key lexical similarities and differences that exist between the two subcorpora. As observed from the table, there are similarities in the use of pronouns, which tend to feature prominently in both corpora. For example, *I* is the most frequent word in the infant corpus whereas *she* appears as the most frequent word in the women subcorpus. This suggests a true reflection of what pertains in the discourse during Verbal Autopsy interview. The *I* appearing as the most frequent word in the infant subcorpus also suggests that the mother of the child is the one who mostly give an account of what led to the death of the child. This narrative begins with the events that surrounded pregnancy. In contrast, the most frequent word in the women subcorpus is *she* referring to the deceased, and *I* is not in the top 20, suggesting the relative giving the account was not as directly involved in events leading to the death. Observing the top 100 words from both corpora further differences can be established. Notably, *labour*, *delivery*, and *baby* are all words that describe the infant corpus, while in contrast words such as *complained*, *accident*, *hypertension*, and *cancer* are words that frequently occur in the women subcorpus.

The Verbal Autopsy corpus and Zipf's law

The Verbal Autopsy corpus word frequency distribution was assessed against Zipf's law, also known as a Power-law distribution, which is found to occur in a range of phenomena including words in human language (Newman 2005). Zipf's law establishes the relationship between the frequencies of words and the rank in which it occurs in the list (Manning and Schutze 1999), where rank number 1 represents the most frequent word in the corpus. This can be expressed mathematically as: R * F = C, where R is the rank of a given word, and F is the frequency of that word and C is a constant that depends on the corpus.

Using Python scripts, all the text was extracted from the XML file into a plain text file. A natural language analysis toolkit called AntConc (Anthony 2005) analysed the entire plain text file to generate word frequencies and their ranks. The output of this was used to plot a graph in Figure 6, which shows Zipfian word distribution in the Verbal Autopsy corpus transformed onto logarithm scales.

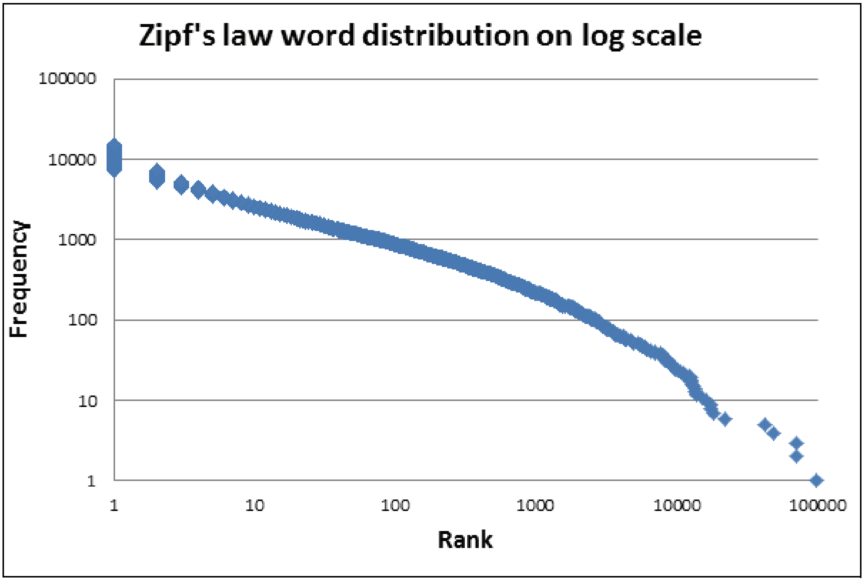


Figure 6: Word frequency distribution of the Verbal Autopsy corpus on log scale

As can be seen the graph has a negative slope, which suggests a word distribution phenomenon expected from Zipf's law. The proportion of words with higher ranks but very low frequency has implications for lexical modelling, as one will have to deal with this high level of lexical sparseness. This characteristic is tested later in the paper.

3.2 Balance and representativeness in sampling

Cause of death groupings

Having a balanced text is a key requirement in corpus linguistics and machine learning tasks. This implies that a sample drawn from any domain of interest must give a balanced representation of that domain. For example, in language research, one might aim to give a balanced representation of text type: written versus spoken. Similarly, within text classification, an important issue to consider is the semantic categories or genre types. Within the context of Verbal Autopsy, one of the semantic categories, which may require a balanced representation, is the cause of death. Figures 7 and 8 show the distributions of these characteristics found in the Verbal Autopsy corpus:

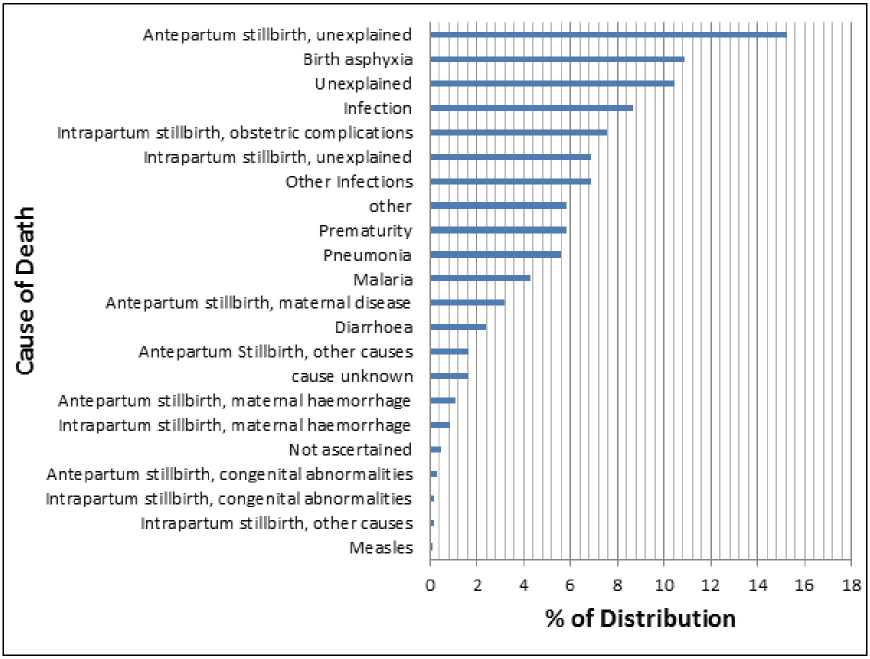


Figure 7: Cause of death distribution among the infant subcorpus

Figure 7 shows cause of death distribution found in the infant subcorpus. As can be seen, there is a high level of imbalance of cause of death categories in both the infant and women subcorpora. *Uncertain* is the commonest category, with over 20 per cent of the Verbal Autopsies that make up the women subcorpus. *Antepartum Stillbirth Unexplained* is the leading cause of death in the infant subcorpus, covering over 15 per cent. Note that the infant subcorpus also has separate cause of death categories *Antepartum Stillbirth, Other Causes*, *Unexplained*, *Other*, *Cause Unknown, Not Ascertained.* To a corpus linguist without medical training, it may not be clear how to distinguish between all categories; we have accepted the semantic tag-set as delivered by the medical experts without attempting to improve their tagging. A significant proportion of these categories are ill defined: *Cause Unknown*, *Other*, *Not Ascertained*, and *Unexplained*. These categories may require special attention by language researchers when carrying out computational modelling.

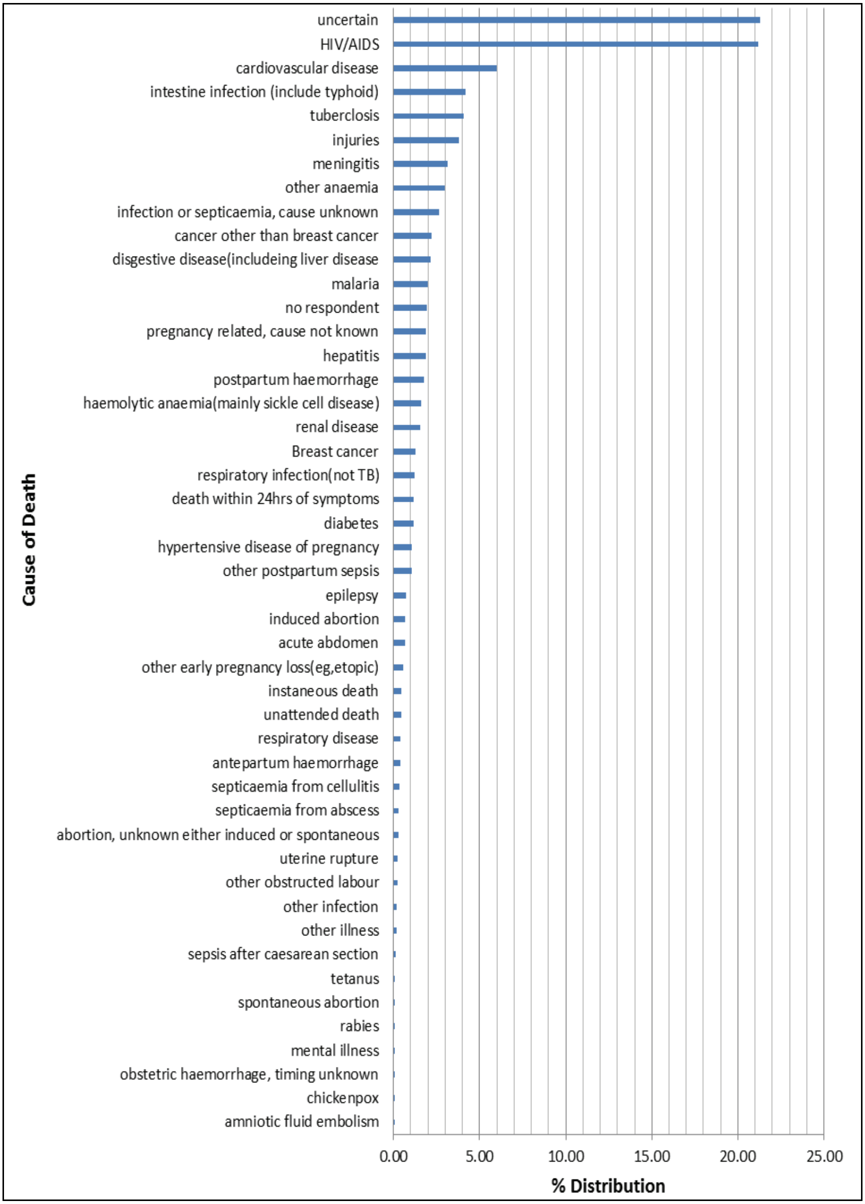


Figure 8: Cause of death distribution among the women subcorpus

Figure 8 shows a rather different distribution of cause of death in the women subcorpus. Two categories, *Uncertain* and *HIV/AIDS* stand out as most significant, each the cause of over 21 per cent of deaths; then there is a long tail of other specific medical conditions, most of which are not found in the infant subcorpus. The vague categories *Unexplained*, *Other*, *Cause Unknown*, *Not Ascertained* do not figure in the women subcorpus; we assume these are covered by *Uncertain*. Figure 8 also illustrates the extent to which the corpus is skewed in terms of the distribution of categories. There are even rare cases of just one instance each in Figure 8: *Obstetric Haemorrhage, Timing Unknown*, *Chickenpox*, and *Amniotic Fluid Embolism.*

Analysis of cause of death groupings

The scheme adopted in assigning the cause of death categories has some form of hierarchy. The infant corpus for example adopted a classification system as illustrated in Figure 9:

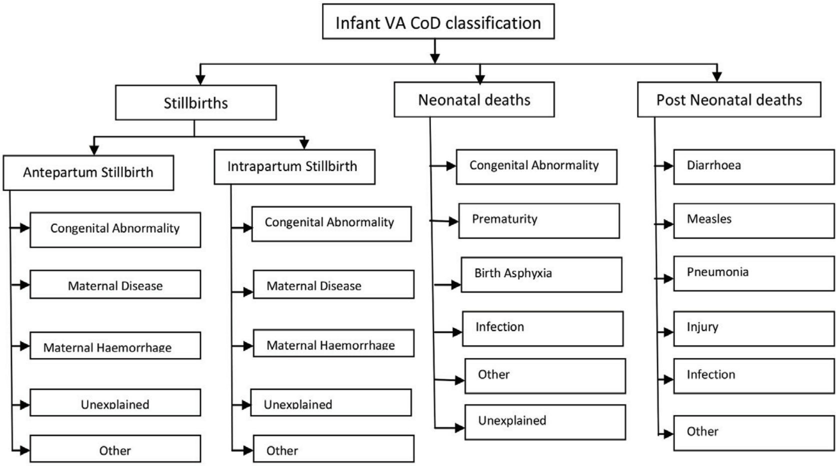


Figure 9: Schematic diagram showing the hierarchy of causes of death adapted from Edmond et al. (2008)

The classification scheme takes into account reported medical symptoms, and the time with respect to delivery in which the death occurred. As seen from Figure 9, *Congenital Abnormality* for example may be classified as either *Antepartum* or *Intra-partum* depending on whether the death occurred before the labour or during labour, as explained by Edmond *et al.* (2008). Such groupings according to stage of delivery do not feature in the women subcorpus cause of death tag-set, except for three causes of death involving haemorrhage at childbirth: *Postpartum Haemorrhage*, *Antepartum Haemorrhage*, *Obstetric Haemorrhage, Timing Unknown*.

It is important to note that the principle of hierarchical groupings of tags is not new to corpus linguistics. This principle has been proposed as a means of dealing with ambiguity in word Part of Speech (PoS) categorisation by many language researchers, e.g. Knowles and Zuraidah Mohd (2003). Atwell *et al.* (1994) proposed this principle in machine learning research on unsupervised learning of word-clusters. It can be deduced from this that the deciding factor of any categorisation scheme for automatic systems is the usefulness of the scheme within the context of the intended application. This has resulted in various PoS-tagging schemes for English language for example. The International Corpus of English (ICE) is an example of a corpus with multiple-level tag set (Greenbaum and Nelson 1996); a PoS-tag consists of a broad category (e.g. noun) and a set of finer-grained subcategories (e.g. common singular noun). Drawing inspiration from the ICE tagging scheme, the semantic categories of the Verbal Autopsy documents demonstrated in Figure 9 could have multiple categories: Time-Of-Death and Type-Of-Death.

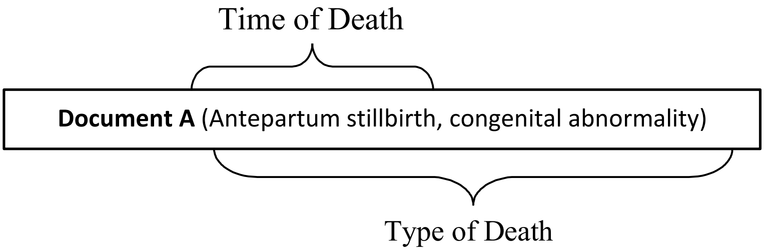

Time of Death
Document A (Antepartum stillbirth, congenital abnormality)
Type of Death

Figure 10: Sample of re-classified Verbal Autopsy document based on hierarchy

This can be illustrated with the Figure 10. The figure indicates a classified Verbal Autopsy document based on its content and the cause of death hierarchy. The proposed scheme as shown in this figure may enable documents to be classified as either *Antepartum Atillbirth* or *Antepartum Stillbirth*, *Congenital Abnormality*. Having these levels of classification has an advantage for machine learning approaches to automatic classification. This is because the accuracy of machine learning models tends to increase with increase in number of training examples (Banko and Brill 2001). In situations where there are rare cases of specific examples of Cause of Death in the training corpus, it may be reasonable to have a Verbal Autopsy document classified at higher level (Time-Of-Death) than the fine grained level (Type-Of-Death).

3.3 Sparseness and lexical diversity

Another property that is exhibited in most natural language text, as a side-effect of Zipf's law is the issue of sparseness (Goweder and De Roeck 2001). A corpus is said to be sparse if many of the words in the corpus are uncommon or unknown. It could also mean that some word-combinations or Ngrams are rare in the corpus (Jurafsky and Martin 2008). The figures shown above suggest a high level of sparseness and imbalance in vocabulary and also in the cause of death categories, which could be problematic when trying to apply machine learning: classifiers work best when features of the data and classification categories are reasonably frequent and balanced (Lakeland and Knott 2004).

An assessment was carried out using word type-token ratio to ascertain the lexical variability that exists in the Verbal Autopsy corpus (Youmans 1991). The word type-token ratio is defined as the number of distinct words in a given text divided by the length of the text, and should have a value between *one* and z*ero*. A ratio approaching *one* is a high ratio; this means that each word occurs only once in the text, and therefore the text is lexically sparse. On the other hand, if the type-token ratio is a small fraction close to 0, then the text has low sparseness; this implies that there are a few words repeated many times in the text, which should result in a closed vocabulary text. To test this, we applied the experiment originally carried out by Yahya (1989) to observe the behaviour of word occurrence patterns in Arabic with respect to English, and further adopted by Goweder and De Roeck (2001). For this experiment, we compare the type-token ratio of samples from our corpus with figures based on equivalent-sized samples from the Brown corpus (Francis and Kučera 1979). Some do argue the type-token ratio experiment should exclude stopwords (Graesser *et al.* 2004). However, as pointed out by Forman (2003), stopwords could be domain specific since what is considered a stopword in one domain could be relevant in another domain. For example the word *during*, which is regarded as a function word but also forms part of an expression *during labour* could be the discriminative word that could predict a document belonging to *Intra-partum* as opposed to *Ante-partum* category. The experiment was therefore carried out with all the words.

Table 3: Type-token ratio derived from the infant subcorpus

Length of text	VA corpus distinct words	VA corpus word ratio	Brown corpus distinct words	Brown corpus word ratio
100	63	0.63	69	0.69
200	109	0.54	124	0.62
400	177	0.44	165	0.41
800	279	0.38	328	0.41
1,600	398	0.24	621	0.38
3,200	596	0.18	871	0.27
6,400	839	0.13	1,361	0.21

In Table 3, the word type-token ratio reveals an interesting characteristic of the Verbal Autopsy corpus. It can be observed from the table that the majority of the values for the ratios are lower compared to the corresponding values found in the Brown corpus, which supports the intuition that the Verbal Autopsy corpus has a (relatively) closed vocabulary. This view seems to hold as the sample text length increases. For example the text length of 6,400 has ratio of 0.13 and 0.21 for Verbal Autopsy and the Brown corpus respectively, which suggests that the Brown corpus tends to be sparser or contains more uncommon words than the Verbal Autopsy corpus. However, it could also be observed that there is still a significant amount of vocabulary variability in the Verbal Autopsy corpus samples. One possible reason for this issue is the variation in which concepts are expressed. For example, the concept *delivery* is expressed in a variety of ways: *baby came out* and *baby landed*.

3.4 Computational complexity of the Verbal Autopsy Language

The quality of the text introduces challenges for automated text analysis and tagging. PoS tagging is a standard annotation added to corpora, and accuracy of the PoS tagging is important in achieving reliable analyses. However, accuracy of PoS tagging is dependent on the quality of the text. We examined this issue by assessing the accuracy of a standard PoS-tagger, from the NLTK Natural Language Tool Kit (Loper and Bird 2002). We trained the NLTK PoS-tagger with a standard PoS-tagged English corpus, the Brown corpus; and this tagger was then used to tag a sample of Verbal Autopsy documents. The output of the tagger was then evaluated by a PoS-tagging expert (Atwell, see Atwell *et al*. 2000, Atwell

2008), who found that the accuracy of the PoS-tagger on the Verbal Autopsy corpus was 88 per cent. This result is much lower and a significant departure from the performance of PoS taggers on general 'well-formed' English text, generally 96–97 per cent (Brants 2000).

Several factors could account for the low performance of the PoS tagger. One possible factor is the language used in the Verbal Autopsy reports. The grammatical structures found in the Verbal Autopsy text differ from those found in standard English text such as news. For example: *Before labour waters, which look clear and without bad scent broke*. is a typical sentence in the Verbal Autopsy narratives. This example clearly shows the grammatical problems present in the text required to be handled by the PoS tagger. Other problems include wrong use of punctuation, and unknown words. For example *intravenous* was not found in the training corpus and therefore resulted in a wrong PoS tag of plural noun being assigned by the PoS tagger as the default tag for words ending with *-s*.

3.5 Choice of formats and encoding standards

There are several choices available as to the format and standards in which a corpus is to be processed and stored. However, the overriding principle is to allow users the flexibility and the freedom to manipulate and access annotation information. A survey conducted by Cohen *et al.* (2005) on corpora formats suggests that none of the non-XML formats met recoverability criteria. In other words, a corpus encoded in any format other than XML can pose difficulties when mapping between the annotations and the original text. XML encoding is the recommended strategy as it enables a standoff storage, where the annotations are stored separately from the original text (Leech 1993). The advent of various software tools and standard libraries in the application programming interfaces of standard programming language such as Java and Python has made *stand-off annotation* with XML a more convenient encoding approach. This justifies the decision to encode the Verbal Autopsy corpus in an XML format.

3.6 Effect of annotation

The Verbal Autopsy corpus lacks linguistic or syntactic annotations; the Cause of Death classification of each Verbal Autopsy is a medical semantic annotation. The findings of a survey suggest that a significant number of biomedical corpora tends to have no linguistic annotations (Cohen *et al.* 2005). This does not rule out the need for linguistically annotated corpora since this information might increase the usefulness as evidenced by the popularity of some corpora over others within the biomedical domain. The GENIA corpus for example has been employed in numerous tasks outside the original purpose for which it was built. GENIA has linguistic annotations in addition to semantic annotations.

4 Experimental results

An experiment was conducted based on the proposed new hierarchical groupings to identify potential keywords for these groupings. These words could be used as features for machine learning. This experiment in other words could be described as feature extraction which is a well-established technique in computational modelling (Worzel *et al.* 2007). The underlying hypothesis of this experiment is that there are words which should be specific to a particular cause of death category. These words could be the symptoms mentioned in the text. This experiment was to illustrate the feasibility of using corpus based methods to analyse and identify keywords in text that could correlate with a given cause of death category. The experiment focused on the most common cause of death categories of infant Verbal Autopsies.

To test this hypothesis, the corpus was divided according to the new cause of death groupings. One large corpus was created, a combination of all cause of death groupings. Additionally, separate subcorpora were created for each category. Using the AntConc software (Anthony 2005), log-likelihood scores (Rayson and Garside 2000) were computed for each cause of death subcorpus using the combined corpus as a reference corpus. The software allocated a log-likelihood score to each word in the subcorpus, with high scores allocated to words that are peculiar to a category and lower scores to words that are common to both specific category and the reference corpus. This procedure was repeated for all the categories.

Table 4: Top 10 key-*words* in 'Time-of-Death' subcorpora

Intra-partum Stillbirth		**Antepartum Stillbirth**		**Neonatal death**		**Post Neonatal death**		**Non Stillbirth Unknown cause**	
Word	**LLH**	**Word**	**LLH**	**Word**	**LLH**	**Word**	**LLH**	**Word**	**LLH**
dead	410.6	dead	485.0	Weak	240.9	he	138.3	bed	77.4
still	277.9	i	386.5	Cry	195.6	healthy	107.6	him	58.5
i	246.6	womb	269.9	breathing	188.3	sent	95.0	adwoa	57.3
out	143.7	movement	205.9	Could	121.5	diarrhoea	86.2	healthy	55.6
coming	95.0	stillbirth	191.9	Died	119.5	bought	73.3	any	54.7
stillbirth	91.9	macerated	189.6	Breath	110.6	hot	73.3	problem	36.2
came	80.4	still	176.7	incubator	108.3	him	71.7	kandege	34.6
already	77.9	already	153.2	Machine	103.9	coughing	71.6	wake	34.0
deliver	63.0	my	109.3	Oxygen	99.3	her	66.5	Slept	33.5
operation	55.5	moving	108.7	Minutes	88.4	she	63.7	without	33.5

Table 4 shows the top ten words obtained from the experiment. As seen from the table, the experimental result supports the hypothesis. All the top ranked words can intuitively be associated to the cause of death categories. For example it is not surprising to see that *dead*, *already*, *coming*, *stillbirth* appeared as the top ranked words for *Intra-partum Stillbirth* but did not appear in the *Non Stillbirth Unknown* cause category. *Dead* could be an indication of a baby not born alive and *coming* could be an indication of *Intra-partum*, suggesting that the baby died during delivery. Similarly, words such as *cry*, *breathing*, *incubator*, *oxygen* being strong indicators for *Neonatal* death suggesting the baby was born alive because it cried, but died later after keeping the baby in an incubator. These words however, did not appear in the *Intra-partum Stillbirth* or *Antepartum Stillbirth* categories.

5 Discussion and conclusion

We have presented in this paper a semantically annotated corpus of 2.5 million words, which is suitable for the development of automatic methods for analysis of Verbal Autopsy and other language research. The method employed in collecting and building the corpus has been presented, including quality measures taken during the annotation process to ensure quality annotation was obtained.

Furthermore, various formal methods have been used to describe and evaluate corpus suitability for language research. The limitations associated with the method employed in building the corpus have also been outlined, and alternative methods have been discussed, which may lead to a more efficient, cost effective and a general purpose Verbal Autopsy corpus for language research.

We have also demonstrated the feasibility of employing corpus based approaches in the analysis of this corpus by identifying keywords associated with a given cause of death category. The experiment focused on the infant sub-corpus and on the high level of cause of death grouping.

The method employed in transcribing the corpus was one among several approaches that could have been explored, for example Optical Character Recognition or speech-to-text recognition. Another approach could have been to electronically capture the corpus at the point of the interview by use of portable devices. With hindsight these approaches could have been more cost effective but the operational difficulties of the environment are challenging and a paper data collection approach with data entry clerks transcribing the results was a pragmatic choice.

As indicated earlier, the annotation approaches employed in this project also had options which could have eased the logistics burden imposed on the management of the process. One of the key challenges associated with the first approach was the difficulty in coordinating the activities to ensure that the processes were conducted in a timely manner. The second model however was relatively less difficult with regard to the coordination of activities. It was however logistically intensive and expensive in terms of the organisation. One approach which could possibly overcome these challenges is the adaptation of an emerging collaborative annotation tool such as GATE (Bontcheva *et al.* 2010). This tools offer the flexibility to allow annotators to work from any location and coordination is also carried out with ease. It however requires internet connection, which is only recently becoming available in the developing world.

The comparison of the lexical diversity between Verbal Autopsy and Brown corpora was based on convenience since the type-token ratio for the Brown corpus was already available to compare with. This may not be the most appropriate comparison considering the content of the Brown corpus is general published American English text it may be interesting to compare these results with a corpus from the biomedical domain. This however does not invalidate our finding about the lexical diversity of the Verbal Autopsy corpus using the type-token ratio between *zero* and *one* as reference range.

The poor performance of the PoS tagging program demonstrates the computational complexity of the language used in the Verbal Autopsy corpus. This raises questions about the robustness of PoS taggers developed and evaluated using standard English corpora such as Brown, LOB, or Wall Street Journal corpora. The performance may suggest the need for revaluation of state-of-the-art PoS algorithms, to assess their ability to deal with a wider range of text with various degrees of computational challenges such as the Verbal Autopsy text. However, our experiment was conducted using the Brown corpus to train the PoS tagger, and potentially accuracy could be improved if we could obtain a Verbal Autopsy corpus annotated with PoS tags to serve as a training set for the PoS tagging process. This may be a useful approach to revaluation and adaption of PoS taggers for Verbal Autopsies.

The high level of imbalance found in the cause of death distribution may be partly due to the methods employed in collecting the samples that form the content of this corpus. The approach adopted was natural as this process replicated a real world scenario. The rare cases may be a reflection of what pertains in the population from which the data was collected. One option, based on background information, may be to target populations with high incidence rate of the rare cases in order to maximise the chances of identifying these rare categories, to make up the numbers needed to achieve a balanced corpus. This may be part of the preparatory phase of the corpus collection process.

The hierarchical classification scheme proposed may be described as a radical way of dealing with the skewedness found in the corpus. However, this approach may be helpful to potential users in providing at least part of thecategory of a given Verbal Autopsy document. For example having information that a child was born alive but died later facilitates the investigation process compared to no information at all.

Considering the fact that over 40 countries employ Verbal Autopsy as an alternative approach to determining cause of death, this corpus creates a new dimension to language research within the biomedical domain. This new area of research could be useful to these countries. A future task could be to draw experiences from this project to build a corpus of multilingual content covering a range of countries that use Verbal Autopsies. We aim to make the corpus available to other researchers in the near future.

References

(Web versions accessed in February 2013.)

Adolphs, Svenja, Brian Brown, Ronald Carter, Paul Crawford and Opinder Sahota. 2004. Applying corpus linguistics in a health care context. *Journal of Applied Linguistics* 1**:** 9–28. http://www.brown.uk.com/publications/adolphs.pdf.

Anthony, Laurence. 2005. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Professional Communication Conference*, 729–737, Tokyo, Japan: IEEE. http://www.antlab.sci.waseda.ac.jp/research/ipcc_2005_anthony_antconc.pdf.

Atwell, Eric, John Hughes and Clive Souter. 1994. AMALGAM: Automatic mapping among lexico-grammatical annotation models. In J. Klavans and P. Resnik (eds.). *The balancing act – combining symbolic and statistical approaches to language,* 21–28. Las Cruces, New Mexico: Association for Computational Linguistics.
http://acl.ldc.upenn.edu/W/W94/W94-0102.pdf.

Atwell, Eric, George Demetriou, John Hughes, Amanda Schriffin, Clive Souter and Sean Wilcock. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal* 24: 7–23. http://icame.uib.no/ij24/atwell.pdf.

Atwell, Eric. 2008. Development of tag sets for part-of-speech tagging. In A. Lüdeling and M. Kytö (eds.). *Corpus linguistics: An international handbook*, Vol. 1, 501–526, Berlin: Walter de Gruyter.

Banko, Michele and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 26–33. Toulouse, France: Association for Computational Linguistics. http://acl.ldc.upenn.edu/P/P01/P01-1005.pdf.

Bontcheva, Kalina, Hamish Cunningham, Ian Roberts and Valentin Tablan. 2010. Web-based collaborative corpus annotation: Requirements and a framework implementation. In *New Challenges for NLP Frameworks*, 20–27, Valletta, Malta: European Language Resources Association. http://gate.ac.uk/sale/lrec2010/teamware/teamware-lrec10.pdf.

Brants,Thorsten. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing,* 224–231. Seattle, Washington: Association for Computational Linguistics. http://www.aclweb.org/anthology-new/A/A00/A00-1031.pdf.

Byass, Peter, Edward Fottrell, Dao Lan Huong, Yamene Berhane, Tumani Corrah, Kathleen Kahn, Lulu Muhe and Do Duc Van. 2006. Refining a probabilistic model for interpreting verbal autopsy data. *Scandinavian Journal of Public Health* 34**:** 26–31.

Byass, Peter, Kathleen Kahn, Edward Fottrell, Mark A. Collinson and Stephen M. Tollman. 2010. Moving from data on deaths to public health policy in Agincourt, South Africa: Approaches to analysing and understanding Verbal Autopsy findings. *PLoS Medicine* 7(8) http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000325.

Campbell, Oona and Reginald Gipson. 1993. *National maternal mortality study, Egypt 1992–93; report of findings and conclusions.*Cairo: Directorate of Maternal and Child Health Care, Ministry of Health and Population.

Candlin, Christopher N. and Sally Candlin. 2003. Health care communication: A problematic site for applied linguistics research. *Annual Review of Applied Linguistics* 23: 134–154.

Chandramohan, Daniel, Gillian H. Maude, Laura C. Rodrigues and Richard J. Hayes. 1994. Verbal Autopsies for adult deaths: Issues in their development and validation. *International Journal of Epidemiology* 23: 213–222.

Cohen, Bretonnel K., Lynne Fox, Philip V. Ogren and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. 2005. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*: *Mining Biological Semantics,* 38–45, Detroit: Association for Computational Linguistics. http://acl.ldc.upenn.edu/W/W05/W05-1306.pdf.

Danso, Samuel, Eric Atwell, Owen Johnson, Guus ten Asbroek, Karen Edmond, Chris Hurt, Lisa Hurt, Charles Zandoh, Charlotte Tawiah, Zelee Hill, Justin Fenty, Seeba Amenga-Etego, Seth Owusu-Agyei and Betty R. Kirkwood. 2011. Verbal Autopsy corpus for machine learning of cause of death. In *Proceedings of Corpus Linguistics 2011 Conference*. Birmingham, UK. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-271.pdf.

Edmond, Karen, Maria Quigley, Charles Zandoh, Samuel Danso, Chris Hurt, Seth Agyei and Betty Kirkwood. 2008. Aetiology of stillbirths and neonatal deaths in rural Ghana: Implications for health programming in developing countries. *Paediatric and Perinatal Epidemiology* 22**:** 430–437.

Forman, George. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* 3**:** 1289–1305. http://jmlr.csail.mit.edu/papers/volume3/forman03a/forman03a_full.pdf.

Fottrell, Edward, Peter Byass, Thomas Ouedraogo, Cecile Tamini, Adjima Gbangou, Issiaka Sombie, Ulf Hogberg, Kaen Witten, Sophiee Bhattacharya, Teklay Desta, Sylvia Deganus, Janet Tornui, Ann Fitzmaurice, Nicolas Meda and Wendy Graham. 2007. Revealing the burden of maternal mortality: A probabilistic model for determining pregnancy-related causes of death from Verbal Autopsies. *Population Health Metrics* 5: 1. http://www.pophealthmetrics.com/content/5/1/1.

Francis, W. Nelson and Henry Kučera. 1979. *Brown corpus manual – revised and amplified.* ICAME http://icame.uib.no/brown/bcm.html.

Goweder, Abduelbaset and Anne De Roeck. 2001. Assessment of significant Arabic corpus. In *Proceedings of Arabic NLP Workshop at ACL/EACL*.Toulouse, France. http://www.elsnet.org/arabic2001/goweder.pdf.

Graesser, Arthur, Danielle S. Mcnamara, Max M. Louwerse and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods* 36**:** 193–202.

Greenbaum, Sydney and Gerald Nelson. 1996. The International Corpus of English (ICE) Project. *World Englishes* 15**:** 3–15.

Harvey, Kevin J., Brian Brown, Paul Crawford, Aiden Macfarlane and Ann McPherson. 2007.'Am I normal?': Teenagers, sexual health and the internet. *Social Science and Medicine* 65**:** 771–781.

Joshi, Rohina, Alan Lopez, Stephen Macmahon, Srinath Reddy, Rakhi Dandona, Lalit Dandona and Bruce Neal. 2009. Verbal Autopsy coding: Are multiple coders better than one? *Bulletin of the World Health Organization* 87**:** 51–57.

Jurafsky, Daniel and James Martin. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. (2nd edition). New Jersey: Prentice Hall.

Kahn, Kathleen, Stephen Tollman, Michel Garenne and John Gear. 2000. Validation and application of Verbal Autopsies in a rural area of South Africa. *Tropical Medicine and International Health* 5**:** 824–831. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_7/sous_copyright/010024482.pdf.

King, Gary, Lu Ying and Kenji Shibuya. 2010. Designing Verbal Autopsy studies. *Population Health Metrics* 8: 19. http://www.pophealthmetrics.com/content/8/1/19.

Kipper-Schuler, Karin, Vinod Kaggal, James Masanz, Philip Ogren and Guergana Savova. 2008. System evaluation on a named entity corpus from clinical notes. In *Proceedings of the 6th LREC Language Resources and Evaluation Conference*, 3001–3007. Marrakech, Morocco: European Language Resources Association.
http://www.lrec-conf.org/proceedings/lrec2008/pdf/764_paper.pdf.

Kirkwood, Betty, Lisa Hurt, Seeba Amenga-Etego, Chalotte Tawiah, Charles Zandoh, Samuel Danso, Chris Hurt, Karen Edmond, Zelee Hill, Guss ten Asbroek, Justin Fenty, Seth Owusu-Agyei, Oona Campbell and Paul Arthur. 2010a. Effect of vitamin A supplementation in women of reproductive age on maternal survival in Ghana (ObaapaVitA): A cluster-randomised, placebo-controlled trial. *The Lancet* 375: 1640–1649.

Kirkwood Betty, Alex Manu, Charlotte Tawiah-Agyemang, Guus ten Asbroek, Thomas Gyan, Ben Weobong, Eric R Lewandowski, Seyi Soremekun, Samuel Danso, Cathrine Pitt, Hanson, Seth Owusu-Agyei and Hill Zelee**.** 2010b. 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* 11: 58. http://www.trialsjournal.com/content/11/1/58.

Knowles, Gerry and Zuraidah Mohd Don. 2003. Tagging a corpus of Malay texts, and coping with 'syntactic drift'. In *Proceedings of Corpus Linguistics 2003 Conference*. Lancanster, UK.
http://ucrel.lancs.ac.uk/publications/cl2003/papers/knowles.pdf.

Lakeland, Corrin and Alistair Knott. 2004. Implementing a lexicalised statistical parser. In *Proceedings of Australasian Language Technology Workshop,* 47–54. Sydney, Australia.
http://aclweb.org/anthology-new/U/U04/U04-1007.pdf.

Lang, Dee. 2007. *Consultant report – natural language processing in the health care industry.* Cincinnati Children's Hospital.

Leech, Geoffrey. 1993. Corpus annotation schemes. *Literary and Linguistic Computing* 8**:** 275–281.

Leopold, Edda and Jorg Kindermann. 2002. Text categorization with support vector machines: How to represent texts in input space? *Machine Learning* 46**:** 423–444.

Loper, Edward and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL–02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics*. Volume 1. Philadelphia, Pennsylvania: Association for Computational Linguistics. http://arxiv.org/pdf/cs/0205028v1.pdf.

Manning, Christopher and Hinrich Schutze. 1999. *Foundations of statistical natural language processing.* Cambridge MA: MIT Press.

McEnery, Tony and Andrew Wilson. 1996. *Corpus linguistics.* Edinburgh: Edinburgh University Press.

Murray, Christopher, Thomas Laakso, Kenji Shibuya, Kenneth Hill and Alan Lopez. 2007a. Can we achieve millennium development goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. *The Lancet* 370**:** 1040–1054.

Murray, Christopher, Alan Lopez, Dennis Feehan, Shannon Peter and Gonghuan Yang. 2007b. Validation of the symptom pattern method for analyzing verbal autopsy data. *PLOS Medicine* 4**:** 1739–1753. http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0040327.

Murray, Christopher, Alan Lopez, Robert Black, Ramesh Ahuja, Said Ali, Adullah Baqui, Lalit Dandona, Emily Dantzer, Vinita Das, Usha Dhingra, Arup Dutta, Wafaie Fawzi, Abraham Flaxman, Sara Gómez, Bernardo Hernández, Rohina Joshi, Henry Kalter, Aarti Kumar, Vishwajeet Kumar, Rafael Lozano, Marilla Lucero, Saurabh Mehta, Bruce Neal, Summer Ohno, Rajendra Prasad, Devarsetty Praveen, Zul Premji, Dolores Ramírez-Villalobos, Hazel Remolador, Ian Riley, Minerva Romero, Mwanaidi Said, Diozele Sanvictores, Sunil Sazawal and Veronica Tallo. 2011. Population health metrics research consortium gold standard Verbal Autopsy validation study: design, implementation, and development of analysis datasets. *Population Health Metrics* 9**:** 27. http://www.pophealthmetrics.com/content/pdf/1478-7954-9-27.pdf.

Newman, M. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46**:** 323–351. http://arxiv.org/pdf/cond-mat/0412004.pdf.

Pestian, John, Christopher Brew, Pawel Matykiewicz, D. Hovermale, Niel Johnson, Bretonnel Cohen and Włodzisław Duch. 2007. Shared task involving multi-label classification of clinical free text. In *Proceedings of BioNLP '07: Biological, Translational, and Clinical Language Processing,* 97–104. Prague: Association for Computational Linguistics. http://acl.ldc.upenn.edu/W/W07/W07-1013.pdf.

Rayson, Paul and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on comparing corpora.* Hong Kong: Association for Computational Linguistics. http://acl.ldc.upenn.edu/W/W00/W00-0901.pdf.

Sachs, Jefrey and John McArthur. 2005. The millennium project: A plan for meeting the millennium development goals. *The Lancet* 365**:** 347–353.

Soleman, Nadia, Daniel Chandramohan and Kenji Shibuya. 2006. Verbal Autopsy: Current practices and challenges. *Bulletin of the World Health Organization* 84**:** 239–245. http://www.scielosp.org/pdf/bwho/v84n3/v84n3a20.pdf.

Usman, Nuhu. 2005. Developed world is robbing Africa of health staff: There are no easy answers. *British Medical Journal* 331**:** 48–49.

World Health Organization (WHO). 2004. *World report on knowledge for better health: Strengthening health systems.* Geneva: World Health Organization. http://www.who.int/rpc/meetings/en/world_report_on_knowledge_for_better_health2.pdf.

Worzel, William, Apit Almal and Duncan MacLean. 2007. Lifting the curse of dimensionality. In *Genetic Programming Theory and Practice* IV, 29–40.

Yahya, Adana. 1989. On the complexity of the initial stages of Arabic text processing. *Great Lakes Computer Science Conference.* Kalamazzo, Michigan, USA.

Youmans, Gilbert. 1991. A new tool for discourse analysis: The vocabulary-management profile. *Language* 67 (4): 763–789.