DOI: http://doi.org/10.4038/ucr.v3i2.69

University of Colombo Review (Series III),
Vol. 3, No. 2, 2022

Pedagogical and research potential of a digital humanities laboratory: Language corpora as open educational resources (OER)

Dinithi Karunanayake[1], Dushyanthi Mendis[2], Lihini Nilaweera[3], Minoli Wijetunga[4] and Sandani Yapa Abeywardena[5]

1. Senior Lecturer, Department of English, Faculty of Arts, University of Colombo
2. Professor, Department of English, Faculty of Arts, University of Colombo
3. Research Assistant, AHEAD-DOR, Department of English, Faculty of Arts, University of Colombo
4. Affiliated Researcher, Digital Humanities Laboratory, Department of English, University of Colombo
5. Affiliated Researcher, Digital Humanities Laboratory, Department of English, University of Colombo

ABSTRACT

Digital humanities (DH) brings together the study and teaching of the humanities using digital technology, and the use of concepts and methodologies from within the humanities to study the digital. DH laboratories are therefore ideally suited to address the needs of the 'Born Digital' generation – the current generation of students who are very comfortable when navigating digital spaces and who process information differently and more quickly than their predecessors. In spite of this, however, DH laboratories are relatively unknown in Sri Lanka. This article, outlining the potential of a DH laboratory as a concept and site for teaching, learning and research, illustrates as a case study, the Sri Lanka English Newspaper Corpus (SLENC). This is a 31.8 million word corpus of Sri Lankan English newspaper articles under the aegis of the DH Lab of the Department of English, University of Colombo, funded by the AHEAD project of the World Bank. Corpora fit in well with the principles of digital humanities, particularly if they are hosted as open educational resources (OERs). We highlight the research potential of the SLENC, and affirm that DH laboratories are spaces that permit the exploration of the intersection between English Studies and the digital and promote research collaboration and community building.

KEYWORDS:
Born digital generation; DHLab; Digital humanities; English Studies; OER; SLENC

Suggested Citation: Karunanayake, D., Mendis, D., Nilaweera, L., Wijetunga, M., and Yapa Abeywardena, S. (2022). Pedagogical and research potential of a digital humanities laboratory: Language corpora as open educational resources (OER). *University of Colombo Review* (New Series III), *3*(2), 05 - 18.

✉ dinithi@english.cmb.ac.lk, dushyanthi@english.cmb.ac.lk, lihininilaweera@gmail.com, minoliwijetunga@gmail.com, sandani.abeywardena@gmail.com

Digital Humanities and English Studies

What is now referred to as digital humanities (DH) is an area of study that was once known as humanities computing (Svensson, 2009). Originally the domain of computational humanists and computational linguists, digital humanities first made itself known in the 1960s (Alvarado, 2012). The field has now evolved to encompass a multitude of areas and subjects, with many researchers, academics and individuals all over the world calling themselves "digital humanists" or "DH-ers". The term reflects an evolution, where the concept is less about the use of computers and more about "a form of humanism" Bobley, 2010, as cited by Kirschenbaum, 2012, p. 6). The term also reflects the broad scope of activities and methodologies that DH-ers are engaged in. While there are those who would argue the negative effects of the shift to the digital, proponents of DH assert that in the 21st century "… if the humanities are to survive and thrive, digital research tools for the imagined future of our various fields must be developed by scholars who possess expertise in both humanistic inquiry and digital technology" (Price Lab for Digital Humanities, University of Pennsylvania). In other words, digital expertise is no longer viewed as being confined to the fields of computer science or information technology.

DH labs are one of the ideal spaces today to permit the exploration of the intersection between English Studies and the digital because, as observed by Fitzpatrick (2012), the Digital Humanities brings together the study and teaching of the humanities using digital technology, and the use of concepts and methodologies from within the humanities to study the digital. Many universities across the world have already begun work in DH labs, creating, collaborating, and testing boundaries. In Sri Lanka, however, the concept is relatively unknown, and untried. In the initial stages of fashioning a DH laboratory that would provide support for effective teaching and learning practices, as well as opportunities for research in an English Studies classroom, the Department of English at the University of Colombo conducted a survey of DH labs around the world. Key areas of focus situated within English Studies were found to be literature and corpora; linguistics and computation; digital performance, storytelling and archiving; and pedagogical interventions. This scope is vast; but it is this very breadth, along with the fluidity that comes from porous boundaries, that has paved the way for DH research to reach unexpected heights – i.e., there are hardly any limits to what can be explored through the digital.

The SLENC Corpus was one of the initiatives of the DH lab at the Department of English, University of Colombo which was itself an outcome of the recognition of the shift to digital modes of teaching and learning. The Department's current undergraduate students are of the "Born Digital" generation, and effective learning as well as learner wellbeing and comfort are dependent on the creation of pedagogical practices informed by this cohort's engagement with the digital. Palfrey & Gasser (2008) identify the "Born Digital" as a generation for whom the digital is not something new, but something that is accepted as an essential part of their lives. They also expound on the "learner" identity of this generation, observing that these Digital Natives process information differently to previous generations. They consume information through informal methods - for example,

instead of reading a newspaper, they are more likely to engage with the current political climate by watching a comedy show. There are three things that students belonging to this generation seek in their information consumption: "speed, accessibility, and how well it has been sorted" (Palfrey & Gasser, 2008, p. 242). In a global context, therefore, the digital has been identified as one of the best ways to engage with these students; in other words, integrating the digital in pedagogy is one way of creating an effective teaching-learning experience.

A multipurpose online platform such as a DH lab also enables the development, hosting and sharing of open educational resources (OER). OERs are digitized materials offered freely and openly for educators, students and self-learners to use and reuse for teaching, learning and research (Wenk, 2010). Atkins et al. (2007, p. 5) state that "At the heart of the movement toward Open Educational Resources is the simple and powerful idea that technology in general and the World Wide Web in particular provide an extraordinary opportunity for everyone to share, use and reuse knowledge." A DH lab therefore is an ideal framework through which to host a resource such as a language corpus, as corpus data can be made available for students, teachers and researchers either as an online (web-based) resource, or as downloadable text files.

This article is structured as follows: It will first describe language corpora, which are repositories of samples of authentic language-in-use, stored electronically in a format that enables retrieval for the investigation and exploration of usage patterns. It will then provide an overview of corpora of Sri Lankan Englishes as known to the research team, followed by a discussion of the SLENC. This will be followed by the results of preliminary corpus-based investigations, conducted using the SLENC, in order to point to the research potential of the corpus. The article concludes with an emphasis on the usefulness of DH labs as teaching and learning spaces, and their extensive potential for facilitating, supporting and furthering teaching, learning and research in English Studies, as well as collaboration within, and between, research communities.

Language Corpora

Building a corpus is inherently a DH project because it involves digital archiving of human-made artifacts that are then processed and interpreted via a plethora of digital methods (Jensen, 2014). However, corpora are not a new phenomenon – the first machine-readable corpus, the Brown Corpus, was completed in the early 1960s (Jensen, 2014). Electronic corpora have now become resources for lexicographers, grammarians, syntacticians, textbook writers, teachers of English as a second language (ESL), students and researchers. Making use of digitized data (the language texts) and employing computational methods involving concordances and statistics software, situates the methodology known as corpus linguistics (CL) directly within the ambit of digital humanities (Jensen, 2014).

Corpora are not randomly assembled samples of language use. A corpus comprises of texts that have been identified and selected according to some predetermined design or parameters – in other words, corpora are "… principled collections of texts" which are differentiated from general text archives which are seen as databases (Jensen, 2014, p. 120).

Owens (2011) observes that the production of a data set requires the corpus compiler to make choices about what and how to collect and how to encode the information.

The basic steps of compiling a corpus have not changed much over the years. Kennedy (1998, pp.70-85) outlines the following 5-step process:

1. Designing the corpus
2. Planning the storage and cataloging system
3. Obtaining copyright or getting permission for the material included in the corpus
4. Text collection or 'capture' (recording, transcribing, keying or scanning)
5. Markup or annotation of the texts

The first step - corpus design - is driven by the purpose/s of a corpus – i.e., the research objectives or functional uses it is expected to fulfill. Kennedy (1998) makes the point that corpus compilers should have a clear idea of what kinds of analyses are likely to be undertaken using the data, and whether these justify the efforts involved in compiling the corpus. A principled method of data collection, based on size, content, speaker information, demographic details, etc. should therefore be adopted and clearly documented. Kennedy also observes that if corpora are to be used for synchronic analyses, the year/s in which the texts were produced or published must be decided on. If the corpus is to be used for diachronic analyses, decisions have to be made about the period the data covers.

The storing of corpus data, which is the second step, is an aspect that has undergone change from the early years of corpus compilation. The first electronic corpora were stored on magnetic tape; in the 1990s and early 2000s CD ROMs were used (Jensen, 2014). Today, text files are typically stored in a zipped folder on a hard drive or via a cloud storage service. Corpus data can also be stored on an internet server and accessed via a webpage, ensuring greater public accessibility (Jensen, 2014). How the data is catalogued in any one of these storage options depends on the design of the corpus.

Obtaining permission from copyright holders to use published data and informed consent from speakers to use speech data (step 3) is an important consideration in any research study, and compiling a corpus is no exception. Some data deemed to be in public domains may not require written consent before it can be included in the corpus, especially if the corpus will not be used for any commercial purposes. Data sourced from online versions of newspapers fall into this category.

The next step of compilation (step 4) termed "capture" by Kennedy (1998, p. 78), is the collection and inclusion of texts, either written or spoken, for the corpus. Today, there is a marked increase in the automatization of the retrieval process with the help of web-crawlers which, as the name suggests, are computer scripts written in a programming language best suited for the purpose, that are trained to crawl the Web to extract data. This extraction of data by a web-crawler is referred to as "scraping". As evidenced by mega-corpora, the process of collecting corpus data by means of scraping the Web allows for an efficient and quick corpus compilation that would have otherwise been a mammoth and time-consuming task.

The final step of mark-up and annotation of a corpus usually happens at the very end, just before the corpus is made available to students or researchers. Mark-up broadly refers to document mark-up and annotations, such as the indication of parts of speech and "features such as paragraphs, fonts, sentences, including sentence numbers, speaker identification, and marking the end of the text" (Reppen et al., 2010, p. 35). Some corpora are not annotated, leaving the interpretation of semantic, grammatical and pragmatic features in the data to the corpus linguist or researcher. Written corpora which are annotated are, however, useful for students, while spoken corpora are almost always annotated, to indicate features such as turn-taking, pauses and laughter.

Once compiled, corpora have many uses which make them a valuable resource in the classroom. A corpus enables us to identify frequencies of words and phrases in the language samples, or registers represented in the corpus. Corpora can reveal patterns of language such as collocations which may not be easily apparent by looking at single or limited samples of language. Corpus data are thus one of the best resources by which to investigate synchronic and/or diachronic language change. Corpora also help us to verify if our intuitions about language use in a given context are correct, which is particularly important for teachers because "… intuitions about language use often turn out to be wrong" (Biber et al., 2002, p. 10). This represents a shift from the 1980s when generative linguistics rejected the evidentiary value of corpora, adopting instead native-speaker intuition as the only true sources of reliable language data (Francis, 1992, cited in Jensen, 2014). Consulting and using corpus data has helped ESL practitioners and lesson materials writers to design English Language Teaching material that mirrors actual language use. This includes aspects of grammar as well as familiarizing students with longer texts to improve reading comprehension or to practice speech. Similarly, designers of English language tests often consult corpora for samples of authentic language use when setting proficiency and diagnostic tests. Corpus data have also helped to identify features of English that are distinctive to a particular country or region, providing support for the existence of World Englishes – i.e., varieties of English around the world that are reflective of local contexts of use.

Corpora of Sri Lankan Englishes

Given their multiple uses, the lack of corpora of Sri Lankan Englishes that can be used as OERs is surprising. This is not to say that corpora have not been compiled, but rather that access to these corpora is limited in some way. Corpora of Sri Lankan Englishes are of two kinds – those compiled by an institution such as a University, or those compiled by individuals for a specific research purpose. Given below are some corpora of Sri Lankan Englishes compiled by institutions and individuals which are known to the authors of this article. These are all relatively small corpora, although size does not necessarily imply a limitation for research purposes.

Table 1. Corpora of Sri Lankan Englishes

Name of corpus	Size	Content	Compiler	Year released
Sri Lanka English Newspaper Corpus (SLENC)	31.8 million words	21.3 million words from the *Daily News* and 10.5 million words from the *Daily Mirror* from 2015-2018	University of Colombo, Sri Lanka	2022
South Asian Varieties of English (SAVE2020)	18 million words	3 million words (approx) of newspaper English from six South Asian countries from 2019-2020	Justus Liebig University, Germany	2021
Sri Lankan Newspaper Corpus 2015	2 million words	1 million words each from the *Daily News* and *Daily Mirror* in 2015	H.C. Keshala	2021
International Corpus of English – Sri Lanka (ICE-SL)	1 million words	400,000 words of written SLE; 600,000 words of SLE speech	Justus Liebig University, Germany & University of Colombo, Sri Lanka	2019
South Asian Varieties of English (SAVE)	18 million words	3 million words (approx) of newspaper English from six South Asian countries from 2001-2008	Justus Liebig University, Germany	2011
Arts Academic Corpus	100,000 words	Academic journal articles published from 2000 - 2010	Shivani R. Ilangakoon	Not released
Corpus of Sri Lankan Newspapers (COSN)	1.6 million words	Articles published from 2004-2006	Mahishi Ranaweera	Not released
SLObits	26,065 words; 120 obituary notices	Obituaries published in 1977 and 2017	Mahishi Ranaweera	Not released

The Sri Lanka English Newspaper Corpus (SLENC)

The Sri Lanka English Newspaper Corpus (SLENC) is the largest and most recently compiled corpus in the list above, with a collection of approximately 31.8 million words, extracted from the online versions of two prominent English newspapers published in Sri Lanka - the *Daily News* and the *Daily Mirror*. The *Daily News* has a circulation of approximately 95,000 copies per day, and the *Daily Mirror*, approximately 76,000. Table 2 provides details of the composition of the SLENC in terms of the contents and size of each sub-corpus.

Table 2. Word counts of SLENC sub-corpora

Year	Daily Mirror (DM)	Daily News (DN)
2015	1,339,157	1,107,218
2016	1,645,797	931,219
2017	3,135,698	9,744,834
2018	4,402,671	9,502,296
Total words	10,523,323	21,285,567

Source: The Digital Humanities Laboratory-Department of English, 2022, p. 5

The decision to compile a corpus of newspaper articles as part of the Department's DH lab initiative was taken in consideration of the following: large amounts of language-in-use data can be accessed relatively easily from the online versions of almost every newspaper published in English. Newspaper articles also occupy an interesting space from the perspective of Sri Lankan Englishes because of the diversity of their authors, and are therefore very likely to contain instances of linguistic variation, some of which might be innovations that could conceivably develop into distinctive features in the future. Popular newspapers, as attested by their high circulation figures, also perform the norm enforcement through the sustained use of a particular variety of a language. In the absence of a comprehensive dictionary, a well-established newspaper can serve as a mechanism of language reference and codification.

The principal function of a corpus is to enable us to see systematic (and hopefully sustained) patterns of language use in a given language, or a language variety, which are not apparent through small or individual samples of written or spoken data. These patterns of use could be evident in the lexis, grammar, lexico-grammar, syntax, semantics or pragmatic uses of a language. All languages are dynamic, in that they are constantly evolving and changing through use, and it is these developments and changes that are of interest to students, language teachers, lexicographers, researchers and linguists who use corpus methodologies such as digital concordance tools and software to investigate corpus data. Our objective in compiling the SLENC was to facilitate all of the above uses and to make easily and freely available a systematically compiled database of language-in-use for anyone interested in investigating Sri Lankan Englishes.

Methodology of corpus compilation

The extraction of data from the online archives of the *Daily News* and *Daily Mirror* was accomplished by the use of web-scrapers in the programming language R. Developing a custom-written code was necessary for this purpose. This was accomplished in collaboration with the Institut für Anglistik of Justus Liebig University, Germany, one of the international partners of the Department of English at the University of Colombo (The Digital Humanities Laboratory-Department of English, 2022). Due to different data storage practices (including meta-data) used in the websites of the two newspapers, two separate codes had to be developed to scrape the required data. At the time of data extraction, we

realized that the *Daily News* website had stored articles using a date-month-year format, while the *Daily Mirror* had done so by category or genre (e.g., Breaking News, News, World News, Political Gossip, Opinion, Features, etc.). As a result, the code for the *Daily News* was developed to scrape by date, while the code for the *Daily Mirror* was developed to scrape by category (The Digital Humanities Laboratory-Department of English, 2022).

Once the extracted data were converted to .txt files to enable searches using free text analysis software, a cleaning process was followed. This included the removal of duplicate text files and the removal of non-informative, machine-generated characters such as trailing commas and additional line breaks. Author by-lines were not removed. A double line separator was added to make the beginning and the end of each article clear. In addition, a conscious effort was made not to include articles written by foreign news agencies as the objective was to compile a corpus of Sri Lankan Englishes. The removal of these articles was done by searching the data for the full name (e.g., Al Jazeera) or the abbreviation used by foreign news agencies (e.g., BBC). The initial stage of removal was done by running a custom-written code to locate the relevant foreign news reports. However, a second stage of removal was found to be necessary after a manual check of the data was done. For this, a second code was written to complete the removal of any remaining foreign news reports (The Digital Humanities Laboratory-Department of English, 2022). A list of foreign news agencies whose news reports were found and removed is given in Appendix.

Research potential of the SLENC

The SLENC data are organized and stored in a way that easily allows investigations of synchronic and diachronic language change. First, the source of the corpus files - i.e., whether they were published in the *Daily Mirror* or *Daily News*, and dates of publication are clearly indicated, and will appear (as given below) when a concordance program is used to search the corpus data. The abbreviation DM is used for the *Daily Mirror* and DN for the *Daily News*:

DM_2022-02-04.txt

DN_2022-02-04.txt

Next, the data files are arranged in directories according to year - i.e., 2015, 2016, 2017 and 2018, (as indicated in Table 2 above). This allows, for instance, the exploration of synchronic language change in a single year, using two sources, the *Daily News* and the *Daily Mirror*. Given that the *Daily News* is published by a state-owned and controlled entity and the *Daily Mirror* by a privately owned media company, there could be differences in stance, opinion, and ideology related to a single significant event or happening, as expressed in language use and discursive constructions. Such differences, if any, can be explored through an analysis of key words in their contexts of use, as well as larger sections of discourse in the corpus data. Editorial choices on what issues to report on and discuss, and what to remain silent on are also an interesting area of study that is likely to reveal the stances of the two newspapers.

Investigating the second type of language change – diachronic features – was identified as a gap in English Studies because most of the newspaper corpora listed in

Table 1 do not cover publication periods long enough to provide evidence of sustained diachronic language change. The SLENC, since it contains articles published between 2015 – 2018, is positioned eight to ten years (in some cases 17 years) after the South Asian Varieties of English (SAVE) corpus which contains data from the same newspapers – the *Daily News* (2001 - 2005) and the *Daily Mirror* (2002 - 2007). The year 2009 falls between the data sets of the SAVE and SLENC corpora and marks the end of Sri Lanka's ethnic war. A comparison of the two corpora could, therefore, possibly yield evidence of diachronic language change, reflecting social realities both during and after the war. When compiling the SLENC, the design of the SAVE corpus was followed in terms of including and excluding texts with precisely this objective in mind. Moreover, neither corpus includes newspaper advertisements. A point to note is that in terms of size, the SLENC is a much larger corpus than SAVE. It is, in fact, the largest corpus of English newspaper data in Sri Lanka that we are aware of at present. It is envisaged, therefore, that the corpus will yield findings about SLEs which are robust and reliable.

From non-volitional to agentive: the case of "disappear" in SLEs

While vocabulary relating to the cessation of Sri Lanka's ethnic war seemed to be the most obvious area of investigation, a change in the form, meaning and use of the verb *to disappear* caught our attention. SLENC data show a change in the use of the verb *to disappear* in the *post-war* period. Typically, *disappear* functions as an unaccusative English verb which denotes a non-volitional action. However, a new form of the verb *disappear* appears in the corpus, as illustrated below:

(1) a Catholic Priest who *was disappeared* thereafter. (DM_2018-05-18.txt)
(2) those who *were disappeared* during the conflict (DM_2015-10-27.txt)

The use of *disappear* as an agentive verb with the copula as shown in (1) and (2) above is not found in the data of two earlier corpora - i.e., ICE-SL (written) and SAVE. The variant use can be attributed to the abduction of persons during and following the period of civil conflict who were subsequently listed as missing or "disappeared". The presence of this variant in both newspaper corpora reveals more than an instance of linguistic innovation. It tells us that, six years after Sri Lanka's war officially ended, violence has not. Linked to the agentive use of *was disappeared*, users of SLEs also appear to be articulating the concept underlying the noun phrase *los desaparecidos* – "the disappeared" – coined in Argentina in the late 1970s. This is evidenced in examples (3) and (4) below, of which there are 169 tokens in total in the SLENC.

(3) Families of *the disappeared* have appeared before commission after commission (DM_2018-03-20.txt)
(4) the fate of *the disappeared* (DM_2016-02-12.txt)

The common factor in the Argentinian and Sri Lankan situations is that the disappearances were not acts without volition – they were forcibly enforced. Furthermore, the coming together of women – mothers, grandmothers and daughters – in solidarity, and in protest against extra judicial abductions in Sri Lanka during the civil war (Perera,

2018) resonates with the 'Madres de Plaza de Mayo' and the 'Abuelas de Plaza de Mayo' movements of Argentina in 1977, where mothers and grandmothers searching for children who had been detained by the military regime gathered in Buenos Aires in silent protests.

Since the 1980s, an estimated 60,000 to 100,000 persons across ethnic and religious communities in Sri Lanka have disappeared (Amnesty International, 2017). The use of disappeared as an agentive verb in news reporting is thus very likely due to the necessity of recognising the agentive nature of these disappearances and expressing the underlying meaning of abduction. As noted by Cronin-Furman & Krystalli (2021, p. 84), "the definition makes clear, a disappeared person is not passively "missing"; to "be disappeared" involves an active verb." The use of *disappear* as an agentive verb is not confined to describing political realities in Argentina and Sri Lanka alone: a few years ago, the headline of a CNN news item about the sudden disappearance for months of a Chinese actress from public view read "Fan Bingbing: Has China's most famous actress been disappeared by the Communist Party?" (CNN, 2018). The underlying similarity here is that China's Communist party is suspected of being the agentive force behind the disappearance of the actress. Unlike the Sri Lankan news reports, however, the body of the CNN article does not use *disappear* in an agentive sense - i.e., the agentive sense does not seem to be normalized as much as it appears to be in the SLENC data. Whether CNN's use of *been disappeared* in the headline is a spreading of the Sri Lankan English usage is difficult to determine at this point, but given the global reach of news reports, coupled with the fact that Sri Lanka has unfortunately been in the news as the country with the world's second-highest number of cases of the disappeared registered with the United Nations Working Group on Enforced or Involuntary Disappearances (Ganguly, 2021), it is not completely out of the realm of possibility. Furthermore, large international news agencies place journalists and reporters in countries all over the world, and because a fairly consistent pattern of language use in a local context can be picked up by an international reporter, it is not inconceivable that the agentive use of *disappear* as found in the discourse of local newspapers is spreading beyond Sri Lanka.

SLENC also carries the potential for research collaboration and community building. It has gained the attention of other corpus-compilers of Sri Lankan Englishes. At least two compilers, H.C. Keshala and M. Ranaweera, have contributed their corpora, The Sri Lankan Newspaper Corpus 2015 and the Corpus of Sri Lankan Newspapers (COSN) respectively, to be housed as resources within the DHLab (https://dhlab.cmb.ac.lk/resources/). This is an illustration of how the DHLab as a hosting platform for the SLENC has proved to be attractive for collaboration in English Studies research both locally and internationally. By hosting resources that are otherwise difficult to locate in one openly accessible, research-oriented platform, the DHLab encourages and empowers individuals who are developing similar resources by giving them more visibility. It can also act as a conduit for researchers, students, and teachers who are likely to use the SLENC in their own investigations.

Conclusion

This article presented the Sri Lanka English Newspaper Corpus (SLENC) as a resource and product of the DHLab of the Department of English, University of Colombo and reflected on the principles within Digital Humanities and their relevance to higher education contexts in Sri Lanka. It argued that the process of compiling the SLENC as an OER addressed the urgency of a scalable and accessible corpus of Sri Lankan Englishes for students, teachers and researchers for the purposes of study and research. It also benefited from expertise and input from experts in corpus linguistics, natural language processing, and web development, both locally and internationally. As a resource, the SLENC is an important addition to the landscape of English newspaper corpora in Sri Lanka, offering an opportunity for teachers, students and researchers to consult a locally compiled language database as opposed to having to rely on those from other countries, particularly, located in the Global North. At present, SLENC has entered the dissemination stage, where its potential for building a research community is becoming apparent. The research findings based on the SLENC data discussed in this paper are preliminary, but clearly show the potential for more in-depth investigations into language use from both structural and discursive perspectives. The DHLab envisions the further expansion of the SLENC, for example, by adding data up to the year 2025, so that the corpus spans a ten year period, which is a reasonable period of time for investigations into diachronic language change. Such an expansion will cover the years of the Covid-19 pandemic and give linguists the opportunity to examine and analyze the morphological processes underlying the creation of innovative Covid-related vocabulary. At the level of discourses and ideologies, a preliminary analysis has already been done on the debates related to the Sri Lankan government's initial cremation only policy for Covid-19 deaths, drawing on a small number of newspaper editorials (Marikar, 2022); and the persuasive rhetoric underlying the State's securitization of the pandemic has been researched and analyzed using conceptual metaphor theory (Amarasinghe, 2021). A freely accessible, well-organized, web-based newspaper corpus, if developed as a scalable resource, thus has the potential to create and generate significant areas of research in English Studies.

Acknowledgements

The DHLab of the Department of English, University of Colombo is supported by the World Bank funded Accelerating Higher Education Development (AHEAD) operation of the Government of Sri Lanka. The authors wish to acknowledge the contribution of R. Perera in locating "the disappeared" as a sociocultural signifier, taking into consideration its performativity in relation to cultural and socio legal contexts, and the contribution of A.S.R Peiris in analyzing the behavior of *disappear* as an agentive verb in SLEs. The authors are also grateful for the comments of two anonymous reviewers which helped to make this article more detailed and structured.

Appendix

Foreign News Agencies whose reports were present in the uncleaned SLENC dataset

adnkronos
Agence France-Presse
Al Jazeera
Anadolu Agency
Associated Press
BBC News
Belarusian Telegraph Agency
BelTA
Bernama
Bloomberg
Bloomberg News
CCTV
CCTV News
CCTV5
CCTV News
China Central Television
Central News Agency
Deccan Herald
Der Spiegel
dpa
Reuters Health
RIA Novosti
Rodong Sinmun
SaPa
Saudi Press Agency
The Canadian Press
The Guardian
Belfast Telegraph
The Telegraph UK
Daily Telegraph
Sunday Telegraph
The Independent
Times of India
United News of India
Xinhua News Agency
Yonhap News Agency
Zee News
Emirates News Agency
Fars News Agency
Hindustan Times
Indo-Asian News Service
Interfax
International Herald Tribune
International Islamic News Agency
Islamic Republic News Agency
ITAR-TASS
Jiji Press
Korean Central News Agency
Korean Herald
Kyodo News
Mehr News Agency
Morning Herald
Philippine News Agency
Press Association
Press Trust of India
Qatar News Agency
Reuters
AAP
AFP
ANI
BBC
CNN
CP
DPA
IANS
ICC
KCNA
MTI
NDTV
PTI
SAPA
TASS
UNI
XINHUA
YONHAP
WION

References

Alvarado, R. (2012). The digital humanities situation. In M. K. Gold (Ed.,) *Debates in the digital humanities* (pp. 50-56). University of Minnesota Press.

Amarasinghe, S. (2021). *Obtain, maintain and legitimize: A conceptual metaphor theory and critical discourse analysis of the discourse on the Sri Lankan government's securitization of the COVID-19 pandemic.* [Unpublished Bachelors' Dissertation]. University of Colombo.

Amnesty International. (2017). *Only justice can heal our wounds.* https://www.amnesty.org/en/documents/asa37/5853/2017/en/

Atkins, D.E., Brown, J.S., & Hammond, A.L. (2007). *A review of the Open Educational Resources (OER) movement: Achievements, challenges, and new opportunities.* The William and Flora Hewlett Foundation. hewlett.org/uploads/files/ReviewoftheOERMovement.pdf.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9–48. https://doi.org/10.2307/3588359

CNN. (2018, 14 September). Fan-Bingbing: Has China's most famous actress been disappeared by the communist party? https://edition.cnn.com/2018/09/14/asia/fan-bingbing-china-celebrity-intl/index.html

Cronin-Furman, K., & Krystalli, R. (2021). The things they carry: Victims' documentation of forced disappearance in Colombia and Sri Lanka. *European Journal of International Relations, 27*(1), 79–101. https://doi.org/10.1177/1354066120946479

Fitzpatrick, K. (2012). The humanities, done digitally. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (NED-New edition, pp. 12–15). University of Minnesota Press. http://www.jstor.org/stable/10.5749/j.ctttv8hq.5

Ganguly, M. (2021, August 25). *Families of Sri Lanka's forcibly disappeared denied justice.* Human Rights Watch. https://www.hrw.org/news/2021/08/25/families-sri-lankas-forcibly-disappeared-denied-justice

Jensen, K.E. (2014). Linguistics and the digital humanities: (Computational) corpus linguistics. *MedieKultur Journal of media and communication research 57,* 115-134.

Kennedy, G. (1998). *An introduction to corpus linguistics.* Longman.

Kirschenbaum, M. (2012). What is digital humanities and what's it doing in English departments? In M. K. Gold (Ed.,) *Debates in the digital humanities* (pp. 3-11). University of Minnesota Press.

Marikar, F.H. (2022). *A critical discourse analysis of Sri Lanka's COVID-19 cremation-only policy as reflected in editorials of English newspapers.* [Unpublished Bachelors' Dissertation]. University of Colombo.

Owens, T. (2011). Defining data for humanists: Text, artifacts, information or evidence? Journal of Digital Humanities 1(1). 6-8.

Palfrey, J. & Gasser, U. (2008). *Born digital: understanding the first generation of digital natives.* Basic Books.

Perera, S. (2018, August 21). *Reproductive justice for the mothers of the disappeared in Sri Lanka.* Resurj. resurj.org/reflection/reproductive-justice-for-the-mothers-of-the-disappeared-in-sri-lanka/

Price Lab for Digital Humanities. (2022). *What We Do.* Retrieved November 30, 2022, from https://pricelab.sas.upenn.edu/about/what-we-do

Reppen, R. (2010). Building a corpus - what are the key considerations? A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistic*s (pp. 31-37). Routledge.

Svensson, P. (2009). Humanities computing as digital humanities. *Digital Humanities Quarterly, 3*(3). Retrieved November 28, 2022, from http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html.

The Digital Humanities Laboratory - Department of English (2022). *Manual to the Sri Lanka English Newspaper Corpus (SLENC).* University of Colombo Press.

Wenk, B. (2010). Open educational resources (OER) inspire teaching and learning. Paper presented at Education Engineering (EDUCON) 2010. IEEE. dx.doi.org/10.1109/EDUCON.2010.5492545