Manual to the Sri Lanka English Newspaper Corpus (SLENC) 2022

Digital Humanities Laboratory
Department of English, Faculty of Arts
University of Colombo

ISBN 978-624-5873-35-7

The compilation of this corpus was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation, a World Bank funded Sri Lankan government operation to support the higher education sector.

Manual to the Sri Lanka English Newspaper Corpus (SLENC)

The Sri Lanka English Newspaper Corpus (SLENC) was compiled by the Digital Humanities Laboratory, Department of English, University of Colombo under the project titled "Situating the Interface of English Studies and the Digital in a Sri Lankan Context: An Evidence-Based Study of a Digital Humanities Research Laboratory in an English Studies Classroom" (2019 - 2022) funded by the World Bank through the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Sri Lankan government.

The SLENC Project Team

Lilani Dias
Nazeefa Musthafa
Lihini Nilaweera
Chinthaka Senanayake
Pawan Wijesinghe
Minoli Wijetunga
Sandani Yapa Abeywardena

Advisors

Dinithi Karunanayake, Department of English,
University of Colombo
Dushyanthi Mendis, Department of English,
University of Colombo
Ruvan Weerasinghe, University of Colombo School of Computing

Technical Expertise

Christoph Wolk, Justus Liebig University, Germany
Nissanka Seneviratne & Jinali Pabasara, Apium Innovations (Pvt) Ltd, Sri Lanka

Cover Design: Thilini Perera

How to cite the Corpus Data

The Sri Lanka English Newspaper Corpus 1.0 (DHLab - Department of English). 2022. Distributed by the Department of English, University of Colombo. URL: https://dhlab.cmb.ac.lk/slenc

How to cite the Manual

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

ISBN 978-624-5873-35-7

Contributing authors

Sandani Yapa Abeywardena, Dushyanthi Mendis and Lihini Nilaweera collaborated in writing the introduction section of the *Manual*.

Dushyanthi Mendis authored the sections describing corpus design, objectives and future directions of the SLENC.

Sandani Yapa Abeywardena contributed to sections on methodology and authored the section on ethics and legal considerations.

Ruhanie Perera contributed the section on the use of ‘disappeared” in contexts external to Sri Lanka.

Lihini Nilaweera and Nazeefa Musthafa collaborated in writing the sections on organization and encoding of data, corpus cleaning and the complete section on web interface design and development.

Esther Surenthiraraj contributed to the section on the use of SLENC data in language tests.

Acknowledgements

Reviewers

The DHLab is grateful for the insightful comments provided by Dinali Fernando and Ruvan Weerasinghe on the content of the *Manual*. Any shortcomings are the responsibility of the authors.

Table of contents

Appendices

Introduction

Corpora are one of the accepted means for the study and investigation of language – be it structural properties or pragmatic uses. Descriptive studies of varieties of English, in particular, have been facilitated by the building of corpora, beginning with the International Corpus of English (ICE) project in the 1990s, which also has a Sri Lankan component - the International Corpus of English-Sri Lanka (ICE-SL) - which was released in 2019. The Sri Lanka English Newspaper Corpus (SLENC) was compiled at the Department of English, University of Colombo as part of a project titled "***Situating the Interface of English Studies and the Digital in a Sri Lankan context: An evidence-based study of a Digital Humanities Laboratory in an English Studies Classroom"*** to facilitate investigations into and studies of Sri Lankan Englishes (SLEs). The project was supported by the World Bank by means of funds awarded on a competitive basis through the Accelerated Higher Education Expansion and Development (AHEAD) program for Development Oriented Research (DOR), aimed at promoting research, development and innovation in institutes of higher education in Sri Lanka.

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 (Fitzpatrick, 2012, p. 13). Digital methodologies allow the investigation of problems that are difficult to find answers to using non-digital methods

(University of Sheffield)[1] due to the large size of the source material. 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)[2]. The broader aim of such expertise would be to ensure that digital tools are "attuned to the commitment of humanities to history, fine-grained analysis, material and cultural specificity, multisensory experience, and the nuances of language and form" (Price Lab for Digital Humanities). Constructing linguistic corpora, compiling digital literary collections and archives (Theimer, 2012), and creating digital performance experiences (Dixon, 2007, p. 3) come within the scope of digital humanities.

[1] https://www.dhi.ac.uk/what-is-a-digital-humanities-project

[2] https://pricelab.sas.upenn.edu/about/what-we-do

1.1 The Sri Lanka English Newspaper Corpus (SLENC)

The SLENC is a database 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*, with a circulation of approximately 95,000 copies per day, represents the voice of the state as it is published by the Associated Newspapers of Ceylon Limited, a government owned corporation. The *Daily Mirror*, on the other hand (circulation of approximately 76,000 copies per day), is published by Wijeya Newspapers Limited, a privately owned media company. Both newspapers are dailies, published from Monday to Saturday. Apart from their popularity, which is evident from the circulation figures cited above, a principal reason for choosing these two newspapers as sources of data for compiling SLENC is that SLENC intentionally follows the design of the Sri Lankan component of the South Asian Varieties of English (SAVE) Corpus (Bernaisch et al., 2011) compiled by the Department of English, Justus Liebig University, Giessen, Germany in order to facilitate diachronic language studies. The goal of the SAVE Corpus is to provide a snapshot of acrolectal written English in six South Asian countries on the assumption that the variety of English found in prominent English newspapers in each country very likely represents the most prestigious variety, and that it is also most likely to be the variety to be eventually codified (Bernaisch et al., 2011, p. 1). Mukherjee (2012) observes that "there is a small but influential minority for whom English is a first language in Sri Lanka. Their usage exerts an enormous normative influence on language in the media" (p. 198). While we cannot be certain that all the texts in SLENC are products of those for whom English

is a first language, it is a fair assumption that much of the writing is representative of an acrolectal variety of Sri Lankan English (SLE).

1.2 Corpus Design

The Sri Lankan component of the SAVE corpus (SAVE-SL) consists of approximately 1.5 million words each from the *Daily Mirror* and *Daily News*. The SLENC is a larger corpus, consisting of 21.3 million words from the *Daily News* and 10.5 million words from the *Daily Mirror*. Another significant difference between the two corpora is that the texts in SAVE-SL were published during the period 2001 - 2007 (Bernaisch et al., 2011, p. 2) while the SLENC texts were published between 2015 - 2018. While a period of less than twenty years is often not sufficient to reveal evolutionary features in a language, it is hoped that SLENC will prove to be useful in discovering at least the beginnings of diachronic linguistic changes in the use of English in Sri Lanka - an objective which will be elaborated on later in this *Manual*.

Because of its larger size, SLENC has been organised by source as well as by year to facilitate research from a variety of perspectives. For each newspaper, texts are organized in directories according to the year of their publication (Table 1). Synchronic studies can therefore be done across the two newspapers.

Table 1: Word counts of 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**

We wish to draw attention to the differences in the total number of words in each sub-corpus because, when doing comparative studies involving frequencies of occurrence of features across texts and registers, the word counts in the research corpora must be comparable; if the number of words in one corpus is substantially greater than the other, raw frequency counts from both corpora are not directly comparable. "Normalization" is a mechanism of adjusting raw frequency counts from texts of different lengths so that they can be compared accurately (Biber, Conrad & Reppen, 1998, p. 263). Because of the difference between the total number of words in the *Daily Mirror* sub-corpus and the *Daily News* sub-corpus (as shown in Table 1 above), researchers who intend to do comparative linguistic studies using the two sub-corpora are advised to convert each frequency to a value per million words, using the normalization formula given below.

Eg:

50 modals/ 4,402,671 x 1,000,000 = 11.356 modals per 1,000,000 words of the DM

50 modals/ 9,502,296 x 1,000,000 = 5.261 modals per 1,000,000 words of the DN

The normalized frequencies of 11 and 5 can now be compared.

1.3 Objectives

The SLENC was compiled with at least five objectives in mind.

a) Facilitating descriptive studies of the use of English in newspapers in Sri Lanka

Although Sri Lankan English has gained recognition from scholars in the World Englishes paradigm as a distinct South Asian variety, many speakers and writers still contest its existence within the country, possibly because Sri Lankan English usages operate below the level of user consciousness. For instance, Fernando (2007), in a survey of a group of teachers of English, found that many of her respondents believed that the phrasal verb "cope up with" was an example of British English usage. Corpora are useful resources which can be used to identify and bring to the level of conscious awareness features of language use that are distinctively Sri Lankan or South Asian. Unfortunately, most corpora of Sri Lankan English compiled up to now are relatively small, such as the ICE-SL of one million words

and the Sri Lankan component of the SAVE corpus which is three million words. However, in spite of these limitations, both corpora have contributed towards the identification of innovations in verb complements (Mukherjee, 2007; Keshala 2017), verb particles (Kumara, 2018), phrasal verbs (Keshala, 2017), prepositional verbs (Dissanayake, 2019) and focus markers (Bernaisch & Lange, 2012) in Sri Lankan English to date; a larger corpus such as the SLENC is likely to reveal more patterns of language use that may also be more robust.

Newspapers are viewed as norm-enforcing mechanisms. Innovations in language use, if sustained over time in newspaper discourse, can become norms or distinctive features of a language variety. This is especially true of newspapers with a long history of publication such as the *Daily Mirror* and *Daily News*. The practice of editorial mediation ensures the minimisation of language errors in newspaper articles to a certain extent; these editorial decisions also contribute to norm enforcement and lend support to the argument that the features of English in the *Daily Mirror* and *Daily News* are representative of an accepted variety of SLE. Corpora compiled using newspaper articles have another important use in descriptive studies of World Englishes. In the absence of formal grammar books or dictionaries, a newspaper archive can perform the function of codification of the features of a regional variety of English, which can be easily accessed through concordance software.

b) Interrogating existing theories and models in the World Englishes paradigm/proposing alternative approaches

One of the most influential models in the World Englishes paradigm is Braj B. Kachru's Concentric Circles Model of the 1980s which classified countries/nations where English is used, into three groups. A more recent model is Edgar Schneider's Dynamic Model which attempts to explain the evolution of English in countries with a history of colonisation by a powerful English-using nation. Both models have weaknesses: Bruthiaux (2003) observes that Kachru's model does not recognise dialectal variation within each of the varieties, and neither does it allow for the complexities that arise in multilingual settings where English is in close contact with local languages; Schneider's model attempts to provide a blueprint for so-called 'evolutionary stages' the English language passes through in a post colonial setting (both in the Inner and Outer Circles), again disallowing for variation and individual growth trajectories. The model also lists criteria for growth and evolution, some of which are unrealistic at best. For example, codification of the variety (usually accomplished through grammar books and dictionaries) is a criterion in Phase 4 of the model; however, before the construction of large-scale corpora, the compilation of dictionaries and grammar books describing lexical and grammatical innovations was reliant on the idiolects of a relatively small team of lexicographers or linguists. As argued above therefore, a corpus can serve as a method of language codification of a new or emerging variety of English, and perhaps even contribute to producing more appropriate theories and descriptions of the evolution of SLEs than we have at present.

c) Identifying and explaining language variation and change in SLEs

The SLENC contains data that allows research into both synchronic variation as well as diachronic language change. Diachronic language change is an aspect of the description and evolution of SLEs that is lacking in the existing research. In some cases, predictions made about trends in SLEs are not quite supported by SLENC data. For instance, Mendis and Rambukwella (2010), noting the occurrence of a variant of the phrasal verb *cope with* (i.e., *cope up with*) in the written component of ICE-SL, comment that *cope up with* "… appears to be making inroads into certain written genres of SLE, and could possibly be a phrasal verb that contributes to the distinctiveness of SLE at some future date" (p. 190).

However, in the SLENC, *cope with* is the more frequently used form, with 378 tokens, while *cope up with* has a frequency of only 27 tokens. While the variant with the extra particle still has a presence, as a percentage, it is a mere 7% of the total occurrences and therefore not significant enough to warrant identification as a distinctive feature or norm of SLEs as reflected in newspaper usage.

In contrast is a change in the use of the verb *to disappear* in the post-civil war period. *Disappear* functions as an unaccusative verb which typically denotes a non-volitional action[3]. However, a new verb form of *disappear* is found in the corpus data, as illustrated below:

[3] The DHLab acknowledges the contribution of ASR Peiris in analyzing the behaviour of *disappear* as an agentive verb in SLEs.

(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 is not found in earlier corpus data of ICE-SL (written) and SAVE. This innovation can be attributed to socio-political realities of the latter period of Sri Lanka's civil conflict, during which time persons were abducted and subsequently listed as missing or "disappeared". A related observation is that users of SLEs also appear to be drawing on the concept underlying the noun phrase *los desaparecidos* ("the disappeared") coined in Argentina in the late 1970s:

(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 between the Argentinian and Sri Lankan situations is that the disappearances were not non-volitional – they were forcibly enforced. Since the 1980s, an estimated 60,000 to 100,000 persons across ethnic and religious communities in Sri Lanka have disappeared. 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; as noted by Cronin-Furman and Krystalli (2021, p. 84), "the definition makes clear, a disappeared person is not passively "missing"; to "be disappeared" involves an active verb." SLENC thus presents opportunities for researchers to go beyond synchronic

linguistic analyses (i.e., the analysis of language use in a limited period or moment) to more interesting and valuable diachronic analyses (i.e., changes in language use over time) which can be attributed to or indexed with social, political or ethno-cultural factors.

d) Using corpus data as a pedagogical resource in language teaching, learning and testing

Pedagogical uses of corpora usually take two main approaches – i.e., indirect applications where teachers use corpus data to create lesson materials, vocabulary lists, glossaries, etc., and direct applications where learners work with the corpus data as an investigative or self-learning activity to understand language structure and patterns of use. The SLENC allows both approaches: the corpus has a variety of genres – news reports, feature articles, editorials, etc. – which teachers can draw on to design English language teaching (ELT) materials. Material for Content and Language Integrated Learning (CLIL) or English for Specific Purposes (ESP) is also available in articles focussed on topics in Business, Banking, Marketing, Politics, Health and Medicine, Environmental issues, etc. published in both newspapers. The second approach (sometimes referred to as data-driven learning, or DDL) is more suited for students who are able to use a concordancer (such as the one available on the DHLab website)[4] to search for keywords and their contextual meanings, collocations, frequencies of use, etc. as a self-learning exercise. An advantage of data-driven learning is that students are exposed to authentic language use rather than language from mediated sources such as

[4] https://dhlab.cmb.ac.lk/slenc-search/

textbooks, grammars, dictionaries and teachers (Thomas 2015, cited in Corino & Onesti, 2019). At a higher level of language consciousness and analysis, the SLENC can be used to train students to identify features of specific genres and registers.

Corpora and corpus data have recently assumed significance in the area of language testing as well. One of the main advantages of using corpus data in language tests is having test tasks reflect genuine language use as much as possible (Hunston, 2002) as opposed to using passages or sentences artificially constructed by test writers.

Using language that reflects the use of local norms in English language tests has several advantages. The most obvious one is fairness to the test-taker, who one has to assume is a user of SLEs. Using test items that reflect patterns of use of a different variety – for instance British or American English – would therefore not be fair. With the focus on varieties of English in the World Englishes paradigm, more and more scholars are arguing for the empowerment of local varieties of English, and a reduction of the hegemony of powerful varieties such as British or American Englishes. Hamid (2014, p. 264) for instance observes that a native speaker model of English can no longer be justified with reference to the sociolinguistic characteristics of the target language use domain. Specifically addressing the use of English in Sri Lankan contexts, Dissanayake (2019) argues for using data from the SAVE corpus in the construction of English language test items as *Daily Mirror* and *Daily News* articles contain examples of the use of the SLE prepositional verbs "discuss about," "comprise of," and "request for". For those who argue that these phrasal verbs are "errors" and that their

use should therefore not be encouraged, studies on SLE phrasal verbs cited earlier in this *Manual* show that far from being errors, these uses are fairly well established in both speech and writing, and are therefore very possibly indicative of developing norms in SLEs.

Data from the SLENC have already been used by Department of English staff in the design of questions (test items) for English language proficiency tests used for recruitment purposes by the University of Colombo as well as by the University Grants Commission. Grammar test items aimed at assessing a test-taker's knowledge of subject-verb agreement, the correct use of tense and aspect, prepositions, conjunctions, and articles in English were constructed drawing on corpus data. Larger passages from the corpus were adapted to construct reading comprehension and cloze passages. As a repository of written texts representing many genres and content areas, the corpus thus functions as a valuable stimulus and source for question generation.

e) An Open Educational Resource (OER)

To date, the Department of English is not aware of systematically compiled, specialised corpora of SLEs which are available online as open educational resources. The SLENC was envisaged as a means of filling this gap, by providing a database of English as it is used in newspapers published in Sri Lanka which would be accessible to students, teachers and researchers alike, without the necessity of paying a fee or requesting access from the compilers. Access to the SLENC data is provided in two ways – by means of a sub-corpus which is searchable online, and the full corpus which can be downloaded from the SLENC homepage.

(See Section 3 of this *Manual* for more information).

2 Methodology

The procedures adopted in compiling, processing, and finalising the corpus included data extraction, data cleaning, and collation of data. These processes are described below.

2.1 Developing the code and extraction of web data

The extraction of data from 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, which was accomplished through collaboration with the Institut für Anglistik of Justus Liebig University, Germany[5], one of the Department of English's international partners.

Two separate codes had to be developed to scrape the required data due to different data storage (including meta-data) practices used in the websites of the two newspapers. When visiting the websites, a viewer can choose to read news articles by category. However, during the early stages of corpus compilation, the project team noted that the manner in which the data was stored in the system on each server was different: the *Daily News* website stored articles using a date-month-year format while the *Daily Mirror* did 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,

[5] The DHLab is grateful for the assistance and guidance of Christoph Wolk of Justus Liebig University, Germany, in the use of R.

while the code for the *Daily Mirror* was developed to scrape by category.

2.2 Corpus compilation

2.2.1 Consistency

Since the SLENC intentionally follows the design of the SAVE-SL corpus in order to facilitate research and study of diachronic language change, efforts were taken to maintain consistency between the content of the two corpora. This required several steps, or stages. First, all available articles in both websites were downloaded with no selection being made, with the exception of advertisements and obituary notices. These two categories were not scraped because they are not included in the SAVE-SL corpus. Other than this selection, no other exclusion of texts was made at this initial stage.

Similar to the SAVE-SL corpus, the two sub-corpora in the SLENC have significantly different categories, reflecting the editorial choices of the two newspapers. While the *Daily Mirror* published articles under 19 categories, the *Daily News* published articles under 11 categories (See Appendix 1).

Furthermore, the SLENC data does not include public comments on news articles. While such comments are a regular feature in online newspaper publications, they were excluded from the corpus on the basis that they are external to the publication. Similarly, images were not included because the corpus is text-based.

A second consideration of consistency arose in relation to ensuring consistency between the two sub-corpora of the SLENC. As the codes used to scrape the data from the two websites differed (due to different information storage approaches) the data scraped was saved differently. While the *Daily News* data was saved based on the date-month-year format, the *Daily Mirror* data was saved based on categories. In order to ensure consistency in the organisation of texts across the two sub-corpora, the date-month-year format was adopted as the standard and the *Daily Mirror* sub-corpus was re-organised according to this format.

2.2.2 Naming Conventions

The text files of the SLENC adopt the following naming convention; DM/DN_YYYY-MM-DD.txt

E.g. DM_2022-02-04.txt
DN_2022-02-04.txt

A single text file contains several articles published on the same day, as the primary criterion of corpus text organisation is the date of publication. Individual news articles in each corpus text file are separated by '==========='. Each article has four containers which function as the header. These containers are visually identifiable by '##' which is a result of the scraping process. Data is recorded in the format of title, supertitle, section and date of publication as shown below. The title and supertitle are extracted from the original article. The title of an article may sometimes have two parts, separated by a line break. In the event of a line break in the

title, the first line is identified as the ‘supertitle’ and the second line is identified as the ‘title’.

E.g.,

```
1          ## Title: Electricians vs. Electrocution
2          ## Supertitle: Getting electricians qualified for safe and quality wiring:
3          ## Section: Features
4          ## Date: Wednesday, March 28, 2018 - 01:00
5          Getting electricians qualified for safe and quality wiring:
6          Electricians vs. Electrocution
7          In 2012, there were 180 deaths from electrocution, most of them owing...
```

(DN_2018-03-28.txt)

2.2.3 Organisation

The directories of the corpus have been arranged first by the sub-corpus (*Daily Mirror* - DM and *Daily News* - DN), and then by year. All text files are stored in sub-directories named by year.

SLENC Dataset 1.0

- **DM_dataset**
  - 2015
  - 2016
  - 2017
  - 2018
- **DN_dataset**
  - 2015
  - 2016
  - 2017
  - 2018

2.2.4 Collation

In both sub-corpora, the number of articles published on a specific date were collated into a single text file. This was done to reduce the number of .txt files and to ensure smoother and quicker searches when using concordance tools such as Antconc. SAVE-SL follows a similar practice in its design.

2.2.5. Encoding

It was noted that the text files have different encoding formats, i.e., UTF-8, ASCII and unidentified encoding formats. The attempt to convert the files to a consistent encoding format like UTF-8/ASCII failed since characters in the files changed into irrelevant encoded characters during the process of conversion.

Please see Researcher Patch 1.0 (Appendix 2) for more information.

2.2.6 Corpus Cleaning

Once formatted, both sub-corpora were cleaned. The processes adopted in cleaning were largely influenced by those adopted by the SAVE-SL corpus team.

Time Period
Articles that were published outside the period of 2015-2018 were removed.

Removal of Articles from Foreign News Agencies

In order to include only articles that offered 'variety-specific data' – i.e., language use representative of SLEs – news reports by foreign news agencies were removed. This process follows the approach adopted in compiling both the SAVE-SL corpus and the SAVE2020 corpus. In the SAVE2020 Corpus Manual, it is noted that foreign news agency reports were removed on the basis that "each national component should only feature material produced by users of the respective variety of English" (Bernaisch et. al, p. 6).

For the process of removing foreign news reports, the Appendices of the SAVE-SL Manual were consulted to obtain abbreviations used by foreign news agencies. Thereafter, the SLENC files were screened for the presence of reports from international news agencies by using both the full name and abbreviation (e.g. Associated Foreign Press and AFP)[6]. The initial stage of removal was done by running a code developed by a software solutions company[7] using python to locate the relevant articles and remove them. A second stage of removal was found to be necessary after running a manual check on the full corpus. For this, a second code[8] was written and run on the data to complete the removal of foreign news reports.

[6] See Appendix 3 for a list of foreign news agencies whose reports were removed from the corpus.

[7] Visit github.com/apium-io/corpus-UOC to learn more about the codes developed by Apium Innovations to help the cleaning process of SLENC.

[8] Visit github.com/NazeefaMu/UoC_Files_Cleaning_2.0 to learn more about the second code.

Removal of Duplicates
Articles which were duplicated during the scraping process were identified and removed by running a code written by Apium Innovations.

Addition of a separator after each article
A separator (as seen in the example below) was added to make the beginning and end of each text clear.

E.g.

Sri Lanka will convert its pension scheme to a contributory scheme like in India, Japan and several other countries, he said. (Yohan Perera)
==========
Title: Little girl dies from burn injuries after clothes catch fire

(DM_2017-01-14.txt)

By-lines
By-lines which identify the author of a news report or feature article were not removed from the corpus data.

Removal of characters irrelevant to linguistic data
The removal of characters such as question marks, trailing commas, and additional line breaks, were mostly done through the system (see images below) using regular expressions.

Question marks (?/??)
E.g., It is believed the jail was overcome after rival drugs gangs fought each other. ?

[solution]
It is believed the jail was overcome after rival drugs gangs fought each other.

(DM_2017-01-19.txt)

Trailing commas (,/„) in the title container
E.g., ,## Title: Govt. to introduce economic agenda:PM

[solution]
Title: Govt. to introduce economic agenda: PM

(DM_2017-01-14.txt)

<U+200B>
E.g., <U+200B>Colombo Uni distances itself from poll survey

[solution]
Colombo Uni distances itself from poll survey

(DM_2015-01-03.txt)

System Development and Source Codes

Please visit the following link to see source code of the database table structure and system developed using PHP and mysql to remove selected irrelevant non-linguistic data. https://github.com/NazeefaMu/UoC_data_processing_system

Fig. 1 -System developed to clean the raw data

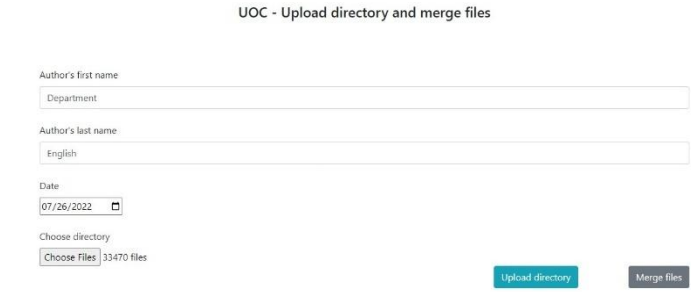

Fig. 2 - The database table structure for the system

2.3 A sample corpus text file

Shown below is an excerpt from a SLENC text file:

```
## Title: Govt. will take policy decision - President
## Supertitle: Tuition classes on Poya Days and Sundays
## Section: Local
## Date: Wednesday, March 28, 2018 - 01:00
Tuition classes on Poya Days and Sundays
Govt. will take policy decision - President

President Maithripala Sirisena said the Government will take a policy
decision on the conducting of tuition classes on Poya Days and Sundays
in future, as per the advice of the Maha Sanga so as to allow students
to receive a Dhamma education.

The President added that the Maha Sanga has been pointing out that the
conduct of tuition classes have had a severe impact on the lives of
students.

President Sirisena was addressing a ceremony to presnt offer the
Sannasa to the position of Chairperson to Ven.Ganthuna Assagi Thera
of the Amarapura Chapter at the BMICH yesterday.

The President offered the Sannasa to Venerable Ganthuna Assagi Thera
while Prime Minister Ranil Wickremesinghe offered the Viginipatha...

==============================

## Title: Gurukula takes lead on first innings
## Supertitle: 30th Battle of Kelaniya
## Section: Sports
## Date: Wednesday, March 28, 2018 - 01:00
30th Battle of Kelaniya
Gurukula takes lead on first innings

Gurukula Kelaniya led Sri Dharmaloka College Kelaniya on the first
innings at the end of day one of the 30th Battle of Kelaniya cricket
encounter played at the R. Premadasa Stadium yesterday.

Gurukula won the toss and electing to field bowled out Sri Dharmaloka
for 107 in 39.3 overs. Madhawa Kavindu top scored with 39. Yushan
Malith took 4 wickets for 14 runs and Lasindu Arosha and Pruthuvi
Rusara                    two                    apiece...
```

(DN_2018-03-28.txt)

3 Web Interface Design and Development

The online interface for SLENC was developed in keeping with the aim to have an open-access educational resource for the study and investigation of Sri Lankan Englishes, as mentioned in section 1.3 above. The two-pronged interface design provides easy access to the full 31.8 million-word SLENC dataset for anyone engaged in corpus-based studies into SLEs, and for anyone looking for samples of authentic language-in-use as material for teaching or language test preparation. Secondly, the corpus home page provides access to a 10 million word sub-corpus (5 million words each from the *Daily Mirror* and the *Daily News*) of the full data set for anyone who is interested in exploring what SLENC has to offer, with the help of a concordance tool developed in-house for the purpose.

The concordance tool is designed to enable a key-word-in-context (KWIC) search in the sub-corpus. The KWIC search generates word frequencies and concordance results in table form via a simple text-field input element – i.e., the search bar in the concordance tool. The tool has been designed and developed as a custom plugin on WordPress and written in HTML, PHP and JavaScript.

The table that contains the KWIC search results allows the user to identify the newspaper from which the article was sourced. The set of results can also be downloaded (if necessary) by the user as an excel file. Additionally, if the user clicks on the filename, they will be redirected to a page containing the keyword and its context in an expanded file view.

The SLENC dataset and the SLENC concordance tool are available on the DHLab website at dhlab.cmb.ac.lk/slenc and dhlab.cmb.ac.lk/slenc-search/ respectively.

The full corpus when downloaded can be searched for KWICs and other patterns of language use with the help of text-analysis software such as AntConc, WordSmith or Voyant Tools.

4 Ethical and Legal Considerations

As digital research ethics is a priority of the DHLab, a key consideration in compiling the SLENC included ensuring ethical research practices were followed in a digital space. This included a consultation of the Sri Lankan law on copyright.

In terms of the Sri Lankan law, news articles of the day are not considered to be protected by copyright. Section 9 of the Intellectual Property Act (2003) provides that "news of the day published, broadcast or publicly communicated by any other means" are not considered "works protected" under the Act. Given that the *Daily News* and *Daily Mirror* news articles constitute "news of the day" which are published and publicly communicated through online media, the data scraped are not protected by copyright under the Act and can be freely reproduced.

In any event, the Act provides for the fair use of works that are protected as well. Section 13 provides that even where a work is protected, for instance in the case of articles published in newspapers on "current economic, political or religious topics", such articles can be reproduced (unless there is an express condition prohibiting such reproduction when the article was first published), as long as its source is clearly indicated, and provided that it does not "conflict with the normal exploitation of the work" nor "unreasonably prejudice the legitimate interests of the author." Therefore, the Act provides for the reproduction of news articles (news of the day as well as articles on current economic, political, or religious topics) without permission from the author or the publisher. Furthermore, the SLENC merely collates material

that already exists in the public domain and which can be discovered by any person using an ordinary search engine, and therefore, does not exploit the work in any measure nor prejudice the legitimate interests of the author nor news outlet. While there were no express conditions which prohibit the reproduction of these news articles, to ensure that ethical research practices are followed, the corpus compilation team also examined the specific policies of each website.

While articulated differently, both websites required credit to be given to the sites. The *Daily News* website carries a "Legal Disclaimer"[9] which outlines the manner in which its material can be used. In particular, the Legal Disclaimer provides that the "Information presented on this website is considered public information (unless otherwise noted in material) and may be distributed or copied for non-commercial (personal, educational, research etc.) purposes." The *Daily News* site, therefore, acknowledges that the information presented is public. The *Daily Mirror* website carries the following: "All content on this website is copyright protected and can be reproduced only by giving the courtesy to 'dailymirror.lk'[10]. Similarly, the *Daily News* website notes that "the use of any material from this website" requires appropriate credit and link to the webpage where the information was taken". The SLENC clearly indicates that the news articles are sourced from the two respective websites, and therefore, assigns appropriate credit to the relevant news outlets in accordance with both their policies.

[9] https://dailynews.lk/content/legal-disclaimer

[10] https://www.dailymirror.lk/

5 Future Directions

In the future, we hope to offer a tagged version of the SLENC. We would also like to compile and offer a specialized corpus based on common genres found in newspapers, for the design of subject-specific ELT materials or for the identification of genre-specific move structure patterns.

List of References

Bernaisch, T., Mendis, D., & Mukherjee, M. (2019): *Manual to the International Corpus of English – Sri Lanka*. Giessen: Justus Liebig University, Department of English.

Bernaisch, T., Koch, C., Mukherjee, J., & Schilk, M. (2011). *Manual for the South Asian Varieties of English (SAVE) corpus.* Giessen: Justus Liebig University, Department of English.

Bernaisch, T., Heller, B., & Mukherjee, J. (2021): *Manual for the 2020-Update of the South Asian Varieties of English (SAVE2020) Corpus*. Version 1.1. Giessen: Justus Liebig University, Department of English.

Bernaisch, T., & Lange, C. (2012). The typology of focus marking in South Asian Englishes. *Indian Linguistics 73*(1-4), 1-18.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics*. Cambridge University Press.

Bruthiaux, P. (2003). Squaring the circles: issues in modeling English worldwide. *International Journal of Applied Linguistics, 13*(2), 159-178.

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

Corino, E., & Onesti, C. (2019). Data-driven learning: A scaffolding methodology for CLIL and LSP teaching and learning. *Frontiers in Education 4*(7). doi: 10.3389/feduc.2019.00007

Dissanayake, A.K. (2019). Sri Lankan University students' and English lecturers' acceptance of selected Sri Lankan English prepositional verbs: pedagogical implications. *CINEC Academic Journal Vol 3*, 68-73.

Fernando, D. (2007). Good English or Sri Lankan English? A study of English teachers' awareness of their own variety. [Unpublished Masters thesis]. University of Reading, Reading, UK.

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

Hamid, M. O. (2014). World Englishes in international proficiency tests. *World Englishes 33*(2), 263-277. https://doi.org/10.1111/weng.12084

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Intellectual Property Act, No. 36 (2003). https://www.nipo.gov.lk/web/images/pdf_downloads/Intellectual_Property_Act_No_36_of_2003.pdf

Keshala, H.C. (2017). *Linguistic innovations, acceptance and norm development. Is written Sri Lankan English forming alternative norms?* [Unpublished Master's thesis]. University of Colombo.

Kumara, M.D.S.S. (2018). *Sinhala and Tamil influence on Sri Lankan English particle use: A corpus-based study on the case of 'for'*. In Premadasa, K.M.G.P., Wijesinghe, W.L.P.K. and Adikari, A.M.T.N. (Eds.), Proceedings of the 10th International Research Conference of the General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka, February 2018. 730-736 http://www.kdu.ac.lk/irc2017/downloads.html

Mendis, D. & Rambukwella, H. (2010). Sri Lankan Englishes. In A. Kirkpatrick (Ed.), *The Routledge Handbook of World Englishes* (pp. 181-196). Routledge.

Mukherjee, J. (2007). Sri Lankan English: Evolutionary status and epicentral influence from Indian English. In K. Stierstorfer (Ed.), *Angliestentag 2007* Munster. 359-368.

Mukherjee, J. (2012). English in South Asia - Ambinormative orientations and the role of corpora: The state of the debate in Sri Lanka. In A. Kirkpatrick & R. Sussex (Eds.), *English as an international language in Asia: implications for language education, vol. I.* Springer. DOI:10.1007/978-94-007-4578-0_12

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/

Appendix 1 Categories retrieved from the *Daily Mirror* and *Daily News* websites in 2019

Daily Mirror categories (retrieved in 2019):

1. News
    - Top Stories
    - Latest News
    - Image News
    - Weird News
    - Budget
2. Opinion
    - Opinion
    - D.B.S. Jeyaraj column
3. Cartoon
4. Features
5. Expose
6. Hard Talk
7. Business
8. Sports
9. Cricket
10. Video
11. Poll
12. Leisure
13. Travel
    - Travel Getaways
    - Travel Tips
    - Business Travel
14. Medicine
15. Happy News
16. Technology
    - Technology
    - Digital Transformation
17. Gossip
18. World News
19. Jyotisha

Daily News categories (retrieved in 2019):

1. Local
2. Political
3. Business
4. Editorial
5. World
6. Law and Order
7. Entertainment
8. T&C
9. Sports
10. Obituaries
11. More

Appendix 2 Researcher Patch 1.0

Avoiding dual hits

The title of each article is included in the container as well as in the body of the article. Therefore, when searching, dual hits of the same text may appear. However, the title in the container (header) can be identified with ## whereas the title in the body would not be preceded by a container identifier.

Encoding issues

The data is mostly encoded in UTF-8, but there may be other unidentifiable encoding formats.

The corpus data is best searched using AntConc on Windows; problems may arise for MacOS users who wish to search the corpus using Antconc.

For users of Windows, we strongly recommend the use of the Windows Notepad 11.2206.17.0 for the optimum viewing of text files in the dataset.

Gibberish

There may be files with mistranslated text and/or unwanted HTML code and/or markup (due to the manner in which the data was scraped), or untranslatable material due to text being in Sinhala or Tamil font. However, we have not noticed any substantial impact on the data because of this.

Appendix 3 Foreign News Agencies whose reports were present in the uncleaned SLENC dataset

adnkronos	Emirates News Agency
Agence France-Presse	Fars News Agency
Al Jazeera	Hindustan Times
Anadolu Agency	Indo-Asian News Service
Associated Press	Interfax
BBC News	International Herald Tribune
Belarusian Telegraph Agency	International Islamic News Agency
BelTA	Islamic Republic News Agency
Bernama	ITAR-TASS
Bloomberg	Jiji Press
Bloomberg News	Korean Central News Agency
CCTV	Korean Herald
CCTV News	Kyodo News
CCTV5	Mehr News Agency
CCTV News	Morning Herald
China Central Television	Philippine News Agency
Central News Agency	Press Association
Deccan Herald	Press Trust of India
Der Spiegel	Qatar News Agency
dpa	Reuters

Reuters Health	AAP
RIA Novosti	AFP
Rodong Sinmun	ANI
SaPa	BBC
Saudi Press Agency	CNN
The Canadian Press	CP
The Guardian	DPA
Belfast Telegraph	IANS
The Telegraph UK	ICC
Daily Telegraph	KCNA
Sunday Telegraph	MTI
The Independent	NDTV
Times of India	PTI
United News of India	SAPA
Xinhua News Agency	TASS
Yonhap News Agency	UNI
Zee News	XINHUA
	YONHAP
	WION