MANUAL TO THE SRI LANKAN COMPONENT OF THE INTERNATIONAL CORPUS OF ENGLISH

ICE-SL

How to cite this manual:
Bernaisch, Tobias, Dushyanthi Mendis & Joybrato Mukherjee (2019): *Manual to the International Corpus of English – Sri Lanka*. Giessen: Justus Liebig University, Department of English.

Justus Liebig University Giessen

Department of English
Otto-Behaghel-Strasse 10B
35394 Giessen, Germany

University of Colombo

Department of English
P.O. Box 1490 Colombo 3,
Sri Lanka

Table of Contents

1. Introduction 3
   - I The ICE project 3
   - II The Sri Lankan ICE component 4
2. Notes on corpus compilation 6
   - I Selection of material 6
   - II Formal criteria 7
   - III Time frame of text sampling 7
   - IV Sociobiographical criteria for authors 8
   - V Additional notes on individual categories 8
   - VI Compilation and annotation procedure 9
   - VII Anonymization of personal names 10
3. Notes on markup for the spoken texts in ICE-SL 12
   - I General notes 12
   - II Content markup 12
   - III Non-corpus material 13
   - IV Normalizing the text 14
   - V Essential, recommended and optional markup for spoken texts 14
4. Notes on markup for the written texts in ICE-SL 15
   - I General notes 15
   - II Typographic markup 15
   - III Content markup 16
   - IV Normalizing the text 19
   - V Essential, recommended and optional markup for written texts 20
5. CLAWS-tagged version of ICE-SL 21
6. Notes on copyright and metadata 22
7. Works cited 23

Appendix 1: CLAWS C7 tagset 24

Appendix 2: Untranscribed Text Markup in ICE-SL Spoken 29

Appendix 3: Umbrella Transcriptions in ICE-SL Spoken 31

1. Introduction

I The ICE project

The *International Corpus of English* (ICE) is a worldwide project initiated by Sidney Greenbaum in the late 1980s aiming at the compilation of a computerized and comparable corpus of international Englishes covering both first- and second-language varieties. It includes spoken and written texts drawn from various genres. Each national/regional component follows an identical structure, which guarantees a relatively high degree of comparability of the individual national ICE components. Each component consists of one million words composed of 500 text samples of 2,000 words. The text categories to be included are defined and so are the selection criteria for speakers/authors and the time frame. The same annotation scheme is applied to each component.

Since the beginning of the ICE project, several national components have been completed, many are still in progress and several new components have been initiated more recently. At the end of 2018, the status of the individual ICE components was as follows:

Complete and available:

Australia
Canada
East Africa
Great Britain
Hong Kong
India
Ireland
Jamaica
New Zealand
Nigeria
Philippines
Singapore
Sri Lanka

In progress:

Bahamas
Fiji
Ghana
Gibraltar
Malaysia
Malta
Namibia
Pakistan
Puerto Rico
Scotland
South Africa
Trinidad and Tobago
Uganda
United States

For more information on the ICE project and the other national components, please consult the official ICE website: <http://ice-corpora.net/ice/>.

II The Sri Lankan ICE component

The idea for a Sri Lankan component of ICE and the first collection of data dates back to the early 1990s. Under the auspices of Christopher Tribble, several Sri Lankan universities collaborated in compiling samples and started annotation. However, for various reasons the undertaking was put on hold for several years.

In 2006, the project was taken up again as a collaboration between the University of Colombo, Sri Lanka, and Justus Liebig University Giessen, Germany. Even though spatially removed, the linguistics section of the Department of English in Giessen with its research focus on South Asian varieties of English had a great interest in the completion of the component and was able to receive the necessary funding to re-launch the project (cf. Mukherjee et al. 2010: 64) and finalise it thanks to financial support from the German National Science Foundation (BE 5812/2-1) in 2019. The material assembled by the previous team was kept where appropriate and data collection and annotation was resumed with an initial focus on the written component. Joybrato Mukherjee (Justus Liebig University Giessen), Dushyanthi Mendis (University of Colombo) and Tobias Bernaisch (Justus Liebig University Giessen) supervised the compilation and were supported by various research and student assistants in the process.

People who worked on the compilation of ICE-SL:

Coordinators:

Joybrato Mukherjee
Dushyanthi Mendis
Tobias Bernaisch

Senior advisor (until 2008):
Christopher Tribble

Research assistants:
Shariya Dilini Algama
Shivanee Illangakoon
Christopher Koch
Menikpura DSS Kumara
Patrick Maiwald
Vivimarie V. Medawattegedera
Melanie Revis
Marco Schilk
Janina Werner

Student assistants:

Sandani Yapa Abeywardena
Radhika Shavindri Atygalla
Roshny Shannon Constantine
Rosette Thilini De Alwis
Ranali Diluka Fernando
Ayudhya Gajanayake
Lisa Gebhard
Justus Grebe
Janine Manishka Gunasekara
Lilly Heidinger
Anne Hoffmann
Christina Hoppermann
Layla Hutz
Avanthi Jayasuriya
Janina Keim
Mirjami Körtvelyessy
Phusathi Liyanaarachchi
Nileptha Julia Magallage
Swasha Fernando

Student assistants (cont.):

Fathima Shazna Muneer
Stephanie Nicole
Lihini Nilaweera
Simone Pauls
Kevin Perera
Karin Puga (Stoklasa)
Janina Petznick
Zarah Rizwan
Marie Schielke
Daniel Schneider
Stefanie Schultz
Atara Senn
Mareike Sich
Saambaviy Sivaji
Aksha Suares
Rebecca Surenthiraraj
Ruth Surenthiraraj
Tamesha Tennakoon
Praveen Dilrukshan Tilakaratne
Thakshala Tissera
Marie-Christine Vogel
Pramodha Weerasekera
Piumi Sakuntala Wijesundara

Although the *ICE Markup Manual for Spoken/Written Texts* (Nelson 2002, henceforth *ICE Markup Manual*) specifying the guidelines for corpus compilation forms the basis for each national component, it still provides some room for interpretation as regards e.g. text allocation in the various genres, markup conventions etc. This may hamper cross-component comparability, which is why we want to document decisions made during the corpus compilation and annotation procedure. In order to ensure that corpus compilation be conducted along similar lines across a number of components to counteract criticism voiced against earlier ICE components, the project members of ICE-SL tried to coordinate principles of corpus compilation and annotation with other ICE teams. Against this background, ICE workshops have been hosted by Marianne Hundt (ICE-Fiji), Joybrato Mukherjee (ICE-Sri Lanka), Dagmar Deuber (ICE-Trinidad & Tobago), Ulrike Gut (ICE-Nigeria), Magnus Huber (ICE-Ghana) and Manfred Krug (ICE-Malta). The main foci of the workshops were homogenization of markup conventions and principles of text collection as well as mutual exchange about shared challenges in corpus compilation and annotation.

For more information on the Sri Lankan component of ICE, please make sure to contact the ICE-SL team at <ice-sl@anglistik.uni-giessen.de>.

2. Notes on corpus compilation

I Selection of material

In the selection of corpus material, the ICE-SL team generally follows the guidelines set by the ICE framework (cf. Nelson 2002; Greenbaum 1996). Deviations from these guidelines are listed and justified here. ICE-SL is subdivided into the categories and subsections documented in Table 1 (the numbers in brackets indicate the number of 2,000-word texts in each category).

SPOKEN (300)	Dialogues (180)	Private (100)	Face-to-face conversations (90) Telephone conversations (10)	S1A-001–090 S1A-091–100
		Public (80)	Classroom Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Legal cross-examinations (10) Business Transactions (10)	S1B-001–020 S1B-021–040 S1B-041–050 S1B-051–060 S1B-061–070 S1B-071–080
	Monologues (120)	Unscripted (70)	Spontaneous commentaries (20) Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)	S2A-001–020 S2A-021–050 S2A-051–060 S2A-061–070
		Scripted (50)	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)	S2B-001–020 S2B-021–040 S2B-041–050
WRITTEN (200)	Non-printed (50)	Student Writing (20)	Student Essays (10) Exam Scripts (10)	W1A-001–010 W1A-011–020
		Letters (30)	Social Letters (15) Business Letters (15)	W1B-001–015 W1B-016–030
	Printed (150)	Academic writing (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)	W2A-001–010 W2A-011–020 W2A-021–030 W2A-031–040
		Popular writing (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)	W2B-001–010 W2B-011–020 W2B-021–030 W2B-031–040
		Reportage (20)	Press news reports (20)	W2C-001–020
		Instructional writing (20)	Administrative Writing (10) Skills/hobbies (10)	W2D-001–010 W2D-011–020
		Persuasive writing (10)	Press editorials (10)	W2E-001–010
		Creative writing (20)	Novels & short stories (20)	W2F-001–020

Table 1: The ICE-SL corpus design.

The audio material for the spoken part of ICE-SL was obtained either by recording speakers directly on-site or by acquiring recordings from public or private sources, e.g. TV/radio stations, YouTube channels, etc. The written texts were obtained through personal contacts in academic and non-academic contexts, through publications publically available on the Sri Lankan market or in libraries and through online publications.

Due to the given framework of genres, texts were chosen randomly, but according to well-defined content-related and formal criteria. The choice of the texts for each genre was guided by previously completed ICE components – particularly ICE-New Zealand, ICE-Great Britain and ICE-India. In addition, these genre categorisations were checked against the judgement of native speakers of Sri Lankan English.

II Formal criteria

Each text in ICE-SL consists of 2,000 words. Many of the text samples are taken from coherent and continuous texts featuring more than 2,000 words. After 2,000 words, the text in the original piece of writing or audio recording is either not transcribed or treated as extra-corpus material and annotated accordingly. Where text samples are shorter than 2,000 words (e.g. student letters, press reports, spontaneous commentaries, broadcast news), several subtexts are joined in a text file to meet the required word count (cf. Greenbaum & Nelson 1996: 5).

In the selection of texts, several features can make a text unsuitable for an ICE component and are therefore also avoided in ICE-SL. They include the following characteristics as laid down by Greenbaum (1996: 3):

(a) Creative writing intended to represent non-standard uses of English
(b) Highly idiosyncratic uses of English as in novels or short stories
(c) Texts with large numbers of formulae, foreign words or lengthy quotations.

However, these criteria are formulated in a fashion open to interpretation. It is unclear what would count as ‘highly idiosyncratic uses of English’, or when ‘large numbers’ of formulae or foreign words are reached. In ICE-SL, these decisions have been made on a case-by-case basis and established via consensus in the team, which is why the interpretation of the above criteria may differ from text to text and in comparison to other national components.

III Time frame of text sampling

Since the ICE-SL project started significantly later than the initial ICE projects, it was not possible to comply with the original time frame of 1990 to 1994 envisaged by the ICE coordinators. Instead, the text samples in ICE-SL date from 2003 to 2009 for the written and from 2010 to 2018 for the spoken data.

IV Sociobiographic criteria for authors

The general rule for speaker inclusion for all ICE components is to only select texts by "adults (over 18) who have received formal education through the medium of English to the completion of secondary level schooling" (Greenbaum & Nelson 1996: 5). Where speakers do not fulfil all the requirements but are still considered highly representative of and/or influential for Sri Lankan English, e.g. news casters without a secondary-level degree, they are included in the corpus.

Overall, an effort is made to ensure the best possible "representation of differences in sex, age, education, occupation, locality, and so on" (Greenbaum & Nelson 1996: 5).However, a certain bias may be found towards younger speakers from the south of Sri Lanka (in particular Colombo) moving in academic contexts. This is due to the high concentration of acrolectal speakers of English in the Colombo area and the regular contact of youths with English via the Internet. Also, even though the Sri Lankan civil war ended in 2009, various areas in Sri Lanka were still difficult to access during the data collection phase of the project. Hence, the corpus might be skewed towards certain subgroups of the population.

A further issue particular to Sri Lanka is the inclusion of contributions from speakers who have spent considerable periods of time outside Sri Lanka. While other ICE components are more restrictive in this respect, the ICE-SL team decided against excluding speakers who spent time abroad. In fact, many CVs of speakers of Sri Lankan English feature more or less extensive stays abroad. It is therefore considered "essential to include these speakers as well since they form an integral and significantly large part of the English speaking community in Sri Lanka" (Mukherjee et al. 2010: 67). Yet, to give a distinct impression of Sri Lankan English, as opposed to English in Sri Lanka, we decided to only include speakers who have spent the majority of their lives in Sri Lanka and have received the majority of their education there.

V Additional notes on individual categories

W1B-001–030 (Letters)

Both for social and business letters, the majority of the data consists of emails. At the time of compilation, emails had already to a large extent replaced handwritten and typed letters and are therefore more representative of present-day Sri Lankan

English letter writing. Even though the original ICE framework deems emails unsuitable for the corpus (cf. Greenbaum 1991: 4), from the perspective of technological and social development handwritten letters have become nearly obsolete nowadays. It is therefore assumed that the language of emails can be compared to the language of letters to a degree that justifies the use of email correspondence in this section.

W2C-001–020 (Reportage)
In this section, special care was given to include articles only by Sri Lankan authors. Articles by press agencies such as AFP, Reuters, etc. were avoided and not included in the corpus.

S1A-001–090 (Face-to-Face Conversations)
In general, we did not begin transcribing face-to-face conversations until after the first fifteen minutes of the conversation were over to avoid issues with the observer’s paradox. During the first fifteen minutes, the speakers grew accustomed to the situation and paid less attention to being recorded afterwards. Yet, there are still some speakers in the corpus who refer to the fact that their conversations are being recorded.

S1A-091–100 (Telephone Conversations)
The texts in this section stem from Skype conversations with the video signal turned off and radio call-in shows. We opted for the inclusion of audio-only Skype conversations since they increasingly replace more traditional telephone conversations and because they are easier to process electronically.

S1B-001–020 (Classroom Lessons)
Sri Lankan classrooms may occasionally not be as interactive as classrooms in other countries where ICE components have been or are collected, but an effort was made to find teachers that engage with their students to be able to include dialogic data in the corpus.

S1B-061–070 (Legal Cross-Examinations) &
S2A-061–070 (Legal Presentations)
In Sri Lanka, legal matters are relatively rarely negotiated in English and recording these in court is not allowed. To ensure comparability with other ICE components, we decided to include material from moot courts, where law students give legal presentations and conduct cross-examinations.

VI Compilation and annotation procedure

For the written data, the processes of corpus compilation and annotation can be divided into four major phases.

(1) Transcription/digitisation: Handwritten texts such as student essays and handwritten letters are orthographically transcribed by the ICE-SL team. For texts available in printed form, text recognition software is used where possible to digitise the data. If the print quality is not suitable for scanning and text recognition, the texts is transcribed by hand as well. Where texts are available in electronic form, they are converted into the required format and directly sorted into the respective category.

(2) Proofreading of transcripts: All texts typed by hand are proofread carefully. Also the data obtained with the help of scanning software is checked for faulty text recognition and corrected accordingly.

(3) Annotation with ICE markup: The data are annotated according to the ICE markup scheme.

(4) Proofreading of markup: Once annotated, the markup of each text is proofread twice for completeness and correctness.

For the spoken data, the processes of corpus compilation and annotation can be divided into these four major phases.

(1) Transcription: Audio recordings are orthographically transcribed in F4 and some markup elements are already inserted at this stage so that the recordings do not have to be revisited at a later markup stage.

(2) Proofreading of transcripts: The transcript of the audio recording and the markup are checked by a different team member. To ensure a consistently high quality of the transcripts, each transcript in the corpus is either transcribed or proofread by a speaker of Sri Lankan English.

(3) Annotation with ICE markup: The remaining ICE markup is added in the proofread transcripts via an R script.

(4) Proofreading of markup: The fully annotated transcription is manually checked for problems and corrected accordingly.

A final semi-automatic check and subsequent correction of the entire corpus concludes corpus compilation and annotation.

VII Anonymisation of personal names

In the non-printed written material of ICE-SL and in all the transcripts of the spoken part, all personal names to the exception of some generally well-known personalities of public life have been anonymised in order to protect the privacy of the authors and other persons mentioned in the texts.

During the process of anonymisation, all personal names were collected on a list, along with the corresponding text unit number. Native speakers of Sri Lankan English classified these names as male or female and made up alternative names according to gender. All names were then changed in the corpus. The altered names

are marked as <@> </@>.
The list of original and changed names is kept confidentially with the ICE-SL team, so that all original names can be retrieved if needed.

3. Notes on markup for the spoken texts in ICE-SL

The markup for the spoken texts in ICE-SL was carried out according to the official *ICE Markup Manual for Spoken Texts*. It is available for download from the official ICE website <http://ice-corpora.net/ice/manuals.htm>

All cases of markup practices with the spoken texts in ICE-SL which are not covered by the official manual or which deviate from the suggested procedure are documented here with references to the corresponding sections of the official *ICE Markup Manual for Spoken Texts*. Further, the official manual leaves several markup categories as optional to the respective teams and this documents which markup categories were used.

I General notes

WORD COUNT

When calculating the word count, the text segments marked with <O>, <X>, <&>, <+>, <indig>, <foreign> or <unclear> as well as unfilled pauses are excluded from the word count.

II Content markup

3.2 OVERLAPPING SPEECH <[> </[> and <{> </{>

Only speech overlapping with speech was marked whereas the ICE manual also includes the possibility of pauses overlapping with speech. When a speaker uttered something while another speaker was pausing, this was treated as a non-overlapping text unit of the speaker.

3.3 ANTHROPOPHONICS

For transcribing anthropophonics, a set of identifiers for untranscribed text (please see Appendix 2) and a set of umbrella transcriptions (please see Appendix 3) were used. Umbrella transcriptions were used for some common, but highly variable realisations such as filled pauses, non-standard *no* or shortened *because*. In some instances, we opted for a certain umbrella transcription due to its pragmatic function in a given context, e.g. when “ah” was used as a backchannel instead of an expression of emotion by a certain speaker, the umbrella transcription *mmh* instead of *ahh* was used. Yet, the corpus may still contain different formal realisations for one function.

Some utterances like *aha* or *oh* are forms whose lexemes can be found in the Oxford English Dictionary (OED). Thus, the corresponding umbrella transcriptions *mmh* or *ahh* were not used respectively, but the orthographic representation proposed by the OED.

Also, instead of using both *uh* and *uhm* for voiced pauses as proposed by the manual, only *uh* was used as an umbrella transcription for voiced pauses.

3.4 Abbreviations, numerals and dates

Instead of spaces, underscores were used between the individual letters of an alphabetism.

e.g. <$A><ICE-SL:S1A-077#37:1:A><[>Yeah yeah</[></{> <}><->he</-> <=>he</=></}> is the original Flash <{><[>like</[> in nineteen-ninety I watched that T_V series also

3.6 Orthographic words <w> </w>

All words containing internal apostrophes and words ending with an apostrophe were given orthographic word markup.

3.7 Foreign words <foreign> </foreign> &
3.8 Indigenous words <indig> </indig>

For the identification of foreign and indigenous words in the spoken part of ICE-SL, the following procedure was followed. If a word form was documented in the *Oxford English Dictionary* online (OED online) or Meyler's (2007) *A Dictionary of Sri Lankan English*, it was not given any markup because the documentation of this form in at least one of the dictionaries indicates that this form is part of Sri Lankan English. In case a form was not listed in these dictionaries, it was checked against the *Gunasena Great Sinhala Dictionary* (Wijayatunga 2012) and the *Tamil Moli Akarathi: Tamil-Tamil Dictionary* (Kathiraiverpillai 2012) and given the <indig> markup if it was documented in any of the two dictionaries. If the form was not documented in any of the reference works for English or the local languages, the form received the <foreign> markup.

3.11 Unclear words <unclear> </unclear>

<unclear>word</unclear> was used for what the transcribers thought was a single word they could not understand and <unclear>several_words</unclear> for a number of words the transcribers did not understand.

III Non-corpus material

4.5 Untranscribed text <O> </O>

A set of untranscribed text markups (please see Appendix 2) was used for anthropophonics, speech that was not relevant and contextually relevant sounds.

4.6 Editorial comments <&> </&>

Instead of using editorial comments to mark noises that are contextually relevant, untranscribed text markup was used (please see III 4.5).

IV Normalizing the text

5.1 Repetitions and hesitations &
5.2 Self-corrections

To facilitate an accurate word count, deletion and normalization were separated via a space.

e.g. <$A><ICE-SL:S1A-077#30:1:A><[>Mmh</[></{> <w>that's</w> true yeah we <}><-><w>don't</w></-> <=>we never know</=></}> what will happen

V Essential, recommended and optional markup for spoken texts

Table 2, taken from the *ICE Markup Manual*, lists all the essential, recommended and optional markup categories. The categories printed in bold are used in ICE-SL.

Essential	Recommended	Optional
Text units	**Incomplete words**	**Normalization**
Subtexts	Mentions	Discontinuous words
Extra-corpus	**Orthographic words**	
Editorial comments	**Changed names**	
Untranscribed text	**Foreign words**	
Unclear words	**Indigenous words**	
Unusable characters	**Quotations**	
Uncertain transcription	**Pauses**	
Speaker IDs	**Overlapping speech**	
	Orthographic words	

Table 2: ICE markup categories used in for the spoken data.

4. Notes on markup for the written texts in ICE-SL

The markup for the written texts in ICE-SL was carried out according to the official *ICE Markup Manual for Written Texts*. It is available for download from the official ICE website <http://ice-corpora.net/ice/manuals.htm>

All cases of markup practices with the written texts in ICE-SL which are not covered by the official manual or which deviate from the suggested procedure are documented here with references to the corresponding sections of the official *ICE Markup Manual for Written Texts*. Further, the official manual leaves several markup categories as optional to the respective teams and this documents which markup categories were used.

I General notes

WORD COUNT

When calculating the word count, the text segments marked as <O> and <}> are included. The text segments marked as <X> are excluded from the word count.

SOURCE MATERIAL

Some texts from section W2D-001 to W2D-010 include a number of spelling errors caused by text recognition software. As these texts were stored in electronic format without back-up scans of the original texts before the project was re-initiated in 2006, it is no longer possible to differentiate spelling mistakes in the original manuscripts from spelling mistakes produced by the text recognition software. For pragmatic reasons, these mistakes have been corrected by means of normative replacement. Even though treated as spelling mistakes, they do not necessarily reflect the original orthography of the author.

II Typographic markup

3.9 UNUSABLE CHARACTERS &XXX;

The following markup strings are used for non-standard SGML characters in addition to the list in Appendix 2 (p. 17) of the *ICE Markup Manual for Written Texts*:

±	&plusminus;
×	&multiply;
≤	&lte;
≥	&gte;
©	&copyright;

ā	&amacron;
Ā	&Amacron;
ḍ	&ddot;
Ē	&Emacron;
í	&iacute;
ī	&imacron;
ļ	&lcedille;
ḷ	&ldot;
ḻ	&lline;
ṃ	&mdot;
Ň	&ncaron;
ņ	&ncedille;
ṇ	&ndot;
ṅ	&ndotabove;
ô	&ocircumflex;
ō	&omacron;
õ	&otilde;
ŗ	&rcedille;
Ś	&Sacute;
ŝ	&scircumflex;
ṭ	&tdot;
Ú	&Uacute;
ū	&umacron;
Ψ	&PSI;

III Content markup

General order of markup elements: Content markup always comes at the beginning of a sentence, followed by the text unit markup.

e.g. <O><text unit>… … …</O>
<X><text unit>… … …</X>
<footnote><text unit>… … …</footnote>

4.1 Headings <h> </h>

Names following article headers are marked as <p>, not <h>.

e.g. <h><ICE-SL:W2C-010#97:4>Truce talks a new beginning - Minister Nimal Siripala</h>
<p><ICE-SL:W2C-010#98:4>Bandula Jayasekara</p>

4.3 Footnotes <footnote> </footnote>

Each footnote receives a text unit markup of its own. If a footnote consists of more than one complete sentence, each sentence counts as one text unit. References in footnotes count as only one text unit, even if they include full stops.

e.g. <footnote><ICE-SL:W2B-017#18:1>1 Geert Hofstede, 1991. Cultures and Organisations: Software of the mind. McGraw-Hill, London</footnote>

Section W2F: footnotes by editors (e.g. explanations of words, facts, etc.) are treated as extra-corpus material (<X> </X>) and are not included in the word count since they were not written by the author.

4.5 Deleted text <del> </del>

Section W1A and W1B: Deleted text in handwritten documents is included in the transcription. However, it is marked as non-corpus material and is not considered in the word count.

e.g. <ICE-SL:W1A-011#54:2>And as Charles Darwin said, <X><del>to be in this</del></X> only the fittest will survive in this world

4.7 Orthographic spaces <space>

The markup <space> is used both for excess and missing space in order to ensure the correct word count.

e.g. *<ICE-SL:W1A-009#40:1>[…] living scattered all   over the country
<ICE-SL:W1A-009#40:1>[…] living scattered <space>all over the country

*<ICE-SL:W1A-009#41:1>[…] classified as IAB schools(schools having Advanced level classes
<ICE-SL:W1A-009#41:1>[…] classified as IAB schools<space> (schools having Advanced level classes

4.8 Orthographic words <w> </w>

All words containing internal apostrophes and words ending with an apostrophe are given orthographic word markup.

4.11 Quotations <quote> </quote>

Quotations of one complete sentence or more and set in quotation marks are marked as follows: <X><quote>"………"</quote></X>
They do not count towards the word count.

Quotations that are embedded in a sentence but could also stand as complete sentences on their own are marked as extra-corpus text.

e.g. <ICE-SL:W2A-003#20:1>As Eskey (1989) says <X><quote>"language is a major problem in second language reading and even educated guessing at meaning is no substitute for accurate decoding”</quote></X> (1989; 97).

<ICE-SL:W1A-011#30:1>So nothing wrong if I say <X><quote>”Internet is the best invention”</quote></X>

Sections W2C (press news reports), W2E (press editorials) and W2F (novels/short stories):
Quotations receive the corresponding markup <quote> </quote> but are NOT marked as extra-corpus text <X> </X>.

4.12 Foreign words <foreign> </foreign> &
4.13 Indigenous words <indig> </indig>

Words from a South Asian origin needed to be classified as either ‘indigenous’ (Sinhala, Tamil), ‘foreign’ (e.g. Hindi, Urdu) or naturalized as Standard (Sri Lankan) English (e.g. the word ‘saree’). Since this distinction can in many cases be ambiguous, the classification of foreign/indigenous underwent several steps:
During the phases of annotation and proofreading, any term which appeared to be foreign or indigenous was checked in the OED online. Terms mentioned in the OED were considered as Standard English and received no markup. Foreign terms not included in the OED were marked accordingly as <foreign>. Terms which could not be assigned clearly received the preliminary markup <in-fo> </in-fo> and were collected in a list.
This list was then given to native speakers of Sri Lankan English to be classified as ‘indigenous’ (i.e. Sinhala or Tamil), ‘foreign’ (any language other than English, Sinhala or Tamil) or Sri Lankan English. The words were then marked accordingly in the corpus. Words which were classified as Sri Lankan English, e.g. if they have indigenous roots, but are now used with English inflections, received no markup, even if they are not listed in the OED.
In the context of specific or technical terminology (e.g. medical or botanic terms), the OED also served as reference. Any words without an entry in the OED were marked as <foreign> </foreign>.

4.14 Captions and graphics

Graphics, tables, etc. are not transcribed. They receive the markup for untranscribed text <O> </O> and a standardized, numbered markup element (table1/diagram1/image1/photograph1). Captions, if applicable, are transcribed and separately marked as untranscribed data <O> </O> after the table, diagram, etc.

e.g. <p><O><ICE-SL:W1A-010#65:1>table1</O><O>Table 1 The problems faced by English teachers, Grades 6-8:</O></p>

Markup for graphics and captions which are embedded in continuous text is inserted after the sentence in which the graphic and caption appears.

4.16 Unclear words <unclear> </unclear>

Words within the <unclear> </unclear> markup are separated by underscores in order to preserve the correct word count.

e.g. <ICE-SL:W1A-011#71:4>[…] which is held in every
<X><del><unclear>one_word</unclear></del></X> year

5.2 Untranscribed text <O> </O>

Mathematical formulae are not transcribed but replaced by the markup for untranscribed text: <O>formula</O>

If several formulae occur within one text, they are numbered (cf. captions and graphics)

IV Normalizing the text

6.1 Misspellings

Several cases are unclear as to whether they count as misspellings and should be normalized. These cases are dealt with in the following manner:

Punctuation around citations is not normalized, even if inconsistent.

e.g. <ICE-SL:W1A-008#80:1>[…] 25% of domestic violence cases are reported. (Daily News, 25 Aug. 1996)
<ICE-SL:W1A-009#11:1>[…] oppressive relations with the dominant society"(Paulston 181).

Hyphenation, if clearly erroneous according to the OED, is corrected by means of normative replacement.

e.g. * <ICE-SL:W1A-010#20:1>[…] from three years to twenty six years
<ICE-SL:W1A-010#20:1>[…] from three years to
<}><->twenty six</-><+>twenty-six</+></}> years

BUT: No normative replacement markup in cases of free variation.

e.g. key words OR
keywords

Section W1B (Social letters and Business letters): Normative replacement was used initially to correct cases of variation in punctuation and capitalization. However, this practice was later considered irrelevant for the corpus and abandoned. Previously normalized instances have been left in the corpus, but normative replacement has not been applied to each case in a consistent manner.

This also applies to phenomena particular to emails, i.e. reduced word forms (such as *c u* for *see you*), particular abbreviations (such as *lol* for *laughing out loud*), the use of emoticons, emphatic prolongation of words (such as *Helloooo*), etc.

V Essential, recommended and optional markup for written texts

Table 3, taken from the *ICE Markup Manual*, lists all the essential, recommended and optional markup categories. The categories printed in bold are used in ICE-SL.

Essential	Recommended	Optional
Text units	**Incomplete words**	**Normalization**
Subtexts	**Deleted text**	Boldface
Extra-corpus	**Footnotes**	Italics
Editorial comments	**Footnote references**	Typeface
Untranscribed text	Marginalia	Roman
Unclear words	Mentions	Underline
Unusable characters	**Orthographic words**	Smallcaps
Uncertain transcription	**Changed names**	Subscript
	Orthographic space	Superscript
	Foreign words	**Line-breaks**
	Indigenous words	Discontinuous words
	Quotations	
	Headings	
	Paragraphs	

Table 3: ICE markup categories used in for the written data.

5. CLAWS-tagged version of ICE-SL

In addition to the standard plain-text version including structural markup, a CLAWS- tagged version of ICE-SL is also available. In this version, the texts have been part-of-speech tagged using the C7 tagset as devised by the University of Lancaster, UCREL (University Centre for Computer Corpus Research on Language). The entire tagset is provided in Appendix 1 and can also be found online via <http://ucrel.lancs.ac.uk/claws7tags.html>.

6. Notes on copyright and metadata

ICE-SL is published for academic purposes. Consequently, permission to use data in the corpus was sought from individual speakers and writers or the respective copyright holders. In case permission to use the data was not provided directly upon the reception of the data, the respective contributors were contacted personally either by email or post with a request for consent to use their text(s). In the case of texts from TV or radio where the individual contributor was not retraceable, the publishers were contacted instead.

Metadata on texts and authors is helpful for studies with sociolinguistic aims. In ICE-SL, the collection of metadata was therefore attempted where possible, but proved to be more achievable in some categories than in others. For the retrieval of metadata, a questionnaire was given to the authors as part of the copyright agreement. The questionnaire covered the following categories of sociobiographic information:

- gender
- age
- occupation
- nationality
- place of birth
- place of residence
- stays abroad and duration
- cultural/ethnic background
- educational background (highest educational degree)
- language skills
- linguistic habits and surroundings

The amount of metadata available differs for the various categories. In the spoken part of ICE-SL, the categories face-to-face conversations (S1A-001–090) and telephone conversations (S1A-091–100) are almost complete because direct contact with the contributors was possible. For other categories such as spontaneous commentaries (S2A-001–020), we used a different consent form as only the publishers and not the speakers were approached. The availability of metainformation for the written part of ICE-SL is similarly mixed.

All the metadata collected are available upon request in spreadsheet format from the ICE-SL team.

7. Works cited

Greenbaum, Sidney (1991): “The development of the International Corpus of English”. In Karin Aijmer, Bengt Altenberg (eds.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 83–92.

Greenbaum, Sidney (1996): “Introducing ICE”. In Sidney Greenbaum (ed.): *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon, 3–33.

Greenbaum, Sidney & Gerald Nelson (1996): “The International Corpus of English (ICE) Project”. *World Englishes* 1(1): 3–15.

Kathiraiverpillai, N. (2012): *Tamil Moli Akarathi: Tamil-Tamil Dictionary*. Chennai: Saratha Publishers.

Meyler, Michael (2007): *A Dictionary of Sri Lankan English*: Colombo: Mirisgala.

Mukherjee, Joybrato, Marco Schilk & Tobias Bernaisch (2010): “Compiling the Sri Lankan component of ICE: Principles, problems, prospects”. *ICAME Journal* 34: 64–77.

Nelson, Gerald (2002): *International Corpus of English - Markup Manual for Spoken/Written Texts*. <http://ice-corpora.net/ice/manuals.htm> [24 September 2018].

Wijayatunga, Harischandra (2012): *Gunasena Great Sinhala Dictionary*. Colombo: Gunasena.

Appendix 1: CLAWS C7 tagset

Taken from the University of Lancaster, UCREL (University Centre for Computer Corpus Research on Language) website (<http://ucrel.lancs.ac.uk/claws7tags.html>).

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that),in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these,those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (‘ or’s)
IF	for (as preposition)

II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)
MC	cardinal number,neutral for number (two, three..)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction,neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)

PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)
REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too)
RGQ	wh- degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)
RGT	superlative degree adverb (most, least)
RL	locative adverb (e.g. alongside, forward)
RP	prep. adverb, particle (e.g about, in)
RPK	prep. adv., catenative (about in be about to)
RR	general adverb
RRQ	wh- general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)
VB0	be, base form (finite i.e. imperative, subjunctive)

VBDR	were
VBDZ	was
VBG	being
VBI	be, infinitive (To be or not... It will be...)
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	Having
VHI	have, infinitive
VHN	had (past participle)
VHZ	Has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)
VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
XX	not, n’t
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A’s, b’s)

NOTE: "DITTO TAGS"
Any of the tags listed above may in theory be modified by the addition of a pair of numbers to it: eg. DD21, DD22 This signifies that the tag occurs as part of a sequence of similar tags, representing a sequence of words which for grammatical purposes are treated as a single unit. For example, the expression *in terms of* is treated as a single preposition, receiving the tags:

in_II31 terms_II32 of_II33

The first of the two digits indicates the number of words/tags in the sequence, and the second digit the position of each word within that sequence.
Such ditto tags are not included in the lexicon, but are assigned automatically by a program called IDIOMTAG which looks for a range of multi-word sequences included in the idiomlist. The following sample entries from the idiomlist show that syntactic ambiguity is taken into account, and also that, depending on the context, ditto tags may or may not be required for a particular word sequence:

at_RR21 length_RR22
a_DD21/RR21 lot_DD22/RR22
in_CS21/II that_CS22/DD1

Appendix 2: Untranscribed Text Markup in ICE-SL Spoken

Description	Tag
Activity (when someone clearly doing something causes a 'pause' in their speech)	<O>activity</O>
Address of Sinhala speaker	<O>addresses_Sinhala_speaker</O>
Commercial break in recording	<O>advertisement</O>
Audience applause	<O>applause</O>
Notable pause in recording	<O>break_in_recording</O>
Speaker coughs	<O>cough</O>
Deleted passage	<O>deleted_passage</O>
Sound of speaker drinking	<O>drinks</O>
Speaker exhales	<O>exhale</O>
Finger clicking	<O>finger_clicking</O>
Speaker groans	<O>groan</O>
Speaker hums	<O>humming</O>
Imitations of someone else's speech	<O>imitation</O>
Speaker inhales	<O>inhale</O>
Radio jingle (also including spoken lines)	<O>jingle</O>
Legal jury comments in	<O>jury_comment</O>
Speaker laughs	<O>laugh</O>
Musical interlude, singing or humming in recording	<O>music</O>
Background noise interrupting the recording	<O>noise</O>
English passage by non-corpus speaker	<O>passage_by_non_Sri_Lankan_speaker</O>
Passage in Sinhala	<O>passage_in_Sinhala</O>
Phone causes speakers to react	<O>phone</O>
Quotations not integrated into text unit/longer than one text unit	<O>quoted_passage</O>
Speaker screams	<O>scream</O>
Speaker sighs	<O>sigh</O>
Speaker sneezes	<O>sneeze</O>

Student asks question	<O>student_questions</O>
Student responds	<O>student_response</O>
Speaker clears her/his throat	<O>throat_clearing</O>
Uncertain number of students chorus an answer or comment	<O>uncertain_chorus</O>; <chorus>…</chorus> when what the students answer in a chorus can be understood
Unclear discussion in parallel to transcribed speech	<O>unclear_parallel_discussion</O>
Speaker whistles	<O>whistle</O>
Speaker yawns	<O>yawn</O>
Two or more unclear words	<unclear>several_words</unclear>
Unclear word	<unclear>word</unclear>

Appendix 3: Umbrella Transcriptions in ICE-SL Spoken

Description	**Transcription**
Exclamation of victory	yay
Exclamation of emotion	ahh
Filled pauses	uh
Backchannelling	mmh
Affirmative responses	yes/yeah/yup (approximations to utterance); mmh
Negative responses	no/nuh (nuh as umbrella transcription for non-standard no); mhm
Shortened *because*	coz
Shortened *you all*	y'all
Okay	okay
Rising intonation, tag question, intensifying discourse particle	huh
Shortened *want to, going to, trying to, kind of, sort of*	wanna, gonna, trynna, kinda, sorta
Exclamation of disgust	urgh
Indigenous question tags	no/ne
Click of the tongue	tch
Silencing someone	sh
‘Ey’ (to draw attention)	<indig>ey</indig>