# Manual to the Written Component of the International Corpus of English – Sri Lanka

# ICE-SL [W200]

**How to cite this manual:**
Körtvelyessy, Mirjami, Tobias Bernaisch, Joybrato Mukherjee and Dushyanthi Mendis (2012): *Manual to the Written Component of the International Corpus of English – Sri Lanka*. Giessen: Justus Liebig University, Department of English.

JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

UNIVERSITY OF COLOMBO

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

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

# Table of Contents

# 1. Introduction

## a. 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 categories are defined, as well as 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. Up to the present day (early 2012), the status of the individual ICE components is as follows:

Complete and available:

Canada (ICE-CAN)
East Africa (ICE-EA - Kenya & Tanzania)
Great Britain (ICE-GB)
Hong Kong (ICE-HK)
India (ICE-IND)
Ireland (ICE-IRE)
Jamaica (ICE-JA)
New Zealand (ICE-NZ)
Philippines (ICE-PHI)
Singapore (ICE-SIN)

In progress:

Australia (ICE-AUS)
Bahamas (ICE-BA)
Fiji (ICE-FJ)
Ghana (ICE-GH)
Malaysia (ICE-MLA)
Malta (ICE-MLT)
Namibia (ICE-NAM)
Nigeria (ICE-NG)
Pakistan (ICE-PK)
South Africa (ICE-SA)
Sri Lanka (ICE-SL)
Trinidad and Tobago (ICE-T&T)
Uganda (ICE-UG)
United States (ICE-USA)

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

## b. The Sri Lankan component of ICE

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 joined 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 Department of English Linguistics 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). The material assembled by the previous team was kept where appropriate, and data collection and annotation was resumed with a focus on the written component. Joybrato Mukherjee (Justus Liebig University Giessen) and Dushyanthi Mendis (University of Colombo) supervised the compilation and were supported by various research assistants and student assistants.

Members of the ICE-SL team past and present:

**Coordinators**:
Joybrato Mukherjee
Dushyanthi Mendis

**Senior advisor (until 2008)**:
Christopher Tribble

**Research assistants**:
Tobias Bernaisch
Shivanee Illangakoon
Christopher Koch
Menikpura DSS Kumara
Patrick Maiwald
Vivimarie V. Medawattegedera
Marco Schilk
Janina Werner

**Student assistants**:
Dilini Algama
Shavindri Attygalla
Ranali Fernando
Manishka Gunasekara
Christina Hoppermann
Mirjami Körtvelyessy
Shazna Muneer
Marie Schielke
Daniel Schneider
Stefanie Schultz
Mareike Sich
Nadine Spahn
Karin Stoklasa
Marie-Christine Vogel

Although the *ICE Markup Manual for 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., which, in turn, may hamper the cross-component comparability of ICE. 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 with other ICE teams.

Against this background, ICE workshops have been hosted by Marianne Hundt (ICE-Fiji), Joybrato Mukherjee (ICE-SL), Dagmar Deuber (ICE-Trinidad & Tobago), Ulrike Gut (ICE-Nigeria), Magnus Huber (ICE-Ghana) and Manfred Krug (ICE-Malta). The main foci of the workshops generally were homogenization of markup conventions and text collection and mutual exchange about shared challenges in corpus compilation and annotation.

At the beginning of 2012, the written component of ICE-SL (ICE-SL [W200]), consisting of 200 texts and accompanied by the present manual, was finalized and made available to the research community. The spoken component of ICE-SL is in progress, but a release date is still to be announced.

For more information on the Sri Lankan component of ICE, please contact the ICE-SL team:
ice-sl@anglistik.uni-giessen.de

# 2. Notes on corpus compilation

## a. Selection of material

In the selection of corpus material, the ICE-SL team generally followed the guidelines set down by the ICE framework (cf. Nelson 2002; Greenbaum 1996). Deviations from these guidelines are listed and justified here. ICE-SL [W200] 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).

| Text categories | | | Codes |
|---|---|---|---|
| **Non-printed** (50) | **Student writing** (20) | Student essays (10)<br>Exam scripts (10) | **W1A** |
| | **Letters** (30) | Social letters (15)<br>Business letters (15) | **W1B** |
| **Printed** (150) | **Academic writing** (40) | Humanities (10)<br>Social sciences (10)<br>Natural sciences (10)<br>Technology (10) | **W2A** |
| | **Popular writing** (40) | Humanities (10)<br>Social sciences (10)<br>Natural sciences (10)<br>Technology (10) | **W2B** |
| | **Reportage** (20) | Press news reports (20) | **W2C** |
| | **Instructional writing** (20) | Administrative writing (10)<br>Skills/hobbies (10) | **W2D** |
| | **Persuasive writing** (10) | Press editorials (10) | **W2E** |
| | **Creative writing** (20) | Novels & short stories (20) | **W2F** |

Table 1: The text categories of ICE-SL

Material for the respective text categories was obtained through personal contacts in academic and non-academic contexts, through publications 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. With the choice of the texts and text types to fill the various sections, previously completed ICE components (in particular ICE-New Zealand, ICE-Great Britain and ICE-India) served as a baseline for text selection. In addition, these categorizations were checked against the judgement of native speakers of Sri Lankan English. This was done to ensure the acceptability of the categorization from a Sri Lankan perspective.

## b. Formal criteria

Each text in ICE-SL [W200] consists of 2,000 words. Many of these text samples are taken from coherent and continuous texts. After 2,000 words, the text following the end of the last sentence to begin in the 2,000-word limit 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 or press reports), several subtexts are added to complete the word count (cf. Greenbaum & Nelson 1996: 5).

In the selection of texts, several text features are considered unsuitable for ICE corpora and therefore avoided. They include the following characteristics, as laid down by Greenbaum (1991: 4):

(a) Creative writing intended to represent nonstandard uses of English
(b) Highly idiosyncratic uses of English, for example in novels or short stories
(c) Large numbers of mathematical or statistical formulae
(d) Large numbers of foreign words
(e) Lengthy quotations from other writers

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 [W200], these decisions have been made on a case-by-case basis and established via a 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.

## c. 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 intended by the ICE framework. Instead, the text samples of ICE-SL [W200] date from 2003 to 2009.

## d. Sociobiographical criteria for authors

Since Sri Lanka is a multilingual country, it is important to clearly delineate criteria as to who qualifies as a contributor of data for ICE-SL [W200]. The general rule for all the 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 fulfill all the requirements but are still considered appropriate for the corpus, e.g. news casters without a secondary level degree, they are included.

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 speakers from the south of Sri Lanka, in particular Colombo, and from an academic context. This is due to the high concentration of acrolectal speakers of English in the Colombo area, and the fact that in the Sri Lankan context the texts for the different genres in the corpus are very often composed by highly educated writers.

A further issue particular to Sri Lanka is the inclusion of contributions from authors who have spent considerable periods of time abroad. While other ICE components are more restrictive in this respect, for instance by only accepting contributors who have been living in the country since the age of 10 (e.g. ICE-New Zealand, cf. Vine 1999: 10), the ICE-SL team decided against such strict criteria. In fact, Sri Lankan English is strongly characterized by influences from other varieties of English, be it through stays abroad or through the media and business contexts. 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).

## e. Additional notes on individual categories

### W1B (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 (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.

## f. Compilation and annotation procedure

The process of corpus compilation can be divided into the four major phases of (1) transcription/digitization, (2) proofreading of transcripts, (3) annotation with ICE markup, (4) proofreading of markup.

(1) Transcription/digitization: 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 digitize the data. If the print quality is not suitable for scanning and text

recognition, the texts need to be 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 need to be proofread carefully. Also the data obtained with the help of scanning software needs to be checked for faulty text recognition and corrected accordingly.

(3) Annotation with ICE markup: Due to the idiosyncrasy of the ICE markup scheme, annotation in ICE-SL [W200] is done manually, i.e. without the help of scripts or automatic annotation.

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

## g. Anonymization of personal names

In the non-printed material of ICE-SL [W200] (i.e. Student writing (W1A) and Letters (W1B)), all personal names have been anonymized in order to protect the privacy of the authors and other persons mentioned in the text.

During the process of compilation, 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 (where possible), 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

The markup for ICE-SL [W200] was carried out according to the official *ICE Markup Manual*. It is available for download as an MS Word file on the official ICE website:
<http://ice-corpora.net/ice/manuals.htm>

In the process of compiling ICE-SL [W200], various decisions needed to be made in terms of annotation procedures, thus gearing it to the Sri Lankan context. This, however, may lead to a certain degree of heterogeneity in comparison to other (earlier) components, which might have adopted different guidelines. This manual is an attempt to document decisions specific to ICE-SL as concisely as possible to facilitate work with ICE-SL [W200] for anybody conducting research with it.

The main problems arose during the annotation process, when cases were not covered by the given *ICE Markup Manual*. This applied, for instance, to special characters unusable in SGML notation, which needed an alternative markup string. It was also unclear how different layers of markup were to be embedded, such as content markup (<O> </O>, <X> </X>, etc.) in combination with text unit markup. A major point of discussion was the degree of normalization that should be applied to the text.

All cases of markup practices in ICE-SL [W200] which are not covered by the original manual or which deviate from the suggested procedure are documented here. The references in brackets indicate the corresponding chapter of the original *ICE Markup Manual*. Further, the original manual leaves several markup categories as optional to the respective teams. This chapter specifies which of these categories are used for ICE-SL [W200].

## a. 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.

## b. General markup

There is no deviation from the manual in terms of general markup. See sample texts (Appendix 1) for the realization of general markup in ICE-SL [W200].

## c. Typographic markup

### UNUSABLE CHARACTERS (3.9)[1]

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*:

| | |
|---|---|
| ± | &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; |

[1] The numbers in brackets refer to the corresponding sections of the *ICE Manual*

## d. 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>

### HEADINGS (4.1)

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>

### FOOTNOTES (4.3)

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.

### DELETED TEXT (4.5)

*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

### ORTHOGRAPHIC SPACES (4.7)

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

### QUOTATIONS (4.11)

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>.

### FOREIGN WORDS <FOREIGN> </FOREIGN> (4.12); INDIGENOUS WORDS <INDIG> </INDIG> (4.13)

Words of 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 words underwent several steps:

During the phases of annotation and proofreading, any term which appeared to be foreign or indigenous was checked in the *Oxford English Dictionary* (OED) online version. 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>.

### CAPTIONS AND GRAPHICS (4.14)

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.

### UNCLEAR WORDS <UNCLEAR> </UNCLEAR> (4.16)

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

### UNTRANSCRIBED TEXT (5.2)

Mathematic 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)

## e. Normalizing the text

### MISSPELLINGS (6.1)

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.

## f. Essential, recommended, and optional markup in written texts (App. 3)

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 [W200].

| **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 2: ICE markup categories

# 4. CLAWS-tagged version

In addition to the standard plain-text version including structural markup, a CLAWS-tagged version of ICE-SL [W200] is also available. In this version, the texts have been tagged according to part-of-speech, using the C7 tagset as devised by the University of Lancaster, UCREL (University Centre for Computer Corpus Research on Language). Samples of texts annotated with CLAWS can be found in Appendix 1, alongside plain text samples. The entire tagset is provided in Appendix 2 and can also be found online under <http://ucrel.lancs.ac.uk/claws7tags.html>.

Please note that any word count calculated on the basis of the CLAWS-tagged version will be inaccurate. In the process of tagging, the ICE markup has been separated from the actual corpus text it is attached to in the plain-text version, so that in the tagged version each item of markup would be counted as a word of its own. Please use only the plain-text version with the ICE markup for any calculation of word counts.

# 5. Notes on copyright and metadata

ICE-SL [W200] is intended to be published for academic purposes. Consequently, permission by the respective copyright holders needed to be obtained for all texts and subtexts and for both published and unpublished material. Each of the contributors was contacted personally either by email or by post with a request for consent to use their text(s). In the case of texts from books, magazines, newspapers and other publications where the individual author was not retraceable, the publishers were contacted instead.

Metadata on texts and authors, especially sociobiographical data, is helpful for studies with sociolinguistic aims. In ICE-SL [W200], 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 (see Appendix 2 for full questionnaire). The questionnaire covered the following categories of sociobiographical 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. The categories Student writing (W1A) and Letters (W1B) are almost complete, because direct contact with the contributors was possible. For other categories such as newspaper articles, sociobiographical information was almost irretrievable.

All the metadata collected are available upon request in an MS Access database from the ICE-SL team.

# 6. 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*. 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.” In: *World Englishes* 1(I). 3-15.

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

Nelson, Gerald (2002): *International Corpus of English - Markup Manual for Written Texts*. <http://ice-corpora.net/ice/manuals.htm> [19 March 2012].

Vine, Bernadette (1999): *Guide to the New Zealand Component of the International Corpus of English (ICE-NZ)*. Wellington: Victoria University.

# Appendix 1: Text samples (excerpts)

## W1A STUDENT ESSAYS

### Plain text with ICE-markup

```
<I><X><p><ICE-SL:W1A-005#1:1>ENG 3246</p>

<p><ICE-SL:W1A-005#2:1>Introduction to Second Language Acquisition</p>

<p><ICE-SL:W1A-005#3:1>Mid Semester Project</p>

<p><ICE-SL:W1A-005#4:1><@>Shirantha Dissanayaka</@></p>
<p><ICE-SL:W1A-005#5:1>02/BA/11924</p>
<p><ICE-SL:W1A-005#6:1>A11924</p>

<p><ICE-SL:W1A-005#7:1>(2338 words)</p></X>




<h><ICE-SL:W1A-005#8:1>Interlanguage</h>

<p><ICE-SL:W1A-005#9:1>Thousands of millions of people are engaged in the
process of learning a second or a third language. <ICE-SL:W1A-005#10:1>The
possible reasons could include migration, academic purpose or as a new
experience. <ICE-SL:W1A-005#11:1>This language <w>learner's</w> target is to
achieve a high language proficiency in their target language. <ICE-SL:W1A-
005#12:1>However they are not gaining the language proficiency overnight. <ICE-
SL:W1A-005#13:1>Depending on the person it would take months or even years.
<ICE-SL:W1A-005#14:1>Not all the learners gain the expected target language
proficiency level. <ICE-SL:W1A-005#15:1>Some people give up, some people stop in
midst of classes while some are gaining the necessary language proficiency.
<ICE-SL:W1A-005#16:1>During the meantime of starting the learning of a new
language to finishing learning it (with reaching the target language
proficiency) language learners produce target language output. <ICE-SL:W1A-
005#17:1>Such output could be called as interlanguage. <ICE-SL:W1A-
005#18:1>Abbreviated with the shortened form IL, interlanguage has become a new
interest in the linguistic studies. <ICE-SL:W1A-005#19:1>A more formal
definition to interlanguage could be taken from Encarta Dictionary
<quote>"intermediary form of a language: a form of language produced by learners
of a second or foreign language, combining features of two or more
languages<space>"</quote> (Microsoft® Encarta® 2006). <ICE-SL:W1A-
005#20:1>Before the exploration of interlanguage errors of second language
learners was seen with a negative perspective. <ICE-SL:W1A-005#21:1>With
developments in error analysis IL errors were considered to be a constructive
force. <ICE-SL:W1A-005#22:1>Through the analysis of errors the linguists were
able to make certain findings.</p>

<p><X><quote><ICE-SL:W1A-005#23:1>"The value of error-making in language
learning was consequently reassessed, with a move away from seeing error as a
purely negative phenomenon. <ICE-SL:W1A-005#24:1>Error analysis became a
valuable tool in the classroom for teachers and researchers. <ICE-SL:W1A-
005#25:1>Various taxonomies were devised to account for certain types of error
(e.g. Dulay and Burt 1974). <ICE-SL:W1A-005#26:1>It was suggested that spoken
and written texts produced different kinds of errors, that there were
differences between grammatical and lexical errors, that it was possible to
construct a gradation of serious and less serious errors."</quote></X></p>
<p><ICE-SL:W1A-005#27:1>(Powell Geraint, P4)</p>

<p><ICE-SL:W1A-005#28:1>In selection of my data source I devised a novel
approach. <ICE-SL:W1A-005#29:1>Instead of collecting new data from an alien
source I used my own diary entries as the data for analysis. <ICE-SL:W1A-
005#30:1>Extracts from diary entries of years 1999, 2001 and <}><->2001</-
><+>2004</+><}> were selected for this purpose. […]
```

## CLAWS-tagged

```
<s>
<X>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W1A-005#1:1>_NULL ENG_NP1 3246_MC </p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W1A-005#2:1>_NULL Introduction_NN1 to_II Second_MD Language_NN1
Acquisition_NN1 </p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W1A-005#3:1>_NULL Mid_JJ Semester_NP1 Project_NN1 </p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W1A-005#4:1>_NULL <@>_NULL Shirantha_NP1 Dissanayaka_NP1 </@>_NULL
</p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W1A-005#5:1>_NULL 02/BA/11924_MC </p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W1A-005#6:1>_NULL A11924_FO </p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W1A-005#7:1>_NULL (_( 2338_MC words_NN2 )_) </p>_NULL
</s>
<s>
</X>_NULL <h>_NULL <ICE-SL:W1A-005#8:1>_NULL Interlanguage_VV0 </h>_NULL
<p>_NULL
</s>
<s>
<ICE-SL:W1A-005#9:1>_NULL Thousands_NNO2 of_IO millions_NNO2 of_IO people_NN
are_VBR engaged_VVN in_II the_AT process_NN1 of_IO learning_VVG a_AT1
second_NNT1 or_CC a_AT1 third_MD language_NN1 ._.
</s>
<s>
<ICE-SL:W1A-005#10:1>_NULL The_AT possible_JJ reasons_NN2 could_VM include_VVI
migration_NN1 ,_, academic_JJ purpose_NN1 or_CC as_II a_AT1 new_JJ
experience_NN1 ._.
</s>
<s>
<ICE-SL:W1A-005#11:1>_NULL This_DD1 language_NN1 <w>_NULL learner_NN1 's_GE
</w>_NULL target_NN1 is_VBZ to_TO achieve_VVI a_AT1 high_JJ language_NN1
proficiency_NN1 in_II their_APPGE target_NN1 language_NN1 ._.
</s>
<s>
<ICE-SL:W1A-005#12:1>_NULL However_RRQV they_PPHS2 are_VBR not_XX gaining_VVG
the_AT language_NN1 proficiency_NN1 overnight_RT ._.
</s>
<s>
<ICE-SL:W1A-005#13:1>_NULL Depending_II21 on_II22 the_AT person_NN1 it_PPH1
would_VM take_VVI months_NNT2 or_CC even_RR years_NNT2 ._.
</s>
<s>
<ICE-SL:W1A-005#14:1>_NULL Not_XX all_DB the_AT learners_NN2 gain_VV0 the_AT
expected_JJ target_NN1 language_NN1 proficiency_NN1 level_NN1 ._.
</s>
<s>
<ICE-SL:W1A-005#15:1>_NULL Some_DD people_NN give_VV0 up_RP ,_, some_DD
people_NN stop_VV0 in_II midst_NN1 of_IO classes_NN2 while_CS some_DD are_VBR
gaining_VVG the_AT necessary_JJ language_NN1 proficiency_NN1 ._.
</s>
<s> […]
```

## W1B LETTERS

### Plain text with ICE-markup

```
<I><p><ICE-SL:W1B-007#1:1><}><->aiyo</-><+><indig>Aiyo</indig></+></}>&dotted-
line; poor you with the papers&dotted-line; <}><->i</-><+>I</+></}> know exactly
how you feel. <ICE-SL:W1B-007#2:1>I remember thinking always, that, that is the
ONE aspect of my job that <}><->i</-><+>I</+></}> HATE! <ICE-SL:W1B-
007#3:1><}><->so</-><+>So</+></}> my sympathies. <ICE-SL:W1B-007#4:1>You and
<}><->i</-><+>I</+></}> both seem to be plagued by baases these days. <ICE-
SL:W1B-007#5:1>Our house is being completed these days (finally). <ICE-SL:W1B-
007#6:1><}><->so</-><+>So</+></}> downstairs and that little verandah thingy
that you have to pass through is being tiled, walls painted, kaparadufyed etc.
about six or seven young men strolling about the place with six different
equally annoying ring tones in their mobile phones which keep going off so often
that <}><->i</-><+>I</+></}> wonder how they find time to paint walls or lay
tiles. <ICE-SL:W1B-007#7:1>Once in a way when <}><->i</-><+>I</+></}> pass
upstairs one of them breaks into song. <ICE-SL:W1B-007#8:1>Really. <ICE-SL:W1B-
007#9:1><}><->so</-><+>So</+></}> annoying. <ICE-SL:W1B-007#10:1>And the dust is
awful. <ICE-SL:W1B-007#11:1><}><->anyway</-><+>Anyway</+></}> <}><->i</-
><+>I</+></}> sometimes shut myself in my room and try to write. <ICE-SL:W1B-
007#12:1>Yes <@>Jack</@> sends me poetry sometimes. <ICE-SL:W1B-007#13:1>He is
really nice no? <ICE-SL:W1B-007#14:1>Met him a couple of times, once after <}><-
>i</-><+>I</+></}> read my poetry somewhere and he came and told me he loves my
stuff, and of course <}><->i</-><+>I</+></}> told him <}><->i</-><+>I</+></}>
really like his stuff, so funny. <ICE-SL:W1B-007#15:1><space>About my poem, yeah
<}><->i</-><+>I</+></}> deliberately put a break there, to kind of show the
break in <quote>'normal'</quote> expectedness kind of. <ICE-SL:W1B-007#16:1>And
also they are <quote>'outside'</quote> of everything. <ICE-SL:W1B-007#17:1>And
yes, <}><->i</-><+>I</+></}> did know<+>.</+></p>

<p><ICE-SL:W1B-007#18:1>He is a paranoid schizophrenic, with homicidal</p>
<p>tendencies sometimes. <ICE-SL:W1B-007#19:1>And mostly lovely when he <}><-
>isnt</-><+><w>isn't</w></+></}> ill.</p>
<p><ICE-SL:W1B-007#20:1><}><->take</-><+>Take</+></}> care and hope you have a
tolerable weekend considering all you are dealing with :)<+>.</+></p></I>



<I><p><ICE-SL:W1B-007#21:2><}><->how</-><+>How</+></}> are you doing&dotted-
line; <ICE-SL:W1B-007#22:2><}><-><@>esther</@></-><+><@>Esther</@></+></}> had
sent me the pictures from that talk with your students at <}><->colombo</-
><+>Colombo</+></}> uni, they were really fun, so sweet of her! <ICE-SL:W1B-
007#23:2><}><->Theres</-><+><w>There's</w></+></}> a nice one with you talking
and me watching. <ICE-SL:W1B-007#24:2><}><->talking</-><+>Talking</+></}> of
which, <}><-><@>saumya</@></-><+><@>Saumya</@></+></}> has written to me saying
she just got into that poem of mine and she thinks <}><->its</-
><+><w>it's</w></+></}> a great idea that it is taught along with mirror. :-)
<ICE-SL:W1B-007#25:2>(<}><->i</-><+>I</+></}> am going to tell her,</p>
<p>that <}><->wasnt</-><+><w>wasn't</w></+></}> my idea, it was yours) <ICE-
SL:W1B-007#26:2>Btw, on the gratiaen front &dotted-line; even <@>Afifa</@> has
submitted! <ICE-SL:W1B-007#27:2>Help!!! <ICE-SL:W1B-007#28:2><}><->all</-
><+>All</+></}> those judges must be</p>
<p>saturated with poetry. <ICE-SL:W1B-007#29:2><}><->sending</-
><+>Sending</+></}> you a recent one (well, <}><->december</-
><+>December</+></}>) <}><->i</-><+>I</+></}> wrote after visiting a friend
(along with his sister) at Angoda. <ICE-SL:W1B-007#30:2>I sent it first to the
sister and she sent it to a friend of hers and she wrote back to me saying it is
<quote>"beautiful"</quote> and also</p>
<p>that she bought more than 10 copies of my book and sent it to her friends
around the world, and that she is looking forward to my next book. <ICE-SL:W1B-
007#31:2>I have never met this lady, but it sure felt good to hear that! <ICE-
SL:W1B-007#32:2>Anyways&dotted-line; hope <}><->alls</-><+><w>all's</w></+></}>
well (inspite of the <w>wombat's</w> return). […]
```

## CLAWS-tagged

```
<s>
<p>_NULL
</s>
<s>
<ICE-SL:W1B-007#1:1>_NULL <}>_NULL <->_NULL aiyo_NNU </->_NULL <+>_NULL
<indig>_NULL Aiyo_NP1 </indig>_NULL </+>_NULL </}>_NULL &dotted-line;_NULL
poor_JJ you_PPY with_IW the_AT papers_NN2 &dotted-line;_NULL <}>_NULL <->_NULL
i_ZZ1 </->_NULL <+>_NULL I_PPIS1 </+>_NULL </}>_NULL know_VV0 exactly_RR how_RRQ
you_PPY feel_VV0 ._.
</s>
<s>
<ICE-SL:W1B-007#2:1>_NULL I_PPIS1 remember_VV0 thinking_VVG always_RR ,_,
that_DD1 ,_, that_DD1 is_VBZ the_AT ONE_MC1 aspect_NN1 of_IO my_APPGE job_NN1
that_CST <}>_NULL <->_NULL i_ZZ1 </->_NULL <+>_NULL I_PPIS1 </+>_NULL </}>_NULL
HATE_VV0 !_! <ICE-SL:W1B-007#3:1>_NULL <}>_NULL <->_NULL so_RG </->_NULL
<+>_NULL So_RR </+>_NULL </}>_NULL my_APPGE sympathies_NN2 ._.
</s>
<s>
<ICE-SL:W1B-007#4:1>_NULL You_PPY and_CC <}>_NULL <->_NULL i_ZZ1 </->_NULL
<+>_NULL I_ZZ1 </+>_NULL </}>_NULL both_DB2 seem_VV0 to_TO be_VBI plagued_VVN
by_II baases_NN2 these_DD2 days_NNT2 ._.
</s>
<s>
<ICE-SL:W1B-007#5:1>_NULL Our_APPGE house_NN1 is_VBZ being_VBG completed_VVN
these_DD2 days_NNT2 (_( finally_RR )_) ._.
</s>
<s>
<ICE-SL:W1B-007#6:1>_NULL <}>_NULL <->_NULL so_RR </->_NULL <+>_NULL So_RR
</+>_NULL </}>_NULL downstairs_RL and_CC that_DD1 little_JJ verandah_NN1
thingy_NN1 that_CST you_PPY have_VH0 to_TO pass_VVI through_RP is_VBZ being_VBG
tiled_VVN ,_, walls_NN2 painted_VVN ,_, kaparadufyed_JJ etc._RA about_RG six_MC
or_CC seven_MC young_JJ men_NN2 strolling_VVG about_II the_AT place_NN1 with_IW
six_MC different_JJ equally_RR annoying_JJ ring_NN1 tones_NN2 in_II their_APPGE
mobile_JJ phones_NN2 which_DDQ keep_VV0 going_VVG off_RP so_RG often_RR that_CST
<}>_NULL <->_NULL i_ZZ1 </->_NULL <+>_NULL I_PPIS1 </+>_NULL </}>_NULL
wonder_VV0 how_RRQ they_PPHS2 find_VV0 time_NNT1 to_TO paint_VVI walls_NN2 or_CC
lay_JJ tiles_NN2 ._.
</s>
<s>
<ICE-SL:W1B-007#7:1>_NULL Once_RR in_II a_AT1 way_NN1 when_RRQ <}>_NULL <->_NULL
i_ZZ1 </->_NULL <+>_NULL I_PPIS1 </+>_NULL </}>_NULL pass_VV0 upstairs_RL
one_MC1 of_IO them_PPHO2 breaks_NN2 into_II song_NN1 ._.
</s>
<s>
<ICE-SL:W1B-007#8:1>_NULL Really_RR ._.
</s>
<s>
<ICE-SL:W1B-007#9:1>_NULL <}>_NULL <->_NULL so_RR </->_NULL <+>_NULL So_RG
</+>_NULL </}>_NULL annoying_JJ ._.
</s>
<s>
<ICE-SL:W1B-007#10:1>_NULL And_CC the_AT dust_NN1 is_VBZ awful_JJ ._.
</s>
<s>
<ICE-SL:W1B-007#11:1>_NULL <}>_NULL <->_NULL anyway_RR </->_NULL <+>_NULL
Anyway_RR </+>_NULL </}>_NULL <}>_NULL <->_NULL i_MC1 </->_NULL <+>_NULL I_PPIS1
</+>_NULL </}>_NULL sometimes_RT shut_VVD myself_PPX1 in_II my_APPGE room_NN1
and_CC try_VV0 to_TO write_VVI ._.
</s>
<s>
<ICE-SL:W1B-007#12:1>_NULL Yes_UH <@>_NULL Jack_NP1 </@>_NULL sends_VVZ me_PPIO1
poetry_NN1 sometimes_RT ._.
</s>
<s> […]
```

## W2D INSTRUCTIONAL WRITING SKILLS/HOBBIES

### Plain text with ICE-markup

<I><h><ICE-SL:W2D-014#61:2>Rock Climbing</h>
<h><ICE-SL:W2D-014#62:2>In the Footholds of Hanuman</h>

<p><O><ICE-SL:W2D-014#63:2>photograph6</O><O>Courtesy of Adventure Asia</O></p>

<p><ICE-SL:W2D-014#64:2>Rock climbing in Sri Lanka goes back to antiquity with the story of Hanuman the monkey-general contained in the Indian epic, the Ramayana. <ICE-SL:W2D-014#65:2>Hanuman searched the island for the princess Sita, abducted by Ravanna, demon-king of Lanka. <ICE-SL:W2D-014#66:2>In doing so Hanuman climbed the peaks of the central massif while dodging the fire arrows of his pursuers. <ICE-SL:W2D-014#67:2>Thankfully, rock climbing today is less stressful.</p>

<p><ICE-SL:W2D-014#68:2>by Jayanthi Kuru-Urumpala</p>

<p><ICE-SL:W2D-014#69:2>Hanging off the edge of a 20-metre vertical rock face, you feel as if you are on top of the world. <ICE-SL:W2D-014#70:2>Not only do you get a <w>bird's</w> eye view of your surroundings, but you also feel a sense of achievement and fulfilment after a successful climb. <ICE-SL:W2D-014#71:2>Although some may argue that <w>it's</w> easier to just find a small footpath and walk to the top, the actual thrill of rock climbing can only be guaranteed if you climb your way to the top. <ICE-SL:W2D-014#72:2>Sounds impossible? <ICE-SL:W2D-014#73:2>Not really. <ICE-SL:W2D-014#74:2>If <w>you've</w> never done it before, take the challenge and give it a try.</p>

<p><ICE-SL:W2D-014#75:2>Before climbing, however, it is imperative that you do some basic stretch exercises to warm up your muscles. <ICE-SL:W2D-014#76:2>Rock climbing is a sport that requires you to use almost all your muscles, including the ones you never knew you had! <ICE-SL:W2D-014#77:2>Make sure you pay extra attention to the muscles on your arms, shoulders, thighs and calves while warming up.</p>

<p><ICE-SL:W2D-014#78:2>Wearing suitable clothes is also important. <ICE-SL:W2D-014#79:2>Jeans should be avoided under any circumstance because they restrict your movement. <ICE-SL:W2D-014#80:2>Long shorts or loose three-quarter pants are recommended as they provide maximum flexibility and protect your knees from getting bruised. <ICE-SL:W2D-014#81:2>Also keep in mind when choosing what you wear that you need to be able to bend your knees easily.</p>

<p><ICE-SL:W2D-014#82:2>You will also have to wear a climbing harness, which will be provided by the adventure company you choose to go climbing with. <ICE-SL:W2D-014#83:2>Make sure your climbing instructor checks your harness before you take off. <ICE-SL:W2D-014#84:2>A loose strap could be disastrous, even fatal. <ICE-SL:W2D-014#85:2>Climbing shoes are an added advantage as they give you a better grip, often making your feet act like extra hands. <ICE-SL:W2D-014#86:2>Choose a pair that fits you well – they should not be too tight or too loose. <ICE-SL:W2D-014#87:2>Most companies offering climbing activities often provide the shoes. <ICE-SL:W2D-014#88:2>However, if they are not available, a light pair of running shoes or sneakers would do just fine. <ICE-SL:W2D-014#89:2>A chalk bag (a small pouch containing chalk powder that can be strapped onto the back of your climbing harness) also comes in handy if you have sweaty fingers.</p>

<p><ICE-SL:W2D-014#90:2>As for safety issues, unless <w>you're</w> a professional climber, make sure that you have someone to belay you while you climb. <ICE-SL:W2D-014#91:2>This basically means that you are attached to a safety rope which is attached to the person belaying you, whose task is to support you in case you fall. <ICE-SL:W2D-014#92:2>So in case you loose your grip and fall while climbing, you do not come crashing to the ground but instead remain suspended in the air. <ICE-SL:W2D-014#93:2>This is often the ideal time to take a good look around you and enjoy the view wherever you are! […]

## CLAWS-tagged

```
<I>_NULL <h>_NULL <ICE-SL:W2D-014#61:2>_NULL Rock_NN1 Climbing_NN1 </h>_NULL
<h>_NULL <ICE-SL:W2D-014#62:2>_NULL In_II the_AT Footholds_NN2 of_IO Hanuman_NP1
</h>_NULL <p>_NULL
</s>
<s>
<O>_NULL <ICE-SL:W2D-014#63:2>_NULL photograph6_FO </O>_NULL <O>_NULL
Courtesy_NN1 of_IO Adventure_NN1 Asia_NP1 </O>_NULL </p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W2D-014#64:2>_NULL Rock_NN1 climbing_NN1 in_II Sri_NP1 Lanka_NP1
goes_VVZ back_RP to_II antiquity_NN1 with_IW the_AT story_NN1 of_IO Hanuman_NP1
the_AT monkey-general_NN1 contained_VVN in_II the_AT Indian_JJ epic_NN1 ,_,
the_AT Ramayana_NP1 ._.
</s>
<s>
<ICE-SL:W2D-014#65:2>_NULL Hanuman_NP1 searched_VVD the_AT island_NN1 for_IF
the_AT princess_NN1 Sita_NP1 ,_, abducted_VVN by_II Ravanna_NP1 ,_, demon-
king_NN1 of_IO Lanka_NP1 ._.
</s>
<s>
<ICE-SL:W2D-014#66:2>_NULL In_II doing_VDG so_RR Hanuman_NP1 climbed_VVD the_AT
peaks_NN2 of_IO the_AT central_JJ massif_NN1 while_CS dodging_VVG the_AT
fire_NN1 arrows_NN2 of_IO his_APPGE pursuers_NN2 ._.
</s>
<s>
<ICE-SL:W2D-014#67:2>_NULL Thankfully_RR ,_, rock_NN1 climbing_NN1 today_RT
is_VBZ less_RGR stressful_JJ ._. </p>_NULL <p>_NULL
</s>
<s>
<ICE-SL:W2D-014#68:2>_NULL by_II Jayanthi_NP1 Kuru-Urumpala_NP1 </p>_NULL
<p>_NULL
</s>
<s>
<ICE-SL:W2D-014#69:2>_NULL Hanging_VVG off_II the_AT edge_NN1 of_IO a_AT1 20-
metre_NNU1 vertical_JJ rock_NN1 face_NN1 ,_, you_PPY feel_VV0 as_CS21 if_CS22
you_PPY are_VBR on_II31 top_II32 of_II33 the_AT world_NN1 ._.
</s>
<s>
<ICE-SL:W2D-014#70:2>_NULL Not_XX only_RR do_VD0 you_PPY get_VVI a_AT1 <w>_NULL
bird_NN1 's_GE </w>_NULL eye_NN1 view_NN1 of_IO your_APPGE surroundings_NN2 ,_,
but_CCB you_PPY also_RR feel_VV0 a_AT1 sense_NN1 of_IO achievement_NN1 and_CC
fulfilment_NN1 after_II a_AT1 successful_JJ climb_NN1 ._.
</s>
<s>
<ICE-SL:W2D-014#71:2>_NULL Although_CS some_DD may_VM argue_VVI that_CST
<w>_NULL it_PPH1 's_VBZ </w>_NULL easier_JJR to_TO just_RR find_VVI a_AT1
small_JJ footpath_NN1 and_CC walk_VV0 to_II the_AT top_NN1 ,_, the_AT actual_JJ
thrill_NN1 of_IO rock_NN1 climbing_NN1 can_VM only_RR be_VBI guaranteed_VVN
if_CS you_PPY climb_VV0 your_APPGE way_NN1 to_II the_AT top_NN1 ._.
</s>
<s>
<ICE-SL:W2D-014#72:2>_NULL Sounds_VVZ impossible_JJ ?_?
</s>
<s>
<ICE-SL:W2D-014#73:2>_NULL Not_XX really_RR ._.
</s>
<s>
<ICE-SL:W2D-014#74:2>_NULL If_CS <w>_NULL you_PPY 've_VH0 </w>_NULL never_RR
done_VDN it_PPH1 before_RT ,_, take_VV0 the_AT challenge_NN1 and_CC give_VV0
it_PPH1 a_AT1 try_NN1 ._. </p>_NULL <p>_NULL
</s>
<s> […]
```

# Appendix 2: 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: e.g. 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 3: Consent form and metadata questionnaire

(see following pages)

| | |
|---|---|
| **Dr Dushyanthi Mendis**<br>Department of English<br>University of Colombo<br>P.O. Box 1490<br>Colombo 3<br>Sri Lanka<br><br>Phone: +94 11 250 0438<br>Email: dushyanthi@english.cmb.ac.lk | **Prof Dr Joybrato Mukherjee**<br>Department of English<br>Justus Liebig University Giessen<br>Otto-Behaghel-Strasse 10 B<br>35394 Giessen<br>Germany<br><br>Phone: +49 641 9930150<br>Fax: +49 641 9930159<br>Email: mukherjee@uni-giessen.de |

# INTERNATIONAL CORPUS OF ENGLISH

Thank you for your willingness to contribute written language data to the International Corpus of English, Sri Lanka Component (ICE-SL). This project is being carried out by the Department of English, University of Colombo (c/o Dr Dushyanthi Mendis), and the Department of English, Justus Liebig University Giessen, Germany (c/o Prof Dr Joybrato Mukherjee) and involves collecting spoken and written material representing present-day Sri Lankan English. More information on ICE can be found on http://ice-corpora.net/ice/. Data from ICE-SL will be used for teaching purposes and for linguistic research on English in Sri Lanka. If you wish your contribution to be anonymous, your name will be treated with the strictest confidentiality and will not be revealed to third parties.

Please read the information below and fill in the blank spaces or tick the boxes as applicable. Please sign and date this form to confirm your willingness to contribute the data. Thank you!

## INFORMATION ON THE MATERIAL CONTRIBUTED

Do you want the material contributed to be anonymised? □ yes
□ no

## SPEAKER-SPECIFIC INFORMATION (STRICTLY CONFIDENTIAL)

### General information

1. Given name(s) and surname: ..........................................................................................

2. Address: ..........................................................................................................

3. Email: ..........................................................................................................

4. Gender: □ male
□ female

5. Year of birth: ..........................................................................................................

6. Occupation: ..........................................................................................................

7. Nationality: ..........................................................................................................

### Places of residence

8. Were you born in Sri Lanka? □ yes (please go to question 9)
□ no (please go to question 10.1.)

9. In which parts of the country did you grow up and how long did you live there?

(place)....................................................................from.....................until......................
(place)....................................................................from.....................until......................
(place)....................................................................from.....................until......................
(please go to question 11.1.)

10.1. In which country were you born? .....................................................................................

10.2. In which country did you grow up?.....................................................................................

10.3. How old were you when you moved to Sri Lanka? .............................................................

11.1. Have you ever lived abroad for more than six months?
□ yes
□ no (please go to question 12)

11.2. Where and when did you live outside Sri Lanka?

(place)....................................................................from.....................until......................
(place)....................................................................from.....................until......................
(place)....................................................................from.....................until......................

**Cultural background**

12. Which ethnic group do you identify with?
□ Sinhalese
□ Tamil
□ Sri Lankan Moor
□ Burgher
□ Other (please specify:..............................................)

**Educational background**

13. What is your **highest** educational qualification?
□ GCE Ordinary Level
□ GCE Advanced Level
□ Bachelor
□ Master
□ Doctorate/PhD
□ Higher National Diploma
□ other ......................................
□ none

**Language skills**

14. What language did you first speak at home?..........................................................................

15. What other languages do you speak fluently? ........................................................................

16. Did you have English-medium instruction until your A-level exam?
□ yes
□ no

**Background information about your linguistic habits and your linguistic surroundings**

17.1. I am in touch with British English via

□ speakers of British English
□ newspapers
□ online material
□ broadcasts
□ literature
□ other: ...................................................………...............
□ nothing

17.2. How often are you in contact with British English?

□ daily
□ a few times a week
□ a few times a month
□ a few times a year
□ other: .....................................................................................

18.1. I am in touch with American English via

□ speakers of American English
□ newspapers
□ online material
□ broadcasts
□ literature
□ other: ...................................................………...............
□ nothing

18.2. How often are you in contact with American English?

□ daily
□ a few times a week
□ a few times a month
□ a few times a year
□ other: .....................................................................................

**CONSENT AGREEMENT**

I hereby give permission that the material contributed may be used for the Sri Lankan component of the International Corpus of English (ICE-SL) for the purpose of linguistic research, teaching and scientific presentations and publications and I agree that a copy of the material described above may be kept permanently in the Department of English of Justus Liebig University Giessen, Germany, and in the Department of English of the University of Colombo, Sri Lanka.

Date: ......................................      Signature: ...........................................................................