The Singapore Corpus

User Manual

Gerald Nelson

April 2002

Contents

Introduction & Credits

It gives me great pleasure to introduce the Singapore component of the International Corpus of English.

The corpus was compiled by members of the Department of English Language & Literature, The National University of Singapore.

The project was initiated by Professor Paroo Nihilani, and was later directed by Dr Ni Yibin, with assistance from Dr Anne Pakir and Dr Vincent Ooi. See Nihilani (1992) and Ooi (1997).

The corpus follows the common design of ICE corpora, details of which may be found on the ICE website, at http://www.hku.hk.english/research/ice/index.htm. More detailed information on ICE may be found in Greenbaum (1990, 1991a, 1991b, 1996).

Queries relating directly to the Singapore corpus should be addressed to Dr Anne Pakir, Department of English Language & Literature, National University of Singapore, Kent Ridge, Singapore 0511. Email:ellannep@nus.edu.sg.

Queries about ICE should be addressed to Dr Gerald Nelson, Department of English, The University of Hong Kong, Pokfulam Road, Hong Kong SAR. Email: ganelson@hkucc.hku.hk.

Gerald Nelson
Hong Kong, April 2002

1. ICE Text Categories and Filenames

The files in the corpus bear filenames corresponding to their classification in the hierarchy of ICE Text Categories. These categories and the corresponding filenames are shown here. On the corpus design, see Leitner (1992), Nelson (1996b).

SPOKEN	**S**
DIALOGUE	**S1**
PRIVATE	**S1A**
Direct Conversations	S1A-001 *to* S1A-090
Telephone Calls	S1A-091 *to* S1A-100
PUBLIC	**S1B**
Class Lessons	S1B-001 *to* S1B-020
Broadcast Discussions	S1B-021 *to* S1B-040
Broadcast Interviews	S1B-041 *to* S1B-050
Parliamentary Debates	S1B-051 *to* S1B-060
Legal Cross-examinations	S1B-061 *to* S1B-070
Business Transactions	S1B-071 *to* S1B-080
MONOLOGUE	**S2**
UNSCRIPTED	**S2A**
Spontaneous Commentaries	S2A-001 *to* S2A-020
Unscripted Speeches	S2A-021 *to* S2A-050
Demonstrations	S2A-051 *to* S2A-060
Legal Presentations	S2A-061 *to* S2A-070
SCRIPTED	**S2B**
Broadcast News	S2B-001 *to* S2B-020
Broadcast Talks	S2B-021 *to* S2B-040
Non-broadcast Talks	S2B-041 *to* S2B-050

WRITTEN	**W**
NON-PRINTED	**W1**
NON-PROFESSIONAL WRITING	**W1A**
Student Essays	W1A-001 *to* W1A-010
Examination Scripts	W1A-011 *to* W1A-020
CORRESPONDENCE	**W1B**
Social Letters	W1B-001 *to* W1B-015
Business Letters	W1A-016 *to* W1B-030
PRINTED	**W2**
ACADEMIC WRITING	**W2A**
Humanities	W2A-001 *to* W2A-010
Social Sciences	W2A-011 *to* W2A-020
Natural Sciences	W2A-021 *to* W2A-030
Technology	W2A-031 *to* W2A-040
NON-ACADEMIC WRITING	**W2B**
Humanities	W2B-001 *to* W2B-010
Social Sciences	W2B-011 *to* W2B-020
Natural Sciences	W2B-021 *to* W2B-030
Technology	W2B-031 *to* W2B-040
REPORTAGE	**W2C**
Press News Reports	W2C-001 *to* W2C-020
INSTRUCTIONAL WRITING	**W2D**
Administrative Writing	W2D-001 *to* W2D-010
Skills & Hobbies	W2D-011 *to* W2D-020
PERSUASIVE WRITING	**W2E**
Press Editorials	W2E-001 *to* W2E-010
CREATIVE WRITING	**W2F**
Novels & Stories	W2F-001 *to* W2F-020

2. markup Symbols in Spoken Texts

<$A>, <$B>, *etc*	Speaker identification
<I>…</I>	Subtext marker
<#>	Text unit marker. Marks the beginning of every "text unit", which corresponds loosely to the orthographic sentence. See Text Unit Numbering.
<O>…</O>	Untranscribed text, eg, <O> speech by George Bush </O>
<?>…</?>	Uncertain transcription
<.>…</.>	Incomplete word(s)
<[>…</[>	Overlapping string
<{>…</{>	Overlapping string set
<,>	Short pause
<,,>	Long pause
<X>…</X>	Extra-corpus text
<&>…</&>	Editorial comment
<@>…</@>	Changed name or word
<quote>…</quote>	Quotation
<mention>…</mention>	Mention, eg, "the word <mention> of </mention>
<foreign>…</foreign>	Foreign word(s)
<indig>…</indig>	Indigenous word(s)
<unclear>…</unclear>	Unclear word(s)

Full details of the markup for spoken texts may be found in the *ICE Markup Manual for Spoken Texts*, which may be downloaded from the ICE website, at http://www.hku.hk/english/research/ice/manuals.htm. See also Nelson (1996a).

3. markup Symbols in Written Texts

<l>...</l>	Subtext marker - marks the beginning and end of each individual sample.
<#>	Text unit marker. Marks the beginning of every sentence and heading. See Text Unit Numbering.
** ... **	Paragraph
<h>...</h>	Heading
<bold>...</bold>	Bold print
<it>...</it>	Italics
** ... **	Underlined text
<smallcaps>...</smallcaps>	Small capitals
<X>... </X>	Extra-corpus text
<quote>...</quote>	Quotation
<foreign>...</foreign>	Foreign word(s)
<indig>...</indig>	Indigenous word(s)
<O>...</O>	Untranscribed material, eg. <O> diagram</O>
<&>...</&>	Editorial comment
<->...</-> <+>...</+>	Misspelled word, followed by its correct spelling, eg. **<->**goverment**</->** **<+>**government**</+>**
<mention>...</mention>	Mention, eg, "the word <mention> *of* </mention>"

Full details of markup for written texts may be found in the *ICE Markup Manual for Written Texts*, which may be downloaded from the ICE website, at http://www.hku.hk/english/research/ice/manuals.htm. See also Nelson (1996a).

4. text Unit Numbering

In written texts, a "text unit" corresponds to an orthographic sentence. Headings, sub-headings, addresses, and captions are also designated as text units.

In spoken texts, a text unit corresponds loosely to the orthographic sentence, though many of them are syntactically incomplete. A change of speaker turn always corresponds to a new text unit.

Each text unit in the corpus has been numbered as shown in this extract:

```
<ICE-SIN:W2A-002#1:1>
<h> Controversial Issues In Curriculum Development </h>

<ICE-SIN:W2A-002#2:1>
By Gan Cheong Eng

<ICE-SIN:W2A-002#3:1>
<h> Background </h>

<ICE-SIN:W2A-002#4:1>
The proliferation of short and long courses in management
studies has not only made difficult the choice of a right
course by students, but also the task of distinguishing a
unique course by administrators and teachers.
```

The numbering scheme is as follows:

ICE-SIN The corpus name, ICE Singapore.

W2A-002 The Text Category, in this case Academic Writing: Humanities. See Text Categories and Filenames.

#1:1, #2:1, #3:1 The text units are numbered in a continuous sequence throughout each text. This is denoted by the *first* number following #.

Some texts are composite (ie they consist of two or more different samples). We refer to these samples as "subtexts". The number following the colon denotes the subtext number. By convention, every text has at least one subtext, so the subtext number is always at least 1.

In spoken texts, the text unit number additionally includes the speaker identification (A, B, C, etc.), e.g.

<ICE-SIN:S1A-001#2:3:A>

This refers to text unit 2, in subtext 3, uttered by speaker A.

5. licence Agreement

International Corpus of English
The Singapore Corpus (ICE-SIN)
Licence Agreement

In the following, "ICE-SIN" refers to "The Singapore Component of the International Corpus of English". The Licensee is the purchaser of the Corpus and agrees to abide by this licence agreement. By placing the CD in the CD-ROM drive of their computer, the Licensee is agreeing to the terms of this licence.

General terms and conditions

The Corpus must be used for **non-profit linguistic research** purposes only. The licence cannot be transferred, lent, or re-sold.

The Licensee agrees not to reproduce or redistribute the ICE-SIN Texts or to use all or any part of the ICE-SIN Texts in any commercial product or service. A copy of the ICE-SIN Corpus may be made for backup purposes.

Copyright in all ICE-SIN Texts is retained by the original copyright holders.

The Corpus may be fully installed onto the Licensee's computer, by copying the relevant files from the CD supplied onto the computer's hard disk.

The Licensee is allowed to make copies of the Corpus on computers within the Institution named in this licence.

The licence entitles all staff and students of the named Institution to make use of the Corpus on these computers.

It is the responsibility of the Licensee to ensure that the Corpus cannot be accessed from outside the named Institution. The licence does not entitle the Licensee to include the Corpus in a public-access internet site.

It is the responsibility of the Licensee to ensure that other users of the Corpus within the named Institution are made aware of the terms of this Licence.

Publications based on the ICE-SIN Corpus may include citations from ICE-SIN Texts only in a way which would be permitted under the fair dealings provision of copyright law.

All publications based on the ICE-SIN Corpus must give credit to the ICE-SIN Corpus and to the Department of English, The National University of Singapore.

The Licensee agrees to cooperate in any future enquiries made by the International Corpus of English or by the Department of English, The National University of Singapore, concerning the use of the ICE-SIN Corpus.

The general terms and conditions apply.

REFERENCES

Greenbaum, Sidney (1990) 'Standard English and the International Corpus of English'. *World Englishes* 9. pp.79-83.

Greenbaum, Sidney (1991a) 'ICE: the International Corpus of English'. *English Today* 28. pp.3-7.

Greenbaum, Sidney (1991b) 'The development of the International Corpus of English'. In: Karin Aijmer and Bengt Altenberg (eds.) *English corpus linguistics. Studies in honour of Jan Svartvik*. London: Longman. pp.83-91.

Greenbaum, Sidney (ed.) (1996) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Leitner, Gerhard (1992) 'International Corpus of English: Corpus design - problems and suggested solutions'. In: Gerhard Leitner (ed.) *New directions in English language corpora: methodology, results, software developments*. Berlin: Mouton de Gruyter. pp.33-64.

Nelson, Gerald (1996a) Markup Systems. In: Greenbaum (1996), pp.36-53

Nelson, **Gerald** (1996b) The Design of the Corpus. In: Greenbaum (1996), pp.27-35

Nihilani, Paroo (1992) 'The International Computerized Corpus of English'. In: Anne Pakir (ed.) *Words in a cultural context*. Singapore: UniPress. pp.84-88.

Ooi, Vincent (1997) 'Analysing the Singapore ICE corpus for lexicographic evidence'. In: Magnus Ljung (ed.) *Corpus-based studies in English*. Amsterdam: Rodopi. pp.245-260.