1

The NIE corpus of spoken Singapore English (NIECSSE)

David Deterding and Low Ee Ling

Introduction

In recent years, much linguistic research has used computer-based corpora of speech, enabling various researchers to access the same data and thereby make direct comparisons of their findings. Furthermore, use of such corpora eliminates the need for each researcher to collect and painstakingly transcribe a completely new set of data. Typical corpora that have been used in the analysis of English include the International Corpus of English (ICE: Greenbaum 1996) and the Machine-Readable Spoken English Corpus (MARSEC: Roach, Knowles, Varadi & Arnfield 1993).

Data from Singapore comprises one component of the ICE corpus, known as ICE-SIN, and some useful findings have emerged from this project (eg Ni & Ler 2000; Ooi 2001). Although ICE-SIN includes a substantial amount of spoken data such as conversations, phone calls, business transactions, demonstrations, and unscripted and scripted speech (Ni & Ler 2000:170), there is a need to collect more spoken data, and this has resulted in the Grammar of Singapore English Corpus (GSEC) funded by the National University of Singapore (NUS). However, as these recordings were made in natural conditions, there is a great deal of extraneous noise and also many unintelligible or untranscribable segments (Zhu 1999). Such recording conditions certainly encourage natural speech, which is valuable for the analysis of patterns of grammar and word usage. However, it means that detailed acoustic/phonetic analysis is often difficult or even impossible. The NIE corpus aims to fill this gap.

Deterding D, Brown A & Low E L (eds) English in Singapore:
Phonetic Research on a Corpus © *McGraw-Hill Education (Asia) 2005*

The main goals of the NIECSSE are:

- to provide high-quality recordings, comparable to those in MARSEC, that are ideally suited for phonetic analysis
- to make the data easily accessible and immediately available to all researchers with no restrictions except for the need to acknowledge the source
- to encourage people from around the world to use data from Singapore English for their research

This paper, which is an updated version of an early description of the corpus (Deterding & Low 2001), gives an overview of the NIECSSE as it exists currently, and as it is distributed on the CD-ROM.

Recording conditions

All recordings were made in the Phonetics Laboratory at NIE. The laboratory is quiet but not soundproofed. In all cases, a high-quality Shure SM48 dynamic microphone was positioned just a few inches from the lips of the subjects. For the interviews, the interviewer's voice is much quieter: it is comprehensible and has been fully transcribed, but the focus of the data is intended to be the subjects, not the interviewer.

All recordings were made directly onto a computer using CSL software from KAY. The sampling rate was 22050, to ensure a high-quality recording. This is exactly half the most common sampling rate for music on compact disks, but it is rather higher than that generally adopted for recorded speech (Hayward 2000:68). It allows an analysis up to the Nyquist frequency of 11025 Hz, which is more than adequate for the complete description of speech.

The recordings were saved using the standard .wav format to ensure that the data can easily be used by other researchers without the need to download any special software.

Data

There are three kinds of data:

- interviews
- a read passage
- instances of extra consonants from informal conversations

These three sets of data will be described separately.

Interviews

A total of 46 Singaporeans have been interviewed, 31 female (labelled F1 to F31) and 15 male (M1 to M15). They are all educated Singaporeans, most of them trainee teachers at NIE. All of them speak English well, and for many of them, English is their best language, although nearly all speak at least one other language fluently. Brief biographical details, including age, ethnic group, languages spoken, relationship with the interviewer, and date of recording, are included with the data of each speaker. A summary of the speakers is shown in Table 1.1.

Table 1.1: Summary of the SgE subjects

Speaker	Ethnic group	First language	Speaker	Ethnic group	First language
F1	Chinese	Mandarin	F24	Chinese	Mandarin
F2	Chinese	Mandarin	F25	Chinese	Mandarin
F3	Chinese	Hokkien	F26	Chinese	English
F4	Malay	Malay	F27	Chinese	Mandarin
F5	Chinese	Cantonese	F28	Chinese	Mandarin
F6	Malay	Malay	F29	Chinese	Mandarin
F7	Malay	Malay	F30	Chinese	Mandarin
F8	Sikh/Chinese	English	F31	Chinese	English
F9	Chinese	Mandarin	M1	Chinese	Mandarin
F10	Chinese	Mandarin	M2	Chinese	Hokkien
F11	Chinese	Hokkien	M3	Chinese	Mandarin
F12	Chinese	Mandarin	M4	Chinese	Mandarin
F13	Chinese	Teochew	M5	Chinese	Mandarin
F14	Chinese	Mandarin	M6	Chinese	Mandarin
F15	Chinese	English	M7	Chinese	Mandarin
F16	Malay	Malay	M8	Chinese	Mandarin
F17	Chinese	Teochew	M9	Chinese	Mandarin
F18	Chinese	Mandarin	M10	Chinese	Mandarin
F19	Eurasian	English	M11	Chinese	Mandarin
F20	Chinese	Hokkien	M12	Chinese	Hokkien
F21	Chinese	Hokkien	M13	Chinese	Mandarin
F22	Chinese	Mandarin	M14	Chinese	Teochew
F23	Chinese	Mandarin	M15	Indian	English

Although the shortage of Malay and especially Indian speakers is unfortunate if we want to make reliable comparisons between the speech of the different ethnic groups in Singapore, the fact that the overwhelming majority of the speakers are Chinese reflects the ethnic make-up of Singapore fairly accurately.

The interviewer is a British lecturer at NIE, and had taught most of the speakers, in some cases for a number of courses, before the interviews. Because they were talking to their lecturer, and also because they were acutely aware that they were being recorded, the speakers will have been using a style of speech among the most formal of their repertoire (Pakir 1991). On the diglossic model of Gupta (1992), the students will have been using their H(igh) variety and not the L(ow) colloquial variety ('Singlish') that is more likely to be found in informal conversations between friends.

Five young British speakers have also been recorded, two female (labelled BF1 and BF2) and three male (labeled BM1 to BM3). Although this is not the main focus of the corpus, it may be useful to provide a comparison: if a feature of speech is claimed to be special to Singapore, it may be valuable to check whether this feature also occurs with British English speakers. These speakers were recorded under exactly the same conditions as the Singaporean speakers.

All interviews lasted for five minutes. At the end of each recording, the subject was asked if any of the material should be deleted, and a few did ask for some of what they had said to be removed. One or two of the interviews are therefore shorter than five minutes.

For ease of access, all the interviews have been cut into segments of between 20 seconds and one minute in length. A suitable break in the interview was chosen for each break, in the hope that each segment represents a reasonably coherent stretch of discourse. All the interviews have been transcribed orthographically, including all filled pauses (eg *er, mmm*) and silent pauses (shown as "..."). In the transcripts, the interviewer is shown as I, and the interviewee (subject) as S. The transcripts are kept in html format along with each speech file, to ensure ease of access.

Read passage

Three of the speakers who participated in the interviews (M1, F1 and F2) also recorded the *North Wind and the Sun* passage in order to investigate speech patterns in a read passage.

Although there is a question about the naturalness of read speech, there are substantial advantages in having prepared texts that allow the

comparison of the same material from different speakers. The intention is to expand this section of the data, with recordings of data specially designed to analyse specific features of pronunciation (such as vowel length, voicing of fricatives, consonant cluster simplification and use of dental fricatives).

Extra consonants

In order to evaluate the incidence of extra final consonants in relatively informal situations, Lim (2003) recorded three of the speakers (F9, F10 and F13) speaking among themselves in fairly casual conversations. The instances of extra final consonants found in these recordings have been extracted and included in the 'extra consonants' section.

Availability

The NIECSSE is available:

- on a CD-ROM accompanying this book
- on a CD-ROM by request from the first author
- on-line at videoweb.nie.edu.sg/phonetic/niecsse/index.htm

Fellow researchers are welcome to use these recordings for research purposes, with suitable acknowledgements.

Acknowledgement

The collection of the corpus, and its analysis, was made possible by the NIE-funded project *Theoretical Speech Research and its Practical Implications* (RI 1/03 LEL, 2004 to 2007) and its predecessor project *An Acoustic Analysis of Singapore English with Special Reference to its Pedagogical Applications* (RP 11/99 LEL, 1999 to 2003).

References

Deterding D & Low E L (2001) 'The NIE corpus of spoken Singapore English (NIECSSE)' *SAAL Quarterly* 56 (Nov 2001):2–5.

Greenbaum S (1996) *Comparing English Worldwide: The International Corpus of English* Oxford University Press.

Gupta A F (1992) 'Contact features of Singapore Colloquial English' in K Bolton & H Kwok (eds.) *Sociolinguistics Today: International Perspectives* Routledge, London and New York, pp 323–45.

Hayward K (2000) *Experimental Phonetics* Longman, Harlow and New York.

Lim S H (2003) 'Extra final consonants in educated Singapore English' Honours academic exercise, National Institute of Education, Singapore.

Ni Y & Ler S L (2000) 'Wordforms and their linguistic and cultural implications: a comparison between two corpora in the International Corpus of English' in A Brown (ed) *English in Southeast Asia 99* National Institute of Education, Singapore, pp 159–77.

Ooi B Y V (2001) 'Upholding standards or passively observing language? Corpus evidence and the concentric circles model' in B Y V Ooi (ed) *Evolving Identities: The English Language in Singapore and Malaysia* Times Academic Press, Singapore , pp 168–83.

Pakir A (1991) 'The range and depth of English-knowing bilinguals in Singapore' *World Englishes* 10(2):167–79.

Roach P, Knowles G, Varadi T & Arnfield S (1993) 'MARSEC: A machine-readable spoken English corpus' *Journal of the International Phonetic Association* 23(2):47–54.

Zhu S F (1999) 'A frame upon which the melodic line is hung: a look at Singapore English intonation forms and functions' Paper presented at the 4th English in Southeast Asia (ESEA) Conference, 22–24 November 1999, National Institute of Education, Singapore.