Corpora of NZ English

Several corpora of New Zealand English are located at Victoria University of Wellington and the University of Canterbury.

Victoria University of Wellington

Linguists at Victoria University of Wellington have been involved in the collection of New Zealand English for four different corpora: two spoken, one written, and a third including both spoken and written data.

Wellington Corpus of Written New Zealand English (WWC)

The Wellington Corpus of Written New Zealand English (WWC) consists of one million words of written New Zealand English collected from writings published between 1986 to 1990. The WWC has the same basic categories as the Brown Corpus of written American English (1961) and the Lancaster-Oslo-Bergen corpus (LOB) of written British English (1961). The corpus also parallels the structure of the Macquarie Corpus of written Australian English (1986). The WWC consists of 2,000 word excerpts on a variety of topics. Text categories include press material, religious texts, skills, trades and hobbies, popular lore, biography, scholarly writing and fiction.

Wellington Corpus of Spoken New Zealand English (WSC)

The Wellington Corpus of Spoken New Zealand English (WSC) comprises one million words of spoken New Zealand English collected between 1988 to 1994. The corpus consists of 2,000 word extracts (where possible) and comprises different proportions of formal, semi-formal and informal speech. It includes both monologue and dialogue categories, as well as broadcast and private material collected in a range of settings. Seventy-five percent of the corpus is informal dialogue. A brief outline of the Wellington Corpus of Spoken New Zealand English (WSC) is provided here: http://www.victoria.ac.nz/lals/research/corpora/wcs.aspx

The New Zealand component of the International Corpus of English (ICE-NZ)

The New Zealand component of the International Corpus of English (ICE-NZ) consists of one million words of spoken and written New Zealand English collected between 1989 and 1994. It includes 600,000 words of speech and 400,000 words of written text. The Wellington Corpus of Spoken New Zealand English and the spoken component of ICE-NZ share 9 categories. Because informal conversational data in particular was so difficult to collect, there is an overlap of 339,530 words (173 files) between the two corpora to achieve economy in data collection.

New Zealand Spoken English Database (NZSED)

The New Zealand Spoken English Database (NZSED) is collecting a representative sample of English spoken in late 20th/early 21st century New Zealand. Researchers and students in phonetics will be able to access digitally stored quality recordings of the speech sounds of New Zealand English in controlled linguistic contexts. NZSED is being constructed in such a way that it is easily accessible to powerful search and statistical procedures such as EMU and R. The recorded materials are based on those recorded for ANDOSL (the Australian National Database of Spoken Language), and include both read material and dialogue. More information on the NZSED is available here: http://www.victoria.ac.nz/lals/staff/paul-warren/nzsed/index.htm

For more information on any of the Wellington corpora above and how to access them contact:

Corpus Manager
Archive of New Zealand English
School of Linguistics and Applied Language Studies
PO Box 600
Wellington
New Zealand

Email: Corpus-Manager@vuw.ac.nz

Diachronic corpus of New Zealand English

This corpus consists of almost 5 1/2 million tokens of running text gathered in six indicator years (1850, 1880, 1910, 1940, 1970, 2000) from newspapers, parliamentary debates and the New Zealand School Journal, and is the largest corpus yet assembled for the study of New Zealand English. This corpus is notable for its diachronic component and its inclusion of a neglected genre (writing for young people), and has already been used for a range of purposes, i.e. studying Maori lexical items, gender and language, and vocabulary in New Zealand English. Six of the sixteen files, containing around two million tokens, can be searched electronically, and are available by contacting john.macalister@vuw.ac.nz.

University of Canterbury

The University of Canterbury holds three further major corpora of New Zealand English. These corpora span almost the entire history of English spoken in New Zealand and are being analysed as part of the Origins of New Zealand English (ONZE) project [insert link to summary on ONZE project). The summaries of the corpora below are taken mostly from Maclagan and Gordon (1999).

Mobile Unit Corpus

The oldest data is contained in the Mobile Unit Corpus, a collection of oral history and other recordings made between 1946 and 1948 by the Mobile Unit of the National Broadcasting Corporation. These were acquired from the Radio NZ Sound Archives in 1989. The 250 or so speakers in this archive were born between 1850 and the early 1900s, most of them from 1860-1890. Many are of the first generation of New Zealand-born angolophones, and their speech gives us information about the very early stages of the development of New Zealand English.

Intermediate Corpus

The Intermediate Corpus contains approximately 130 recordings of speakers born between 1890 and 1930, most of them from 1900-1925. Sixty-nine of the speakers were interviewed by oral historian Rosemary Goodyear between 1989 and 1995. The remainder of the corpus comes from a variety of sources, including 55 recordings collected by various ONZE researchers in the 1990s. Many of the latter recordings feature speakers who are descendants of Mobile Unit interviewees.

Canterbury Corpus

The Canterbury corpus consists of recordings of speakers born between 1935 and 1980. This collection began in 1994 and has been added to every year since. The recordings involve the speaker reading a list of 200 words and participating in a conversation with the interviewer. The corpus is designed to include a balanced number of participants differing on the following variables: occupation (non-professional or professional), gender (men and women), and age (young adults (20-30) and middle-aged (45-60)). The word list is designed to elicit features of New Zealand pronunciation of special interest, such as front vowels, closing diphthongs, the possible merger of EAR and AIR, the vocalisation of /l/, the pronunciation of grown (etc.) as one or two syllables, the pronunciation of the TOUR vowel and the increasing trend for 'th' to be pronounced as /f/.

References

Bauer, Laurie. 1994. Introducing the Wellington Corpus of Written New Zealand English. *Te Reo* 37: 21-28.

Bauer, Laurie 1993. Manual of Information to Accompany the Wellington Corpus of Written New Zealand English. Wellington: Department of Linguistics, Victoria University of Wellington.

Holmes, Janet, Bernadette Vine and Gary Johnson 1998. The Wellington Corpus of Spoken New Zealand English: a Users' Guide. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Maclagan, Margaret A. and Elizabeth Gordon. 1999. Data for New Zealand social dialectology: the Canterbury corpus. New Zealand English Journal 13: 50-58.

Vine, Bernadette. 1999. A word on the Wellington corpora. *New Zealand English Journal* 13: 59-61.

Warren, Paul. 2002. NZSED: building and using a speech database for New Zealand English. *New Zealand English Journal* 16: 53-58.

Julia de Bres
Research Fellow
School of Linguistics and Applied Language Studies
PO Box 600
Wellington
New Zealand
julia.debres@vuw.ac.nz