Data for New Zealand social dialectology: the Canterbury Corpus.

Margaret Maclagan and Elizabeth Gordon

University of Canterbury

Introduction

In 1988 Janet Holmes and Allan Bell published a paper in *Te Reo*, 'Learning by experience: notes for New Zealand social dialectologists.' Here they set out the methodology that they developed in Wellington for the major research project known informally within New Zealand as the Porirua Project (elsewhere as the Wellington Social Dialect Survey). They discuss the issues of controlling and eliciting social variables, and set out their schedules including the demographic questionnaire, the stimulus material for eliciting formal speech (fill the gap exercises, questions on lexical items, a reading passage, and two word lists) and informal speech (questions which would be most likely to elicit spontaneous conversation). This initial plan was later modified but the resulting data collected through the project has been of great value, not just for its initial purpose — the description of a corpus of New Zealand English (NZE) — but as a source of data for other more narrowly based research projects: HRTs, narrative, Maori English, ear/air etc. The Porirua Project has had a major influence on sociolinguistic research in New Zealand over the past decade and will continue to be influential in the future. The data gathered has now been supplemented by that collected in the Wellington Corpus of Spoken New Zealand English (Holmes et al, 1998) so that even more material is now available.

Spoken data at the University of Canterbury

Research into NZE at the University of Canterbury began in a less structured way and much has been connected to student research or to the teaching of specific university courses (Gordon and Maclagan, 1995). The holdings at Canterbury now consist of three major corpora which span almost the entire history of English spoken in New Zealand. The oldest data is contained in the Mobile Unit Corpus (see Lewis, 1996), a collection of oral history and other recordings made between 1946 and 1948 by the Mobile Unit of the National Broadcasting Corporation. 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 anglophones, and their speech gives us information about the very early stages of the development of NZE. This corpus has undergone partial analysis in a project at the University on the origins of NZE (ONZE). (See Gordon 1998, 1999; Gordon & Trudgill 1999; D. Maclagan 1998; Trudgill 1997, 1999; Trudgill & Lewis 1999; Trudgill et al. 1998; Watson et al. 1998.)

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. Work on the data in the Intermediate Corpus is still ongoing, and neither the results of analysis nor a fuller description of the archive has yet been published.

The third corpus, and the subject of this paper, is the Canterbury Corpus, which consists of recordings of approximately 300 speakers who were born between 1935 and 1980. Our aim here is to describe the Canterbury Corpus and provide sufficient information about the data contained in it so that other researchers will be able to judge whether accessing the corpus would be useful for their own purposes.

Background to the Canterbury Corpus

This corpus grew out of two specific Linguistics courses on New Zealand English at the University: a B.A. Honours course which was set up in 1992, and a third year course which started in 1994. The initial planning for the courses was done in conjunction with Professors Lesley and James Milroy who were visitors under the auspices of the Canterbury Visiting Fellowship in 1992. The Milroys suggested replicating the kind of field work courses that they had taught in Britain where students recorded both word lists and casual speech and wrote a report on field work methodology. Although the original word list devised in 1992 has now been considerably modified, the general structure for research as suggested by the Milroys has remained the underlying basis for the collection of data in the Corpus. Students in the third year course are required to record two New Zealand speakers selected according to specific criteria. The speakers read a prepared word-list and are engaged in conversation, with a minimum of half an hour recorded, of which ten minutes are then transcribed orthographically. Students choose whether to record the word list at the beginning, during or at the end of the interview.

Structure of the Canterbury Corpus

Selection of Speakers

The students work in groups of four and each member of a group is asked to collect two recordings so that collectively the group covers the eight categories in the speaker quota sample:

Young adults (approx. age 20-30)

Non-professional (Manual/unskilled)		Professional	
Male	1	Male	1
Female	1	Female	1

Middle-aged speakers (approx. age 45-60)

Non-professional (Manual/unskilled)		Professional	
Male	1	Male	1
Female	1	Female	1

Some students choose to record two speakers where only one social variable contrasts (e.g., sex), while others wish to maximise the contrasts in their data. They are required to discuss and define the categories in the speaker quota sample and to reflect on the ways in which their choice of speaker affects the comparisons available to them. There are now over 30 subjects in each of the 8 speaker categories. Although most subjects are from the Canterbury region, there are some students who were able to record speakers from their home towns in various parts of New Zealand.

Identifying Social Class

The definition of social class categories in New Zealand immediately raises problems. The use of the terms 'manual/unskilled' and 'professional' suggested by the Milroys has always been the source of considerable dissatisfaction among students but alternatives such as 'working class' or 'middle class' were considered to be equally unsatisfactory. The original categories suggested have been retained, now coded 'non-professional' and 'professional' but their limitations are readily acknowledged. In order to assign social class ratings, all the speakers in the corpus are coded according to both occupation and education.

For occupation we follow other New Zealand social scientists and use the Elley-Irving Index of Social Stratification (1985) together with the revised version prepared by the NZ Ministry of Education (1990). While this index has the advantage of comparability and consistency, it has become somewhat dated. For example, the occupation of 'counsellor', now a fairly common type of employment for women, is not included. Another disadvantage is that, as the data in the Index are based on census information, groups such as farmers, who are able to claim a low income and high expenses, are rated as social class 4 or 5. This would suggest that private girls' boarding schools are largely populated by people of a low income and social class, when the reality is clearly different.

We devised a 6 point education scale as follows:

PhD/Higher tertiary	1
Tertiary degree	2
Trade certificate/diploma	3
Sixth-Seventh form qualification	4
School Certificate only	5
No secondary school qualification	6

When the two scales are combined and divided by 2, the highest rank is still 1 (professional occupation such as doctor, lawyer, university lecturer with a PhD or a higher tertiary degree), and the lowest rank remains 6 (someone without any educational qualification in menial occupations such as domestic cleaner). One of the advantages of combining the two scales is that it can differentiate between those who have obtained highly paid positions through educational qualifications, and those who have achieved such positions without much formal education. Likewise it can identify someone in a low paid job who nevertheless has high educational qualifications. When a sample of 250 speakers from the Canterbury Corpus was assessed on the combined rating scale, the professional speaker groups had mean ratings between 2 and 2.5 and the mean for the non-professional speaker groups ranged between 4.5 and 5.5. Even though New Zealanders are not comfortable with rankings of social class, the professional and non-professional categories used in the Canterbury Corpus would appear to be tapping into two different sections of society.

Phonological variables

Subjects are asked to read a word-list, including the number at the beginning of each line (see Appendix A). This gives us about 200 words to analyse per speaker. The word list has been designed to elicit features of New Zealand pronunciation which are of special interest such as the front vowels, the 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/.

Subjects do not regard the numbers at the start of each line as being part of the word list. Their reading of the numbers therefore gives more informal pronunciations of variables such as the DRESS vowel (line 5) which can be compared with the numbers 7 and 10, and therefore helps to counteract the style shifting expected in word list reading. The effect is particularly clear for line 26 which is designed to elicit flapped /t/. Speakers who use very few flapped /t/ in line 26 often use flaps in numbers such as 13, 14, 30, and 31.

Casual Speech

One advantage of using students to collect casual speech is the incredible variety of the contacts they are able to make. Interviewees range from convicted criminals currently serving prison sentences or just released, proprietors of massage parlours, and those currently unemployed, through to lawyers, doctors and the owners of various businesses. The variety is much greater than the authors could have obtained by using conventional sociolinguistic contact techniques such as friend of a friend (Milroy, 1987).

The students record at least 30 minutes of casual speech and transcribe 10 minutes. They are given explicit guidelines about the format for these transcriptions so that all those held in the Canterbury Corpus use the same conventions (see Appendix B). An extract from a transcription follows:

Sample transcript (extract)

is it alright for Joe to lick his . scar?

it is as long as he doesn't do it *too much too roughly* a little bit . you're sitting on a wee nerve there darling hey -- a little bit's fine because it's um keeps it clean and and their tongue's quite sterile **mmm** dogs' tongues . but if he did it a lot um . he has to have one of those bucket things on his head because .

mm keep his eyes .

because if they lick too often the tongue's rasping and it gets it quite raw . **mm** and . you note that cut that's the main cut . and he's sort of got it like that and there's another one round here

is that his cut over from . his past operations?

same side . it's not on the same scar you can still see is that stopped or is it still going?

still going I think

thankyou thankyou--*[unclear]* down here there's another one he took it out and with . there **yeah** two incisions but got it out in one piece I don't know how he could . he must have cut the bone . perhaps loosened it away and then *cut out the bigger one*

right . makes sense

but this the um the second . operation he had which was is a major one too . he had to wear a . a bucket after . several days because he was licking it a lot and it was getting quite raw - um . and so hopefully he won't do that cos those bucket things are annoying they . for them . they put one on on last time . that looked a bit like an upside down lampshade . have you ever seen one?

yeah those *big circular column things*

oh oh right and it was quite big and when we brought him home if he didn't . if he wasn't careful . and aimed himself directly through a doorway . it was so big it would hit on the **yeah** . on the door frame

All orthographic transcriptions and their associated tape recordings are kept available for research in a database in the Linguistics Department. Students in the third year course do a second practical assignment on a topic of their choice and several students have already used the Canterbury Corpus database for projects (see Appendix C). The second projects are also kept available for consultation in the Department.

Synchronic and diachronic data in the Canterbury Corpus

The Corpus has grown every year since 1994 as each new class of New Zealand English students collect their data. The question now arises as to how long new data can be added to the Corpus without adversely affecting the validity of the speaker groups. In order to allow changes in apparent time to be observed, the Corpus is structured to give two age groups: 20–30 and 45–60 years. However, if new speakers continue to be added over too long period, we will strike a diachronic problem in that there is the risk of language changes being caught up in what are supposed to be synchronically uniform groups. So far we have collected data over a period of 6 years

without encountering any apparent problems. We hope to be able to collect data for a period of 10 years and then reevaluate the position. There is a built-in safety check. All recordings are dated, so we can always collate data according to the year collected as well as according to the age of the speaker. If data is collected for 10 years, we will then be able to compare the material collected during each 5 year period and see whether any trends become apparent.

Future work

Several papers using material from the Corpus have already been published (see Appendix D). To date, more work has been done with the word list data than with the casual speech. The material in the corpus holds the potential for a great deal more analysis and investigation of NZE. For example, we have not yet examined the developing affrication of /tr/ and /str/, the extent of /l/-vocalisation or the increasingly central pronunciation of the /ʊ/ in *good*. Nor have we done more than a preliminary investigation of the increasing substitution of /f/ for /θ/.

The corpus is structured to enable various comparisons to be made. As well as the obvious comparisons of age, sex and social class, comparisons can now be made between speakers of similar ages recorded several years apart. We have yet to study the casual speech data in detail or to systematically compare word list and casual speech for many of the variables already studied from word list data alone.

As we have worked on the Mobile Unit Corpus we have been struck by the number of times speakers who were born in the 1800s show instances of changes that are usually regarded as part of modern NZE. We have coined the term *embryonic variants* (Gordon & Trudgill, 1999) for these early examples of changes that later became widespread. As time passes, it will become clear that examples of embryonic variants of future changes have been captured in the recordings of the Canterbury Corpus. Because we are contemporaneous to the data, we are not always able to tell which unusual pronunciations are merely idiosyncratic and which are actually early instances of changes that will become much more common in the future.

Conclusion

The three corpora available at the University of Canterbury provide spoken material touching on almost the whole history of English in New Zealand. Most of the data is also available in written transcriptions. The Mobile Unit corpus provides us with data from almost the beginnings of English in this country and the Intermediate Corpus provides material for people born early in the 20th century. Both contain a large number of speakers, but neither is structured, so that it is not always possible to find speakers who fit a particular category. The Canterbury Corpus provides approximately 300 speakers with at least 30 speakers in each of 8 carefully selected categories. Together the three corpora constitute a unique archive of spoken and transcribed data on New Zealand English, which, it is hoped, will take its

place alongside the valuable material already available at other centres of research.

Acknowledgement

We wish to thank Gillian Lewis for her helpful comments on an earlier version of this paper.

References

Elley, W. B. & Irving, J.C. 1985. The Elley-Irving Socio-Economic Index: 1981 census revision. New Zealand Journal of Educational Studies. 20: 115-128.

Gordon, E. 1998. The origins of New Zealand speech: the limits of recovering historical information from written records. English World-Wide 19 (1): 61-85.

Gordon, E. in press 1999. Embryonic variants in New Zealand English sound changes. Proceedings of the Sixth NZ Language and Society Conference, Victoria University of Wellington, June 1998. To be published in a special edition of Te Reo.

Gordon, E. & Maclagan, M. 1995. Making a virtue of necessity: combining teaching and research in the study of New Zealand English. New Zealand English Newsletter 9: 27-31.

Gordon, E. & Trudgill, P. in press, 1999. Shades of things to come: Embryonic variants in New Zealand English sound changes. English World-Wide 20.

Holmes, J. & Bell, A. 1988. Learning by experience: notes for New Zealand social dialectologists. Te Reo: Journal of the Linguistic society of New Zealand 31: 33-41.

Holmes, J., Vine, B., & Johnson, G. 1998. The Wellington Corpus of Spoken New Zealand English: a User's guide. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Lewis, G. 1996. The origins of New Zealand English: a report on work in progress. New Zealand English Journal 10: 25-30.

Maclagan, D. 1998. /H/-dropping in early New Zealand English. New Zealand English Journal 12: 34–42.

Milroy, L. 1987. Language and Social Networks 2nd edition. Oxford: Basil Blackwell. First edition published 1980.

New Zealand Ministry of Education. 1990. Derivation of Elley-Irving Codes from Census Occupations. Wellington: Unpublished manuscript.

Trudgill, P. 1997. The chaos before the order: New Zealand English and the second stage of new dialect formation. In E.H. Jahr (ed.) Historical Sociolinguistics. Berlin: Mouton de Gruyter.

Trudgill, P. 1999. A Southern Hemisphere East Anglian: New Zealand English as a resource for the study of 19th century British English. In P. Lucko and U. Carls (eds.) Festschrift for Klaus Hansen. Berlin: Humboldt University, 1999.

Trudgill, P. & Lewis, G. in press 1999. A window on the past: New Zealand evidence for the phonology of 19th-century English English. American Speech.

Trudgill, P, Gordon, E. & Lewis, G. 1998. New-dialect formation and Southern Hemisphere English: the New Zealand short front vowels. Journal of Sociolinguistics 2: 35-51.

Watson, C., Maclagan, M., & Harrington J. 1998. Acoustic evidence for vowel change in New Zealand English, presented at Laboratory Phonology VI, York, July. Published version available at: /http:/www.shlrc.mq.edu.au/~watson/labphon/lab.nze.html

Appendix A: Word List

1. hit hid hint
2. boot booed boo tune dune (i.e. sand-dune)
3. bird curt burn
4. bat bad back bag ban
5. bet bed beck beg Ben
6. but bud buck bug bun
7. bark barn path laugh dance
8. bought bored born bore
9. book good put
10. beat bead beak bean
11. loud lout how cow town

12. tie tied tight pie pine
13. hay bay bait paid pay pain take
14. moat mow mowed moan
15. beer bear here hair ear air
16. spear spare shear share cheer chair
17. hid had hard hoard who'd hood head heard heed hud hod
18. groan grown moan mown throne thrown
19. weather which whether witch when wine while whine
20. ten shed add yes end bed
21. doll dole dull
22. school full wool will pool well
23. fill filling fall falling fool fooling four
24. milk child railway cold
25. ferry fairy herring hearing
26. city letter fatter ladder scatter better batter Peter
27. tour pour sure sewer skewer cure poor
28. street train tree dream
29. mother father nothing something
30. think thin with toothbrush breathe clothe beneath
31. milk silk sulk gold
32. Ellen Alan

Appendix B: Transcription guidelines

Type the **interviewer's utterances in bold;** the interviewee's in plain type.

Start each major utterance on a new line, but not small feedback responses like 'mmm' or 'yeah'. These can simply be inserted where appropriate.

Break the transcript up into more easily readable chunks by inserting blank lines after each pair of utterances, or when the interviewee changes topic in a lengthy turn.

Use no capital letters except for proper nouns and 'I'.

Use a minimum of conventional punctuation:

> Use question marks, especially if the grammatical structure does not indicate a question, but the intonation does, as in the second utterance here:
>
> do you have a cat? you have a cat?
>
> Don't use commas or fullstops to indicate clauses and sentences.
>
> Instead, use the following conventions to indicate pauses:
>
> | | | |
> |---|---|---|
> | fullstop . | | = very short hesitation |
> | dash | - | = hesitation |
> | two dashes | -- | = long hesitation |

Use these fillers where appropriate:

mmm um er [for schwa] ahh

Use conventional spelling most of the time, but where necessary use colloquialisms such as 'yeah' 'gonna' 'cos' 'gotta' 'dunno' etc. if this is what was said.

Where you can't decipher, type *[unclear]* or ***[unclear]*** , depending on who is speaking. Use this method of brackets and italics when you wish to record a comment, such as *[another person enters the room]* or *[laugh]*.

When speech is overlapping, indicate this with italics.

Finally, don't tidy up the speech. Leave in the repetitions, fillers and errors.

Appendix C: List of projects which make use of data in the Canterbury Corpus.

1994 - Stephanie Davis — Standard / Non Standard English.
1994 – M J Burden — The /ʊə/~/ɔ/ distinction. An analysis of the word 'tour'. Is there a variation in pronunciation. If so, is there a class/age/sex distinction?
1994 - Amanda Chapman — /f/ and /v/.
1994 - Gabrielle Tavener — NZ English /oʊn/ : /oʊən/.
1995 - Gillian Galbraith — /oʊn/ : /oʊən/.
1995 - Sarah Macann — /aʊ/.
1995 - Andrea Martin — Grown part ii - the continuing saga of the /groʊn/ evolution.
1995 - Janine Romano — /ɔ/ and /uə/ and /ʊə/
1996 - Andrea Benfell — Study of the palatalisation of coronal consonants occurring before the GOOSE vowel.
1996 - Angela Sumner — The occurrence of High Rising Terminals in New Zealand English
1997 - Rachel Rowlands — An examination of tense usage in professional and non professional speakers of New Zealand English
1998 - Melissa Kennedy — Semantic and syntactic variation of *real/really*.

Appendix D: Papers which make use of data in the Canterbury Corpus.

Gordon, E. 1998. The origins of New Zealand speech: the limits of recovering historical information from written records. English World-Wide 19(1): 61-85.

Gordon, E. in press 1999. Embryonic variants in New Zealand English sound changes. Proceedings of the Sixth NZ Language and Society Conference, Victoria University of Wellington, June 1998. To be published in a special edition of Te Reo.

Gordon E., & Maclagan, M. A. 1995. Making a virtue of necessity: Combining teaching and research in the teaching of New Zealand English. New Zealand English Newsletter, 9: 27-31

Gordon, E. & Maclagan, M. A. 1995. The Changing sound of New Zealand English. The New Zealand Speech-Language Therapists' Journal. 50: 32-40

Gordon, E. & Trudgill, P. in press, 1999. Shades of things to come: Embryonic variants in New Zealand English sound changes. English World-Wide. 20.

Maclagan, M. A. in press, 1999. Women and language change in NZE: the case forconsidering individual as well as group data. Proceedings of the Sixth NZ Language and Society Conference, Victoria University of Wellington, June 1998. To be published in a special edition of Te Reo

Maclagan, M. A. 1998. Diphthongisation of /e/ in NZE: a change that went nowhere? New Zealand English Journal, 12: 43-54

Maclagan, M. A. & Gordon, E. 1996. Women's role in sound change: the case of two New Zealand closing diphthongs. New Zealand English Journal, 10: 5-9

Maclagan, M. A. & Gordon, E. 1996. Out of the AIR and into the EAR: Another view of the New Zealand diphthong merger. Language Variation and Change. 8:125-147

Maclagan, M. A. & Gordon, E. 1998. How GROWN grew from one syllable to two. Australian Journal of Linguistics, 18: 5-28

Maclagan, M. A., Gordon, E. & Lewis, G. 1998. Women and sound change: conservative and innovative behaviour by the same speakers. Language Variation and Change 10:19-41

Trudgill, P, Gordon, E. & Lewis, G. 1998. New-dialect formation and Southern Hemisphere English: the New Zealand short front vowels. Journal of Sociolinguistics 2: 35-51