Te Reo 37 (1994):21-28

Introducing the Wellington Corpus of Written New Zealand English

Laurie Bauer
Victoria University of Wellington

I am pleased to be able to announce the publication of the Wellington Corpus of Written New Zealand English.[1] The corpus (along with the manual describing it) is now available either from us at Victoria University or through the International Computer Archive of Modern English (ICAME) at the University of Bergen. In this paper, I shall give some brief details of the background to the corpus (most of which have already been described in Bauer 1993), discuss briefly the New Zealand characteristics of the corpus, and give a few simplistic examples of points which can be extrapolated from it.

1. Background

Many New Zealand linguists will recall Derek Davy's plenary session at the 6th New Zealand Linguistics Conference in Wellington in 1985, where he urged that a corpus of New Zealand English should be established. A number of us from Victoria took his proposals seriously enough to start thinking about creating a New Zealand corpus to parallel such overseas corpora as the Brown, the LOB and — perhaps most importantly — the Macquarie corpora.[2] In fact, we were more ambitious than that, in that we set out to collect not one corpus of New Zealand English, but two: one written, one spoken. I took nominal control of the written corpus, Janet Holmes took nominal control of the spoken corpus. The spoken corpus is still making excellent progress, but it is the written corpus which has been completed first — mainly because it is easier to collect written material. Although I say I took nominal control of this project, I should like to acknowledge the help and support of Wellington linguists such as Allan Bell, Janet Holmes, Graeme Kennedy, and Chris Lane in carrying it out, not to mention the host of research assistants who did most of the actual work and whose names are listed in the preface to the manual. I should

---

[1]This paper was first presented to the 10th New Zealand Linguistics Conference at the University of Waikato in August 1993.

[2]For the Brown corpus, see Kucera and Francis 1967; for the Lancaster-Oslo-Bergen (LOB) corpus, see Johansson *et al.* 1978; for the Macquarie corpus see, e.g., Green and Peters 1991. The Freiburg-LOB corpus referred to later in this paper is under development by Prof Dr Christian Mair in Freiburg-im-Breisgau. Sections A, B and C were made available to me by the Freiburg team, and I should like to thank them for this.

also like to acknowledge the funding agencies, Victoria University's Internal Grants Committee and the NZVCC University Grants Committee.

The idea with the corpus of written English was to make it as much like the LOB corpus as possible. There were a number of reasons for this. Firstly, one of the reasons for having a New Zealand corpus was to allow comparison with British and American forms of English, and the Brown and the LOB corpora made this possible if we followed their format. Secondly, when we began we knew that the Macquarie corpus was being constructed according to the same pattern, and comparison with Australian English seemed even more relevant for New Zealand than comparison with varieties from further afield. Because the Australian corpus is based on the year 1986, the New Zealand one is as well, although for some categories, especially fiction, it was necessary to go beyond 1986 to collect sufficient texts. This does have the disadvantage, of course, that the New Zealand corpus presents English from 25 years later than that in the Brown and LOB corpora. The New Zealand corpus has the same number of texts as the LOB corpus, and except for the fiction, they are in the same categories. Where fiction was concerned, we could not match the Brown and LOB categories (categories such as westerns and detective fiction, for instance) and we simply made a single undifferentiated category of fiction.

Further comparisons with other varieties of English in a rather different selection of categories will be made possible with the International Corpus of English (ICE) being organised by Sidney Greenbaum at University College, London. At Victoria University we are currently working on the New Zealand contribution to that corpus, funded by the Foundation for Research into Science and Technology.

2. Problems

Collecting a corpus of this type is fraught with difficulties. I shall only discuss a few of the most important ones here — problems which may help others doing this kind of work or explain certain features of the corpus.

2.1 Conformity to common standards

When a project of this type runs over several years and employs as many different workers as this one did, it becomes extremely difficult to ensure that everyone conforms to all of the same standards, even where these are written out. In most instances we solved this problem by having everything proof-read by two or more people, by discussing problems that arose in a plenary forum with all the workers employed at a given time, and by attempting to keep written records of what was decided on any given point. On the whole I believe this worked. I note however that line-lengths still differ by a few characters in different parts of the corpus. It would not surprise me to learn that there are other inconsistencies despite our best efforts. Also, when I use the corpus I still find typographical errors in it, though I hope there are not too many of these.

2.2 What it means to be New Zealand

It was obvious to us right from the beginning that we would not be able to guarantee that every part of every text had been written by people who were born and bred in New Zealand. In the case of press reports, for instance, we simply did not have access to information on who the original writers were, let alone on where those people came from. We took the view, however, that this was not crucial; we thought that if the piece had been written and published in New Zealand, this in itself would be some guarantee of the 'New Zealand-ness' of the piece. In a few places (especially in Section J, learned writing) we accepted articles which had been written by New Zealanders but published overseas, on the grounds that this was the norm for such material, but on the whole we tried to make the New Zealand origin of material criterial. For press stories, for instance, we did not collect the international news syndicated by Reuters or UPI, but took local news and NZPA reports. Although we avoided some authors we knew to be of overseas origin (Phil Mann, Colleen Reilly) there were far more that we did not know about. All this ignores the fact that there are, in any case, major problems in deciding who should count as a New Zealander for linguistic purposes; where accent is concerned, a New Zealand passport is not enough (see Bauer 1991).

Inevitably we missed some. In the last few weeks of the work on the corpus, we discovered that some of our authors were of overseas origin or that a particular piece of work, written by a New Zealander, had been heavily edited by an editor of non-New Zealand origin. Here we were able to find replacement texts, but there may still be other cases we do not know about. While we would not wish to claim that such texts 'do not matter', in that they do not affect the status of the sample as New Zealand writing, we would stress that such pieces are generally invisible against the background of other New Zealand writing. To collect a 'pure' New Zealand sample — as Janet Holmes has discovered with the spoken corpus material — involves a huge amount of extra work, and it is not clear that the extra time and money could be justified.

2.3 Copyright

All the material in the corpus is copyright to someone else. Collecting permission from authors and publishers to put material in the corpus was probably the major problem we faced. Even though many authors were delighted and even flattered that their work had been selected to be part of the sample of New Zealand English, and even though only two refusals were received for the hundreds of texts involved, it was simply enormously time-consuming to track down people and to get them to sign appropriate releases when we found them. In some cases where we could not find any trace of the copyright holders at all, we had, in the end, to give up. For instance, there were writers' workshops whose members disbanded immediately after the publication of the works they had produced. There

were authors for whom we could not find a New Zealand address and whose publishers did not have addresses for them either (even, in one case, when they had another work by the same author in press!). For others attempting similar work, I cannot say strongly enough that the only way to cope with this is to over-collect material from every category, and then use the material for which you first manage to get copyright clearance. We are now doing this with the ICE corpus material, and even then it is not always straightforward to get permission.

3. Some results

Simply collecting the material for the corpus has been such a major project for seven years or so, that at times I have lost sight of the aim of all of this, namely to be able to describe New Zealand English better at the end of it all. Accordingly, very little work actually using the corpus has been done yet, although Sigley (1993) indicates that it has great potential. Below, I present a few rather superficial, but nonetheless interesting, findings from the corpus.

3.1 How do you write a date?

We are all aware that there are conflicting ways of writing dates. Quirk *et al.* (1985: 396) imply that the order *14(th) May(,) 1993* is British, while the order *May 14(th)(,) 1993* is American, corresponding to the order of numerals in dates written as *7/2/93*, which can mean 7th February (British) or 2nd July (American). Matters are not that simple, however. In New Zealand the string *7/2/93* would normally mean 7th February (except where computer programs confuse us), but there is variation in the way in which the date is written when it is not simply given as a string of figures.

First of all, let us consider what the corpus shows where no year is given, but simply the date and the month. There are three possible axes of variation here. Firstly, do you write the word *the* as in *the first of May*? The answer is, not usually. Only four such examples occur in the corpus, three of them in the order given above, one in the order *May the first*. If you use this order you can spell out the date in words (one example) but need not (three examples). I shall ignore these four examples in my subsequent figures.

The second possible axis of variation is whether you actually include the *th* (or *st, nd, rd*) after the date. The answer again is, not generally. Only six such examples occur in the corpus, three with the order *14th May*, three with the order *May 14th*. I shall also ignore these in subsequent figures.

The major axis of variation is whether the date precedes or follows the month, that is, whether we write *14 May* or *May 14*. At first glance, the answer seems fairly clear-cut.

Date format in the entire corpus	
14 May	May 14
51	159

That is, New Zealand English appears to have adopted the supposed American way of doing these things. However, a closer analysis reveals that these figures have been skewed by the way the newspapers do things. If we look at some selected sections of the corpus, we find a rather different pattern emerging:

Date format in selected sections of the corpus		
Section	14 May	May 14
Press (A, B, C)	4	130
Religion (D)	4	2
Skills and Hobbies (E)	5	17
Popular Lore (F)	9	3
Biography, Belles Lettres (G)	14	7
Miscellaneous (H)	13	0
Totals of these categories	49	159

In other words, the vast majority of the citations using the *May 14* format come from the press, and outside the press, the *14 May* format is rather more common. Another conclusion which is worth entertaining is that date format in written texts depends more on editorial policy than on general usage.

Now let us consider what happens when a year is included in the date as well as the date and the month. Again, there is variation as to whether *th* is used or not (usually it is not: only ten dates in this category use the *th*, and they are fairly evenly spread between the *14th May* (4) and the *May 14th* (6) formats). There is extra variation possible here, in that there may or may not be a comma before the year. The general rule is *14 May 1993* with no comma, but *May 14, 1993* with a comma. Only seven examples with no *th* break this general rule. Again, however, the break-up between the various sections of the corpus is instructive:

Date format including year in different sections of the corpus		
Section	14 May 1993	May 14, 1993
Press (A, B, C)	0	27
Religion (D)	2	0
Skills and Hobbies (E)	15	2
Popular Lore (F)	13	1
Biography and Belles Lettres (G)	32	0
Miscellaneous (H)	19	0
Learned and Scientific (J)	18	1
Fiction (K, L)	2	2
Totals	101	33

Here, although the press sample is internally consistent, it is much clearer that it is out of line with what is happening elsewhere in the publishing community.

3.2 Different from/to/than

As in the LOB corpus, but not as in the Brown corpus, the New Zealand corpus shows people preferring *different to* to *different than*. The corresponding figures from the three corpora and from the press section only of the Freiburg-LOB corpus of British English based round 1986 are presented in the table (which shows, incidentally, that very few tokens of *different* are followed by any of these particles).

Particle used with *different*: comparison of four corpora				
Corpus	total tokens of *different*	from	to	than
Wlgtn	367	47	8	2
LOB	367	34	7	1
Brown	281	39	0	6
FLOB	38	4	3	0

There is nothing here that is particularly surprising.

3.3 Transitivity

There are several verbs which differ in transitivity in different varieties of English (see Bauer 1987). Of these, *protest* is considered here as being one for which some data is available. First note that the noun *protest* always takes a preposition such as *against* or *about*, since *the protest of* is followed by the person protesting rather than the thing protested against. Nonetheless, the choice of preposition may be of some interest.

Prepositions following the noun *protest* in different corpora

Corpus	against	about	at	over
Wlgtn	4	1	1	0
LOB	5	1	0	1
Brown	6	0	0	1
FLOB	0	1	1	1

There is a suggestion of diachronic change in British English here which would be worth following up with a larger corpus. However, the major point of comparison is what happens with the verb *protest*. The results for this are presented in the next table. Here it can be seen that New Zealand English provides a mid stage between British and American English, but that there is some sign of a change in British English since the time of the material in the LOB corpus.

Transitive and intransitive uses of the verb *protest* in different corpora

Corpus	trans.	against	about	at	over
Wlgtn	4	2	2	1	0
LOB	0	9	0	0	0
Brown	5	3	0	0	0
FLOB	1	2	2	0	0

3.4 Typographical errors

The research assistants who entered and proof-read the articles listed approximately 675 typographical errors in the text. While this number may sound high, when one considers that it is spread over 111,000 lines of text, much of which was written and set at speed, it seems a very modest number. These typographical errors range from missed commas, omitted letters or wrong capital letters to neologisms such as *secuirty* for *security*. Of these 675, however, 115 (or about one sixth) involved the misuse of apostrophes. Interestingly, very few of these errors occurred in the Press sections, which implies that professional writers can be trained to avoid such errors. The three most common categories were the use of a plural noun for a plural possessive noun (40 instances, including one *childrens*), the use of a plural for a singular possessive (23 instances) and the use of *its* for *it's* (19 instances, as opposed to 5 in the other direction). Here is an area in which prescriptivists can have a field day!

Although most of the typographical errors are clearly no more than that, there are some of them which may be interpreted as being rather more indicative of current trends. For example, *woman* appears for *women* twice, even in written material of the kind sampled here (in lines A36 and K39 of the corpus), and there is a single instance (K95) of *bought* for *brought*, and a single instance (G50) of *a* for *an*. The last examples might not be significant, but given that the same features can be heard regularly

in current spoken New Zealand English, they look as though they are at least worthy of comment.

4. Conclusion

The main focus of this report is to introduce the Wellington Corpus of Written New Zealand English, and announce its general availability. I have presented some fairly superficial comments on things that can be discovered using the corpus, but I hope that these will merely spur people on to use it to discover weightier things. The Department of Linguistics at Victoria University of Wellington would appreciate receiving copies of publications deriving from work using this new corpus. We foresee fairly rapid developments in the study of New Zealand English using this corpus, and would like to be able to keep other researchers informed about what work has been done.

References

Bauer, Laurie. 1987. 'Approaching New Zealand English Grammar', *New Zealand English Newsletter*, 1, 12-15.

---------- 1991. 'Who speaks New Zealand English?' *ICE Newsletter* 11.

---------- 1993. 'Progress with a corpus of New Zealand English and early results'. *Corpus-Based Computational Linguistics*, ed. by Clive Souter & Eric Atwell, 1-10. Amsterdam & Atlanta: Rodopi.

Green, Elizabeth and Pam Peters. 1991. 'The Australian Corpus project and Australian English', *ICAME Journal*, 15, 37-53.

Johansson, Stig, Geoffrey N. Leech and H. Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English for Use with Digital Computers.* Oslo: Department of English, University of Oslo.

Kucera, Henry and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English.* Providence RI: Brown University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London and New York: Longman.

Sigley, Robert. 1993. 'Relative clauses in New Zealand Fiction'. Paper presented at the 10th New Zealand Linguistics Conference, University of Waikato, August 1993 and unpublished terms paper, Victoria University of Wellington.