Introducing a New Zealand newspaper corpus

John Macalister

Victoria University of Wellington

Electronic corpora are a valuable tool in language study. Corpus-based study allows us to make statements about what is probable within a language, rather than merely possible. It also provides us with empirical data against which to test our impressions and observations. In recent years, researchers into New Zealand English have been in the fortunate position of having two computer-readable corpora to draw on, the *Wellington Corpus of Written New Zealand English* (Bauer, 1993) and the *Wellington Corpus of Spoken New Zealand English* (Holmes et al., 1998). These two one-million-word corpora present a comprehensive sample of New Zealand English of the early 1990s. However, as language is forever changing, there will always be scope for the creation of further corpora, whether as a representation of a language variety or to capture the characteristics of a particular domain of usage.

The purpose of this paper, therefore, is to introduce a recently-created corpus of New Zealand English, drawn from newspapers published in the year 2000. The corpus was originally designed as part of the database[1] for the last of six indicator years[2] in a study of lexical change in written New Zealand English from 1850 to the present.

1. Issues in creating the corpus

1.1 Why newspapers?

The first question that needs to be dealt with is: why should newspapers be expected to represent the language used in 2000, or, indeed, at any particular time? Bell (1991: 3–4) presents several reasons for studying media language, among which is the simple fact of its being there, and elsewhere points out that 'media usage influences and represents people's use of and attitudes towards language in a speech community' (Garrett & Bell, 1998: 3). Newspapers, in other words, speak to their readers in their readers' language. If they did not, they would be unlikely to prove commercially viable, and would not long survive. While, as Fowler (1991: 40) warns, '[i]t is *not* being claimed that the newspaper copies the language which its readership does actually use in private life' (original emphasis), newspapers can, however, be expected to provide a robust representation of the language in use at a given time.

[1] The remainder of the database for 2000 was contributed by the *School Journals* and by the *New Zealand Parliamentary Debates*.

[2] The other five indicator years were 1850, 1880, 1910, 1940, and 1970.

1.2 How many newspapers?

One of the potential dangers in using media language is that of being overwhelmed by the sheer quantity of available data. Just one newspaper, published six days a week for 52 weeks, would produce an enormous number of words. Fortunately, however, it is not necessary to take the universe of publication. A representative sample of the universe can be obtained by creating a 'constructed week' (Jones & Carter, 1959), a procedure established by Stempel (1952), who examined the representativeness of different-sized samples of issues of a newspaper, with samples being 'selected by making a random start and taking every Nth issue for the entire year. The size of the gap was chosen so that the day of the week moved up one with each choice' (1952: 333).

Using this approach, four separate constructed weeks were assembled, one for each of four daily morning newspapers, each with a different, randomly selected, starting date.

1.3 Which newspapers?

One of the key issues that Kennedy (1998: 60) identifies concerning the validity and reliability of corpus-based research is 'whether that corpus can serve the purposes for which it was intended'. This corpus drawn from New Zealand newspapers is intended to represent written New Zealand English, rather than, say, the Wellington variety of New Zealand English. There needed, therefore, to be some balance between metropolitan and provincial newspapers, and between North and South Island newspapers. A further consideration, for this corpus was created as part of a study in diachronic variation in the New Zealand English lexicon, was that the newspapers selected should have continuity of publication since at least 1880. In the event, a constructed week was created for each of the following daily newspapers, each with a long history of publication:

- the *New Zealand Herald*
- *The Dominion* [3]
- the *Otago Daily Times*
- the *Wanganui Chronicle*

1.4 How much of the newspapers?

Newspapers do, of course, consist of a variety of material (for a full discussion see Bell, 1991: 12–15, from which much of the following is derived). The two dominant genres are news and advertising. In the advertising genre, language is generally used either creatively and unusually or formulaically. For this reason, advertising language was excluded from this corpus, and the focus was on the news genre.

---

[3] The history of newspapers in Wellington is slightly more complicated than that of the other main centres. The Wellington Independent ran from 1845 – 1874, when it was purchased by the New Zealand Times, which paper continued the Independent's serial numbering. The New Zealand Times co-existed with The Dominion, which was established in 1907, for 20 years until being incorporated into the latter in 1927.

The news genre can be sub-divided into three broad categories:

- service information i.e. lists – sports results, tv programmes, share prices, weather forecasts, and so on
- opinion i.e. editorials, letters to the editor, columns, reviews
- news

For the purposes of this research, the sub-category of service information was excluded in its entirety, whereas the opinion and news sub-categories were included.

News can be further sub-divided into four categories:

- hard news - local/national/international
- feature articles
- special topic news, such as sports, business, information technology but not politics
- headlines, bylines, captions.

All four of these categories were included in the corpus, and were assigned to one of seven sub-genres (see Section 2 below).

One qualification was applied to the inclusion of the content in the corpus. As it is intended to represent written New Zealand English, a working definition of New Zealand English, as it appears in the print media, needed to be applied. This is a problem which has vexed other corpus compilers at different times. For example, Holmes (1995: 6–7) discusses the stringent criteria that were established for defining a New Zealander in creating the *Wellington Corpus of Spoken New Zealand English*. Similarly, while Bauer (1993) does not explicitly address this question, and the *Wellington Corpus of Written New Zealand English* does contain a handful of texts that almost certainly originated from abroad (see, e.g., Texts A07, E21 (1993: 28, 60)), such texts are very few and it is clear that most overseas-sourced/edited stories were excluded as not representing written New Zealand English.

With newspapers, it is necessary to recognise that the news stories that appear are derived from a number of sources locally, nationally and internationally, and that different newspapers may treat the same story differently. The differences may be as simple as omitting a paragraph or two because of space limitations, or as subtle as rephrasing. There has been considerable discussion of these and other, related, phenomena in the field of media discourse (Bell, 1991; Fowler, 1991; Fairclough, 1995; Bell & Garrett, 1998). However, for the purposes of this corpus New Zealand English was defined as any piece of news or opinion that was written in New Zealand about New Zealand. This automatically excluded almost all international news, and all feature articles sourced from overseas, where distinctly New Zealand words were not expected to be found. Thus, for example, reports from Fiji about a coup attempt were excluded, but news stories about the New Zealand government's response to that crisis were not.

Care was taken to exclude advertising features and advertorial content, many of which have the appearance of news but belong in the advertising genre. Fortunately, most such content is clearly signalled within a newspaper.

2. Corpus description

The corpus that resulted from applying this approach totals 819,495 word tokens and consists of 28 computer files. Seven sub-genres have been created within the corpus, the sub-genres being:

- New Zealand News (national, regional, local)
- Politics
- Business
- Sport
- Features and Columns (including lifestyle pages)
- Editorial (including letters to the editor)
- Reviews (films, books, television, &c)

Thus, it is possible to access the corpus in three ways: in its entirety as a representation of written New Zealand English, by regional sub-corpus, or by specific sub-genre.

The regional composition is shown in Table 1 and the sub-genre composition in Table 2.

Table 1: Regional Composition of the New Zealand Newspaper Corpus

	NZH	WC	DOM	ODT	
Tokens	252,922	124,829	212,657	229,087	**819,495**

The smallest contribution to the corpus was from the smallest population centre, Wanganui, with broadly equal contributions from Auckland, Wellington and Dunedin.

Table 2: Sub-genre Composition of the New Zealand Newspaper Corpus

	NZN	Pol	Bus	Spo	FC	Ed	Rev	
Tokens	298,328	30,449	122,518	144,706	131,897	48,530	43,067	**819,495**

There are, therefore, four major sub-genres (New Zealand News, Business, Sport, Features and Columns) and three minor sub-genres (Politics, Editorial, Reviews).

3. A demonstration of the Corpus's potential

A corpus of this size is going to be of most obvious use in researching high frequency lexical items. To yield useful information about low frequency types would require far larger corpora, along the lines of the British National Corpus.

The primary motivation for the creation of this corpus was to yield insights into the presence and uses of Maori words in current New Zealand English (see Macalister, 2001, for a discussion of the use of *kiwi*, for example). To illustrate this potential, an examination of the presence of *Maori* (including the plural form *Maoris* and the hybrid *Maoridom*) was carried out by region and by sub-genre, as shown in Tables 3 and 4. *Maori* was selected for this demonstration as it is the highest frequency Maori word generally found in New Zealand English (Kennedy & Yamazaki, 1999). Its presence is likely to provide an insight into the visibility of Maori issues and the audibility of Maori voices in the media.

Table 3: Distribution of Maori *by Region*

Newspaper	No. of Tokens	% of Types
NZH	134	0.05
WC	164	0.13
DOM	167	0.08
ODT	94	0.04

A moment's reflection would suggest that the ratio of Maori to non-Maori in the population centres would be mirrored by the relative frequency of *Maori* in the regional sub-corpora. As can be seen in Table 3, the type *Maori* is most likely to be found in the Wanganui media, least likely in Dunedin, which would appear to accord with expectations based on demographic factors. However, it is slightly surprising to note that the Wellington newspaper has a higher overall occurrence than the Auckland paper. An explanation for this may be found in the fact that political discourse may be a more significant feature in the capital city than elsewhere. Certainly the occurrence of *Maori* is most frequent in the Politics sub-genre, as Table 4 illustrates.

Table 4: Distribution of Maori *by Sub-genre*

Sub-genre	No. of Tokens	% of Types
Politics	109	0.36
Editorial	118	0.24
New Zealand News	188	0.06
Features, Columns	90	0.07
Reviews	20	0.05
Sport	21	0.015
Business	13	0.01

A further analysis of the data shows that, while *Maori* occurs at a roughly similar rate in the Auckland, Wellington and Dunedin newspapers' Politics sub-genres, the size of that sub-genre in the *Dominion* is between two and three times that of the sub-genre in either the *New Zealand Herald* or the *Otago Daily Times*, as shown in Table 5.

Table 5: Relative size of Politics sub-genres

	Size of Politics sub-genre	**% of Regional corpus**
NZH	5,291 tokens	2.09
DOM	12,901 tokens	6.06
ODT	6,563 tokens	2.86

This is a corpus that reflects 'the size and shape of the documents from which it is drawn' (Sinclair, 1991: 19, who also discusses the advantages of such an approach). The explanation for the higher frequency of *Maori* in the *Dominion* is, therefore, to be found in its greater concentration on political news, as befits the political centre of the country, rather than in demography.

The overall conclusion to be drawn from this brief foray into the New Zealand Newspaper Corpus is that, based on the presence of the type *Maori*, visibility and audibility for Maori people are likely to be greatest in Wanganui, and least in Dunedin, and greater in Wellington than in Auckland. Further investigation of the corpus, for example of the distribution of Maori personal names, could be undertaken to test this finding.

It is also worth commenting that the type *Maori* is most likely to be found in the two sub-genres (Politics, Editorial) that are likely to reflect contention and controversy. One interpretation of this linguistic observation may be that it echoes the extent to which the place of Maori within contemporary New Zealand society remains a focus of debate and attention.

4. Conclusion

The hope is that researchers will find this corpus a useful addition to the pool of computer-readable data available for the study of New Zealand English. A small demonstration has been provided to illustrate the potential for lexical analysis and the insights that can be gained.

The editors of the four newspapers that contributed to the corpus have kindly allowed the data to be used for the purposes of linguistic research. A copy of the New Zealand Newspaper Corpus may, therefore, be obtained by contacting the writer at the New Zealand Dictionary Centre, Victoria University of Wellington, P. O. Box 600, Wellington.

Bibliography

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

Bell, Allan 1991. The Language of News Media. Oxford: Blackwell.

Fairclough, Norman 1995. Media Discourse. London: Edward Arnold.

Fowler, Roger 1991. Language in the News: Discourse and Ideology in the Press. London: Routledge.

Garrett, Peter & Allan Bell 1998. Media and Discourse: A Critical Overview. In Allan Bell and Peter Garrett (eds) Approaches to Media Discourse. Oxford: Blackwell: 1-20.

Holmes, Janet 1995. The Wellington Corpus of Spoken New Zealand English: A Progress Report. New Zealand English Newsletter 9: 5–8.

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

Jones, Robert L. & Roy E. Carter Jr., 1959. Some procedures for estimating 'news hole' in content cnalysis. Public Opinion Quarterly 23: 399–403.

Kennedy, Graeme 1998. An Introduction to Corpus Linguistics. London: Longman.

Kennedy, Graeme & Shunji Yamazaki 1999. The Influence of Maori on the New Zealand English Lexicon. In John M. Kirk (ed), Corpora Galore: Analyses and Techniques in Describing English. Amsterdam: Rodopi: 33–44.

Macalister, John 2001. The Transformation of the Kiwi. English in Aotearoa 43: 20-22.

Sinclair, John 1991. Corpus Concordance Collocation. Oxford: Oxford University Press.

Stempel, Guido. H. III 1952. Sample size for classifying subject matter in dailies. Journalism Quarterly 29: 333–334.