TOWARDS A CORPUS OF EARLY WRITTEN NEW ZEALAND ENGLISH – NEWS FROM *EREWHON*?

Marianne Hundt: Englisches Seminar, Universität Zürich, Plattenstrasse 47, CH-8043 Zürich, Switzerland, <m.hundt@es.uzh.ch>

Abstract

The history of New Zealand English is well attested. Previous studies focussed on the development of the New Zealand accent (Gordon et al. 2004) and are based on spoken data. Written data would enable linguists to study the emergence of standard New Zealand English (NZE) and differential change in this variety vis à vis British and American English. The present article discusses the requirements that such a diachronic corpus of written NZE should meet and presents a case study on the use of the progressive. The data from the Corpus of Early New Zealand English (CENZE) show that the frequency with which the progressive is currently used in NZE is a very recent development that is unlikely to be attributable to influence from Irish English (IrE) during the colonial period.

1. Introduction

Erewhon is the title of a novel by Samuel Butler that was published in 1872. Reading the title of this novel backwards provides a clue for the literary genre. Moreover, on the basis of Butler's biography and the descriptions of landscapes, this utopia has been localised in the south island of New Zealand. However, *Erewhon* turns out not to be the sought-after Atlantis but a place that

allows Butler to project some shortcomings of Victorian England. The hero of the novel, a farmer called Higgs, therefore leaves *Erewhon* disillusioned in a hot air balloon. The connection between Butler's novel and the potential corpus of Early New Zealand English (NZE) is that it forms part of a collection of early New Zealand texts that were digitized at the *New Zealand Electronic Text Centre* at Victoria University of Wellington (VUW). I learnt about their efforts to digitize texts during a stay as a visiting professor at VUW and contacted the head of the electronic Text Centre, Alison Stevenson, who made their texts available to me. Additional material was obtained from the National Library (namely the *Proceedings der Royal Philosophical Society of New Zealand*) and from the world-wide-web (mainly newspaper texts and an early letter collection).

Obviously, a collection of digitized texts is not automatically a corpus. As Biber et al. (1998:4) point out, a corpus is a "[…] large and principled collection of natural texts." How one might get from a collection of digitized texts to a representative corpus of early New Zealand English and why such a corpus might be useful for linguists will be the topic of this article.

Apart from the utopian (or rather dystopian) tenor, Butler's novel also makes use of satirical elements, so I might be permitted to quote Fillmore's (1992: 35) caricature of two extremist approaches to the study of language at some length:

> ARMCHAIR linguist:
> He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations.)
>
> CORPUS linguist:
> He has all of the primary facts he needs, in the form of a corpus, of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations.)

The lesson to be learnt from this satirical description is that a corpus should always be the answer to a linguistic query, that is a means to an end and not an end in itself. So what are potential research questions that a corpus of early New Zealand texts might enable us to answer? One area of research might be differential language change, i.e. the development of NZE vis à vis other national varieties of English such as British (BrE) and American English (AmE).[1] In this case, our corpus of early New Zealand texts would have to be compiled in a way that would make a comparison with existing historical corpora possible. Obviously, NZE grammar is not categorically distinct from BrE or AmE. What makes NZE grammar distinct is mostly a question of preference for certain grammatical options available in global English. At this level, NZE grammar may actually be rather 'exotic'. One example of the currently exotic state of NZE vis à vis varieties such as Australian English (AusE) or AmE is in the use of the progressive form, e.g. *John* ***is texting*** *a message to his girlfriend with his new mobile*, which is used much more frequently in NZE than in other native varieties of English (see Collins 2009 and Hundt and Vogel 2011). In this paper, I will therefore present a case study on the use of the progressive in what constitutes the nucleus of a corpus of early New Zealand texts.

In part two of this paper, I will briefly comment on previous research on the diachronic and regional developments of the progressive in English. In section three, I will focus on the steps involved in moving from a text database to a corpus, as well as the challenges and limitations that such a project involves. Part four will present results from a study on progressive constructions in early NZE as well as historical BrE and AmE texts.

2. The progressive – historical and regional developments

The origins of this grammatical construction can be traced back to Old English times, but even in Shakespeare's writing it had not become obligatory (see Polonius' question *What do you read my lord?* (Act II, Scene ii) which is not a question about Hamlet's reading habits). It is only during the nineteenth century that the progressive becomes more frequently used (Strang 1982, Smitterberg 2005, Kranich 2008). The progressive is still spreading in the twentieth century (Mair and Hundt 1995, Smith 2005, Leech et al. 2009), but there is relatively little regional difference between BrE and AmE (Leech et al. 2009: 122). NZE turns out to be quite exotic because New Zealanders use

the progressive much more frequently than people in the UK or the US. Hundt (1998: 75) provides empirical evidence of a regional difference between northern and southern hemisphere varieties; more recently, Collins (2009) has used a subset of the ICE corpora to show that usage in NZE is actually significantly different from AmE and BrE but also from AusE (see Table 1).

Table 1: Progressives across four Englishes (approx. 120,000 words per variety; from Collins 2009: 116)[2]

	NZ	AUS	US	GB
speech	57.7%	71.8%	76%	69.5%
writing	42.3%	28.2%	24%	30.5%
N	894	753	626	660

A particularly interesting finding is that New Zealanders use the progressive – a construction that is typical of colloquial, spoken English – much more frequently in written language than Americans, Britons or Australians. Our study on progressives in student writing shows that New Zealanders actually use the progressive with a similar frequency to some people who have learnt English as a second (ESL) or foreign (EFL) language (see Figure 1).[3]

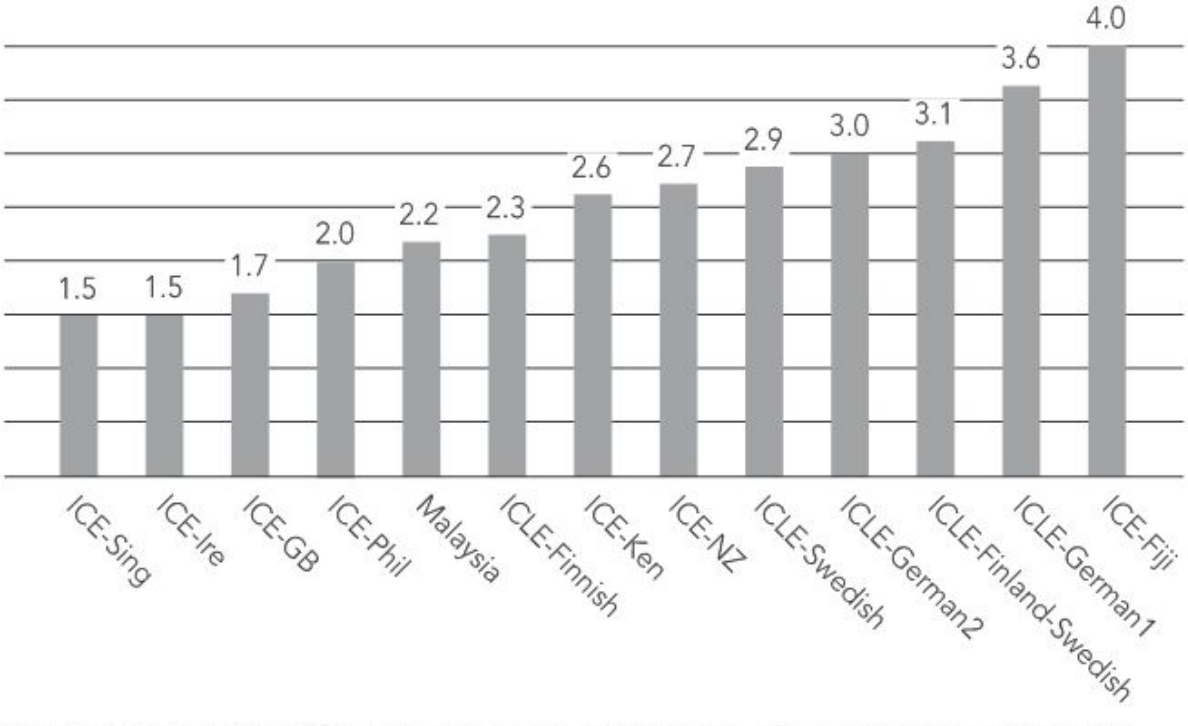

Figure 1: Normalized frequencies (per 1,000 words) of progressives across ENL, ESL and EFL corpora (student writing) – from Hundt and Vogel (2011: 154)

A look at Figure 1 shows that students in New Zealand use the progressive almost as frequently as Swedish-speaking students in Finland and more frequently than students in Kenya. This regional difference is also confirmed for printed academic texts (see Hundt and Vogel, 2011: 155). Incidentally, students from Ireland use the progressive with a lower frequency than that attested in essays collected for the British component of the ICE corpus.

It is important to note in this context that the similarities between NZE, on the one hand, and ESL and EFL usage, on the other hand, mainly concern the frequency with which the progressive is used. Less obvious are structural similarities in the use of the construction, such as the combination of a progressive with stative verbs like *be* or *love*; these are considered typical of non-native speaker usage, but are also occasionally found in BrE or AmE, for instance.[4] The following examples illustrate instances in second-language varieties of English in Kenya, Singapore, Fiji or the Philippines where writers of BrE or AmE are more like to use a simple form (emphasis added throughout):

1. This essay *will be discussing* six factors why women have to work for empowerment. (ICE-Fiji w1a-015)
2. It spread due to movement of laborers. It *is being used* now in Zambia as a language of education. (ICE-Ken w1a-003)
3. Whereas in the 2nd article, it says that the economy *is fast rising* ever since the Ramos Administration started. (ICE-Phil w1a-011)
4. However, according to Hume, there is not guarantee that just because nature *has been uniformly functioning* in the past, it will continue to do so always. (ICE-Sing w1a-014)

Example (3) is particularly interesting because the present progressive is used in a context in which we would expect to see the present perfect or a present perfect progressive in BrE. Interestingly, it bears a striking resemblance to a chance finding from NZE: While holidaying in New Zealand, I came across a notice in the shared bathroom facilities on a camping ground which asked parents to accompany young children to showers and toilets. The reason given was "*We are experiencing too many accidents of late.*" From the perspective of a speaker of BrE, this sentence is unusual because it combines a present progressive with an adverbial that normally combines with the present perfect (most likely a simple present perfect).[5] However, there was no obvious non-

native influence on this sign and, furthermore, the native speakers of NZE that were asked to comment on the sentence did not find this usage unusual. It is possible that some native speakers of English may extend the progressive to contexts of perfective marking because the past progressive is also used in a similar way, namely as a marker of recent past: "*Tom, you were just telling me that in all you had nine students going down there*" (COCA:CNN_Morning, 1997; quoted from Bergs and Pfaff 2009). Moreover, Fraser Gupta (2006: 104f.) found that the progressive is occasionally used by inner-circle speakers (mainly in the US) following expressions such as *This is the first time I ...* . In other words, the extended use of the progressive in New Zealand English (both in terms of its frequency and some of its functions) fits with Gachelin's (1997: 43) claim that the extended use of the progressive in New Englishes may eventually lead to long-term change in the English language as a whole: "Its generalization […] may herald what will be World English usage in the next century."

The grammaticalisation and spread of the progressive construction from Old English onwards are well documented (see, e.g., Denison 1993: 371–410). The historical details are not relevant to our discussion here. There is one aspect, however, that is worthwhile mentioning, namely the possible influence of language contact with speakers of Gaelic (see e.g. Keller 1925 or Filppula and Klemola 2012). Gaelic has a periphrastic construction combining the verb *be*, a preposition and a verbal noun that is used to refer to ongoing events, including the possibility of combining with some stative verbs like *living* (Ronan, 2001: 50). In other words, possible influence from Gaelic would open up a wider functional range of the progressive construction in a contact variety of English such as IrE. Language contact and/or dialect contact might thus also have played a role in the spread of the progressive construction in NZE. There is language contact with non-native speakers of English in the colonial and post-colonial context. But it is also possible that contact between speakers of different regional varieties of English might be the reason why the progressive is used so frequently in present-day NZE. The most likely sources of regional dialectal influence in the development of NZE would be Scottish and Irish English (see Bauer 1994, 1997 or Gordon et al. 2004).[6] McCafferty and Moreno (2010, ms.) have investigated, among other things, the use of the progressive in a diachronic corpus of IrE letters that provides valuable comparative data.

A diachronic corpus of early New Zealand writing would allow us to verify whether the frequent use of the present progressive in current NZE is the result

of language contact with (a) non-native speakers of English or (b) speakers of other regional dialects of English like IrE. We would also be able to show whether the progressive was used more frequently in New Zealand than in BrE from the early colonial days or whether the prolific use of the construction in NZE is a more recent development. The two aspects are connected in so far as an early (colonial or immediately post-colonial) dating of the phenomenon would speak for influence from regional dialects whereas recent spread is more likely to have been supported by contact with speakers of English as a second language.

The question is what a diachronic corpus should look like that might allow us to test these hypotheses. Holmes (1994: 27) described the ideal scenario for a study of recent change in New Zealand English. The same requirements also apply to the earlier periods of NZE in comparison with its 'parent' variety, BrE, or other relevant corpora of English as a first language:

> The ideal situation […] would appear to be to use two corpora constructed on parallel principles at […] different points in time. Assuming that any variation identified can be reasonably attributed to language change over time, rather than to, say, topic differences or stylistic differences between the corpora […]. Unfortunately, no such parallel corpora exist for New Zealand English.

In the following, I will discuss how we might build such a parallel corpus from existing digitized texts.

3. From electronic text collection to corpus

3.1 Existing diachronic corpora

As pointed out in the introduction, a corpus is not simply a collection of texts but one that has been based on sampling principles. For a corpus of early written New Zealand English, sampling with criteria that will make the corpus comparable with existing diachronic corpora of reference varieties like British, American and Australian English is advisable because this will minimise 'cost'. Suitable diachronic corpora that sample these varieties are COOEE (a *Corpus of OZ Early English*) and ARCHER (*A Representative Corpus of Historical English Registers*).[7]

COOEE contains texts from the years 1788–1900, including both speech-like texts and private letters alongside more formal text types such as official announcements by the government. However, the corpus is not publicly

available and can therefore not be used for comparative research. But it provides important methodological input for the compilation of a corpus of early New Zealand writing, as we will see.

ARCHER is a diachronic corpus of British and American texts from the middle of the seventeenth to the end of the twentieth century. The corpus is divided into sub-corpora of fifty-year periods. It provides comparative data on the two varieties for a number of speech-like and written registers (drama, fiction, medical, scientific and legal writing, newspapers, journals and diaries, private letters, sermons). Individual samples consist of approximately 2,000 words (sometimes comprising more than one text, for instance in the category 'newspapers'). The target for each text category, register and sub-period is a total of 10 samples (i.e. approximately 20,000 words). ARCHER is currently the best available corpus for comparative studies on differential change in varieties of English in the late Modern period. It thus provides a very suitable sampling frame for a corpus of early New Zealand writing.

3.2 A corpus of Early New Zealand writing: challenges

Even with a suitable, ready-made sampling frame, the compilation of a corpus of early New Zealand texts is far from straightforward. The main challenges are to (a) determine the criteria for including a text in the corpus and (b) to cope with the available spread of text types, and (c) to evaluate whether the diachronic cuts that the ARCHER sampling frame provides also provide helpful sub-samples for a diachronic corpus of early New Zealand writing. These questions will be addressed in the following sections.

3.2.1 Who qualifies as a New Zealander?

Corpus compilation in the colonial and early post-colonial context faces the problem that it is not easy to determine when an immigrant becomes a New Zealander and thus eligible to be considered as an author whose texts should be included in the corpus. Even the compilers of the spoken corpus of New Zealand English collected in the 1990s asked themselves this question:

> Who should be allowed to contribute to the corpus? [...] It is a particularly vexatious problem for colonial societies where large sections of the community are immigrants. At what point does an immigrant become a New Zealander? (Holmes, Vine and Johnson 1998: 23)

Bauer (1991, unpaginated) speculates on the (socio-)linguistic processes that may have affected the language of immigrants, and that grammar, in

particular, is likely to have remained relatively stable even after a lengthy stay in the colony:

> Britons (or Australians or Americans ...) arriving in New Zealand may consciously or unconsciously adapt their speech to use particular vocabulary items, but they are unlikely to be even subconsciously aware of the statistical trends in the usage of particular grammatical patterns. We must therefore predict that they are unlikely to make appropriate changes to these aspects of their speech, even after lengthy residence. Now, it might be that this supposition is false, and that they do adapt appropriately after sufficient length of time.

This is a speculation and the only way of settling the matter would be longitudinal data on the development of the grammar of individual immigrants, i.e. the kind of evidence that we are unlikely ever to be able to collect for previous periods. Conscious or unconscious adoption of grammatical features is not that unlikely to occur, though.[8] Rissanen (1984: 418f.) argues that the language of people who migrated to America even after having received their education in Britain is a good source for the study of Early American English: "The people producing these texts [from the 1640s, M.H.] had spent their youth and acquired their education in England, but they had lived in America for a number of years […]." Likewise, anyone who migrated to New Zealand as a child or young adult would be a good informant for an emerging variety of NZE.

In addition to migration to New Zealand, New Zealand-born authors might also leave the country and spend time in another English-speaking country and thus adopt (grammatical) features from a different regional variety of English. The criteria that were applied in the compilation of the spoken and written corpora of New Zealand English in the 1990s considered both possibilities (immigration to and temporal emigration from New Zealand): only speakers who had been resident in New Zealand at age 10, who had spent less than ten years outside of the country, and who had returned at least a year before the text to be included in the corpus was produced. For several reasons, such strict selection criteria are difficult if not impossible to apply in the collection of a corpus of early written New Zealand texts. Why this is the case can be illustrated with the biographies of some authors that were included in the text database compiled by the *New Zealand Electronic Text Centre*. I will briefly summarize some biographical facts and then comment on their relevance to corpus compilation.

- Hon. James Coutts Crawford (1817–1889) was born in Scotland, arrived in New Zealand (via Australia) in 1939 at the age of 22 but returned to England twice between 1841 and 1857. He died 1889 in London.
- Walter Buller (1838–1906) was born in Hokianga as the son of a missionary and is thus, by birth, a true New Zealander; but he travelled to Europe, too, in 1870. He gave a paper to the *Philosophical Society* before his journey, though, which makes this particular text a clear candidate for inclusion in the corpus.
- Edwin Fairburn (1827–1911), New Zealand-born and the son of early immigrants, like Crawford and Buller, also travelled to Europe. But even though we know that he went to Germany and Austria we do not have information on how long he stayed there.
- Richard Treacy Henry (1845–1929) was born in Ireland. Aged 6, he migrated to Australia in 1852 with his parents and thus spent his formative years in the southern hemisphere (if not in New Zealand itself). In 1874, aged 29, he moved from Australia to New Zealand. His biography is typical of some migrants in so far as they did not necessarily arrive directly from the British Isles but sometimes via Australia (see Gordon et al., 2004: 44f.), one fact that has been taken to explain the close historical connection between the two varieties.

For the early colonial period, biographies such as those of Buller and Fairburn are quite rare since most migrants arrived in New Zealand as young adults. Most of the authors included in the text database of the *New Zealand Electronic Text Centre* did not spend their lives exclusively in the colony, even after they had arrived there. This is a complication that Fritz (2007: 65f.) also faced in the compilation of COOEE; he concluded that

> AusE developed […] from the dialects and sociolects the immigrants spoke and wrote. Therefore all English texts in early Australia are valid sources. None is inherently better than the other.

But there are also clear criteria for excluding certain authors. Samuel Butler (1835–1902), the author of *Erewhon*, is one of them. He was born in England and only spent five years of his life in New Zealand (between 1859 and 1864), working on a sheep farm near Christchurch; he published a couple of articles in a local newspaper, among them one entitled 'Darwin Among the Machines'

(1863). *Erewhon*, however, was only published on his return to England. It is probably the descriptions of landscapes in the novel that are so obviously related to his stay in New Zealand that lead to the novel being included in the database of the *New Zealand Electronic Text Centre* (but in a later edition, namely from 1927). The fact that Butler spent five years of his life in New Zealand is not enough to qualify him as an author of emerging New Zealand English in the colonial period.

Another potentially problematic case is Katherine Mansfield (1888–1923). She was born in New Zealand, spent her formative years in the country and also received most of her schooling in Wellington. Between 1902 and 1906 she attended school in London and returned to New Zealand for a short while afterwards; she died of tuberculosis in France aged only 35. The texts that were included in the corpus of early New Zealand writing were written in Europe, but the fact that she was born in New Zealand and spent her childhood and early youth there make her a New Zealand author. Fritz would have excluded her as an eligible source as he only included texts that were produced in Australia, New Zealand or on Norfolk Island in COOEE (2007: 66).

There are a few additional complications that do not allow us to apply the same strict criteria in the collection of corpora of early colonial and post-colonial writing as we would apply in the compilation of a corpus of current English. First of all, the names of authors for individual texts are not necessarily known (e.g. in the case of newspaper articles that are published without the author's name). But even if the name of the author is known we do not necessarily have any biographical background information. A lot of potential contributors to a corpus of early NZE were simply not well-known or important enough to be included in biographical sources. If we were to include only those authors where biographical information is available this might even skew the data included in the corpus by giving preference to well-known informants who are likely to be of a relatively high social background.

Second, because a lot of the material comes from published sources, we can never rule out some editorial influence and thus the editor's linguistic background as an additional layer in the text. This is an aspect that Bauer (1991, unpaginated) has also pointed out as a potential source of non-authentic language use even for corpora of written post-colonial NZE: "[...] it is impossible to avoid speakers who are not technically speakers of New Zealand English [...] the problem is likely to be greatest in the print media [...]." In a paper that investigates personal letters from an edited collection (Hundt, forthcoming) I was able to demonstrate that such editorial influences

are more likely to affect aspects of orthography but largely seem to leave morpho-syntactic variables unaffected.

3.2.2 Availability of text types

The readily available digital early New Zealand texts obviously do not perfectly match the sampling frame of the ARCHER corpus. And even when there are texts for a particular text category, there is not necessarily enough material to fill a sample (ten times 2,000 words) that would match the ARCHER framework. At other times, the available material allows for sampling at 30-year intervals within or across two ARCHER sub-periods. An overview of the number of words in a first version of the CENZE corpus is given in Table 2.

Table 2: Availability of early New Zealand texts according to the ARCHER sampling frame

	1800–49	1850–99	1900–49	1950–99
Drama	—	—	—	—
Fiction	—	✓	✓	—
Medical	—	—	—	—
Scientific	—	✓	✓	✓
Legal	—	—	—	—
Newspapers	✓	✓	✓	✓
journals & diaries	—	—	—	—
private letters	✓	✓	✓	—
Sermons	?	?	?	?

New Zealand only became a crown colony in 1840, and it is therefore not surprising that, with the exception of letters from emigrants and newspapers, no material is available for the first half of the nineteenth century. The large gaps in the second half of the twentieth century is due to copyright restrictions: the *New Zealand Electronic Text Centre* mostly digitized texts that are not subject to copyright restrictions.[9]

The sampling for private letters beyond those by early settlers in the 1840s is somewhat problematic, too, since the letter collections included in the material of the *New Zealand Electronic Text Centre* are from two authors only (one for the 1860s, the other for the 1920s). As pointed out previously,

some of the earliest texts are letters by emigrants to New Zealand that were published soon after New Zealand had become a crown colony to advertise the new colony to potential settlers in GB; this material was digitized by the University of Auckland and is publicly available on the internet (for a more detailed discussion of these data, see Hundt forthcoming).[10]

The text category 'sermons' (religious writing) turned out to be problematic for a different reason. ARCHER samples mostly sermons (i.e. persuasive texts) for this register. The material digitized by the *New Zealand Electronic Text Centre* consists of texts that describe the mythology of the Maori, exclusively, and are therefore not suitable as a parallel source of texts for the CENZE.

Historical newspapers were not digitized by the *New Zealand Electronic Text Centre* but by the national Library of New Zealand (for the years 1839–1920).[11] Over eight million individual articles can be downloaded from the library's webpage. However, the texts were OCRed[12] but not manually post-edited. This means that for each article to be included in the corpus, the facsimile of the original print version has to be consulted and the texts have to be corrected manually before they can be included in the corpus. Finally, narrative prose texts had to be supplemented by additional material beyond that available from the *New Zealand Electronic Text Centre.*

3.2.3 Which sub-periods and how many?

In terms of diachrony, the ARCHER sampling frame uses 50-year periods. Individual samples are spread more or less evenly[13] across this time span resulting in a continuous coverage of the material. However, it might be easier to demonstrate diachronic change if the sampling keeps to more narrowly defined sampling points. In previous research on recent grammatical change, sampling points at approximately 30-year intervals have proven useful as this roughly corresponds to the distance between two generations of speakers (see Leech et al., 2009 and Hundt and Leech, 2012). For registers that are well attested early on, it might even make sense to sample at 20-year intervals. One problem, as we will see, is that not enough material is available to fill the text categories of the ARCHER sampling frame in whatever chronological grid is adopted: with nine registers and ten samples of about 2,000 words each, every diachronic sample would require 180.000 words worth of text – a somewhat ambitious goal.

3.3 A first Corpus of Early New Zealand English (CENZE)

With all the limitations discussed in this section, what will a corpus of early New Zealand texts look like? Table 3 shows that we are still far from the goal of 180.000 words per diachronic sample.

Table 3: Number of words in CENZE corpus according to the ARCHER sampling frame (registers rather than diachronic cuts)

	1800–49	1850–99		1900–49		1950–99	TOTAL
drama	—	—		—		—	—
fiction	—	20.969		20.855		—	41.824
medical	—	—		—		—	—
scientific writing	—	*1870s*		*1900s*	*1930s*	*1960s*	
		20.266		14.390	20.776	20.429	75.861
legal texts	—	—		—		—	—
newspapers	*1840s*	*1860s*	*1880s*	*1920s*	*1940s*		
	20.180	20.437	20.372	21.215	20.401	—	102.606
journals & diaries	—	—		—		—	—
private letters	*1840s*	*1860s*		*1920s*			
	20.364	20.790		20.709		—	61.863
sermons	?	?		?		?	—
Total	40.544	102.835		118.346		20.429	282.154

The registers with the best diachronic coverage are personal letters and newspaper texts. Newspaper texts are available from 1839 onwards – initially from the *New Zealand Gazette and Wellington Spectator*, only. For the very early colonial years, we will probably not be able to move beyond what is currently available: the settlers had other immediate concerns than to write novels or scientific treatises soon after their arrival in the new country. And even though they were likely to have gone to church, archiving early sermons in those days was not a priority, either. Later gaps in the coverage of the ARCHER registers (e.g. fiction from the first half of the twentieth century) are more likely to be filled. In some cases, the representativeness of the texts is not particularly good (see the problems discussed in relation to private letters in the 1860s and 1920s discussed in section 3.2.2). All in all, the first CENZE is a bit of a patchwork affair. Nonetheless, it can fruitfully be used to monitor

the development of some grammatical patterns. In this paper, I will use the progressive as a case study.[14]

4. Case study: The progressive in CENZE

In order to allow for comparability with previous ARCHER-based studies (Hundt 2004a, 2004b), the same criteria were applied for the definition of the linguistic variable. Using WordSmith Tools, I searched for combinations of the auxiliary *be* with a present participle (allowing for material to occur between auxiliary and participle). In a second step, all non-progressives were manually removed from the concordances, including instances where the participle has adjectival rather than verbal function (e.g. *This news is shocking* or *His countenance was repulsive and forbidding*) and examples with participles that function as an apposition rather than as part of the verb phrase (e.g. *He was at home, repairing the roof*) (see Hundt 2004a: 56). Similarly, patterns where *be* was a copula followed by a gerund were excluded manually from the concordance (e.g. *Consequently what is called keeping the length of arc constant is really allowing it to become slightly longer than the desired length*, […] ARCHER 1925angu.s7b). Instances with two participles (e.g. *A deadly bark beetle is attacking and killing many hickories*, ARCHER 1932FeltS7a) were only counted once. As in Hundt (2004a), instances of *going to* as a future time expression were excluded from the datasets.

Table 4 gives the results, both in terms of absolute frequencies and normalized (per 10,000 words). Normalization is necessary to enable comparison across the differently sized sub-corpora and to facilitate comparison with previous research. The data from the CENZE corpus have been supplemented with searches in the written part of the Wellington Corpus of NZE for the second half of the twentieth century to obtain data for the last sub-period sampled in ARCHER.

Not surprisingly, the progressive occurs with different normalized frequencies in different registers. It is most frequent in private letters (a text type that was found to have relatively high frequency of other colloquial patterns in previous studies, see e.g. Smitterberg, 2005: 77f.). As far as diachronic developments are concerned, however, the letters data might not be a reliable indicator because the material from the 1860s and 1920s are not representative samples (one author only in each of the sub-periods).

Influence from IrE in the letters is unlikely if we compare the results

Table 4: Progressives in CENZE (normalized frequencies per 10,000 words in brackets; figures in square brackets give normalized frequencies from the Wellington corpus)

	1800–49	1850–99		1900–49		1950–99
fiction	—	54 (25.8)		82 (29.3)		[53]
scientific writing	—	*1870s* 8 (3.9)		*1900s* 23 (16.0)	*1930s* 4 (1.9)	*1960s* 8 (3.9)
newspapers	1840s 18 (8.9)	1860s 35 (17.1)	1880s 43 (21.1)	1920s 57 (26.9)	1940s 41 (20.1)	[43.2]
private letters	1840s 88 (43.2)	1860s 45 (21.6)		1920s 106 (51.2)		—
Total	106 (26.1)	185 (18.0)		313 (26.4)		—

from CENZE with those from McCafferty and Moreno (2010, ms.). In the 1830s letters in *CORIECOR* (Corpus of Irish English Correspondence), the progressive occurs with a frequency of only 41.8 occurrences per 10,000 words. Moreover, MacCafferty and Moreno include instances of *be going to* (Kevin McCafferty, p.c.) which were excluded from my counts. In other words, the progressive is used more frequently in the early New Zealand letters in the 1840s than in a contemporaneous collection of IrE letters. It is also quite frequent in newspapers and fictional writing. The register with the lowest occurrence of progressives is the most formal one represented in CENZE, namely scientific writing.

The two data points available from fictional writing do give evidence of an increase of progressives across time. In the fiction sample from ICE-NZ that Collins (2009: 116) analysed, progressives are used with a frequency of 122 per 10,000 words and thus significantly more often than in the first half of the twentieth century. More reliably, the newspaper evidence shows that progressives become more frequent in New Zealand English between the early colonial days and the first half of the twentieth century, even though there is a decrease between the 1920s and 1940s.

How does the development of the progressive in NZE compare with its spread in BrE and AmE? We will look at the two text types with the best diachronic coverage in CENZE, science and newspaper reportage.

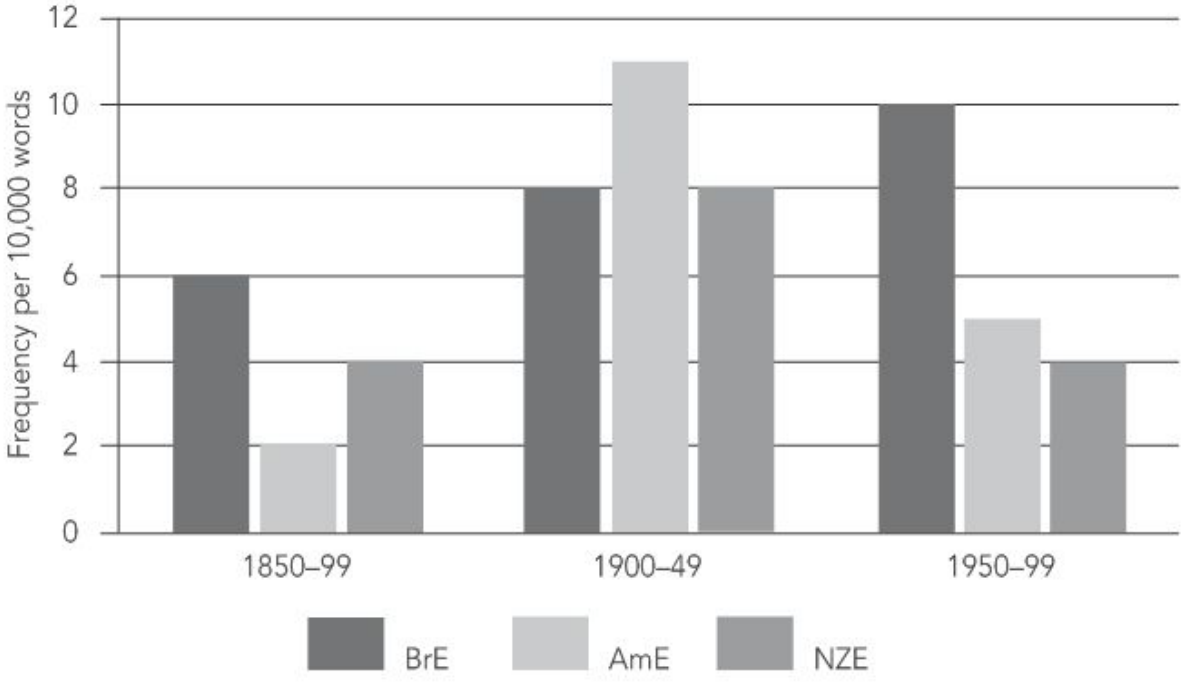

Figure 2: Progressives in the science sub-corpus – ARCHER vs. CENZE

In scientific writing, the text frequency of progressives is rather low, so the figures have to be interpreted rather cautiously. Nevertheless, we see an increase over time in the BrE part of ARCHER on the one hand, whereas in AmE and NZE, the peak in the first half of the twentieth century is followed by a decline. With the overall text frequency of progressives in scientific writing being so low, there is an obvious risk of individual samples having a skewing effect on the results. This seems to be the case for the New Zealand texts from the beginning of the twentieth century: most progressives are found in only two out of the ten samples. One of these samples is from a text written by the Irishman Richard Treacy Henry, whose parents had migrated to New Zealand via Australia (see 3.2.1). The author of the other text with a higher-than-average frequency of progressives is an Englishman who arrived in New Zealand aged 26. In other words, only one of the two authors has an Irish background, thus making language contact with a variety of IrE as the sole reason for a frequent use of the progressive rather unlikely. The significantly higher frequency of progressives in New Zealand academic writing that Collins (2009: 116) and Hundt and Vogel (2011: 154f.) observe must thus be a recent development. Further evidence from this comes from a comparison of the ARCHER and CENZE data with evidence from the ICE corpora: ARCHER samples scientific writing from 1975 (BrE) and 1954–1997 (AmE), CENZE from the 1960s; the ICE corpora, on the other hand, are comprised

of material that was collected from the 1990s onwards. This diachronic bias does not seem to play a role for BrE, with the (natural) science sub-samples in ARCHER and ICE-GB yielding comparable normalized frequencies of progressives at 10 and 7 occurrences per 10,000 words, respectively. The difference between the CENZE and ICE-NZ data, on the other hand, shows that the progressive has increased significantly in New Zealand academic writing towards the end of the twentieth century: the science texts in CENZE yield 4 progressives per 10,000 words, whereas those in ICE-NZ yield 31 progressives per 10,000 words.

Let us now turn to the diachronic development in a text type where progressives are used more frequently: newspaper texts. Figure 3 plots the diachronic developments in ARCHER and CENZE. Even though the sub-periods in CENZE are different from those in ARCHER, the overall diachronic trend becomes clear: Early New Zealand newspapers have a comparable relative frequency of progressives as we find in the newspaper texts included in ARCHER.

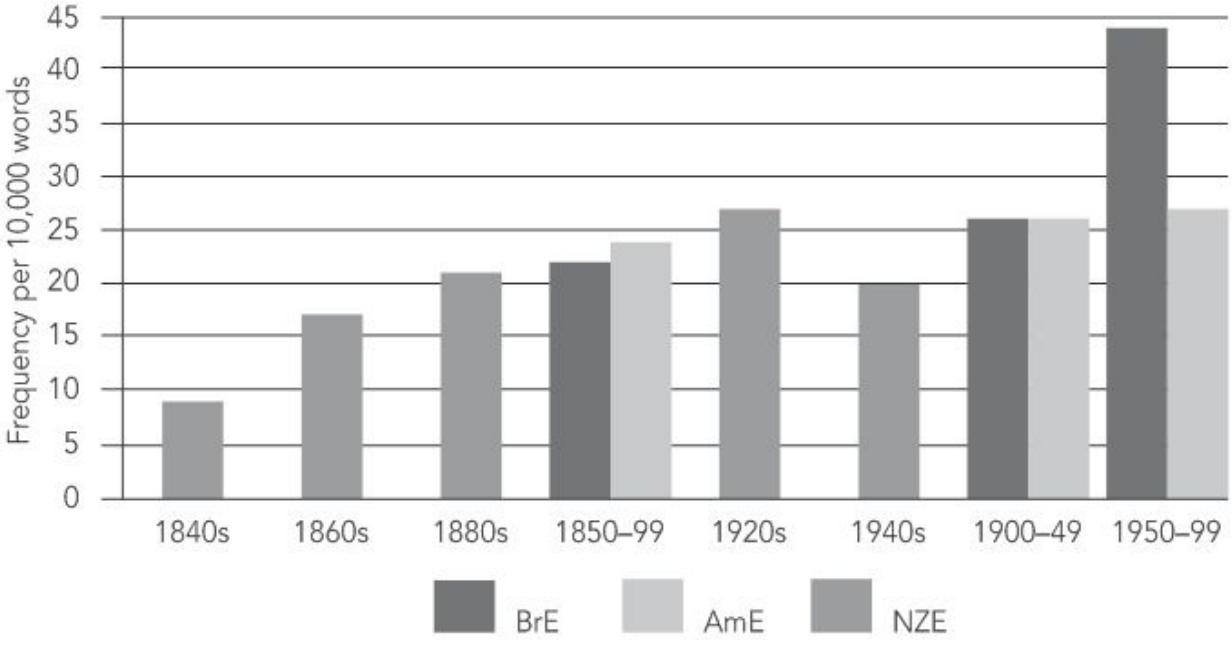

Figure 3: Progressives in newspaper writing – ARCHER and CENZE

Figure 4 compares late twentieth-century newspaper data from ARCHER with evidence from the corresponding sections in the Brown family of corpora and the Wellington Corpus of Written New Zealand English, which was compiled from texts published in the late 1980s (see Hundt, 1998: 75f.). These results, together with those from CENZE, again suggest that the frequent use of the

progressive in NZE is more likely to be due to recent change rather than an earlier predilection of New Zealanders to use the progressive.

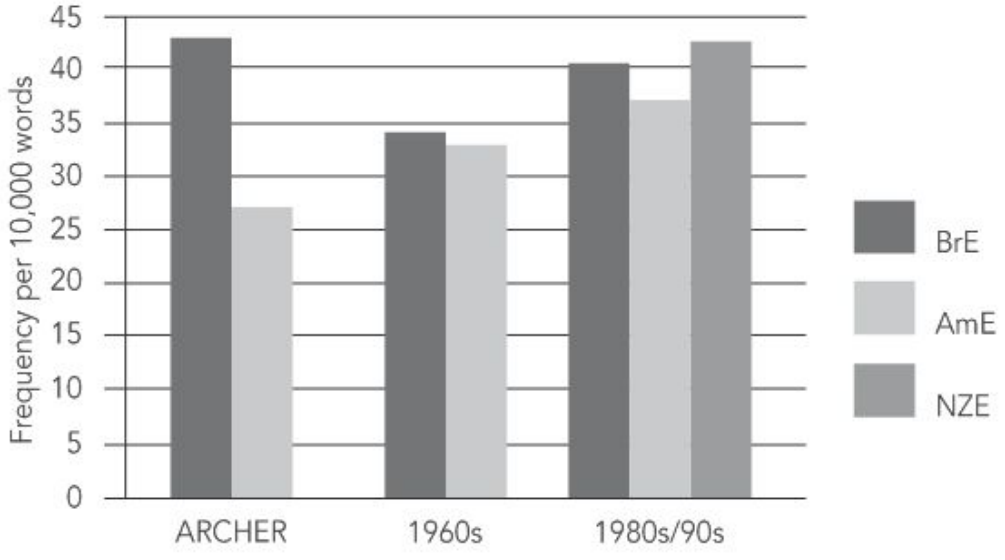

Figure 4: Progressives in 20th-century newspaper writing: ARCHER and the Brown family

Figure 4 also shows that we have to be cautious when we compare results from different corpora. ARCHER samples only reportage whereas the Brown family includes reportage, editorials and reviews. A sub-sample from FLOB comparable in size and composition to the ARCHER texts (national newspapers rather than provincial; a cross-section of different kinds of news) yields 42.2 progressives per 10,000 words and thus a slightly higher normalized frequency than the overall sample (40.4 progressives/10,000 words), which includes editorials and reviews. The topic also appears to play a role, with society news containing more progressives than political news or sports reportage. Furthermore, a sub-sample of 1990s British provincial newspaper reportage yields a much higher normalized frequency of progressives at 57.6 per 10,000 words. Thus, the composition of samples is particularly important for small diachronic corpora that comprise only about 20,000 words per register and period. The press section of the Brown-type corpora with a total of 88 samples and approximately 176,000 words may thus produce somewhat more robust results than the newspaper texts in ARCHER.

To sum up, the more frequent use of the progressive in current NZE is a recent development that is likely to date to the second half of the twentieth century. Dialect contact with IrE during the early days of the development of NZE is an unlikely source of the more frequent use of the progressive in New

Zealand today, both in terms of the diachronic developments as well as the evidence from individual authors in this small-scale study.

5. Conclusion and outlook

Despite the availability of digitized texts from the early colonial period and later stages in the history of New Zealand, compiling a corpus of early New Zealand writing is not as easy and straightforward a task as one would hope. The question is whether our brief visit to the *Erewhon* of historical corpus linguistics in New Zealand has discouraged us to the extent that we simply want to board that hot air balloon and leave. Contrary to the fears of the inhabitants of Butler's *Erewhon*, technical evolution has not lead to the development of machines that think and act for themselves. The Cyborgs of computational linguistics are still not even a remote possibility on our horizon. There is still a lot of manual labour involved in the compilation of historical corpora. The germ of a historical corpus of early New Zealand texts described in this article could be developed into a more representative corpus with additional data. For the category of letters, this would probably mean the inclusion of handwritten documents that are hopefully to be found in some archives. But even though the texts included in my embryonic corpus of early New Zealand writing do not yet amount to a representative sample of the emerging written variety in colonial and post-colonial New Zealand, what is available so far can be used to test hypotheses on relatively frequent grammatical patterns, such as the progressive. The case study has also shown that results from relatively small sub-samples have to be treated with particular caution, especially if findings from different corpora are compared. The evidence from CENZE suggests that the progressive was not used significantly more frequently in early New Zealand writing than in comparable texts from Britain and the US. Instead, data from the Brown-family of corpora and components of the International Corpus of English indicate that New Zealanders seem to have moved ahead of other ENL speakers and writers in the use of the progressive quite recently.

Notes

1 See Hundt (2009a) on differential change in BrE and AmE.
2 The differences are prove significant at $p \leq 0.001$ in a chi-square test. Note that Hundt (1998: 75), using newspaper data from the Brown family of corpora, only, did not find a significant difference between Australian and New Zealand English. Both are ahead of British and American English in the growing use of the progressive in that study, indicating that text type is an important factor to consider in the study of progressives.
3 Note that not all ESL varieties use the progressive more frequently than it occurs in BrE: SingE, for instance, has an even lower incidence of progressives.
4 For its use in an advertising campaign, based on a Timberlake song, see http://en.wikipedia.org/wiki/McDonald's_advertising (accessed 15th February 2010).
5 For similar use of the progressive in IrE, see McCafferty and Moreno (2010, unpaginated).
6 Note that Gordon et al. (2004) investigate the development of the New Zealand accent rather than developments in the grammar of the variety.
7 For COOEE, see Fritz (2007). Background information on ARCHER can be found in Biber, Finegan and Atkinson (1994). For information on the different versions of ARCHER, see Yáñez Bouza (2011). The comparative data used in this paper come from material to be included in the forthcoming version of the corpus (ARCHER 3.2), which provides broader coverage of AmE than previous versions of the corpus did. Information on COOEE and ARCHER is also available from http://www.helsinki.fi/varieng/CoRD.
8 For second dialect acquisition (accent), see Tagliamonte and Molfenter (2007). Some principles described in this article might also apply to the acquisition of grammatical preferences in a new or evolving dialect.
9 Narrative prose from the second half of the twentieth century was not digitized by the *New Zealand Electronic Text Centre* for reasons of copyright. This gap in the corpus could be filled relatively easily because these texts are available either in print or as samples in existing corpora, such as the Wellington Corpus of Written New Zealand English or the New Zealand component of the ICE corpus.
10 The letters were digitized by the Early New Zealand Books project at the University of Auckland, New Zealand. They can be found at http://www.enzb.auckland.ac.nz/document//1843_-_Letters_from_Settlers_and_Labouring_Emigrants (last accessed 17.01.2011).
11 http://paperspast.natlib.govt.nz/cgi-bin/paperspast
12 OCR stands for 'optical character recognition' and thus is shorthand for 'automatic digitization of text'.
13 Occasionally, sampling for an individual register diverges from this sampling principle: scientific British texts in the twentieth century, for example, stem from the years 1925 and 1975 (for the two sub-periods) only.
14 Hundt and Szmrecsanyi (2012) use the same corpus to investigate animacy as a determinant of grammatical variation in NZE vis à vis BrE and AmE. In

Hundt (forthcoming), I focus on a broader range of potentially non-standard constructions in early New Zealand letters (the focus there is on the 1840s material, only). Hundt (in preparation), finally, investigates the use of relativizers in restrictive vs. non-restrictive relative clauses in the science part of the corpus.

References

Bauer, Laurie. 1991. 'Who speaks New Zealand English?' *ICE Newsletter* 11. (unpaginated)

Bauer, Laurie. 1994. 'English in New Zealand.' In Robert W. Burchfield (ed). *The Cambridge History of the English Language. Volume V: English in Britain and Overseas. Origins and Developments*. Cambridge: Cambridge University Press. 382–429.

Bauer, Laurie. 1997. 'Attempting to trace Scottish English influence on New Zealand English.' In Edgar W. Schneider (ed). *Englishes Around the World: Studies in Honour of Manfred Görlach. Volume 2. Caribbean, Africa, Australasia*. Amsterdam and Philadelphia: Benjamins. 257–272.

Bergs, Alexander and Maike Pfaff. 2009. 'I was just reading this article. Is the perfect of the recent past on its way out?' Paper presented at the SEU Symposium *Current Change in the English Verb Phrase*, 14 July 2009.

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. 'ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers.' In Udo Fries, Gunnel Tottie and Peter Schneider (eds). *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research and Computerized Corpora. Zürich 1993*. Amsterdam: Rodopi. 1–13.

Biber, Douglas, Susand Conrad and Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Collins, Peter. 2009. 'The progressive in English.' In Pam Peters, Peter Collins and Adam Smith (eds). *Comparative Studies in Australian and New Zealand English Grammar*. Amsterdam: Benjamins. 115–123.

Fillmore, Charles J. 1992. '"Corpus linguistics" or "computer-aided armchair linguistics"?' In Jan Svartvik (ed). *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82, Stockholm 4–8 August 1991*. Berlin and New York: Mouton de Gruyter. 35–60.

Filppula, Markku and Juhani Klemola. 2012. 'Celtic and Celtic Englishes.' Bergs and Laurel J. Brinton (eds). *English Historical Linguistics. An International Handbook. (HSK 34-2)*. Berlin: Mouton de Gruyter. 1687–1703.

Fritz, Clemens W.A. 2007. *From English in Australia to Australian English. 1788–1900*. Frankfurt: Peter Lang.

Gachelin, J.M. 1997. 'The progressive and habitual aspects in non-standard Englishes.' In Edgar W. Schneider (ed). *Englishes Around the World 1. General Studies, British Isles, North America. Studies in Honor of Manfred Görlach.* Amsterdam and Philadelphia: Benjamins. 33–46.

Gordon, Elizabeth, Lyle Campbell, Jennifer Hay, Margaret Maclagan, Andrea Sudbury and Peter Trudgill. 2004. *New Zealand English. Its Origins and Evolution*. Cambridge: Cambridge University Press.

Holmes, Janet, Bernadette Vine and Gary Johnson. 1998. *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: Victoria University.

Holmes, Janet. 1994. 'Inferring language change from computer corpora: Some methodological problems.' *ICAME Journal* 18: 27–40.

Hundt, Marianne. 1998. *New Zealand English Grammar. Fact or Fiction?* Amsterdam and Philadelphia: John Benjamins.

Hundt, Marianne. 2004a. 'The passival and the progressive passive – a case study of layering in the English aspect and voice systems.' In Hans Lindquist and Christian Mair (eds). *Corpus Approaches to Grammaticalisation in English*. Amsterdam and Philadelphia: Benjamins. 79–120.

Hundt, Marianne. 2004b. 'Animacy, agency and the spread of the progressive in eighteenth- and nineteenth-century English.' *English Language and Linguistics* 8(1): 47–69.

Hundt, Marianne. 2009a. 'Colonial lag, colonial innovation, or simply language change?' In Günter Rohdenburg and Julia Schlüter (eds). *One Language, Two Grammars: Morphosyntactic Differences between British and American English*. Cambridge: CUP. 13–37.

Hundt, Marianne. 2009b. 'Global feature – local norms? A case study on the progressive passive.' In Lucia Siebers and Thomas Hoffmann (eds). *World Englishes – Problems, Properties and Prospects*. Amsterdam and Philadelphia: Benjamins. 287–308.

Hundt, Marianne. 2012. 'Varieties of English: Australian/New Zealand English.' In Alexander Bergs and Laurel J. Brinton (eds). *English Historical Linguistics. An International Handbook*. (HSK 34-2). Berlin: Mouton de Gruyter. 1995–2012.

Hundt, Marianne. Forthcoming. 2013. 'Heterogeneity vs. Homogeneity.' To appear in Anita Auer, Dominic Watts and Daniel Schreier (eds). *Letter Writing and Language Change*. Cambridge: Cambridge University Press.

Hundt, Marianne. In preparation. Relatives in scientific English: Variation across time and space. To appear in Winnie Cheng and Franca Poppi (eds). *The Three Waves of Globalization. Proceedings from the 2011 CLAVIER conference*. Amsterdam and Philadelphia: Benjamins.

Hundt, Marianne and Katrin Vogel. 2011. 'Overuse of the progressive in ESL and learner Englishes – fact or fiction?' In Joybrato Mukherjee and Marianne Hundt (eds). *Second-language varieties and learner Englishes*. Amsterdam and Philadelphia: John Benjamins. 145–166.

Hundt, Marianne and Geoffrey Leech. 2012. 'Small is beautiful – on the value of standard reference corpora for observing recent grammatical change.' In T. Nevalainen & E. Traugott (eds.). *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 175–188.

Hundt, Marianne and Benedikt Szmrecsanyi. 2012. 'Animacy in early New Zealand English.' *English World-Wide* 33(3): 241–263.

Keller, W. 1925. 'Keltisches im englischen Verbum.' In *Anglica: Untersuchungen zur Englischen Philologie, Alois Brandl zum Siebzigsten Geburtsgage überreicht*, In *Sprache und Kulturgeschichte*. (Palaestra, 147). Leipzig: Mayer and Müller. 55–66.

Kranich, Svenja. 2008. *The Progressive in Modern English. A Corpus-Based Study of Grammaticalization and Related Changes*. PhD Dissertation, Freie Universität Berlin.

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Changes in Contemporary English: A Grammatical Study*. Cambridge: CUP. (Chapter 6, 'The Progressive' by Nicholas Smith).

Mair, Christian and Marianne Hundt. 1995. 'Why is the progressive becoming more frequent in English? – A corpus-based investigation of language change in progress.' *Zeitschrift für Anglistik und Amerikanistik* 43 (2): 111–122.

McCafferty, Kevin and Carolina P. Amador Moreno. 2010. '*A Corpus of Irish English Correspondence* (CORIECOR). A tool for studying the history and evolution of Irish English'. Unpublished ms.

Ronan, Patricia. 2001. 'Observations on the progressive in Hiberno-English.' In John M. Kirk and Dónall P. Ó Baolill (eds.). *Language Links. The Languages of Scotland and Ireland.* Belfast: Cló Ólscoil na Banriona. 43–58.

Rissanen, Matti. 1984. 'The choice of relative pronouns in seventeenth-century American English.' In Jacek Fisiak (ed). *Historical Syntax*. Berlin: Mouton de Gruyter. 419–35.

Smith, Nicholas. 2005. *A Corpus-Based Investigation of Recent Change in the Use of the Progressive in British English*. Unpublished PhD thesis, Lancaster University.

Smitterberg, Erik. 2005. *The Progressive in 19th-century English: A Process of Integration*. Amsterdam: Rodopi.

Strang, Barbara M.H. 1982. 'Some aspects of the history of the *be + ing* construction.' In J. Anderson (ed). *Language Form and Linguistic Variation: Papers Dedicated to Angus McIntosh*. (Current Issues in Linguistic Theory 15). Amsterdam: John Benjamins. 427–74.

Tagliamonte, Sali A. and Sonja Molfenter. 2007. '"How'd you get that accent?" Acquiring a second dialect of the same language.' *Language in Society* 36: 649–675.

Yáñez Bouza, Nuria. 2011. 'ARCHER past and present (1990–2010).' *ICAME Journal* 35: 205–36.