

Inferring language change from computer corpora: Some methodological problems¹

Janet Holmes, Victoria University of Wellington

Abstract: As the number and size of computer corpora grow, linguistic researchers are increasingly using them to study changes in language over time. Comparing usage at one point in time with usage at a later or an earlier period seems a stunningly simple and Sausurreanly impeccable method of studying language change. Needless to say the reality is rather different. This paper identifies some of the methodological problems encountered in using computer corpora to describe changes in sexist usages in New Zealand English (NZE) over a twenty-five year period.

1. Which corpora?

The ideal situation for comparing corpora as a basis for studying language change would appear to be to use two corpora constructed on parallel principles at two 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, then comparing two similarly constructed corpora seems to offer a relatively straightforward method of identifying at least lexical changes which have occurred during the period between the two collection dates.

Unfortunately, no such parallel corpora exist for New Zealand English. The first one million word corpus of written New Zealand English, the Wellington New Zealand Corpus (WNZC), has only just been completed (Bauer 1993). It has been constructed to parallel the Lancaster-Oslo/Bergen (LOB) Corpus in size and composition, and it records printed material from 1986. There is no earlier parallel corpus of New Zealand English with which it can be compared.

Parallel corpora collected at two different points in time are being constructed for British English, however. At the University of Freiburg, Christian Mair is compiling a 1991 corpus to match the 1961 LOB Corpus as closely as possible. The 'Press' component of this Freiburg

Corpus has been completed (Sand and Siemund 1992), and work is underway on other sections of the corpus. Even such an apparently straightforward goal is by no means unproblematic, and Sand and Siemund (1992: 120) document some of the difficulties they have encountered in trying to match LOB newspaper texts as closely as possible.

In the absence of an exactly parallel set of texts from two different points in time for New Zealand English, is it impossible, then, to contemplate a corpus-based diachronic study of NZE? One possibility is to compare the data from the 1986 WNZC with the data from the 1961 LOB Corpus. The texts have been carefully matched, so the comparison involves similar text types. But, it also involves comparing very different populations in terms of social and linguistic variables. New Zealand is much more homogeneous on these dimensions than Britain.

While there are very real problems in assuming that a written corpus collected in Britain in 1961 can be usefully compared with a written corpus collected in New Zealand in 1986, the assumption is not as unreasonable as it might at first appear. Firstly, it is clear that Britain still had considerable influence on New Zealand in the 1950s and 1960s, not least through the amount of British printed material to which New Zealanders were exposed. Phillips (1991: 184-88) documents ways in which Britain was a major influence on New Zealand intellectual life at this time. Britain was the primary source of international news, both printed and broadcast. Newspapers in the 1950s and 1960s ran a large proportion of British stories, and editorials discussed 'British topics'. About half the books reviewed in the *New Zealand Listener* in 1957 were from Britain, and even New Zealand authors were published in London and wrote for a British audience. By 1986, however, these patterns had changed dramatically. Home-grown New Zealand programmes and writers were systematically displacing this large British component in the New Zealand broadcast and print media. So, for example, by the mid-1980s, 65% of the books reviewed in the *New Zealand Listener* were New Zealand books, and New Zealand authors were being published locally for local audiences (Phillips 1991: 184-88). In other words, while Britain was clearly a major influence on the New Zealand media at the time when the LOB Corpus was constructed, things had changed by 1986. It seems reasonable to assume, then, that such changes might be reflected in a parallel New Zealand Corpus constructed twenty-five years later.

A second reason for considering that a comparison between the the 1961 LOB Corpus and the 1986 WNZC could provide a reasonable basis for inferring language change in New Zealand English is that the kinds of written texts selected for inclusion in the corpora are likely to be less influenced by social variables than spoken texts, or than written texts deliberately selected to illustrate such cultural diversity as exists between the two communities (see, for example Leech and Fallon's (1992) cultural comparison of Britain and the U.S.A. using the parallel LOB and Brown corpora). New Zealand linguistic usage in 1961, at least in the area of written discourse, seems likely to have closely resembled that of British usage of the same era. At the very least, it seems safe to assume that, compared to Britain, New Zealand would not have been leading language change in written English lexis or morphology in 1961.

On the basis of these assumptions, I undertook a comparison of the 1986 WNZC with the 1961 LOB Corpus as a basis for some inferences about language change in New Zealand lexical usages such as *chairman* vs *chairperson* and morphological usages such as *author* vs *authoress* over the twenty-five year period.

2. The relevant universe of discourse: What to count?

Any corpus analysis, including studies of change over time, raises the issue of what to count. Most research in corpus linguistics cites the frequencies of an item in a corpus of a fixed size (eg 40 occurrences of *cottage* in a one million word corpus). Cooper (1984), for example, used a 500,000-word corpus of American newspapers, magazines, periodicals, and the *Congressional Record* to study 'androcentric generics'. He reported his results in the form of the number of instances of a particular item per 5000 words of text. There are many studies which use similar measures.

Given corpora which are similarly constructed and of identical size, it is possible to compare the frequencies of particular items. We can then discover that the forms *businessman* and *businessmen* together occur 24 times in the one million word Brown Corpus of written American English compared to only 6 instances in the comparably constructed British LOB Corpus. Or that *chairman* occurs 107 times in LOB compared to 68 times in the Brown Corpus (Hofland and Johansson 1982).

In comparing the 1961 British LOB Corpus with the similarly constructed 1961 American Brown Corpus, Hofland and Johansson (1982) go beyond simple frequency lists, however, to provide statistical infor-

mation on the relative significance of frequency differences in the corpora being compared. They report the coefficient of difference between the corpora, and a chi-square value indicating where the difference between the frequency scores is significant at various levels. Since even a million word corpus is not large enough to reliably reflect differences in vocabulary between British and American English, they advise that it is probably safest in making comparisons to restrict attention to items which show significant differences of frequency. This is one way of taking account of the fact that differences in frequencies of items which occur relatively rarely may be simply due to chance.

When sociolinguists began analysing variation in texts, they described the problem as one of identifying the relevant 'universe of discourse' or 'envelope of variability' (Lavandera 1978) for any particular item. When analysing phonological variation, sociolinguists treat variant pronunciations as alternative ways of 'saying the same thing'. The sociolinguistic analysis involves establishing the proportion of alternative variants used by any individual or group in particular styles, and so on. The number of actual occurrences of a particular variant is compared to its potential number of occurrences. So, for example, initial /h/ in words such as *house* is generally regarded as having two variants [h] and zero. Analysis will thus involve considering the number of [h] realisations and the number of zero realisations, as a proportion of the total number of possible realisations of /h/ in the data. The total number of places where /h/ occurs is sometimes called the total envelope of variability and it is an important way of ensuring that the analysis has integrity.

There has been a great deal of debate about what constitutes the total universe of discourse or envelope of variability outside phonology (Labov 1978, Lavandera 1978, Sankoff 1988). While morphological constructions are relatively unproblematic (eg alternative ways of expressing the *-ed* regular past tense or the morpheme *-ing* can be satisfactorily identified), the relevant parameters are much less obvious when one considers syntactic variables. Linguists may agree that *she's the one I love* and *she's the one that I love* can be reasonably categorised as alternative ways of saying the same thing, but is that the case for the active and passive forms of an utterance? It can be argued, for instance, that there is an important semantic difference between *I was overcharged* and *you overcharged me*. Lexical choices present similar problems – only more so. It is difficult to see how one could analyse the selection of one word rather than another as a choice from a restricted set of alternative

ways of saying the ‘same’ thing. Firstly, it would be virtually impossible to establish the complete set in many instances; and secondly it is very likely that the choice could be regarded as a motivated selection of a specific item precisely in order to express a subtly **different** meaning.

What are the implications of this approach for corpus research? There are many cases where it may not be possible or useful to examine all the possible choices from which a writer or speaker selected a specific word such as *cottage* or *book*. Similarly, in examining the use of generics, such as *he* and *man* it would be interesting to identify all the alternatives from which the language user has selected. Interesting, but often not feasible, unfortunately. It is simply impossible to identify the full range of stratagems used by writers or speakers who deliberately choose to avoid using forms such as generic *man* or *he*. Even in the case of forms such as *fireman* and *chairman*, where it would seem more feasible, it is difficult to be confident that all possible alternatives have been identified. The total set for *chairman*, for instance, includes at least *chairwoman*, *chairperson*, *chair*, *convenor*, *convener*, and perhaps *coordinator*, *director*, and *president* in some contexts. And doubtless there are other possibilities.

On the other hand, it seems possible that for some items at least, the notion of a total envelope of variability or universe of discourse is a potentially useful one. There are surely some cases where lexical choices are both restricted and totally specifiable, and where it is thus possible to relate the choice of one item to all other alternatives. This was my line of reasoning until I began actually trying to use such an approach in identifying sexist and non-sexist usages in corpus data. The process of analysing the relative frequencies of comparable forms in different corpora at different points in time produced yet another raft of problems.

2.1 Counting *Ms* usage

The choice between *Mrs*, *Miss*, *Ms* as honorific before a woman’s surname would seem, at first sight, to be an ideal candidate for analysis in terms of choice from a restricted set. There are just three alternatives which appear to constitute the total envelope of variability so it appears to make sense to report the frequencies of one in relation to the others. In principle, wherever *Mrs* or *Miss* occur, *Ms* could be used instead, so, more precisely, it would seem reasonable to report the frequency of *Miss* vs *Ms* and *Mrs* vs *Ms* since these are the specific alternatives available in any instance. In other words for the form *Ms* we can attempt

to compare its actual occurrence with its potential for occurrence – something which is very difficult with most lexical items.²

In practice, things are not so simple. Examining such usages involves looking at ways women are addressed and referred to in relatively formal contexts. There are obviously a number of forms of address and reference which have been excluded. Restricting the analysis to written corpora, some women may appropriately be referred to by a title such as *Dr* or *Professor*, and there is at least one additional alternative which seems relevant in an assessment of changes over time in alternative forms of reference for women. Where a woman is referred to by honorific plus surname, in some contexts it would be possible to use simply first name plus surname: eg. ‘a harder-line monetarist policy than Mrs Margaret Thatcher ever dared’ (WNZC B13 176). Clearly there is no way that computer-programmes can identify zero variants.

Nor is it useful to simply report the relative frequencies of *Miss* vs *Ms* and *Mrs* vs *Ms* without taking account of whether these represent genuine alternatives in context. This can be clearly demonstrated by describing the different results which would be produced if such factors are ignored.

There were no relevant instances of *Ms* in the 1961 LOB Corpus. Table 1 provides the numbers of *Mrs*, *Miss* and *Ms* forms in the WNZC. It is based on what would be published in a word frequency list, except that I have, of course, omitted instances of the verb *miss*, the disease *MS*, and the abbreviation *ms* for manuscript.

Table 1: All *Miss/Mrs/Ms* forms in WNZC

	No.	%
<i>Ms</i>	16	4.1
<i>Mrs</i>	233	59.7
<i>Miss</i>	141	36.2
	<hr/> 390	<hr/> 100

It appears from Table 1 that *Ms* was a relatively infrequent form in New Zealand English in 1986. In fact, however, it was not as infrequent as this simple frequency table suggests. Further analysis of these forms in context, revealed that many of the instances of *Mrs* and *Miss* in Table 1 were not instances where *Ms* was a genuine alternative. Consider the following examples:

- | | |
|--|-------------------------|
| 1) You in trouble Miss? | Address term. |
| 2) Hasty Miss. | Name of a horse. |
| 3) Miss Universe contest, Mrs World | Title of a competition. |
| 4) The Prime of Miss Jean Brodie | Title of a book/film. |
| 5) <i>Miss, Ms</i> and <i>Mrs</i> are the choices available. | Citation forms |

In such cases *Ms* could not be substituted for *Miss* or *Mrs*. In an analysis comparing a choice between genuine alternatives, such forms must be deleted.

It was also clear when analysing the forms in context that they were very differently distributed in the various genres or discourse types in the WNZC. In particular, while *Ms* occurred almost exclusively in the press section (there was one instance in the fiction category, and one instance in the Miscellaneous (Foundation Reports) category), *Mrs* and *Miss* were most frequent in the fiction or biography categories.³ In other words, this evidence too suggests that the forms in Table 1 are not genuine alternatives. Clearly the issue of whether *Ms* can be classified as an alternative to *Miss* and *Mrs* in fiction or biography is a very thorny one. For example, authors no doubt select between alternatives for their characters, not in terms of their own ideology, but in terms of the ideology they wish readers to attribute to the characters; biographers must take account of historical accuracy, and are thus likely to hesitate to give a person the title *Ms* if she never used it herself, and so on.

Such considerations led to a comparison of the relative frequencies only of forms which appeared to be genuine alternatives, in just the press section of the WNZC. Table 2 provides the results.

Table 2: Relevant *Miss/Mrs/Ms* forms from WNZC Press section (i.e. sections A, B, C)

	No	%
<i>Ms</i>	14	14.7
<i>Mrs</i>	77	81.1
<i>Miss</i>	4	4.2
	95	100

The results are very different from those in Table 1 in that the instances of *Miss* and *Mrs* in WNZC have dropped dramatically. And

perhaps most interesting is the fact that the relativity of *Ms* and *Miss* has been reversed. *Ms* is used more than three times as often as *Miss* in the newspaper material, though *Mrs* remains by far the most frequently occurring form.⁴ This interesting result would not have emerged without a consideration of the forms in context in order to identify genuine alternatives within the relevant universe of discourse.

2.2 Counting sexist suffixes

It is clear from the discussion of *Ms* that it is crucial to examine forms in context in order to be confident that genuine alternatives are involved. A similar point can be made in relation to comparisons between forms at different points in time, though the analysis in such cases is even more complicated. In comparing sexist usages such as *-ess* and *-ette* suffixes in the 1961 LOB Corpus with those in the 1986 WNZC, for instance, for the purpose of deducing change, it was necessary to identify only those forms in the data which could reasonably have been avoided. It seems reasonable to include in a comparison between the two corpora, instances such as *authoress*, *actress* or *millionairess* used to refer to contemporary women in the press section, for instance. In such cases, the writer had available the alternative unmarked forms *author*, *actor* and *millionaire*. Similarly a form such as *brunette* used as a noun to refer to a woman could (and should!) have been avoided. But it would be misleading to infer from a reference to a *suffragette* in a historical document, or to a *governess* in a biography, for instance, that such forms were in current usage. Historical as well as referential accuracy requires that these forms be used.

Nor can items referring to individuals, such as the *Duchess of York*, provide information on changing usage in the area of sexist suffixes. There are no genuine non-sexist alternatives to such forms. However, once the existence of genuine alternatives is used as a criterion for inclusion of relevant forms, the analytical pathway becomes much more slippery. Most language users might agree that there is no genuine current alternative to the form *princess*, for instance, certainly as a title, but even in a fairy story. But what about *waitress* and *heroine*? Where does one draw the line between 'avoidable' and 'unavoidable' sexist suffixes, and on what basis? A form such as *hero*, for example, might once have been considered inappropriate for a female referent, yet it is used explicitly to refer to a woman in the WNZC. In other words, a form which formerly would have been used only to refer to men, is now being used to include or refer to women – clear evidence of

semantic change. Since language change obviously includes changes in the semantic features of forms such as *hero* and *waiter*, as well as the substitution of one form for another, decisions about what to count are crucial in providing an accurate picture of change in progress. So while it is clear that an analysis of the number of sexist forms at two different points in time should compare only instances where sexist usages could have been avoided and a non-sexist form could have been substituted, it is also clear that making such judgments is not always easy.

2.3 What is a pseudo-generic?

As a final example of the problems of what to count in corpus research, I will discuss the issue of what qualifies as an instance of generic *man*. Using corpus data to infer language change in the area of sexist and non-sexist usages, it would seem relatively straightforward to compare the frequency of instances of generic *man* in corpora collected at different points in time. As you might expect by now, the analysis turned out to be far from straightforward. Consider the following pairs of examples:

- (1) A 35-year-old man was killed in a collision
- (2) The man, 35, whose name was not made public
- (3) fearful of what man can do to man
- (4) The reason for this slaughter she lays at the feet of man.
- (5) Very important we get the right sort of man for Commodore, someone who can represent us properly.
- (6) 'Back in old Greymouth, though,' Red went on 'a man'd be sitting in the pub with a schooner under his nose...'

While it is clear that the referent in (1) and in (2) is 'a specific male person', and that the referent in (3) and in (4) is 'generic man' or 'humankind', it is not at all clear whether the referent in (5) and (6) is specifically male or rather 'any human being'. *Man* in (1) and (2) is not a generic; *man* in (3) and (4) would be widely recognised as a form intended to have generic reference, but classification of *man* in (5) and (6) as a generic or not depends on a wide range of potentially relevant factors, some of which are impossible to identify. What did the writer intend, for instance? What would readers deduce? Different readers would almost certainly make different deductions depending on such factors as their age, their level of education, and their political

stance in relation to feminism. A feminist could object and say these are clearly cases where *person* should have been used, but the context strongly suggests the speaker had 'male' in mind. There is, moreover, a range of psychological research demonstrating that sentences using forms such as (so-called) generic *man* prime readers to perceive mainly male images (e.g. Moulton, Robinson and Elias 1978, Ng 1990, 1991, Schneider and Hacker 1973, Silveira 1980, Wilson and Ng 1988). Consequently, it is now rather difficult to identify uncontroversial instances of 'generic' *man* or to be confident of isolating tokens of the use of generic *man* which could have been avoided.

Instances which may have been intended as generic could be challenged by feminists as misleading in that they inappropriately suggest 'male human being' to many readers. They could be considered **pseudo-generics**: i.e. forms which claim to be generic while in fact suggesting 'male'. Examples include phrases such as 'the man in the street', 'as good as the next man', 'the tax man', and so on. Using the following two criteria I classified such phrases as generics in a broad sense:

- (1) Using 'person' would not lose relevant referential information about the referent.
- (2) Using 'person' would make it clear that this referent could be female, and that the context does not *require* a masculine referent.

Applying these criteria involves constantly making judgments about what constitutes 'relevant information' and what constitutes a case where a masculine referent is required, as opposed to more appropriate. The following were classified as pseudo-generics, but it is likely not all readers would confirm this judgment, particularly when a phrase is considered a well-established idiom or cliché.

one man one vote
to a man
the white man's burden
a man's home is his castle
no-man's land
to the last man
We could have a man on the moon in six months
families are starving because a man cannot get a job

Drawing such distinctions is obviously problematic. One person's intended generic may be another's clearly male referent. One consequence is what feminist linguists have called 'slippage',⁵ not only in people's usage, but also in interpreting their meaning accurately for purposes of corpus analysis. Ideally the analysis requires a panel of judges representing the community whose language the corpus purports to represent. But although it is difficult, this kind of analysis is quite crucial in studying language change. The subtle changes in perceptions about who can appropriately be assumed to be included in terms such as *man* represent change in progress. If one generation recognises that 'the man in the street' could be a woman, while the next considers he must be male, then the pseudo-generic status of such forms is at the cutting edge (imprecise and frayed though it may be) of change in progress.

Table 3: Generic *man* from press sections

	WNZC Corpus	LOB Corpus
<i>man</i> (total instances)	134	151
generic <i>man</i> (‘humankind’)	11 (8.2%)	15 (9.9%)
pseudo-generic <i>man</i> (‘person’)	10 (7.5%)	32 (21.2%)

Table 3 makes it clear that there were relatively few instances of generic *man* meaning ‘humankind’ in either the 1961 LOB Corpus or in the 1986 WNZC. Moreover, the percentage of such forms does not appear to have altered over the period. There was, however, a clear difference in the frequency with which pseudo-generics occurred in the two corpora. They were three times as frequent in the LOB Corpus compared to the WNZC, suggesting that there has been a reduction in the frequency of such forms over time.

3. Conclusion

The prospect of using corpus data to infer language change over time is an exciting one. It is clearly possible to make suggestive and interesting comparisons between the frequencies of items in corpora of similar size and composition which have been constructed at different points in time.

Using a study of sexist and non-sexist usages for exemplification, I have highlighted in this paper some of the problems that can arise in making such comparisons, and in particular the importance of considering context in comparing forms which are in the process of change.

When examining the introduction and spread of newer non-sexist forms such as *chairperson* or *Ms*, for instance, it is important to be aware of the alternatives available to language users in the contexts of use. *Ms* cannot be considered a possible substitute for the name of a character in fiction, for instance. A language user who dislikes such forms but nevertheless wishes to avoid sexist usages may use avoidance strategies which are difficult to identify using frequency counts.

The same points apply to the avoidance of marked sexist forms such as *authoress*, *actress* and *hostess*. It is crucial to check that the forms represent genuine examples of contemporary usage, rather than justifiable references to the way women's roles were referred to in earlier historical periods, for instance. It is important to count only forms where alternatives are genuinely available. There is no reasonable way of avoiding the *-ess* form in referring to the *Princess of Wales* or the *Duchess of York*. And it is also important to document evidence that unmarked forms such as *hero* and *manager* do occur with female referents. This is evidence of change in progress.

Finally I have discussed some of the problems of deciding what counts as an instance of a form when the form in question is undergoing a process of semantic change. At the beginning of this century the meaning and reference of generic *man* was relatively unproblematic. It was widely used to include both male and female referents. Increasingly, however, the form *man* is being interpreted as signalling only a male referent. The problems of deciding what counts as a generic usage is very problematic. In comparing usages in different corpora it became clear that at least two different categories could be identified: a clearly generic usage where 'humankind' was the referent, and a pseudo-generic usage where 'person' could reasonably be substituted, whatever the sex of the writer's intended referent. It is the latter usages that are most interesting from the point of view of language change in progress. But it is equally in this area that analytical decisions are most subjective and most difficult. Discovering that using corpus data to study language change raises as many theoretically and methodologically interesting questions as it answers should not be surprising. That after all is the nature of any research.

Notes

1. I would like to thank Laurie Bauer for his useful comments on an earlier draft of this paper.
2. Note that while the choice between *Miss* and *Mrs* is generally a referential one, ie. usually (though not always) based on marital status, the choice between *Ms* and *Mrs/Miss* is a sociolinguistic one, though as Pauwels (1987) demonstrates its exact sociolinguistic significance is not universally agreed upon.
3. Note too that press usage simply reflects the editorial policies of the newspapers at the time the data was collected. Vivienne Holt's (1988) study of the policies of a variety of New Zealand newspapers in the area of non-sexist usages makes it very clear that (mainly male) editorial gate-keepers determine the relative frequencies and even the appearance of forms such as *chairperson* and *Ms* in the New Zealand press.
4. There were, incidentally, only 4 instances of *Ms* in the press section of the 1991 Freiburg Corpus referred to above (three of which referred to the same person), suggesting that New Zealand usage is ahead of British in this area.
5. 'Slippage' is when an utterance starts off as a generic but slips into masculinity during its progress: eg. 'The lack of vitality is aggravated by the fact that there are so few able-bodied young adults about. They have all gone off to look for work, leaving behind the old, the disabled, the women and the children' (cited in Cameron, in press: 11, from *The Sunday Times*).

References

- Bauer, Laurie 1993. Progress with a Corpus of New Zealand English and early results. In Clive Souter and Eric Atwell (eds), *Corpus-based computational linguistics*. Amsterdam and Atlanta, Ga: Rodopi.
- Cooper, Robert L. 1984. The avoidance of androcentric generics. *International Journal of the Sociology of Language* 50:5-20.
- Hofland, Knut and Johansson, Stig 1982. *Word frequencies in British and American English*. London: Longman.

- Holt, Vivienne 1988. How New Zealand newspapers name people. *Occasional Papers in Language and Linguistics* 1:9-18.
- Labov, William 1978. Where does the sociolinguistic variable stop? A reply to B. Lavandera. *Texas Working Papers in Sociolinguistics* 44 Austin TX: Southwest Educational Development Laboratory.
- Lavandera, Beatriz 1978. Where does the sociolinguistic variable stop? *Language in Society* 7,2:171-82.
- Leech, Geoffrey and Roger Fallon 1992. Computer corpora - what do they tell us about culture? *ICAME Journal* 16:29-50.
- Moulton, J., G. M. Robinson and C. Elias 1978. Psychology in action: Sex bias in language use: 'neutral' pronouns that aren't. *American Psychologist* 33:1032-1036.
- Ng, S.H. 1990. Androcentric coding of *man* and *his* in memory by language users. *Journal of Experimental Social Psychology* 26: 455-464.
- Ng, S.H. 1991. Evaluation by females and males of speeches worded in the masculine, feminine or gender-inclusive reference form. *International Journal of Applied Linguistics* 1, 2:186-97.
- Phillips, Jock 1991. New Zealand and the Anzus alliance: Changing national self-perceptions, 1945-88. In Richard W. Baker (ed) *Australia, New Zealand and the United States: Internal Change and Alliance Relations in the ANZUS States*. New York: Praeger. 183-202.
- Sand, Andrea and Rainer Siemund 1992. LOB-30 years on ... *ICAME Journal* 16:119-122.
- Sankoff, David 1988. Sociolinguistics and syntactic variation. In Frederic Newmeyer (ed) *Linguistics: The Cambridge Survey IV*. Cambridge: Cambridge University Press. 140-61.
- Schneider, J.W. and S.L. Hacker 1973. Sex role imagery and the use of the generic 'man' in introductory texts. *The American Sociologist* 8, 1:12-18.
- Silveira, Jeanette 1980. Generic masculine words and thinking. *Women's Studies International Quarterly* 3:165-178.
- Wilson, E. and S. H. Ng 1988. Sex bias in visual images evoked by generics: A New Zealand study. *Sex Roles* 18:159-169.