Corpus Linguistics and Cultural Difference in Canada

Margery Fee
University of British Columbia

In 1997, Janice McAlpine and I finished *The Canadian Guide to English Usage* (Oxford). When my predecessor, W.C. Lougheed, handed the project over to me in 1987, a desktop computer with a 10-megabyte hard drive was cutting edge technology. The Strathy Unit at Queen's University in Kingston, Canada, where the project was housed, was producing its corpus with a Kurzweil flatbed scanner that cost $40 000, a Tallgrass external hard drive the size of a breadbox that cost about as much as a new Volkswagen, and search software written by clever amateurs. Over the next few years, technology raced ahead, and we were fortunate to be able to augment our corpus substantially. We ended up with 16 million words of selected book, magazine and newspaper text as well as 163 million "extra" words of newspaper text (the latter from CD-ROMs of Canadian newspapers donated by Southam Press and the Toronto *Globe and Mail*). The corpus is still being expanded. Technology had made it possible for us to use real examples of Canadian English to back up our claims about Canadian usage rather than just guessing; that is, we were able, because of our use of a corpus, to be descriptive rather than prescriptive. Even late in the project, however, it was difficult for us to deal with grammatical structures because our corpus was not parsed; we were using as our search tool a cheap piece of software called Gofer (Microlytics) that looked for word-strings at about a megabyte a minute, much faster than the more complex tools designed for academics. Gofer made it possible to find data to allow us to analyse most of the major issues in usage, as well as many distinctive Canadian features, spelling in particular.

We are now beginning a revision, and I expect we will find a lot has changed in how we are able to augment and search our corpus. Of course, the most obvious change is our ability to search the internet as a whole and sites within it. As Angus J. Kennedy points out, since the mid-1990s, the internet has transformed how we live, producing "the closest thing yet to an all-encompassing snapshot of the human race. Never before have our words and actions been so immediately accountable in front of such a far-reaching audience" (516-17). Academics in language and literary studies are still coming to terms with how the internet and the New Media can provide our disciplines with new data and new approaches to it. Only recently has it been possible to find cheap useful software tools to support such research (see Rundell, Lancashire). One irony of the change is that while we had to work hard to get enough text into the corpus, the internet provides us with too much. It will never be a good substitute for a structured corpus. Ours used the 15 subject categories of the million-word pioneer Brown/LOB corpora with the addition of feminism and computing, as well as aiming for national coverage in the newpaper text category. Since we were looking mainly for the uncommon words and expressions that are the focus of usage debates, rather than high-frequency words and grammatical structures, we needed a larger corpus so as to find meaningful numbers of examples to assess.

Corpora are not used in much language study in Canada. Noam Chomsky's methodology is dominant in the few departments of linguistics in Canadian universities. As you know, he focuses on abstract formal structures at or below the level of the sentence, and these sentences are typically made up by someone (usually the researcher) who speaks English as a first language. Its drive is to discover universals originating in similarities in the human brain rather than to find differences based on culture. Most people who use corpora in Canada are working in lexicography, dialectology or rhetoric, rather than linguistics. The "cultural difference" in my title, then, refers to differences between academic interests in Europe and North America, differences among disciplines, primarily linguistics and literature, as well as different levels of linguistic nationalism within and between these places.

These cultural differences explain why huge resources in the form of digitized text which I will describe later, are under-used by Canadian researchers. Ironically, at least as much work is done on Canadian English in Europe as in Canada itself. North America has fallen behind, partly because Europe faces the practical needs of a vast English-language teaching project (see Pennycook)

and partly because of the acceptance of broader and more integrated approaches to language study. Indeed, initiatives from the UK have inspired some projects in North America. Before I left Queen's in 1993, Nancy Belmore and Sabine Bergler (both of Concordia University in Montreal) and I began work collecting material for a corpus that includes spoken and written Canadian English as part of a world-wide project called the International Corpus of English (ICE) based at University College London (www.ucl.ac.uk/english-usage/ice/). A similar project described in 1998 by Hongyin Tao and Linda R. Waugh of Cornell University is again joint with a UK university, Cambridge (http://www.bol.ucla.edu/~ht37/cv.htm#research.). Their corpus will include examples of Canadian spoken English and is part of a large pedagogical project. Several other projects based in the United States include Canada including the Telsur Project (http://www.ling.upenn.edu/phono_atlas/home.html) at the University of Pennsylvania, headed by William Labov, working to produce a phonetic atlas.

The Canadian nationalism that supported projects such as the *Dictionary of Canadianisms on Historical Principles*, edited by Walter S. Avis and published in 1967, has since changed focus somewhat and the big academic lexicographical projects are bilingual French-English projects (see the Bilingual Canadian Dictionary Project, headed by Roda Roberts, University of Ottawa, http://www.uottawa.ca/vr-recherche-research/perspectives/v2n1/) and Quebec French dictionaries (see, for example, *Le grand dictionnaire terminologique*, http://www.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp). These are projects that usually require grant support, because even the desk dictionaries in English and French that that include only the most common Canadianisms are expensive propositions, and these are designed for an even smaller market. An earlier nationalist focus on differences between Canadian and American or British English is now rather dated in a world where varieties are seen as defined as complex wholes ("englishes"), rather than by a few odd forms that are seen as "mistakes" or at least as exceptions to the implicit norms of British or American English. However, language difference is still a hot topic in Quebec. There was a fevered reaction to the small entry I wrote in the usage guide on Quebec English. News that I thought it could be described as a variety in its own right hit the front page of the *Montreal Gazette* (the only English daily newspaper in the city). It became clear in the many radio and television interviews that followed that the only feature of popular interest in the whole book was this point about language, not surprising in Montreal, a city where the slightest shift in language use between French and English receives serious

attention. Corpus linguistics is key to discovering where contact between English and French in Quebec has led to differences that prove that language policies favouring French in that province have actually had a widespread impact on English (just as favouring English in the years before 1976 when the Parti Qubcois came to power had the opposite effects). For example, English words that are obsolete or rare in Canadian English elsewhere appear in Quebec English because they are similar to high-frequency French words. Thus, for example, the word "fête" is by far more frequent in Quebec English than in Canadian English outside the province (Fee). Pamela Russell, at the University of Sherbrooke in Quebec, has been funded by that province to produce a corpus of Quebec English and more of these sorts of "contact phenomena" will doubtless be discovered in her work.

It is regrettable that linguistics as a discipline in North America has been slow to take up new approaches that take advantage of what has become a ubiquitous computer technology. Less surprising is the absence of such an approach in the work of literary scholars. Few have required the computer for more than word processing, library research and email, and the idea that their work could be described as lacking because of its failure to take account of the world of discourse which includes literary texts has not really been hammered home. Yet few among us still adhere to New Critical formalism or structuralism, which has theoretical parallels with Chomsky's preference for the study of stripped-down invented text. Most of us now would agree that meaning is made in context and reality is socially constructed. What once was a huge theoretical divide between literary and other sorts of text has diminished, in theory, if not in practice. Michael Stubbs sums it up: "all texts make intertextual references. In what they include or omit, all texts make assumptions about their readers or listeners. Texts are shaped by prior texts, by repetitions or by being oriented to routines of conventions. Therefore, all texts are inherently historical" (Stubbs 34). If this is so, literary texts should be studied as part of a much broader range of texts. For an example of this sort of approach, inspired by the neo-Firthian theories of M.A.K. Halliday, see Terry Threadgold's work on the historical incident that inspired the Australian novel *The Chant of Jimmy Blacksmith,* by Thomas Keneally (also a feature movie, 1978). Instead of looking at this novel in isolation, she compared it to contemporary newspaper accounts of the incident and other fictional accounts, showing how all these quite different texts situated Aboriginal people in similar ways. Most literary analyses that use a corpus tend to focus on the statistics, dropping the 'old-

fashioned' interpretive approach. In my view findings from corpus work should be integrated into the overall study, as Threadgold does (even though she did not use a corpus to make her findings). Another even earlier work that shows the potential for literary corpora is Caroline Spurgeon's *Shakespeare's Imagery* (1935); she was able to use concordances produced laboriously by hand to do her analyses. Now, of course, such a concordance (keyword in context list/ KWIC) can be produced in seconds for any digitized text. For those interested in Canadian literary texts in digital formats, several useful resources exist: the Digital Library of Canada, held by the National Library and National Archives of Canada, Early Canadiana Online, produced by the Canadian Institute for Historical Microreproductions, and the Canadian Literature Archive at the University of Manitoba, for example (see resource list below for more).

Online digital text collections now can supplement our literary work. Obviously, even simple keyword searches can produce useful comparative information. Anyone can now do this kind of research from anywhere, using the internet. (Although this means it is no longer necessary to read and take notes of the whole work, I still argue that to do both is preferable, since to use keyword searches only extracts data from its context in the same way that I have been criticizing in formalist and positivist methodologies). Another way to examine this material is to put it into a concordance program, which produces collocational sets, what J.R. Firth called "actual words in habitual company" (qtd. Stubbs 173). This allows one to trace changes in local cultural meanings. For example, the word *Aboriginal* was rarely, if ever used to refer to indigenous peoples in Canada until the promulgation of the 1982 Constitution Act.

If we accept that "social institutions are always supported by text types" (Stubbs 59), and that if we want to be clearer about what exactly does get lost in translation or over time, corpus linguistics is an essential tool. Stubbs notes that "there are patterns that contribute to the meaning of texts that are not open to direct observation" (92). Thus all readings are usefully supplemented by computer analysis and, where possible, contextualization in a body of text that provides a corpus for that particular study. He argues that particular keywords have "microhistories" within a single text (92)—in a short text, like a poem, we can easily see this process at work. In longer texts, like novels, we should ask ourselves why we do not think of scanning them so that we can follow up collocations and shifts in the use of keywords. Further, we know that novels emerge out of discourse and select their meaning from a much broader array of possibilities. This is the juncture where we could compare the treatment of

particular stereotypes in fiction and other genres. Obviously, the computer can't do all the work, but it can help. For example, I found it instructive to enter the keywords "language," "Indian" and "abstract" into Early Canadiana online. This quickly produced a set of notions about Aboriginal languages still familiar today—for example, that Aboriginal languages have many concrete expressions for material phenomena, but not for abstract moral concepts like good and evil. For example, from William Elliott Griffis, *Sir William Johnson and the Six Nations* (New York: Dodd Mead, 1891): "Unaccustomed to abstract reasoning, the Indian was perforce obliged to draw the imagery of thought entirely from the environment of his life on land and water. Hence, his speech superabounded with metaphors" (47). That these pompous and mistaken ideas about language were produced by men still in the early stages of language learning would have been funny if the consequences for the Aboriginal peoples hadn't been quite so genocidal. (For the corrective view, see George Lakoff and Mark Johnson's *Metaphors We Live By*). The findings of such searches of historical texts can also be instructively compared to contemporary ones.

In the Transculturalisms project supported by the International Council for Canadian Studies, a team led by Sneja Gunew (see http://transculturalisms.arts.ubc.ca/), focused on the variety of terms demarcating cultural difference and culturally sensitive issues—terms like *metissage*, *hybridity*, and *mixed race*. These are keywords that the methods of corpus linguistics could usefully examine. Sometimes, terms like these draw a near-absolute line between two cultures. For example, in Canada the word *separatist* is rare in Quebec, where *sovereignist* and *independentist* are preferred. I suspect that most English-speakers outside Quebec would not be able to even suggest a term other than *separatist* for someone who hopes for independence for Quebec from Canada. Another example comes from a recent visit to the University of Southern California; what we had heard of in Canada as the "Watts riots" (1965, Los Angeles) was occasionally described there as the "Watts uprising," a shift in word choice which shifted the media focus on the dangers of a violent urban underclass to the just political aspirations of an oppressed minority. However, a Google search on 3 June 2004 revealed 6490 hits for "Watts riots" and only 524 for "Watts uprising." Ironically, the first site listed for the first search is at the University of Southern California, located near Watts (http://www.usc.edu/isd/archives/la/watts.html); many of the hits on "Watts uprising" refer to a book by Gerald Horne, *Fire This Time: The Watts Uprising and the 1960's* (Charlottesville & London: University Press of Virginia, 1995). What might appear to be

minor distinctions, ones that might not even register clearly in a dictionary, carry whole complex histories of political struggle along with them.

The point I want to leave you with is that if we believe that discourse makes and remakes social reality, we need to study more discourse, to move beyond the limits that human perception and memory can attain and to use the technology that has, in only the last 10 years or so, magically appeared on our own desks.

Works Cited

Avis, Walter, et al. *A Dictionary of Canadianisms on Historical Principles*. Toronto: Gage, 1967.

Fee, Margery, "Frenglish in Quebec English Newspapers." *Papers from the 15th Annual Meeting of the Atlantic Provinces Linguistic Association*. Ed. Wm. J. Davey. Sydney, NS: APLA, 1992. 12-23.

Fee, Margery and Janice McAlpine. *The Canadian Guide to English Usage*. Toronto: Oxford UP, 1997.

Kennedy, Angus J. *The Rough Guide to the Internet*. London: Rough Guides, 2002.

Lakoff, George and Mark Johnson. *Metaphors We Live By*. Chicago : U of Chicago P, 1980.

Lancashire, Ian. *Using TACT with Electronic Texts: A Guide to Text-analysis Computing Tools, version 2.1*. New York: Modern Language Association, 1996.

Pennycook, Alastair. *English and the discourses of colonialism*. London: Routledge, 1998.

Rundell, Michael. "The Corpus of the Future and the Future of the Corpus." Paper del. at conference "New Trends in Reference Science," Exeter, 29 March 1996. http://www.ruf.rice.edu~barlow/futcrp.html

Russell, Pamela. "An Investigation of Lexical Borrowings from French in Quebec English." *The Twenty-Third LACUS Forum 1996*. Ed. Alan K. Melby. Chapel Hill, NC : Linguistic Association of Canada and the United States, 1997. 429-39.

Spurgeon, Caroline. *Shakespeare's Imagery and What It Tells Us*. Cambridge: Cambridge UP, 1935.

Stubbs, Michael. *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell, 1996.

Threadgold, Terry. "Stories of Race and Gender: An Unbounded Discourse." *Functions of Style*. Ed. David Birch and Michael O'Toole. London : Pinter, 1988. 169-204

Resource List for Canadian English in Digital Format

(sites accessed in August 2003; thanks to Cynthia Genest for her help compiling this list)

• Bilingual Canadian Dictionary. www.dico.uottawa.ca
A large corpus of Canadian texts in English and French, which can be used to compile other Canadian dictionaries, to produce language teaching materials, and to undertake comparative linguistics research.

• Canada's Digital Collections. Government of Canada.
http://collections.ic.gc.ca
Maintained by young Canadians under government sponsorship.

• Canadian Literature Archive. Department of English. University of Manitoba.
http://www.umanitoba.ca/canlit
Since 1994, this aim of this website has been to reprint electronically out of print and out of copyright works of Canadian fiction, poetry and drama.

• Canadian Poetry. University of Toronto Library.
http://www.library.utoronto.ca/canpoetry/
Contains examples of poetry by Canadian authors, as well as bio-bibliographical information.

• CBC Archives. Canadian Broadcasting Corporation.
http://www.cbc.ca

• Early Canadiana Online. Canadian Institute for Historical Microreproductions (CIHM), Canadian Heritage and the National Library of Canada http://www.canadiana.org/english/
Currently the site contains over 1.4 million pages of printed material contained in more than 8400 volumes. Covers the period from first contact to early 20th century. Includes government publications, records of the Hudson Bay Company and the Jesuits.

• National Library of Canada and National Archives of Canada
http://collection.nlc-bnc.ca/e-coll-e/about-e.htm
The Electronic Collection of the National Library of Canada consists of Canadian books and periodicals published online. It includes more than 10, 510 titles and more than 43, 299 serial issues published by both the commercial publishing sector and the government publishing sector.

• Parliament of Canada http://www.parl.gc.ca
Debates of Senate and the House of Commons. (Hansard).

• Supreme Court of Canada http://lexum.umontreal.ca/csc-sec/en/index.html
Judgements of the Supreme Court of Canada since 1985.

• Strathy Language Unit, Queen's University, Kingston, Canada http://post.queensu.ca/~strathy/

• Text Encoding Initiative http://www.tei-c.org
Indexes projects using SGML/TEI markup language, including the Thomas Raddall Electronic Archive Poject, housed at Dalhousie University.