Corpora: Capturing language in use

'Language cannot be invented; it can only be captured.' (Sinclair 1997: 31)

Alexandra D'Arcy
University of Victoria

to appear in W. Maguire and A. McMahon (eds.), *Analysing variation in English: What we know, how we know it, and why it matters*. Cambridge: Cambridge University Press.

WHAT'S OUT THERE?

The purpose of this section is to give a sense of existent English corpora. It is impossible, however, to provide a complete overview. There are simply too many and corpus construction projects (public and private) are likely to continue *ad infinitum*. But it is also the case that many public corpora come at a cost, literally. The International Computer Archive of Modern and Medieval English (ICAME) collection, for example, costs 3,500 NOK for an individual user license (at the time of writing, roughly equivalent to 685 USD, 345 GBP, or 440 EUR).[1] For those without the necessary funds (i.e. most students), these fees present the ultimate barrier.

Corpora: Public and free of charge

In what follows, some public and free corpora are briefly outlined. The list is by no means exhaustive; it is simply intended as a starting point for students interested in variation in English.

Dialect atlases

Dialect atlases are an excellent source of data for studying variation and most university collections include at least one (some have more than 100). Both the geographic representation and the historical time depth of English dialect atlases allow for innumerable investigations of lexical, phonological, and phonetic variation across time and space. The most recently published is the *Atlas of North American English* (Labov et al. 2006), but for a point of historical comparison one can also find the *Linguistic Atlas of New England* (Kurath et al. 1939-1943). Online, there is the *Linguistic Atlas Projects*, a portal to a number of atlas projects in the United States (e.g. AFAM, LAGS, LAMSAS, etc.). You can also access the *Dialect Topography Project* (Chambers 1994), which investigates words (and their pronunciation) used both in Canada and in regions of the United States that border Canada.

The Oxford Text Archive

The Oxford Text Archive is a repository for literary and linguistic resources. Most of the holdings are in text format, but some audio and video files are archived as well. All the texts can be accessed for free simply by submitting your email address (used to send the link to the text of interest), but for those marked 'restricted' users are required to register before the resource can be downloaded.

[1] Some corpora can be purchased for a nominal fee for classroom use (e.g. BNC Baby, a four million-word subset of the BNC); the full BNC can be searched online for no cost using the interface created by Mark Davies, http://corpus.byu.edu/. This site also provides links to COCA (the Corpus of Contemporary American English) and the TIME Magazine corpus, among others.

Text and speech-based corpora

Among traditional text-based and speech-based corpora, there are a few that can be accessed via the Internet for non-profit academic research.[2] In most cases, a password is required, obtainable by downloading the appropriate access request form and/or licensing agreement.

- *Brown University Standard Corpus of Present-Day American English*

Via a guest account (as opposed to purchasing a membership), the full text of the Brown corpus can be accessed through the Linguistic Data Consortium, LDC Online. Guests can also access an indexed collection of Arabic, Chinese and English newswire text, the Switchboard and Fisher collections of telephone speech, and the American English Spoken Lexicon.

- *Buckeye Natural Speech Corpus*

The Buckeye corpus is a sociolinguistically stratified corpus of unmonitored casual conversations from Columbus Ohio. It includes data from 40 speakers (male and female, over 40-years-old and under 30-years-old) in text and audio format. The materials can be accessed for research and teaching purposes after submitting a completed license agreement.

- *Corpus of Early Ontario English, pre-Confederation section*

CONTE-pC is a diachronic, text-based corpus of early Canadian English with three genres (newspaper texts, diary entries, letters). It is similar in design to ARCHER (A Representative Corpus of Historical English Registers), enabling comparisons with other historical varieties of English (see Dollinger 2008: 99-119). At the time of writing, CONTE-pC is in the final proof-reading stage but once complete it will be available through the Oxford Text Archive.[3] Period covered: 1776 to 1849.

- *International Corpus of English*

The ICE corpora include both written and spoken texts. Of the eight completed regional corpora (thirteen others are currently under construction), five can be accessed free of charge through the ICE site: East Africa, Hong Kong, India, Philippines, and Singapore.

- *Newcastle Electronic Corpus of Tyneside English*

NECTE is a public dialect corpus from Tyneside (Allen et al. 2007). It consists of two synchronic corpora, one from the late 1960s and one from the early 1990s. The materials are available in a variety of formats (digitized audio, standard orthographic transcription, phonetic transcription, POS-tagged) and may be accessed by students (undergraduate and postgraduate), academics, and members of the public for *bona fide* research purposes (e.g. class projects, research) upon submitting the access request form.

- *Santa Barbara Corpus of Spoken American English*

The SBC contains naturally occurring discourse from across the US (e.g. Alabama, California, Montana, New Mexico, Washington, etc.). Most of the conversations are face-to-face interactions, but some record other modes of discourse such as telephone conversations, lectures, medical interactions, and narratives of personal experience.[4] The SBC can be purchased in CD or DVD format from the LDC or the transcripts and their corresponding audio files can be downloaded from TalkBank.

---

[2] For descriptions of, and accessibility details for, other public specialized corpora, see many of the contributions in Beal et al. (2007a).

[3] Before that time, researchers can access CONTE-pC by individual request.

[4] Summaries of the sixty discourse segments in the SBC can be found at http://www.linguistics.ucsb.edu/research/sbcorpus_summaries3.html

- *Scottish Corpus of Texts and Speech*

The SCOTS corpus contains written and spoken texts of Scots and Scots English, and includes a handful of Scottish Gaelic texts as well. The corpus covers the period 1945 to 2007, though most of the spoken texts (which are synchronised with the audio recordings) were recorded after 2000. After agreeing to the terms and conditions outlined on the site at http://www.scottishcorpus.ac.uk/termsandconditions.html, SCOTS can be searched online at no charge to the user.

The World-Wide-Web

Finally, the Web itself can be a corpus and there are search engines available for this purpose. Two in particular have been designed to retrieve linguistic data from the Web: WebCorp (Renouf 2003, Morley 2006) and GlossaNet (Fairon 2000). GlossaNet is an automated service that monitors the websites of more than 100 newspapers. Once a search item, dates, and intervals are specified, GlossaNet applies the queries and the results are e-mailed to the user in the form of a concordance (a display of the search item with its surrounding context). Because GlossaNet builds new corpora every day, downloading current editions of newspapers, it is a dynamic corpus. WebCorp is more versatile. It can 'piggy-back' on existent search engines (Google, Altavista, Metacrawler) and is not limited to newspapers. A basic search will result in concordance lines of the query item, but the program also has a built-in suite of tools that enable a number of more advanced searches like pattern matching (e.g. *is * nice* will match *is so nice, is really nice, is very nice*, etc.) or specifying the target domain (e.g. the New Zealand academic domain, *ac.nz*, the BBC website, *bbc.co.uk*, the Canadian government, *gc.ca*).

WHERE TO GO NEXT

The classic primer in corpus linguistics is Sinclair (1991), while the contributions in Beal et al. (2007a) represent the state of the art on specialized corpora. Key readings in sociolinguistic data collection are Labov (1972c), Sankoff and Sankoff (1973), and Milroy (1987), and more currently, Milroy and Gordon (2003) and Tagliamonte (2006a). Poplack (1989) is foundational for issues surrounding sociolinguistic corpus construction and data handling; for careful discussion of text-based corpus construction, see Meyer (2002). On representativeness in data sampling and corpus construction see Sankoff (2005) and Biber (1993). A good starting point is Francis and Ku?era (1964), the companion to the *Brown Corpus*, which established the model for subsequent corpora projects. On annotation in text corpora, see Leech (1993a). For discussion of the issues involved in representing speech in writing, see Ochs (1979), Macaulay (1991a) and Tagliamonte (2007). Kennedy (1998) provides a history of English corpus linguistics and a summary of key research in the field. A valuable resource for those interested in the bourgeoning field of web-based corpus studies is Hundt et al. (2007). Biber et al. (1999) is an excellent reference grammar based on corpora representing British and American English; it is a good place to start when looking for possible project ideas. Online, David Lee's *Bookmarks for Corpus-Based Linguists* (Lee 2001–) is an invaluable resource for all corpus-related issues.

LIST OF ABBREVIATIONS

ACE	Australian Corpus of English (*aka* Macquarie Corpus of Written Australian English)
AFAM	African American and Gullah Project
ANC	American National Corpus
ARCHER	A Representative Corpus of Historical English Registers
BNC	British National Corpus
Brown	Brown University Standard Corpus of Present-Day American English
CELEX	Dutch Centre for Lexical Information
CLAWS	Constituent Likelihood Automatic Word-tagging System
COCA	Corpus of Contemporary American English
COLT	The Bergen Corpus of London Teenage Language (now part of the BNC)
CONTE-pC	Corpus of Early Ontario English, pre-Confederation section
FLOB	Freiburg-Lancaster-Oslo-Bergen Corpus
Frown	Freiburg-Brown Corpus of American English
ICAME	International Computer Archive of Modern and Medieval English
ICE	International Corpus of English
LAEME	Linguistic Atlas of Early Middle English
LAGS	Linguistic Atlas of the Gulf States
LAMSAS	Linguistic Atlas of the Middle and South Atlantic States
LAOS	Linguistic Atlas of Older Scots
LDC	Linguistic Data Consortium
LOB	London-Oslo-Bergen Corpus
LSWE	Longman Spoken and Written English Corpus
MICASE	Michigan Corpus of Spoken Academic English
NECTE	The Newcastle Electronic Corpus of Tyneside English
ONZE	Origins of New Zealand English
SBC	Santa Barbara Corpus of Spoken American English
SCOTS	Scottish Corpus of Texts and Speech
SEU	Survey of English Usage
TEI	Text Encoding Initiative

WEBSITES

Site	http(s):
ANC	//americannationalcorpus.org/
Bookmarks for Corpus-Based Linguists	//devoted.to/corpora
BNC	//www.natcorp.ox.ac.uk/
Buckey Corpus	//buckeyecorpus.osu.edu/
CELEX	//www.ru.nl/celex/
Dialect Topography	//dialect.topography.chass.utoronto.ca/
GlossaNet	//glossa.fltr.ucl.ac.be/
ICAME	//icame.uib.no/
ICE	//www.ucl.ac.uk/english-usage/ice/
LDC	//www.ldc.upenn.edu/ *For a guest account:* //online.ldc.upenn.edu/login.html *To access Brown:* //secure.ldc.upenn.edu/intranet/
LAEME	//www.lel.ed.ac.uk/ihd/laeme1/laeme1.html
LAOS	//www.lel.ed.ac.uk/research/ihd/laos/laos.html
Linguistic Atlas Projects	//us.english.uga.edu/
MICASE	//lw.lsa.umich.edu/eli/micase/index.htm
NECTE	//www.ncl.ac.uk/necte/
ONZEminer	//www.ling.canterbury.ac.nz/jen/onzeminer/
Oxford Text Archive	//ota.ahds.ac.uk/
SCOTS	//www.scottishcorpus.ac.uk/
TalkBank	//talkbank.org/ *For SBC text files:* //talkbank.org/data/Conversation/ *For SBC audio:* //talkbank.org/media/Conversation/SBCSAE/
TEI	//www.tei-c.org/index.xml
Transcriber	//trans.sourceforge.net/en/presentation.php
WebCorp	//www.webcorp.org.uk

BIBLIOGRAPHY

Allen, Will, Beal, Joan, Corrigan, Karen, Maguire, Warren, and Moisl, Hermann 2007. 'A linguistic 'time capsule': The Newcastle Electronic Corpus of Tyneside English', in Beal et al. (eds.), Vol. 2, 16-48.

Andersen, Gisle 1997. '*They gave us these yeah, and they like wanna see like how we talk and all that*: The use of *like* and other discourse markers in London teenage speech', in Ulla-Britt Kostinas, Stenström, Anna-Brita and Karlsson, Anna-Malin (eds.), *Ungdomsspråk i Norden*. MINS 43. Stockholm: Stockholms universitet, Institutionene für nordiska språk. 83-95.

Andersen, Gisle 1998. 'The pragmatic marker *like* from a Relevance-Theoretic perspective', in Andreas Jucker and Ziv, Yael (eds.), *Discourse Markers: Descriptions and Theory*. Amsterdam: John Benjamins. 147-170.

Andersen, Gisle 2001. *Pragmatic Markers and Sociolinguistic Variation.* Amsterdam: John Benjamins.

Avery, Peter, Chambers, J.K., D'Arcy, Alexandra, Gold, Elaine, and Rice, Keren (eds.) 2006. *Canadian English in the Global Context.* Special issue of *Canadian Journal of Linguistics* 51.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. 1995. *The CELEX Lexical Database* (CD-ROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Baron, Naomi 2004. 'See you online: Gender issues in college student use of instant messaging', *Journal of Language and Social Psychology* 23: 397-423.

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

Bauer, Laurie 2002. 'Inferring variation and change from public corpora', in Chambers et al. (eds.), 97-114.

Beal, Joan C., Corrigan, Karen P. and Moisl, Hermann L. (eds.) 2007a. *Creating and Digitizing Language Corpora*. 2 volumes. Hampshire and New York: Palgrave Macmillan.

Beal, Joan C., Corrigan, Karen P. and Moisl, Hermann L. 2007b. 'Taming digital voices and texts: Models and methods for handling unconventional synchronic corpora', in Beal et al. (eds.), Vol.1, 1-16.

Biber, Douglas 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas 1993. 'Representativeness in corpus design', *Literary and Linguistic Computing* 8: 243-257.

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

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, and Finegan, Edward. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Buchstaller, Isabelle and D'Arcy, Alexandra 2009. Localized globalization: A multi-local, multivariate investigation of quotative *be like*. to appear in *Journal of Sociolinguistics*.

Chambers, J.K. 1994. 'An introduction to Dialect Topography,' *English World-Wide* 15: 35-53.

Chambers, J.K., Trudgill, Peter and Schilling-Estes, Natalie (eds.) 2002. *The Handbook of Language Variation and Change*. Malden and Oxford: Blackwell.

Cheshire, Jenny 1999. 'Taming the vernacular: Some repercussions for the study of syntactic variation and spoken grammar', *Cuadernos de Filología Inglesa* 8: 59-80.

Collins, Peter C. 2005. 'The modals and quasi-modals of obligation and necessity in Australian English and other Englishes,' *English World-Wide* 26: 249-273.

D'Arcy, Alexandra 2001. *Beyond Mastery: A Study of Dialect Acquisition*. Unpublished MA thesis. St. John's: Memorial University of Newfoundland.

D'Arcy, Alexandra 2005a. *Like: Syntax and Development*. Unpublished doctoral dissertation. Toronto: University of Toronto.

D'Arcy, Alexandra 2005b. 'The development of linguistic constraints: Phonological innovations in St. John's', *Language Variation and Change* 17: 327-355.

D'Arcy, Alexandra 2007. '*Like* and language ideology: Disentangling fact from fiction', *American Speech* 82: 386-419.

D'Arcy, Alexandra 2008. 'Canadian English as a window to the rise of *like* in discourse', *Anglistik: International Journal of English Studies* 19: 125-140.

Davies, Mark 2002–. *Online Corpora*, http://corpus.byu.edu/.

Dollinger, Stefan 2008. *New-Dialect Formation in Canada*. *Evidence from the English modal auxiliaries*. Amsterdam: Benjamins.

Denbo, Seth, Haskins, Heather, and Robey, David 2008. *Sustainability of Digital Outputs from AHRC Resource Enhancement Projects*. Report to the Arts and Humanities Research Council. December 2008. Available online at http://www.ahrcict.rdg.ac.uk/activities/review/sustainability.htm.

Facchinetti, Roberta, Krug, Manfred, and Palmer, Frank (eds.) 2003. *Modality in Contemporary English*. Berlin and New York: Mouton de Gruyter.
Fairon, Cédrick 2000. 'GlossaNet: Parsing a web site as a corpus', *Lingvisticae Investigationes* 22: 327-340.
Ferrara, Kathleen, Brunner, Hans and Whittemore, Greg 1991. 'Interactive written discourse as an emergent register', *Written Communication* 8: 8-34.
Francis, W. Nelson and Kučera, Henry 1964. *A Standard Corpus of Present-day Edited American English*. Providence: Brown University.
Fromont, Robert and Hay, Jennifer 2006. ONZE Miner: Development of a browser-based research tool. ms, Department of Linguistics, University of Canterbury. Available online at http://www.ling.canterbury.ac.nz/jen/onzeminer/
Garside, Roger 1987. 'The CLAWS word-tagging system', in Garside, Roger, Leech, Geoffrey, and Sampson, Geoffrey (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman. 30-41.
Gordon, Elizabeth, Campbell, Lyle, Hay, Jennifer, Maclagan, Margaret, Sudbury, Andrea, and Trudgill, Peter 2004. *New Zealand English. Its Origins and Evolution*. Cambridge: Cambridge University Press.
Gordon, Elizabeth, Maclagan, Margaret, and Hay, Jennifer 2007. 'The ONZE corpus', in Beal et al. (eds.), Vol.2, 82-104.
Greenbaum, Sidney 1992. 'A new corpus of English: ICE', In Svartvik, Jan (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter. 171-183.
Haslerud, Vibecke and Stenström, Anna-Brita 1995. 'The Bergen Corpus of London Teenage Language', in Leech, Geoffrey, Myers, Greg and Thomas, Jenny. (eds.), *Spoken English on Computer*. London: Longman. 235-242.
Holmes, Janet 1996. 'The New Zealand component of ICE: Some methodological challenges', in Greenbaum, Sidney (ed.), *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press. 163-181.
Hundt, Marianne, Nesselhauf, Nadja and Biewer, Carolin (eds.) 2007. *Corpus Linguistics and the Web*. Amsterdam and New York: Rodopi.
Johansson, Stig, Leech, Geoffrey, and Goodluck, Helen 1978. *Manual of Information to Accompany the Lancaster-Oslo-Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Department of English, Oslo University.
Kallen, Jeffrey and Kirk, John 2007. 'ICE-Ireland: Local variations on global standards', in Beal et al. (eds.), Vol.1, 121-162.
Kennedy, Graeme 1998. *An Introduction to Corpus Linguistics*. London and New York: Longman.
Kirk, John 1992. 'The Northern Ireland Transcribed Corpus of Speech', in Leitner, Gerhard (ed.) *New Directions in Language Corpora*. Berlin and New York: Mouton de Gruyter. 65-74.
Kretzschmar, William A., Anderson, Jean, Beal, Joan C., Corrigan, Karen P., Opas-Hänninen, Lisa Lena, and Plichta, Bartlomiej 2006. 'Collaboration on corpora for regional and social analysis', *Journal of English Linguistics* 34: 172-205.
Krug, Manfred 2000. *Emerging English Modals: A Corpus-based Study of Grammaticalization*. Berlin and New York: Mouton de Gruyter.
Kurath, Hans 1972. *Studies in Area Linguistics*. Bloomington: Indiana University Press.
Kurath, Hans, Hanley, M., Bloch, B., and Lowman, G. S., 1939-1943. *The Linguistic Atlas of New England*. 3 volumes in 6 parts. Providence: Brown University Press.
Kytö, Merja 1996. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*, 3rd edition. Helsinki: Department of English, University of Helsinki.
Kytö, Merja, Rudanko, Juhani, and Smitterberg, Erik 2000. 'Building a bridge between the present and the past: A corpus of 19th-century English', in *ICAME Journal* 24: 85-97. Available online at http://gandalf.aksis.uib.no/journal.html.
Labov, William 1963. 'The social motivation of a sound change', *Word* 19: 273-309.
Labov, William 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics [2nd edition, 2006, New York: Cambridge University Press].
Labov, William 1969. 'Contraction, deletion, and inherent variability of the English copula', *Language* 45: 715-762.
Labov, William 1972a. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.
Labov, William 1972b. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
Labov, William 1972c. *The Design of a Sociolinguistic Research Project*. Report of the Sociolinguistics Workshop, Central Institute of Indian Languages.

Labov, William 1984. 'Field methods of the project on linguistic change and variation', in Baugh, John and Sherzer, Joel (eds.), *Language in Use: Readings in Sociolinguistics*. Englewood Cliffs: Prentice-Hall. 28-54.
Labov, William, Ash, Sharon and Boberg, Charles 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.
Laing, Margaret and Lass, Roger 2007. *A Linguistic Atlas of Early Middle English, 1150-1325*. Edinburgh: The University of Edinburgh.
Lee, David 2001–. *Bookmarks for Corpus-based Linguists*, http://devoted.to/corpora.
Leech, Geoffrey 1993a. 'Corpus annotation schemes', *Literary and Linguistic Computing* 8: 275-281.
Leech, Geoffrey 1993b. '100 million words of English', *English Today* 9: 9-15.
Leech, Geoffrey 2000. 'Modality on the move: The English modal auxiliaries 1961-1992', in Facchinetti, Krug and Palmer (eds.), 223-240.
Ling, Rich and Baron, Naomi 2007. 'Text messaging and IM: Linguistic comparison of American college data', *Journal of Language and Social Psychology* 26: 291-298.
Macaulay, Ronald 1991a. ''Coz it izny spelt when they say it': Displaying dialect in writing', *American Speech* 66: 280-291.
Macaulay, Ronald 1991b. *Locating Dialect in Discourse: The Language of Honest Men and Bonnie Lassies in Ayr*. New York: Oxford University Press.
Macleod, Catherine, Ide, Nancy, and Grishman, Ralph 2000. 'The American National Corpus: A standardized resource for American English', *Proceedings of the Second Language Resources and Evaluation Conference*. Athens: Greece. 831-836.
MacWhinney, Brian 2000. *The CHILDES Project: Tools for Analyzing Talk.* 3rd ed. Mahwah: Erlbaum.
Meyer, Charles 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
Meyer, Charles 2004. 'Can you really study language variation in linguistic corpora?', *American Speech* 79: 339-355.
Meyer, Charles and Nelson, Gerard 2006. 'Data collection', in Aarts, Bas and McMahon, April (eds.), *The Handbook of English Linguistics*. Malden: Blackwell Publishing. 93-113.
Miller, Jim and Weinert, Regina 1995. 'The function of *like* in dialogue', *Journal of Pragmatics* 23: 365-393.
Milroy, James 1992. *Linguistic Variation and Change*. Oxford: Blackwell Publishers.
Milroy, Lesley 1987. *Observing and Analysing Natural Language*. Oxford: Blackwell Publishers.
Milroy, Lesley and Gordon, Matthew 2003. *Sociolinguistics. Method and Interpretation*. Malden and Oxford: Blackwell Publishing.
Morley, Barry 2006. 'WebCorp: A tool for online linguistic information retrieval and analysis', in Renouf, Antoinette and Kehoe, Andrew (eds.), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi. 283-296.
Nelson, Gerald 1996. 'Markup systems', in Sidney Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon. 36-53.
Ochs, Elinor 1979. 'Transcription as theory', in Ochs, Elinor and Schieffelin, Bambi (eds.), *Developmental Pragmatics*. New York, Academic. 43-72.
Poplack, Shana 1989. 'The care and handling of a mega-corpus: The Ottawa-Hull French Project', in Fasold, Ralph and Schiffrin, Deborah (eds.), *Language Change and Variation*. Amsterdam and Philadelphia: John Benjamins. 411-451.
Poplack, Shana 2007. 'Foreword', in Beal et al. (eds.), Vol. 1, ix-xiii.
Poplack, Shana and Tagliamonte, Sali A. 2001. *African American English in the Diaspora: Tense and Aspect*. Oxford: Blackwell Publishers.
Poplack, Shana, Walker, James, and Malcolmson, Rebecca 2006. 'An English 'like no other'?: Language contact and change in Quebec', in Avery et al. (eds.), 185-213.
Preston, Dennis 1985. 'The li'l abner syndrome: Written representations of speech', *American Speech* 60: 328-336.
Preston, Dennis 2000. 'Mowr and mowr bayud spellin': Confessions of a sociolinguist', *Journal of Sociolinguistics* 4: 614-621.
Quirk, Randolf 1968. 'The Survey of English Usage', in Quirk, Randolf, *Essays on the English Language: Medieval and Modern*. London: Longman.
Renouf, Antoinette 1993. 'A word in time: First findings from the investigation of dynamic text', in Aarts, Jan, de Haan, Pieter and Oostdijk, Nelleke (eds), *English Language Corpora: Design, Analysis and Exploitation. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992*. Amsterdam: Rodopi. 279-288.

Renouf, Antoinette 2003. 'WebCorp: Providing a renewable data source for corpus linguists', in Granger, Sylviane and Petch-Tyson, Stephanie (eds.), *Extending the Scope of Corpus-based Research: New Applications, New Challenges*. Amsterdam: Rodopi. 39-58.

Romaine, Suzanne and Lange, Deborah 1991. 'The use of *like* as a marker of reported speech and thought: A case of grammaticalization in progress', *American Speech* 66: 227-279.

Sankoff, David 1988. 'Sociolinguistics and syntactic variation', in Newmeyer, Frederick (ed.), *Linguistics: The Cambridge Survey*. Cambridge: Cambridge University Press. 140-161.

Sankoff, David 2005. 'Problems of representativeness', in Ammon, Ulrich, Dittmar, Norbert, Mattheier, Klaus, and Trudgill, Peter (eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*, 2nd edition. Vol. 2. Berlin: Mouton de Gruyter. 998-1002.

Sankoff, David and Sankoff, Gillian 1973. 'Sample survey methods and computer-assisted analysis in the study of grammatical variation', In Darnell, Regna (ed.), *Canadian Languages in their Social Context*. Edmonton: Linguistic Research Inc. 7-63.

Schneider, Edgar 2002. 'Investigating variation and change in written documents', in Chambers et al. (eds.), 67-96.

Shastri, S.V. 1988. 'The Kolhapur Corpus of Indian English and work done on its basis so far', *ICAME Journal* 12: 15-26.

Siegel, Muffy 2002. '*Like*: The discourse particle and semantics', *Journal of Semantics* 19: 35-71.

Simpson, Rita C., Briggs, Sarah L., Ovens, Janine and Swales, John M. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor: The Regents of the University of Michigan.

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

Sinclair, John 1992. 'The automatic analysis of corpora', in Svartvik, Jan (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm 1991*. Berlin: Mouton de Gruyter. 379-397.

Sinclair, John 1997. *English Grammar* (Collins COBUILD English Grammar). London: HarperCollins.

Smith, Nicholas 2003. 'Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English,' in Facchinetti, Krug and Palmer (eds.), 241-266.

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

Tagliamonte, Sali A. 1998. '*Was/were* variation across the generations: View from the city of York', *Language Variation and Change* 10: 153-191.

Tagliamonte, Sali A. 2002. 'Comparative sociolinguistics', in Chambers et al. (eds.), 729-763.

Tagliamonte, Sali A. 2006a. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Tagliamonte, Sali A. 2006b. 'So cool, right?: Canadian English entering the 21st century', in Avery et al. (eds.), 309-331.

Tagliamonte, Sali A. 2007. 'Representing real language: Consistency, Trade-offs, and thinking ahead!', in Beal, Corrigan and Moisl (eds.), Vol. 1, 205-240.

Tagliamonte, Sali A. and D'Arcy, Alexandra 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language* 85: 55-105.

Tagliamonte, Sali A. and Denis, Derek 2008. 'Linguistic ruin? LOL! Instant messaging and teen language', *American Speech* 83:3-34.

Tagliamonte, Sali A. and Smith, Jennifer 2006. 'Layering, competition and a twist of fate: Deontic modality in dialects of English', *Diachronica* 23: 341-380.

Underhill, Robert 1988. 'Like is like, focus', *American Speech* 63: 234-246.

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

Weinreich, Uriel, Labov, William and Herzog, Marvin 1968. 'Empirical foundations for a theory of language change', in Lehmann, Winfred P. and Malkiel, Yakov (eds.), *Directions for Historical Linguistics*. Austin: University of Texas Press. 97-188.

Williamson, Keith 2008. *A Linguistic Atlas of Older Scots, Phase 1: 1380-1500*. Edinburgh: University of Edinburgh.

Wynne, Martin (ed.) 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/.