International Corpus of English (CANADA)

John Newman (john.newman@ualberta.ca)
Department of Linguistics, University of Alberta

Celebration of the 30th Anniversary of the Social Sciences and Humanities Research Council of Canada, November 26 2008

Introduction

The International Corpus of English (ICE) is a global project which will collect samples of varieties of English as it is spoken and written around the world. Twenty research teams around the world are preparing electronic corpora of their own national or regional variety of English. The corpus will allow researchers to study variation in the use of the English language, e.g., Canadian English, Australian English, Singapore English.

ICE-CANADA is the Canadian component of ICE and the Department of Linguistics at the University of Alberta is the home of ICE-CANADA.

INTERNATIONAL CORPUS OF ENGLISH

Brief history

ICE began in 1990 with the intention of documenting English as used in the early 1990's. The project is based at University College, London, UK and is coordinated by Professor Gerald Nelson at The Chinese University of Hong Kong.

Most of the material for ICE-CANADA was collected in the early 1990's under the direction of Nancy Belmore at Concordia University, Montreal. The Strathy Language Unit, Queen's University, also contributed data. John Newman arranged for all data to be brought to the University of Alberta and took responsibility for further development of the corpus in January 2006.

Research teams in the following countries are co-operating in this project:

Australia	Ireland	Philippines
Canada	Jamaica	Singapore
East Africa	Malaysia	South Africa
Fiji	Malta	Sri Lanka
Great Britain	New Zealand	Trinidad & Tobago
Hong Kong	Nigeria	USA
India	Pakistan	

Design of the corpus

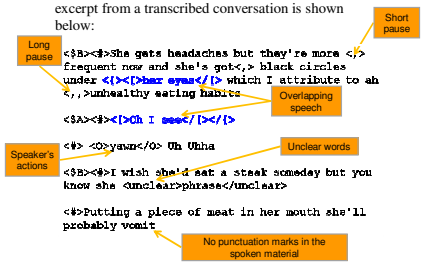
Each national corpus in ICE follows the same overall design: 500 "texts" of approximately 2,000 words each – a total of approximately one million words. "Texts" can be based on written sources (40%) or transcribed spoken language (60%). Table 1 summarizes the design of the spoken material.

Private dialogues 200,000	Conversations Phone-calls	180,000 20,000
Public dialogues 160,000	Class lessons Parliamentary debates Broadcast interviews Broadcast discussions Cross-examinations Business transactions	40,000 40,000 20,000 20,000 20,000 20,000
Unscripted monologues 140,000	Commentaries Unscripted speeches Demonstrations Legal presentations	40,000 60,000 20,000 20,000
Scripted monologues 100,000	Broadcast news Broadcast talks Non-broadcast talks	40,000 20,000 20,000

Table 1. Design of the spoken component of an ICE corpus. Numbers represent the word count for each category.

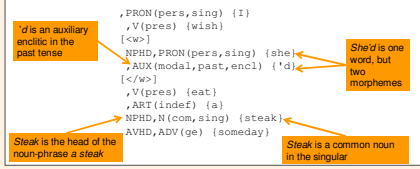
Transcription

Texts are marked up in SGML to encode sentence and paragraph boundaries, foreign words, pauses, etc. An excerpt from a transcribed conversation is shown below:



Tagging

All files will be tagged for part of speech (pronoun, noun, etc.), syntactic features (singular, plural, etc.), and grammatical relations. Here is an example of automatic tagging (using the TOSCA tagger) of *I wish she'd eat a steak someday*:



Research in progress

Core and periphery of world Englishes

Nelson (2006) has initiated a project identifying "core" and "periphery" of the lexicon and morphology of world Englishes, using completed ICE corpora. Core words and grammatical features occur in all the ICE corpora; peripheral ones only occur in some of the corpora:

Items in all 6 corpora **absolute core**	Items in 5 corpora	Items in 4 corpora	Items in 3 corpora	Items in 2 corpora	Items in 1 corpus only **absolute periphery**
zealand	*yup*	*yuri*	*zia*	*zoe*	*zuzana*
zero	*yuppie*	*zaire*	*zig-zag*	*zoned*	*zwerewa*
zinc	*yvonne*	*zeal*	*zilch*	*zoning*	*zwezda*
zone	*z*	*zebra*	*zimbabwe*	*zool*	*zvs*
zones	*zip*	*zeros*	*zipped*	*zoologists*	*zworykin*
zoo	*zoology*	*zero-sum*	*zonal*	*zoomed*	*zyx*
zoological	*zoom*	*zig*	*zonation*	*zooms*	*zyxton*
zooming	*zoos*	*zulu*	*zurich*	*zubin*	*zz*

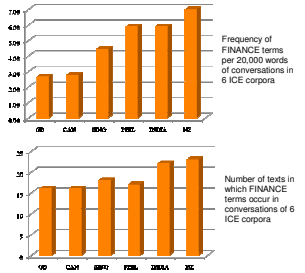
Table 2. Examples of words in the core and periphery of world Englishes using Hong Kong, Great Britain, New Zealand, India, Singapore, and Philippines ICE corpora. The completion of ICE-CANADA will allow Canadian English to be included in this kind of study too. *Table from Nelson (2006: 121).*

The *eh* tag

Columbus (2008a, 2008b) has compared the frequency of use of the *eh* tag in Canadian English with New Zealand English using ICE, as in *He's funny eh I wish I had videoed that when he was singing*. *eh* is regarded as distinctive in conversational style of both varieties. However, the relative frequency of *eh* differ in the two varieties: New Zealand *eh* occurs 595 times while Canadian *eh* occurs 143 times in the 200,000 words of private dialogues in each corpus. *eh* is much more distinctive in New Zealand English than in Canadian English. Columbus has also found different preferences in the way *eh* is used in the two varieties.

Content analysis

Columbus & Newman (in preparation) is investigating topic preferences in the face-to-face conversations of ICE-CANADA and the other ICE corpora. The authors are using sets of domain-specific vocabulary (finance, education, sport, etc.) as search words to establish the relative frequencies of these topics. A pilot study, using a set of 15 common FINANCE terms, suggests that ICE-CANADA contains relatively little conversation about FINANCE:



Future development

The collection and transcription of materials for ICE-CANADA is expected to be completed in 2009. Some necessary data was missing from the material originally collected in the 1990's, e.g., audio recordings of Parliamentary debates, sports commentaries, and broadcast talks. Locating and obtaining the audio for these genres has caused some delays. The guidelines for ICE require that all material should have been produced, originally, in the 1990's.

When completed, ICE-CANADA, along with the other ICE corpora, will be made available to bona fide language researchers. Future development may include conversion of the corpus to XML and uploading of the corpus to the world-wide web. ICE-CANADA would then be available for study as an online searchable website. This phase of development will depend upon additional funding being awarded to the project.

The ICE project offers a unique opportunity to study the state of world Englishes at one point in time. It is only through the acceptance of a common design for the corpora and common guidelines for the transcription and markup of the corpora that a meaningful quantitative comparison of varieties of English can be undertaken.

Literature cited

Columbus, G. (2008a). "Ah lovely stuff, eh?" On invariant tag meanings and usage across three varieties of English...plus one. Paper presented at Methods XIII, University of Leeds, Leeds, United Kingdom, August 11-15, 2008.

Columbus, G. (2008b). Nice day, eh? A comparative look at Canadian and New Zealand *eh*. Paper presented at ACOL (Alberta Conference on Linguistics), Banff Springs, Alberta, October 24, 2008.

Columbus, G. & J. Newman (in preparation). A comparative analysis of discourse topics through lexical frequency. Paper to be presented at the workshop on Corpus, Colligation, Register Variation, Deutsche Gesellschaft für Sprachwissenschaft 2009.

Nelson, G. (2006). The core and periphery of world Englishes: A corpus-based exploration. *World Englishes* 25.1: 115–129.

Acknowledgments

Research on ICE-Canada at the University of Alberta is supported by SSHRC award # 410-2007-0388, a Killam Cornerstones award, funding from the Faculty of Arts and the Department of Linguistics, and the CFI-funded digital humanities infrastructure TAPoR (Text Analysis Portal for Research).