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.

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 (60%) or transcribed spoken language (40%). Table 1 summarises the design of the spoken material.

<table>
<thead>
<tr>
<th>Text type</th>
<th>Number of texts</th>
<th>Conversational phone calls</th>
<th>Public dialogues (80,000)</th>
<th>Unscripted monologue (48,000)</th>
<th>Scripted monologue (100,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focus dialogues</td>
<td>20,000</td>
<td>10,800</td>
<td>28,800</td>
<td>28,800</td>
<td>28,800</td>
</tr>
<tr>
<td>Public dialogues</td>
<td>80,000</td>
<td>48,800</td>
<td>48,800</td>
<td>28,800</td>
<td>28,800</td>
</tr>
<tr>
<td>Unscripted mono</td>
<td>48,000</td>
<td>48,800</td>
<td>28,800</td>
<td>28,800</td>
<td>28,800</td>
</tr>
<tr>
<td>Scripted mono</td>
<td>100,000</td>
<td>48,800</td>
<td>28,800</td>
<td>28,800</td>
<td>28,800</td>
</tr>
</tbody>
</table>

Table 1: Design of the spoken component of an ICE corpus. Numbers are estimates from each country.

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.

<table>
<thead>
<tr>
<th>Language</th>
<th>Core words</th>
<th>Periphery words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>10,000</td>
<td>2,000</td>
</tr>
<tr>
<td>French</td>
<td>8,000</td>
<td>3,000</td>
</tr>
<tr>
<td>German</td>
<td>6,000</td>
<td>4,000</td>
</tr>
<tr>
<td>Spanish</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Italian</td>
<td>4,000</td>
<td>6,000</td>
</tr>
</tbody>
</table>

Table 2: Core and periphery words in ICE corpora.

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, audio advertisements.

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 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.

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 Strathclyde 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 collaborating 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

Transcription

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

```
<participant id="away" name="Away"> Away</participant>
<participant id="here" name="Here"> Here</participant>

Away: How are you doing today?
Here: Not bad, thanks.
Away: What did you do today?
Here: I went to the library.
Away: What did you do there?
Here: I read a book.
Away: What kind of book?
Here: A mystery novel.
```

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 had out of a steak-sandwich:

```
I wish she had out of a steak-sandwich.
```

<table>
<thead>
<tr>
<th>Tagging</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>wish</td>
<td>wish</td>
</tr>
<tr>
<td>she</td>
<td>had</td>
<td>had</td>
</tr>
<tr>
<td>out</td>
<td>of</td>
<td>of</td>
</tr>
<tr>
<td>a</td>
<td>steak-sandwich</td>
<td>a steak-sandwich</td>
</tr>
</tbody>
</table>

Acknowledgments

Research on ICE-CANADA at the University of Alberta is supported by SSHRC, award #406-2007-0386, a Killam Commonwealth award, funding from the Faculty of Arts and the Department of Linguistics, and the CFI-funded digital humanities infrastructure: TAPoR (Text Analysis Portal for Researchers).

Literature cited


