Using corpora as an innovative tool to compare varieties of English around the world: the International Corpus of English story

Josef Schmied
English Language & Linguistics
Chemnitz University of Technology
http://www.tu-chemnitz.de/phil/english/Schmied
Motivation and structure

1. The (hi)story of ICE
   - beginnings and concepts (ICE0)
   - ICE problems

2. The present-day status of ICE
   - resources
   - work in progress

3. Case studies: modalities

4. The future of ICE (ICE2)
   - changing communication patterns 1989 - 2009
   - new practical opportunities
   - new theoretical challenges

Conclusion
1. The (hi)story of ICE

a personal account

1.1 Beginnings and concepts

May 1988 ICAME Birmingham:
Schmied: “Compiling a Corpus of East African English”
discussion on Brown/LOB categories?
more sociolinguistics variable (gender, status, age, 1st language, etc.)
Oct. 1988 proposal:
Greenbaum, Sidney. “A proposal for an international computerised corpus of English. World Englishes 7, 315
1.1 ICE0 concepts

Greenbaum (1988: 31):

1. to sample standard varieties from other countries where English is the first language, for example Canada and Australia

2. to sample national varieties from countries where English is an official additional language, for instance India and Nigeria; and

3. to include spoken and manuscript English as well as printed English.
1.1 Discussing the corpus design

Schmied (1990):
“Corpus-linguistics and the nativization of English”. *World Englishes* 9, 255-268

“corpus-compilation paradox”:
A “national” corpus should contain culture-specific text(type)s, but we can only identify them through corpus analysis
1.2 ICE problems

corpus-compilation:
1) funding (e.g. ICE-US, ICE-Nigeria)
2) adaptations in corpus compilation: technology and culture
3) copy-right for distribution
4) corpus processing:
   annotation and parsing

corpus-application:
1) manuals for restriction (interpretation)
2) query (WordSmith or AntConc) – statistics individual solutions
1.2.2 Adaptation

representativeness vs. comparability

McEnery/Wilson 1996 Edinburgh U.P.
http://www.lancs.ac.uk/fss/courses/ling/corpus/Corpus2/2FRA1.HTM

4 characteristics of the modern corpus:
- Sampling and representativeness
- Finite size
- Machine-readable form
- A standard reference
Appendix 6: List of written texts from Tanzania (word count)

**Informational: Learned**
- Humanities: W2A001T – W2A010T 20.172
- Natural Sciences: W2A021T – W2A027T 20.114
- Technology/Agriculture/Environmental dev.: W2A031T – W2A040T 20.148
  - Total: 80.585

**Informational: Popular**
- General: W2BGEN1T – W2BGEN8T 13.789
  - Total: 80.752

**Informational: Reportage**
- Splash: W2C001T – W2C0010T 20.018
- Reportage/Features: W2C011T – W2C020T 20.139
  - Total: 40.157

**Instructional**
- Administrative/Regulatory: W2D001T – W2D010T 20.120

**Persuasive**
- Institutional: W2E001T – W2E010T 20.078
- Personal Column: W2E011T – W2E020T 20.125
  - Total: 40.203

---

**from ICE- East Africa handbook**

Motivation | ICE History | ICE status | Case studies | ICE2future | Conclusion |
---|---|---|---|---|---|
8/40 |
1.2.2 Spoken text categories in ICE corpora

SPOKEN (300)

Dialogues (180)
- Private (100)
  - Face-to-face conversations (90)
  - Phonecalls (10)
- Public (80)
  - Classroom Lessons (20)
  - Broadcast Discussions (20)
  - Broadcast Interviews (10)
  - Parliamentary Debates (10)
  - Legal cross-examinations (10)
  - Business Transactions (10)

Monologues (120)
- Unscripted (70)
  - Spontaneous commentaries (20)
  - Unscripted Speeches (30)
  - Demonstrations (10)
  - Legal Presentations (10)
- Scripted (50)
  - Broadcast News (20)
  - Broadcast Talks (20)
  - Non-broadcast Talks (10)
## 1.2.2 ICE categories and ICE-EA/Ke/Tz

<table>
<thead>
<tr>
<th>Category</th>
<th>ICE</th>
<th>Ke</th>
<th>Tz</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPOKEN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dialogue</td>
<td>180</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>(written as spoken)</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>100</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Direct conv.</td>
<td>90</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Distanced conv.</td>
<td>10</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>80</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td><strong>WRITTEN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Press editorials</td>
<td>10</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Institutional</td>
<td>--</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Personal columns</td>
<td>--</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
1.2.4 Textual Markup

In written texts, features of the original layout are marked, including sentence and paragraph boundaries, headings, deletions, and typographic features.

Spoken texts are transcribed orthographically, and are marked for pauses, overlapping strings, discourse phenomena such as false starts and hesitations, and speaker turns.

The markup manual is available here.
1.2.4 Wordclass Tagging

ICE texts are automatically tagged for wordclass by the ICE Tagger, developed by Sean Wallis at the Survey of English Usage, University College London. This assigns wordclass tags to each lexical item in the corpus. The tagset has been developed especially for ICE, and is largely based on Quirk et al (1985) *A Comprehensive Grammar of the English Language.*

e.g. Each PRON(univ,sing) of PREP(ge) these PRON(dem,plu) is V(cop,pres) the ART(def) responsibility N(com,sing) of PREP(ge) one NUM(card,sing) person N(com,sing)
1.2.4 Syntactic Parsing

Every sentence in the corpus is analysed at phrase, clause, and sentence level, and the analysis is shown in the form of a parse tree:

```
Each of these is the responsibility of one person
```
## 2. The present-day status of ICE

resources:

1. WWW for ICE and Ice-corpora
2. corpora available
3. publications
The International Corpus of English (ICE) began in 1990 with the primary aim of collecting material for comparative studies of English worldwide. Twenty research teams around the world are preparing electronic corpora of their own national or regional variety of English. Each ICE corpus consists of one million words of spoken and written English produced after 1989. For most participating countries, the ICE project is stimulating the first systematic investigation of the national variety. To ensure compatibility among the component corpora, each team is following a common corpus design, as well as a common scheme for grammatical annotation.

**News 23 October 2009**

The annual ICE Workshop was held in Cebu, The Philippines, on 21 October 2009. The programme is here; progress reports and a report on the round-table discussion will be available soon.
Sample Sound Files @ ICE-corpora.net

Sample sound files are available on this website for the following ICE components:

- Australia
- East Africa (Kenya & Tanzania)
- Great Britain
- Hong Kong
- India
- Jamaica
- The Philippines

The sound files are MPEG layer-3 .mp2 files.
INTERNATIONAL CORPUS OF ENGLISH

East Africa (Kenya & Tanzania) (ICE-EA)@ICE-corpora.net

The East African component of ICE includes data from Kenya and Tanzania. The corpus was first released in lexical form in 1999. Release 1 is available on CD-ROM as part of the ICAME Collection of English Language Corpora (2nd edn). Release 2 (2002) includes a version for use with WordSmith and full documentation. It is available here.

Contact:
Professor Dr Josef Schmied
English Language & Linguistics
Technische Universität
P. O. Box 954
D-09009 Chemnitz
Germany
Email: josef.schmied@phil.tu-chemnitz.de

Kenya
In the multilingual Kenyan context, which has an estimated 40 indigenous languages, English and Kiswahili are co-official languages. English is used in the business sector, in parliamentary debates and politics, education, the civil service, the legal system, the armed forces, the police, and the media, as well as in international communication. Population: 28 million.

Tanzania
Tanzania has an estimated 100 indigenous languages, but Kiswahili is the main official language. English is widely used in areas such as banking, tourism, commerce, the media, and in post-primary education. It is also the language of international communication. Population: 32 million.

Reading

Links
Online Newspapers
East Africa Standard
Africa Online

Online Radio Stations
Africa FM

© 2009 The ICE Project

Project website
http://www.tu-chemnitz.de/phil/english/realeafrica/

Listen to East African English
A selection of sound files from the corpus
2.1 ICE webpage: Corpus design

- 500 files à 2,000-word texts in specific categories
- The texts in the corpus date from 1990 or later. The authors and speakers of the texts are aged 18 or above, were educated through the medium of English, and were either born in the country in whose corpus they are included, or moved there at an early age and received their education through the medium of English in the country concerned.
- The corpus contains samples of speech and writing by both males and females, and it includes a wide range of age groups. The proportions, however, are not representative of the proportions in the population as a whole: women are not equally represented in professions such as politics and law, and so do not produce equal amounts of discourse in these fields. Similarly, various age groups are not equally represented among students or academic authors.
### 2.2 Currently available ICE corpora

<table>
<thead>
<tr>
<th>1st Language (ENL)</th>
<th>2nd Language (ESL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great Britain</td>
<td>East Africa* (Ke/Tz)</td>
</tr>
<tr>
<td>Ireland*</td>
<td>Hong Kong*</td>
</tr>
<tr>
<td>Jamaica*</td>
<td>India*</td>
</tr>
<tr>
<td>New Zealand</td>
<td>The Philippines*</td>
</tr>
<tr>
<td></td>
<td>Singapore*</td>
</tr>
</tbody>
</table>

* free as download
2.3 Publications
3. ICE case studies: modalities

hypotheses for comparing ICE corpora:
- English auxiliaries are very unevenly distributed in actual language usage
- epistemic use is more frequent than deontic
- ENL varieties: “American innovativeness, British conservativism and Australian independence from both” (Collins 2009)
- ESL varieties have a smaller number of modal auxiliaries than ENL varieties: ICE-T has smaller frequencies than ICE-K generally and for epistemic usage specifically
Modal auxiliaries in Tanzania and Kenya
### Case study: Modal auxiliaries in ICE-GB, ICE-Phil and ICE-EA (-K/ -T)

<table>
<thead>
<tr>
<th>Modal auxiliary</th>
<th>Total per corpus</th>
<th>Per million words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICE-GB</td>
<td>ICE-Phil/10</td>
</tr>
<tr>
<td>can</td>
<td>3574</td>
<td>425</td>
</tr>
<tr>
<td>could</td>
<td>1635</td>
<td>130</td>
</tr>
<tr>
<td>may</td>
<td>1219</td>
<td>120</td>
</tr>
<tr>
<td>might</td>
<td>693</td>
<td>45</td>
</tr>
<tr>
<td>must</td>
<td>687</td>
<td>55</td>
</tr>
<tr>
<td>shall</td>
<td>222</td>
<td>30</td>
</tr>
<tr>
<td>should</td>
<td>1117</td>
<td>100</td>
</tr>
<tr>
<td>will</td>
<td>2841</td>
<td>505</td>
</tr>
<tr>
<td>would</td>
<td>3037</td>
<td>270</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>15025</td>
<td>1680</td>
</tr>
</tbody>
</table>
Core modal aux per million words in ICE-GB, -Phil, -K, -T
Results:
Deontic and epistemic modals compared
Summary modal auxilaries

- hypotheses confirmed:
  English in Kenya has developed further towards a New English variety
  than English in Tanzania

- exceptions (like shall) can be explained:
  English in Tanzania is the more formal variety
  (informal texts are expressed more often in Kiswahili)

- further analyses:
  - simple distinction epistemic vs. deontic
    is not always detailed enough:
      e.g. lexeme-specific cases like habitual or historic would
  - n-grams / collocations, negation, etc.
4. The future of ICE (ICE2)

1) a broader corpus basis:
   - diachronic corpus: cf. Brown family
   - larger monitor-corpus: web-based

2) more comparative studies:
   - applied issues:
     - acknowledgment of new norms
     - replacing the native speaker
   - theoretical issues:
     - dynamic model: reanalysis
     - New English sub-categories: deleters vs. preservers
4.0 Changes in international communication since 1989

- “global” communication through internet, esp. WWW, chats, blogs replaces snail-mail / letters, ?? or additional categories?
- English as a “global” language:
  EIL = English as an International Language
  esp. European Union, China
  but “(secondary) education through the medium of English”?
4.1 Diachronic changes to ICE categories

ICE-EA 1990 – 2010:
replacing categories: email
full 1 Million words each

integrated in a larger monitor corpus
4.1.2 Web monitor corpus: keywords approach

**Business letters:** Dear, Yours sincerely/faithfully/truly/etc, invoice, memo, fax, bank, account, financial, enquiries, thank you for, manager, secretary, order, PO Box, date, I enclose, enclosed, I look forward to, c.c.

**Popular natural science:** popular, everyday, environment, diet, disease, plants, animals, reptiles, medicine, health, birds, fish, whales, conservation, zoos, natural history, green issues” rainforest, everything you need to know about..., Guide to..., made easy, global warming, wildlife, botanical, ozone layer.

**Administrative writing:** policy, regulations, procedures, guide, benefits, grants, entitlements, Guide to..., University calendar, safety, register/registration, code of conduct, license/licensing, health regulations, FAQ
4.1.2 Webcrawler

- Customisable, to exclude unwanted files, e.g. images, sounds, movies, .exe. Customised settings can be saved in an “options” file, [icelite.opt]
- Fast: can download entire websites in a relatively short time (depending on the size of the site)
- Stable: it never crashed, even when the download was aborted.
- Can be run ‘in the background’, and won’t interfere with other processes.
- Can be run overnight, and will safely switch off your PC.
- Inserts **time & date accessed** in each downloaded file.

courtesy Nelson, Gerry 2009 ICE-light
4.1.2 Workflow for Monitor Corpus

Use Google Advanced Search to identify major English-language sites in each domain

Use HTrack to download sites

Select texts, and record details in a spreadsheet

Targeted search to fill gaps, using Keywords

courtesy Nelson, Gerry 2009 ICE-light
4.2.1 Applied issues

- Comparative studies and a comparative database help decide questions of usage and norm.
- Quantitative comparisons allow more gradient usage decisions than native speakers.
4.2.2 Theoretical issues: dynamic model - new categorisation?

- evolutionary stages:
  - foundation
  - exonormative stabilisation
  - nativisation
  - endonormative stabilisation
  - differentiation

ICE-K > ICE-T; ICE-Sgp > ICE-My
4.2.2 Theoretical issues: deleters vs. preservers?

Mesthrie/Bhatt 2008:90

“One such broad dichotomy involves varieties that favour deletion of elements and those that disfavour it. In this regard the differences between Sgp Eng (especially amongst those with Chinese substrates) and African varieties are striking.”

(91): *Come what may (come).*

*He made me (to) do it.*

*As you know (that) I am from the Ciskei.*

EAsia vs. Africa?

ICE-Phil/-Sgp vs. ICE-EA/-ZA/-Nig
5. Conclusion

ICE-corpora are a good basis for empirical “national” and comparative corpus work “East-Africanism” (marked <ea>):

- in grammar (modality),
- lexicon (*matatu*)
- morphology/idiomaticity (*grass roots*)
- etc.
MOTIONS
REVISION OF WAGE GUIDELINES

Mr. Anyona: Mr. Speaker, Sir, I beg to move the following Motion:

THAT, in view of the recent economic changes which have affected prices of basic consumer goods, services and rents, and in view of the effects of liberalization process of the economy, this House resolves that:

... 

Mr. Speaker, Sir, the liberalization of the economy has set in motion a very high rate of inflation. The prices of all basic commodities like sugar, maize, meal, bread, and milk have soared. The house rents and transportation charges have also soared. If you look around town in the morning and afternoon, you would find that almost the entire population of this town walk to and from their places of work. This is because they are not able to afford the cost of transportation. They cannot afford the Matatu and Bus fares, so they can only walk to work.

matatu pl. ~s N ‘collective taxi’ in EAfr., esp. Kenya usu. licensed for fixed routes of public transport, but flexible, they leave when ‘full’; infamous for reckless driving and overcrowding; etym. <Sw. three, orig. 3 Shs fare coll.: agent driver, tout, operator, passenger; loc. park, stand, stage, stop, Pin, on board a ~; V enter, board.
**ICE-EA/ ESL**

idioms are less fixed/ more flexible

<table>
<thead>
<tr>
<th>Kenya</th>
<th>Tanzania</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Σ</td>
<td>written</td>
</tr>
<tr>
<td>grassroots</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>grass roots</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>grass root</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Σ</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>
**Current issues**

- Can results of corpus-linguistic analyses of “real usage” help to decide choices of norm and standards on a national and international basis?  
  YES
- Can “objective” corpus-linguistic resources replace “subjective” native speaker intuition as a neutral international standard?  
  YES
- Can corpus analyses add the cognitive dimension to variety formation?  
  YES


Using corpora as an innovative tool to compare varieties of English around the world: the ICE story

This contribution presents the story of the International Corpus of English (ICE) from the earliest discussions in 1989 to the current project applications and the big issues of the future.

The ICE teams in the past have always worked independently, so that some corpora were finished early, others were processed with the help of taggers and parsers, others were given up and restarted several times. Although modern computing facilities and processing tools (like AntConc) have made simple data analyses relatively easy for everyone, the most challenging innovations are possible in corpus collection. Nowadays, the WWW can facilitate some data collection and offers new options for multimedia corpora. This will make future corpus compilation and exploitation a real challenge.

The big issues in applying ICE corpora in theory and practice are, for instance:

- Can results of corpus-linguistic analyses of “real usage” help to decide choices of norm and standards on a national and international basis?
- Can corpora justify national varieties of English all around the world?
- Can corpus-linguistic resources replace the native speaker as a neutral international standard?
- Can corpus analyses add the cognitive dimension to variety formation?

This contribution will illustrate that the International Corpus of English can, despite some shortcomings, be used as a very innovative tool in exploring geographic varieties for researchers of every level.