The Story of ICE

= measuring, evaluating & interpreting culture-specific variation in a global sociolinguistic context
= “collecting & analysing stratified samples of official Englishes world-wide”

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

- The (hi)story of ICE1 (1989-2009)
  - beginnings and concepts
  - ICE problems
- The present-day status of ICE
  - resources
  - work in progress
- Case study: modalities
- The future of ICE (ICE2)
  - changing communication patterns 1989 -2009
  - new practical opportunities
  - new theoretical challenges
1 The (hi)story of ICE (1989-2009)

a personal account

1.1 Beginnings and concepts

May 1988 ICAME Birmingham:
Schmied, Josef. "Compiling a Corpus of East African English"
discussion of Brown/LOB categories?
more sociolinguistics variables (gender, status, age, 1st language, etc.)

October 1988 proposal:
Greenbaum, Sidney. "A proposal for an international computerised corpus of English" World Englishes 7, 315
1.1 ICE 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
1.2.1 Adaptation

representativeness vs. comparability

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)

#### PRINTED

<table>
<thead>
<tr>
<th>Category</th>
<th>Range</th>
<th>Total Word Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Informational: Learned</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities</td>
<td>W2A001T – W2A010T</td>
<td>20.172</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>W2A011T – W2A020T</td>
<td>20.151</td>
</tr>
<tr>
<td>Natural Sciences</td>
<td>W2A021T – W2A027T</td>
<td>20.114</td>
</tr>
<tr>
<td>Technology/Agriculture/Environmental dev.</td>
<td>W2A031T – W2A040T</td>
<td>20.148</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td></td>
<td><strong>80.585</strong></td>
</tr>
<tr>
<td><strong>Informational: Popular</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities</td>
<td>W2B001T – W2B010T</td>
<td>20.133</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>W2B011T – W2B020T</td>
<td>20.223</td>
</tr>
<tr>
<td>Natural Sciences</td>
<td>W2B021T – W2B024T</td>
<td>6.542</td>
</tr>
<tr>
<td>Technology/Agriculture/Small Industry</td>
<td>W2B031T – W2B040T</td>
<td>20.065</td>
</tr>
<tr>
<td>General</td>
<td>W2BGEN1T – W2BGEN8T</td>
<td>13.789</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td></td>
<td><strong>80.752</strong></td>
</tr>
<tr>
<td><strong>Informational: Reportage</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splash</td>
<td>W2C001T - W2C0010T</td>
<td>20.018</td>
</tr>
<tr>
<td>Reportage/Features</td>
<td>W2C011T - W2C020T</td>
<td>20.139</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td></td>
<td><strong>40.157</strong></td>
</tr>
<tr>
<td><strong>Instructional</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administrative/Regulatory</td>
<td>W2D001T – W2D010T</td>
<td>20.120</td>
</tr>
<tr>
<td><strong>Persuasive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institutional</td>
<td>W2E001T – W2E010T</td>
<td>20.078</td>
</tr>
<tr>
<td>Personal Column</td>
<td>W2E011T – W2E020T</td>
<td>20.125</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td></td>
<td><strong>40.203</strong></td>
</tr>
</tbody>
</table>
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><strong>private</strong></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><strong>public</strong></td>
<td>80</td>
<td>100</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></td>
</tr>
<tr>
<td>personal columns</td>
<td>--</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
1.2.3 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)
1.2.5 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:
2 The status of ICE today

Resources ICE1:
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 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.
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 CDROM 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 904
D-09009 Chemnitz
Germany
Email: josef.schmied@phil.tu-chemnitz.de

Dr. Diana Hudson-Etule
Sprachenzentrum
Universität Bayreuth
D-95440 Bayreuth
Germany
Email: diana.hudson-etule@uni-bayreuth.de
diana.hudson-etule@t-online.de

Project website
Http://www.tu-chemnitz.de/phil/english
/res/eastrica/

Kenya
In the multilingual Kenyan context, which has an estimated 40 indigenous languages, English and Kikuyu 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: 29 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
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

1st Language (ENL)
- Great Britain
- Ireland*
- Jamaica*
- New Zealand

2nd Language (ESL)
- East Africa* (Ke/Tz)
- Hong Kong*
- India*
- The Philippines*
- Singapore*

* free as down-load
2.3 Publications
<table>
<thead>
<tr>
<th>Country</th>
<th>Medium</th>
<th>Text-type</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenya</td>
<td>written</td>
<td>printed</td>
<td>private / conversations</td>
</tr>
<tr>
<td>Tanzania</td>
<td>written as spoken</td>
<td>non-printed</td>
<td>private / telephone calls</td>
</tr>
<tr>
<td>India</td>
<td>spoken</td>
<td>dialogue</td>
<td>public / class lessons</td>
</tr>
<tr>
<td>New Zealand</td>
<td></td>
<td>monologue</td>
<td>public / broadcast discussions</td>
</tr>
<tr>
<td>Philippines</td>
<td></td>
<td>parliamentary / debates</td>
<td>public / broadcast interviews</td>
</tr>
<tr>
<td>Singapore</td>
<td></td>
<td>legal / cross-examinations</td>
<td>public / parliamentary debates</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>public / legal cross-examinations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>public / business transactions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unscripted / spontaneous commentaries</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unscripted / unscripted speeches</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unscripted / demonstrations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unscripted / legal presentations</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>scripted / broadcast news</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>scripted / broadcast talks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>scripted / non-broadcast talks</td>
</tr>
</tbody>
</table>

http://www.tu-chemnitz.de/phil/english/Schmied
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
# Table: Modal auxiliaries in ICE-GB, ICE-Phil and ICE-EA (-KT/-TZ)

<table>
<thead>
<tr>
<th>Modal auxiliary</th>
<th>Total per corpus</th>
<th>Per million words</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICE-GB</td>
<td>ICE-Phil</td>
<td>ICE-KE</td>
</tr>
<tr>
<td>can</td>
<td>3574</td>
<td>425</td>
<td>2212</td>
</tr>
<tr>
<td>could</td>
<td>1635</td>
<td>130</td>
<td>1485</td>
</tr>
<tr>
<td>may</td>
<td>1219</td>
<td>120</td>
<td>1143</td>
</tr>
<tr>
<td>might</td>
<td>693</td>
<td>45</td>
<td>249</td>
</tr>
<tr>
<td>must</td>
<td>687</td>
<td>55</td>
<td>652</td>
</tr>
<tr>
<td>shall</td>
<td>222</td>
<td>30</td>
<td>159</td>
</tr>
<tr>
<td>should</td>
<td>1117</td>
<td>100</td>
<td>1155</td>
</tr>
<tr>
<td>will</td>
<td>2841</td>
<td>505</td>
<td>2011</td>
</tr>
<tr>
<td>would</td>
<td>3037</td>
<td>270</td>
<td>1496</td>
</tr>
<tr>
<td>Total</td>
<td>15025</td>
<td>1680</td>
<td>10562</td>
</tr>
</tbody>
</table>
Modal auxiliaries in Tanzania and Kenya

<table>
<thead>
<tr>
<th>Motivation</th>
<th>ICE History</th>
<th>ICE status</th>
<th>Case studies</th>
<th>ICE2future</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

![Bar chart showing comparison between Kenya and Tanzania for different modal auxiliaries.](chart.png)
Core modal auxiliaries per million words in ICE-GB, -Phil, -KE, -TZ
Results ICE-EA: Deontic and epistemic modals compared

- **Background**
- **Concepts**
- **Examples**
- **Hypotheses**
- **Results**
- **Interpretation**
- **Conclusion**

<table>
<thead>
<tr>
<th>Modals</th>
<th>Deontic</th>
<th>Epistemic</th>
</tr>
</thead>
<tbody>
<tr>
<td>can</td>
<td>60%</td>
<td>77%</td>
</tr>
<tr>
<td>could</td>
<td>77%</td>
<td>86%</td>
</tr>
<tr>
<td>may</td>
<td>77%</td>
<td>80%</td>
</tr>
<tr>
<td>might</td>
<td>89%</td>
<td>83%</td>
</tr>
<tr>
<td>must</td>
<td>11%</td>
<td>34%</td>
</tr>
<tr>
<td>shall</td>
<td>20%</td>
<td>17%</td>
</tr>
<tr>
<td>should</td>
<td>23%</td>
<td>14%</td>
</tr>
<tr>
<td>will</td>
<td>73%</td>
<td>27%</td>
</tr>
<tr>
<td>would</td>
<td>40%</td>
<td>17%</td>
</tr>
</tbody>
</table>
Summary modal auxiliaries

- **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 HTTrack 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.
S1B058HK

**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** /ˈmætətə/ n ‘collective taxi’ in E Afr., 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>25</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**
References


<table>
<thead>
<tr>
<th>Site</th>
<th>Title</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>br-intK.txt</td>
<td>S1BO46K</td>
<td>many reasons apart of what I've said &lt;A&gt; When these &lt;ea/&gt;matatu drivers and the touts are right in town they switch ...</td>
</tr>
<tr>
<td></td>
<td>S1BINT2K</td>
<td>before he waits for the bus to stop or the &lt;ea/&gt;matatu stop she jumps off from this bus This is the ... taking care you should not until you arrive &lt;O&gt; &lt;$A&gt; &lt;ea/&gt;matatu drivers have a habit of blocking buses here in Nairobi ...</td>
</tr>
<tr>
<td>br-newsK.txt</td>
<td>S2B012AK</td>
<td>said the injured policeman was rushed to hospital by a &lt;ea/&gt;matatu driver Ministry of Information and Broadcasting staff have been urged ...</td>
</tr>
<tr>
<td></td>
<td>S2B017K</td>
<td>hundred and forty-five Two children were knocked down by a &lt;ea/&gt;matatu along Mombasa Malindi road last Friday The driver sustained minor ...</td>
</tr>
<tr>
<td>br-talkK.txt</td>
<td>S2B031K</td>
<td>risk divorce or worse Half of the passengers in the &lt;ea/&gt;matatu were women and most of them were not amused by ... the people has demonstrated by the women's boycott of the &lt;ea/&gt;matatu not only enjoy the music they hear but also listen ...</td>
</tr>
</tbody>
</table>
Collocate Finder

What is a Collocate Finder?

Select a corpus: ICE_all

Please enter a node word:
Left Span  Node Word  Right Span
4  matatu  4

Cutoff 1

Stopwords

All stopwords are excluded from the results by default. Please check the boxes beside any stopwords you want to include:

- abbreviation
- adjective
- adverb
- article
- conjunction
- contraction
- determiner
- interjection
- metric unit
- noun
An improved version of the original WebCorp, designed to search the web for concordances in real time.

**Search term:**

Enter a word, phrase (no quotes necessary) or pattern

**See the Guide for an explanation of the options**

<table>
<thead>
<tr>
<th>Search Engine:</th>
<th>Case Options:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Case Sensitive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output Format:</th>
<th>Web Addresses (URLs):</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML</td>
<td>Show for concordance lines</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concordance Span:</th>
<th>Number of Pages to Retrieve:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OR

<table>
<thead>
<tr>
<th>Full sentences?</th>
<th>One concordance line per web site</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Site Domain / Country:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Works with Google, AltaVista, Ask and Live Search)</td>
</tr>
</tbody>
</table>
English Web Form

Search for the following phrases:

Phrase 1:   Search in the following domains:
Phrase 2:   Domain 1:
Phrase 3:   Domain 2:
Phrase 4:   Domain 3:
Phrase 5:   Domain 4:
Phrase 6:   Domain 5:
Phrase 7:   Domain 6:
Phrase 8:   Domain 7:

total sites search phrase:  +the
<table>
<thead>
<tr>
<th></th>
<th>absolute:</th>
<th>relative:</th>
<th>intrasite share:</th>
</tr>
</thead>
<tbody>
<tr>
<td>ke10,600,000</td>
<td>36,800</td>
<td>0.347 %</td>
<td>100 %</td>
</tr>
<tr>
<td>tz940,000</td>
<td>12,400</td>
<td>1.319 %</td>
<td>100 %</td>
</tr>
<tr>
<td>ug1,390,000</td>
<td>721</td>
<td>0.052 %</td>
<td>100 %</td>
</tr>
<tr>
<td>uk735,000,000</td>
<td>1,770</td>
<td>0 %</td>
<td>100 %</td>
</tr>
<tr>
<td></td>
<td>ug</td>
<td>absolute: 721</td>
<td>relative: 0.052 %</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td>1,390,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uk</td>
<td>absolute: 1,770</td>
<td>relative: 0 %</td>
</tr>
<tr>
<td></td>
<td>735,000,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Search Results**

<table>
<thead>
<tr>
<th></th>
<th>relative percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ke</td>
<td>0.347 %</td>
</tr>
<tr>
<td>tz</td>
<td>1.319 %</td>
</tr>
<tr>
<td>ug</td>
<td>0.052 %</td>
</tr>
<tr>
<td>uk</td>
<td>0 %</td>
</tr>
</tbody>
</table>

[black and white print view]
research hypotheses:

a) female users use more standard forms of English than male users because they are better educated
b) second-language users are less strict in their idiomaticity than first-language users of English
c) new English forms can be found in pronunciation, lexicon and even idioms
d) upper classes bring innovation into national Englishes
**methodology/data collection questions:**

Is the WWW a corpus?
Is a newspaper archive a corpus?
Is a collection of university student essays a corpus?
How can you include “spokenness” from the WWW?