The Language Commons Wiki

Ed Bice (Meeedan.net)
Steven Bird (University of Melbourne / University of Pennsylvania)
Kurt Bollacker (The Long Now Foundation)
Gary Simons (SIL International)
Laura Welcher (The Long Now Foundation – Rosetta Project)
Outline of this talk:

- Why build a wiki of all human language?
- How will the wiki work?
- A working prototype
- Questions and the future
Why build a wiki of all human language?
Top Ten languages by Native Speakers (Millions)

Mandarin
English
Hindi
Spanish
Arabic
Portuguese
Bengali
Russian
Japanese
German

“Long Tail” of Languages

Half the world population speaks one of 10 languages (>1%)

Most everyone else speaks one of 300 languages (4%)

5% of the world speaks one of 6,500 languages (95%)
Top 10 Languages in the Internet
millions of users

- English: 478
- Chinese: 384
- Spanish: 137
- Japanese: 96
- French: 79
- Portuguese: 73
- German: 65
- Arabic: 50
- Russia: 45
- Korean: 37
- All the rest: 290

Estimated Internet users are 1,733,993,741 for September 30, 2009
Copyright © 2009, Miniwatts Marketing Group
Ethnologue

- 6900 language descriptions
- 41k language names
- ISO 639-3 codes
- statistical summaries
- ethnologue.com
Open Language Archives Community

- digital / non-digital
- federated search
- page per language
- language-archives.org

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Archives</td>
<td>41</td>
</tr>
<tr>
<td>Archives with Fresh Catalogs</td>
<td>17</td>
</tr>
<tr>
<td>Archives with Five-star Metadata</td>
<td>9</td>
</tr>
<tr>
<td>Number of Resources</td>
<td>86042</td>
</tr>
<tr>
<td>Number of Resources Online</td>
<td>40193</td>
</tr>
<tr>
<td>Distinct Languages</td>
<td>7663</td>
</tr>
<tr>
<td>Distinct Linguistic Subfields</td>
<td>29</td>
</tr>
<tr>
<td>Distinct Linguistic Types</td>
<td>3</td>
</tr>
<tr>
<td>Distinct DCMI Types</td>
<td>9</td>
</tr>
<tr>
<td>Average Elements Per Record</td>
<td>17</td>
</tr>
<tr>
<td>Average Encoding Schemes Per Record</td>
<td>7.6</td>
</tr>
<tr>
<td>Average Metadata Quality Score</td>
<td>7.6</td>
</tr>
<tr>
<td>Last Updated</td>
<td>2010-06-10</td>
</tr>
<tr>
<td>Known Integrity Problems</td>
<td>74414</td>
</tr>
</tbody>
</table>
World Atlas of Language Structures

- 2650 languages
- 140 features (e.g. word order)
- 40 authors
- 6,100 references
Rosetta Project

- ~ 2,500 languages
- 100,000+ text pages
- Audio and video recordings
Large Scale Language Preservation Efforts

- Papua New Guinea
- >800 languages
- 100 voice recorders
- record, transcribe
Proposal: All-Language Wiki

An aggregation and discovery portal for information and resources on all 6,900 human languages.

For use by:

• language speakers
• educators
• researchers
• general public
This will be the most comprehensive, accurate and accountable information available for any human language, available in a single location.

Wiki Content

One page per language:

- **structured information** from dozens of external sources
- Further curated by:
  - user-contributed material
  - expert commentary

Linked index pages providing **taxonomic navigation** through 3,900 language families and subgroups.
How will it work?

Why can’t we just use the current Wikipedia?
Lexington County, South Carolina

From Wikipedia, the free encyclopedia

Lexington County is a county located in the U.S. state of South Carolina. In 2000, its population was 216,014. The 2008 Census estimate placed the population at 248,518. Its county seat is Lexington.

Lexington County is part of the Columbia, South Carolina, Metropolitan Statistical Area.

Contents
- Geography
  - Adjacent Counties
- Demographics
- Cities and towns
- Transportation
- References

Geography

According to the U.S. Census Bureau, the county has a total area of 758 square miles (1,963 km²), of which, 699 square miles (1,811 km²) of it is land and 59 square miles (152 km²) of it (7.74%) is water.

Adjacent Counties

- Richland County, South Carolina - east
- Orangeburg County, South Carolina - southeast
- Calhoun County, South Carolina - southeast
- Aiken County, South Carolina - southwest
- Sumter County, South Carolina - west
- Newberry County, South Carolina - northwest

Demographics

As of the census of 2000, there were 215,014 people, 83,240 households, and 59,849 families residing in the county. The population density was 309 people per square mile (119/km²). There were 90,978 housing units at an average density of 130 per square mile (50/km²). The racial makeup of the county was 84.18% White, 12.63% Black or African American, 0.34% Native American, 1.05% Asian, 0.04% Pacific Islander, 0.79% from other races, and 0.98% from two or more races. 1.92% of the population were Hispanic or Latino of any race.

There were 83,240 households out of which 35.50% had children under the age of 18 living with them, 56.60% were married couples living together, 11.60% had a female householder with no husband present, and 28.10% were non-families. 22.50% of all households were made up of individuals and 6.90% had someone living alone who was 65 years of age or older. The average household size was 2.56 and the average family size was 3.01.

In the county, the population was spread out with 26.10% under the age of 18, 8.30% from 18 to 24, 31.60% from 25 to 44, 23.80% from 45 to 64, and 10.20% who were 65 years of age or older. The median age was 36 years. For every 100 females there were 94.50 males. For every 100 females age 18 and over, there
Current Wikipedia Article Structures

Free Form Text

Lexington County, South Carolina

Lexington County is a county located in the U.S. state of South Carolina. In 2000, its population was 216,014. The 2008 Census estimate placed the population at 248,518. Its county seat is Lexington. Lexington County is part of the Columbia, South Carolina, Metropolitan Statistical Area.

Geography

According to the U.S. Census Bureau, the county has a total area of 758 square miles (1,963 km²), of which, 699 square miles (1,811 km²) of it is land and 59 square miles (152 km²) of it (7.74%) is water.

Adjacent Counties

- Richland County, South Carolina - east
- Orangeburg County, South Carolina - southeast
- Calhoun County, South Carolina - southeast
- Aiken County, South Carolina - southwest
- Saluda County, South Carolina - west
- Newberry County, South Carolina - northwest

Demographics

As of the census of 2000, there were 216,014 people, 83,240 households, and 59,849 families residing in the county. The population density was 309 people per square mile (119/km²). There were 96,978 housing units at an average density of 130 per square mile (50/km²). The racial makeup of the county was 84.18% White, 12.63% Black or African American, 0.34% Native American, 1.02% Asian, 0.04% Pacific Islander, 0.79% from other races, and 0.98% from two or more races. 1.92% of the population were Hispanic or Latino of any race.

There were 83,240 households out of which 35.50% had children under the age of 18 living with them, 56.60% were married couples living together, 11.60% had a female householder with no husband present, and 28.10% were non-families. 22.50% of all households were made up of individuals and 6.90% had someone living alone who was 65 years of age or older. The average household size was 2.56 and the average family size was 3.01.

In the county, the population was spread out with 26.10% under the age of 18, 8.30% from 18 to 24, 31.60% from 25 to 44, 23.80% from 45 to 64, and 10.20% who were 65 years of age or older. The median age was 36 years. For every 100 females there were 94.50 males. For every 100 females age 18 and over, there
Lexington County, South Carolina

Lexington County is a county located in the U.S. state of South Carolina. In 2000, its population was 216,014. The 2008 Census estimate placed the population at 248,518. Its county seat is Lexington.

Lexington County is part of the Columbia, South Carolina, Metropolitan Statistical Area.

Contents

1. Geography
   1.1 Adjacent Counties
   2. Demographics
   3. Cities and towns
   4. Transportation
   5. References

Geography

According to the U.S. Census Bureau, the county has a total area of 758 square miles (1,963 km²), of which, 699 square miles (1,811 km²) of it is land and 59 square miles (152 km²) of it (7.74%) is water.

Adjacent Counties

- Richland County, South Carolina - east
- Orangeburg County, South Carolina - southeast
- Calhoun County, South Carolina - southeast
- Aiken County, South Carolina - southwest
- Saluda County, South Carolina - west
- Newberry County, South Carolina - northwest

Demographics

As of the census of 2000, there were 216,014 people, 83,240 households, and 59,849 families residing in the county. The population density was 309 people per square mile (119/km²). There were 90,978 housing units at an average density of 130 per square mile (50/km²). The racial makeup of the county was 84.18% White, 12.63% Black or African American, 0.34% Native American, 1.05% Asian, 0.04% Pacific Islander, 0.79% from other races, and 0.98% from two or more races. 1.92% of the population were Hispanic or Latino of any race.

There were 83,240 households, out of which 35.50% had children under the age of 18 living with them, 56.60% were married couples living together, 11.60% had a female householder with no husband present, and 28.10% were non-families. 22.50% of all households were made up of individuals and 6.90% had someone living alone who was 65 years of age or older. The average household size was 2.56 and the average family size was 3.01.

In the county, the population was spread out with 26.10% under the age of 18, 8.30% from 18 to 24, 31.60% from 25 to 44, 23.80% from 45 to 64, and 10.23% who were 65 years of age or older. The median age was 36 years. For every 100 females, there were 94.50 males. For every 100 females age 18 and over, there

KB
Current Wikipedia Article Structures
Structured Schemas (Templates/Infoboxes)
These structures are insufficient for researchers, who:

- Need access to data en masse.
- Need strong citation structures.

(also relevant to WikiSpecies)
**Problem:** Structures are display oriented, not semantic, and are used inconsistently

**Example:** Language Dialects

<table>
<thead>
<tr>
<th>Language family</th>
<th>Sino-Tibetan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Siitic</td>
</tr>
<tr>
<td></td>
<td>Chinese</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dialects</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jin</td>
</tr>
<tr>
<td></td>
<td>Huizhou</td>
</tr>
<tr>
<td></td>
<td>Wu</td>
</tr>
<tr>
<td></td>
<td>Hunanese</td>
</tr>
<tr>
<td></td>
<td>Jiangxinese</td>
</tr>
<tr>
<td></td>
<td>Hakka</td>
</tr>
<tr>
<td></td>
<td>Yue (including Cantonese-Taishanese)</td>
</tr>
<tr>
<td></td>
<td>Ping</td>
</tr>
<tr>
<td></td>
<td>Shaojiang</td>
</tr>
<tr>
<td></td>
<td>Northern Min</td>
</tr>
<tr>
<td></td>
<td>Eastern Min (including Fuchow)</td>
</tr>
<tr>
<td></td>
<td>Central Min</td>
</tr>
<tr>
<td></td>
<td>Pu Xian</td>
</tr>
<tr>
<td></td>
<td>Southern Min  (including Amoy, Taiwanese)</td>
</tr>
</tbody>
</table>

The major dialect groups are:

**Magrebi Arabic**
- Magrebi Arabic includes Moroccan Arabic, Algerian Arabic, Libyan Saharan Arabic, Tunisian Arabic, and Libyan Arabic, and is spoken by around 45 million North Africans in Morocco, Western Sahara, Algeria, Tunisia, Libya, Niger, and Western Egypt. It is often difficult for speakers of Middle Eastern Arabic varieties to understand. The Berber influence in these dialects varies in degrees.¹⁴

**Egyptian Arabic**
- Egyptian Arabic, spoken by around 76 million in Egypt. It is one of the most understood varieties of Arabic, due in large part to the widespread distribution of Egyptian films and television shows throughout the Arabic speaking world. Closely related varieties are also spoken in Sudan.

**Gulf Arabic**
- Gulf Arabic, spoken by around 94 million people in Arab states and eastern Saudi Arabia.
- Iraqi Arabic, spoken by about 29 million people, with significant differences between the Arabian-like dialects of the south and the more conservative dialects of the north. Closely related varieties are also spoken in Iran, Syria, and Turkey.
- North Mesopotamian Arabic, spoken by around 7 million people in northern Iraq, northern Syria and southern Turkey.

**Levantine Arabic**
- Levantine Arabic includes North Levantine Arabic, South Levantine Arabic, and Cypriot Arabic. It is spoken by almost 35 million people in Lebanon, Syria, Jordan, The Palestinian territories, Israel, Cyprus, and Turkey. It is also called Mediterranean Arabic.

**Other**
- Other varieties include:

---

**List of dialects of the English language**

From Wikipedia, the free encyclopedia

This article does not cite any references or sources. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (December 2008)

This is a list of dialects of the English language. Dialects are linguistic varieties which differ in pronunciation, vocabulary and grammar from each other and from Standard English (which is itself a dialect).

Dialects can be usefully defined as "sub-forms of languages which are, in general, mutually comprehensible".¹⁶ British linguists distinguish dialect from accent, which refers only to pronunciation. Thus, any educated English speaker can use the vocabulary and grammar of Standard English, but different speakers use their own local words for everyday objects or actions, regional accent, or Received Pronunciation, which within the U.K. is considered an accent distinguished by class rather than by region. American linguists, however, include pronunciation differences as part of the definition of regional or social dialects. The combination of differences in pronunciation and use of local words may make some English dialects almost unintelligible from one region to another.
**Problem:** Historical diffs cannot be organized or searched by structure
e.g.: Diffs of Chinese Language

| (cur) = difference from current version, (prev) = difference from preceding version. m = minor edit, ← = section edit, ← = automatic edit summary |
| (latest) View (newer 50 | older 50) | (20 | 50 | 100 | 250 | 500) |
| **Compare selected revisions** |
| (cur) (prev) | 10:45, 21 June 2010 | Symane (talk | contribs) m (63,712 bytes) (Spoken Chinese) (undo) |
| (cur) (prev) | 07:46, 19 June 2010 | DuncanBCS (talk | contribs) m (63,712 bytes) (Unreferenced) (undo) |
| (cur) (prev) | 04:55, 19 June 2010 | 112.213.215.104 (talk) (63,851 bytes) (undo) |
| (cur) (prev) | 09:45, 18 June 2010 | Crescent8 (talk | contribs) m (63,712 bytes) (undo) |
| (cur) (prev) | 03:25, 16 June 2010 | Xqbot (talk | contribs) m (63,700 bytes) (robot Removing: zu:IShayina) (undo) |
| (cur) (prev) | 17:38, 11 June 2010 | BilCat (talk | contribs) m (63,718 bytes) (Reverted edits by Mr.Right.Perfect (talk) to last version by TheLeopard) (undo) |
| (cur) (prev) | 17:14, 11 June 2010 | Mr.Right.Perfect (talk | contribs) m (63,774 bytes) (undo) |
| (cur) (prev) | 03:28, 7 June 2010 | TheLeopard (talk | contribs) (63,718 bytes) (Grammar) (undo) |
| (cur) (prev) | 17:15, 6 June 2010 | Stevertigo (talk | contribs) m (63,721 bytes) (undo) |
| (cur) (prev) | 07:49, 6 June 2010 | KwamiKagami (talk | contribs) (63,723 bytes) (add dialects to info box) (undo) |
| (cur) (prev) | 12:42, 5 June 2010 | Beime (talk | contribs) (63,167 bytes) (Reverted 1 edit by 91.178.127.87; The Chinese languages are generally mutually unintelligible. (TW)) (undo) |
| (cur) (prev) | 11:08, 5 June 2010 | 91.178.127.87 (talk) (63,165 bytes) (undo) |
| (cur) (prev) | 06:52, 5 June 2010 | 72.190.75.130 (talk) (63,167 bytes) (Eduction) (undo) |
| (cur) (prev) | 19:07, 3 June 2010 | ILV TW (talk | contribs) m (53,168 bytes) (undo) |
| (cur) (prev) | 18:13, 2 June 2010 | Courcelles (talk | contribs) m (63,141 bytes) (Reverted edits by 142.24.53.237 (talk) to last revision by Bjmullan (HQ)) (undo) |
| (cur) (prev) | 18:12, 2 June 2010 | 142.24.53.237 (talk) (empty) (Writing) (undo) |
| (cur) (prev) | 18:12, 2 June 2010 | Bjmullan (talk | contribs) m (63,141 bytes) (Reverted 2 edits by 142.24.53.237 identified as vandalism to last revision by Wavelength. (TW)) (undo) |
| (cur) (prev) | 18:11, 2 June 2010 | 142.24.53.237 (talk) (61,227 bytes) (Language and nationality) (undo) (Tag: section blanking) |
Problem: Contributors tend to not name their sources.

Example: Where did the contributor get the fact that there are 31 million Gan speakers?
Problem: Wikipedia article structure promotes consensus rather than supporting simultaneous differing viewpoints when none is dominant

Example: Taxonomic organization
A working prototype

Our example of a possible solution using distributed sources
The Rosetta Project is a global collaboration of language specialists and native speakers working to build a publicly accessible digital library of human languages.

Participate and support The Rosetta Project >>

Featured Content

- Internet Archive
  - The Rosetta Project collection is now available at the Internet Archive.

- Rosetta in Freebase
  - Help us build an open dataset of language information in Freebase.

- ESA Rosetta Mission
  - This ESA craft carries a Rosetta Disk. Where in the solar system is it now?

Archive Statistics

- 26,218 language topics in Freebase
- 4,766 documents in the Internet Archive
- 78,280 downloads as of December 2009

Spotlight Item

- A Complete Grammar of Esperanto
  - View the Book
All Language Metadata Now in Freebase

- Rosetta Base: over 10,000 languages and linguistic entities linked by language family relationship
- All data is linked to other kinds of data in Freebase
- We have rectified ~1500 Wikipedia pages about human languages to our data set
Please help!

• We plan to build The Language Commons Wiki – an aggregation and discovery portal for information and resources on all 6,900 human languages – and we need your help!

• Here are just some of the questions we have:
  ○ Does this satisfy the needs of researchers, native speakers of rare languages, and students?
  ○ What should the relationship between the existing Wikipedia and the Language Commons be? (e.g. can it just be a source too?)
  ○ How do we introduce this new way of editing a wiki article to editors?
  ○ Who do we need to talk to? Who should we get involved?
Thank you!

Questions / Comments / Ideas:

Laura Welcher
laura@longnow.org