Multilingual Lexicography with a Focus on Less-Resourced Languages:
Data Mining, Expert Input, Crowdsourcing, and Gamification

Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era
Language Resources and Evaluation Conference (LREC 2014)
26 May, 2014 - Reykjavik, Iceland
• **~7000 languages spoken worldwide** (Ethnologue 2014)
  • Top 5: Excellent ICT resources
    • English + FIGS (French, Italian, German, Spanish)
  • ~30: Significant ICT resources (speech recognition, machine translation, etc)
    • Many official EU languages + Russian, Arabic, Chinese, Japanese...
    • Hindi, Bengali, Punjabi, Indonesian, Swahili... (100 million+ speakers)
• **~80: More than 10,000,000 speakers**
  • <100: Basic ICT resources (corpora, machine-readable dictionaries, POS taggers, morphological analyzers, parsers, etc)
• **~300: More than 1,000,000 first language speakers**
• **~1500: More than 100,000 first language speakers**
• **~2000: Threatened – 10,000 to 100,000 speakers**
• ~2000: Some existing data (wordlists, Bibles, recordings)
• **~3000: Endangered**
  • Fewer than 3000 speakers
  • Not being passed to next generation – extinct within 100 years
  • Embedded human cultural patrimony

Multilingual Lexicography with a Focus on **Less-Resourced Languages**:

Workshop on Collaboration and Computing for **Under-Resourced Languages** in the Linked Open Data Era

Language Resources and Evaluation Conference
Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era
Language Resources and Evaluation Conference (LREC 2014)
26 May, 2014 - Reykjavik, Iceland
kamusi is Swahili for dictionary
Comprehensive linguistic data for every language
Comprehensive linguistic data for every language

• For people
• For machines
In service since 1994 (originally at Yale Council on African Studies)
International NGO since 2009
• Registered non-profit in USA and Switzerland (and soon Tanzania)

Academic Home since 2013:
EPFL - Swiss Federal Institute of Technology in Lausanne
LSIR - Distributed Systems Information Laboratory
1. Project objectives
2. Acquiring data
3. Crowdsourcing innovations
1. Project objectives
2. Acquiring data
3. Crowdsourcing innovations
light
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries based on spelling are impossible
why multilingual dictionaries based on spelling are impossible
impossible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
light (energy)
light (lamp)
light (weight)
light (pale)

how Kamusi concept linking makes a multilingual dictionary possible
light

English  noun

**Definition:** A device that provides illumination by burning electricity, fuel, or wax.

**Terminology:** general

| Plural   | lights   |

**My Languages**
- *itara* in Kirundi
- *taa* in Swahili
- *lumière* in French
- *lumină* in Romanian

**Other Languages**
- *taala* in Luganda
licht

Dutch zelfstandig naamwoord - noun (Dutch) het-woord, onzijdig (het)

Definition: Voorwerp dat licht produceert, gevoed door elektriciteit of door verbranding van was of een andere brandstof.

Examples:
We gaan slapen. Doe jij het licht (de lichten) uit?

meervoud onzijdig (de) lichten

Other Languages
- lámpara in Spanish
- bugum in Mampruli
- -tara in Kinyarwanda
- بَنٰ in Urdu
- oborabu in Gusii
- Licht in German
light (lamp)
monolingual data
## depth: kamusi (unlimited)

<table>
<thead>
<tr>
<th>Key Information</th>
<th>Word Forms &amp; Origins</th>
<th>Translations &amp; Concepts</th>
<th>Related Intralanguage</th>
<th>Usage Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Phonetic IPA</td>
<td>Gloss</td>
<td>Synonyms</td>
<td>Example 1</td>
</tr>
<tr>
<td>Lemma</td>
<td>Alt Spelling</td>
<td>Gloss Language</td>
<td>Antonyms</td>
<td>Source</td>
</tr>
<tr>
<td>Headword</td>
<td>Tone Spelling</td>
<td>Relation Type</td>
<td>Spawn</td>
<td>Audio</td>
</tr>
<tr>
<td>Part of Speech</td>
<td>Alt Script</td>
<td>Science Taxon</td>
<td>Family</td>
<td>Translation</td>
</tr>
<tr>
<td>Morphemes</td>
<td>Pronunciations</td>
<td>Term Set</td>
<td>Misspellings</td>
<td>Trans Lang</td>
</tr>
<tr>
<td>Definition</td>
<td>Pron Geo-Tag</td>
<td>Tags</td>
<td></td>
<td>Example X</td>
</tr>
<tr>
<td>Def Trans</td>
<td>Dialect</td>
<td>Semantic Sets</td>
<td></td>
<td>Source</td>
</tr>
<tr>
<td>Def Trans Lang</td>
<td>Word Geo-Tag</td>
<td>Image</td>
<td></td>
<td>Audio</td>
</tr>
<tr>
<td>Usage Note</td>
<td>Etym Term</td>
<td>Concept Links</td>
<td></td>
<td>Translation</td>
</tr>
<tr>
<td>Cultural Note</td>
<td>Etym Lang</td>
<td></td>
<td></td>
<td>Trans Lang</td>
</tr>
<tr>
<td>Special Note</td>
<td>Etym Note</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
growth: a “living” dictionary
growth: a “living” dictionary
growth: a “living” dictionary
growth: a “living” dictionary
Edit Dictionary Term Licht

Concept
Term (lemma): licht
Term Language: Dutch
Headword: Licht
Part of Speech: zelfstandig naamwoord - noun
Part Attribute: het-woord, onzijdig (het)

Morpheme Information
Collected Morphemes
- meervoud mannelijk (de)
- meervoud vrouwelijk (de)
- enkelvoud mannelijk de-woord (de-m)
- enkelvoud onzijdig het-woord (het)
- enkelvoud vrouwelijk de-woord (de-v)

Definition
Definition of the term in the language to which it belongs:
Voorwerp dat licht produceert, gevoed door een elektriciteit of door verbranding van was of een andere brandstof.
and more...

- Morphemes and inflections
- Attributes
- Behaviors
- Grouping and ranking
- Agglutination parsing
- Grammatical bridges
- Family relations
- Etymology and temporality
- Geography
- And more...
hooks
hooks

Open Language Resources
1. Project objectives
2. Acquiring data
3. Crowdsourcing innovations
acquiring data

- For people
- For machines
acquiring data

- With people
- With machines
acquiring data

- From existing data
- From experts
- From the crowd
Challenges of existing data for LRLs:

- Paltry
  - Few data items
  - Few data elements
- Poorly structured
- Inconsistently structured
- No OCR training
- IP restrictions

And then:
- Non-disambiguated

**Solution – Merging Engine:**

- Get data into consistent fields
- Human review for sense alignment
acquiring data • experts

Challenges of language specialists for LRLs:

- Limited # of languages
- Experts have limited time
- Expertise costs money
  - 10,000 entries = 1 year of labor

Expert Tool – Edit Engine:

- Easy
- Consistent
- Comprehensive
acquiring data • the crowd

Challenges of crowdsourcing for LRLs:

• Eliciting specific data
• Keeping things simple
• Validating data without “ground truth”
  • Minor errors
  • Wrong data
• Malicious users
  • Lots of ways to make mischief
• Sparking participation
• Crowd size
• Maintaining participation

Solutions:
• Fidget Widget
• Facebook Game
1. Project objectives
2. Acquiring data
3. Crowdsourcing innovations
crowdsourcing

Purposes:

- Gather original data
- Manipulate existing data
- Validate existing data
- Validate new data
crowdsourcing

Methods:

- Targeted questions
- Short questions
- Engaging games
- Motivational incentives
crowdsourcing

Tools:

- Play to Pay
- Fidget Widget (mobile app)
- Facebook Game
crowdsourcing

Fidget Widget - Logic:

• Single word answers
• Thumbs up/ down
• Longer answers for advanced users
Fidget Widget - **Logic:**

- Single word answers
- **Thumbs up/ down**
- Longer answers for advanced users
Fidget Widget – Logic:

- Single word answers
- Thumbs up/ down
- Longer answers for advanced users
crowdsourcing

Fidget Widget - Issues:

- Synchronization
  - Offline in LRL settings
- Security
  - User accounts
- Competing answers
  - Noise vs. signal
- Publication
  - Deciding a quality threshold
crowdsourcing
Gamification (Fall 2014)

Facebook
• Where the people are
• Where their friends are

The Games
• Pyramid of points for winning answers
• Team points for fastest languages
• Difficulty points for stumpers
Crowdsourcing

Contribution analysis

- Good answers – build trust
- Bad answers – change the game
- Malicious answers – remove the user’s history
languages

5 year goal:
100 languages worldwide

words

5 year goal:
10,000 parallel terms in each language

data

5 year goal:
100,000,000 searches/ month

people

5 year goal:
1,000,000 registered users
Network: http://www.linkedin.com/in/martinbenjamin
Write: martin.benjamin@epfl.ch
Meet: http://meetme.so/kamusi