The Kam4D Linguistic Knowledge Graph: Putting Smurfs, Ducks, Lemurs, and Party Terms to the Service of African Languages

Martin Benjamin

SADiLaR: DH Colloquium – 17 November, 2021
South African Centre for Digital Language Resources
kamusi is Swahili for dictionary
Goal: A complete matrix of human expression across time and space

- As a knowledge resource
- As a data resource
- As a basis for any-to-any translation
In service since 1994 - originally at Yale Council on African Studies
International NGO since 2009
- Registered non-profit in 🇺🇸 and 🇨🇭

Academic Home since 2013:
EPFL - Swiss Federal Institute of Technology in Lausanne
First at LSIR - Distributed Systems Information Laboratory
Now at the Swiss EdTech Collider
ACALAN (Intergovernmental language agency for 55 member states of the African Union):
Platform for African Language Empowerment development partner
Kamusi Jargon – essential terms for Kam4D

Lemur =
- Lemma
- Lemmatic form
- Dictionary form
- Canonical form
- Citation form

Party term =
- Multiword Expression
- MWE

- Lemurs and Party Terms
- Smurfs and Ducks
- Costumes and Wardrobes
Kamusi Jargon – essential terms for Kam4D

SMURF =
Spelling/ Meaning
Unit Reference

• Lemurs and Party Terms
• Smurfs and Ducks
• Costumes and Wardrobes

DUCKS = Data
Unified Concept
Knowledge Set
The Kam4D Linguistic Knowledge Graph: Putting Smurfs, Ducks, Lemurs, and Party Terms to the Service of African Languages

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects
1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects
Data: **WORDS** that have been **digitized** in a way that can be used within computer processes
This is a WORD
Lemur =
- Lemma
- Lemmatic form
- Dictionary form
- Canonical form
- Citation form
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
Party term =
- Multiword Expression
- MWE
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
why multilingual dictionaries were impossible
The Kam4D Linguistic Knowledge Graph: Putting Smurfs, Ducks, Lemurs, and Party Terms to the Service of African Languages

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects
1. More problems with linguistic data!
2. The Kam4D solution
3. Kamusi Labs projects

- 2000 languages in Africa
- Very few have any digital resources
- What has been digitized has not been harmonized as interoperable data
The Kam4D Linguistic Knowledge Graph: Putting Smurfs, Ducks, Lemurs, and Party Terms to the Service of African Languages

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects
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
Lab Server Server

- 2,099,419 Smurfs
- 122 Languages

Production

- ~138,000 Ducks
- 44 Languages
Kam4D - kamu.si/kam4d

• 4D = Four Dimensional
  • Time is the fourth dimension - capacity to treat language change and historical languages

• Graph database structure for a complete matrix of human expression across time and space
  • the structure is realistic; the final goal is an impossible aspiration

• Molecular lexicography design
light
meaning
shape
place
sound
time
relationships

recommended reading:

> kamu.si/kam4d
<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portuguese</td>
<td>homem</td>
<td>uma pessoa adulta que é masculina (ao contrário de uma mulher)</td>
</tr>
<tr>
<td>isiZulu</td>
<td>indoda</td>
<td>umuntu wesilisa oseqinile rje</td>
</tr>
<tr>
<td>English</td>
<td>adult</td>
<td>an adult person who is male (as opposed to a woman)</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td></td>
</tr>
<tr>
<td></td>
<td>man</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>homo</td>
<td>any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage</td>
</tr>
<tr>
<td></td>
<td>human</td>
<td></td>
</tr>
<tr>
<td></td>
<td>human</td>
<td>being</td>
</tr>
<tr>
<td></td>
<td>man</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>Word</td>
<td>Definition</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Portuguese</td>
<td>homem (n)</td>
<td>Definition: uma pessoa adulta que é masculina (ao contrário de uma mulher)</td>
</tr>
<tr>
<td>isiZulu</td>
<td>indoda (n)</td>
<td>Definition: umuntu wesilisa oseqinile nje</td>
</tr>
<tr>
<td>English</td>
<td>adult male (n)</td>
<td>Definition: an adult person who is male (as opposed to a woman)</td>
</tr>
<tr>
<td></td>
<td>man (n)</td>
<td>Example: there were two women and six men on the bus.</td>
</tr>
<tr>
<td>Portuguese</td>
<td>homem (n)</td>
<td>Definition: qualquer membro vivo ou extinto da família Hominidae caracteriza-se por uma inteligência superior, articular o discurso e erigir a carruagem</td>
</tr>
<tr>
<td>isiZulu</td>
<td>umuntu (n)</td>
<td>Definition: ilunga eliphilayo nelingasaphili lomndeni wabantu</td>
</tr>
<tr>
<td>English</td>
<td>homo (n)</td>
<td>Definition: any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage</td>
</tr>
<tr>
<td></td>
<td>human (n)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>human being (n)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>man (n)</td>
<td></td>
</tr>
</tbody>
</table>
Kamusi Jargon – essential terms for Kam4D

- Lemurs and Party Terms
- Smurfs and Ducks
- Costumes and Wardrobes

Wardrobe =
A set of forms (a.k.a. inflections) that might be used by a smurf

Costume =
A single form (a.k.a. inflection) that might be used by a smurf
DRIVE

drives, drove, driving, driven

meaning

time

relationships

shaping

place

sound
DRIVE

drives, drove, driving, driven

meaning

shape

place

time

relationships

sound
compel someone
hit a golf ball
operate a car
herd animals
drives, drove, driving, driven
compel someone
meaning
shape
place
time
relationships
sound

Recommended reading: [kamu.si/kam4d](http://kamu.si/kam4d)
Molecular lexicography design
Molecular lexicography designers
The Kam4D Linguistic Knowledge Graph: Putting Smurfs, Ducks, Lemurs, and Party Terms to the Service of African Languages

1. The problem with linguistic data
2. The Kam4D solution
3. Kamusi Labs projects
1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects...
Kamusi Labs projects

1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects…
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
how Kamusi makes a multilingual dictionary possible
Gathering Data for African Languages

- Duck Duck Kamus – aligning existing datasets
- Crowdsourcing games for new microdata
- GOLDdigger – engine for expert editors
Kamusi Labs projects

1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects…
SlowBrew Assisted Translation

- User selects their meaning on the source side (predisambiguation)
  - Users can suggest missing senses
SlowBrew Assisted Translation

- User selects their meaning on the source side (predisambiguation)
  - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
  - Party Terms are treated as Smurfs in Kam4D
  - Separated expressions easily conjoined (unlike NMT)
**SlowBrew Assisted Translation**

- User selects their meaning on the source side (predisambiguation)
  - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
  - Party Terms are treated as Smurfs in Kam4D
  - Separated expressions easily rejoined (unlike NMT)

**Smurfs and Ducks**

**Kam4D**

[kamu.si/kam4d](http://kamu.si/kam4d)

- **SlowBrew**

She drove everyone in her class at school up the wall last night

- **drove** ➔ **drive**
  - through Kam4D “costumes”
  - NLP toolkits not available for most languages
  - “drive” triggers search for downstream party terms

- **annoy**

- **Social stratus**
  - **School room**
  - **Style**

- **Group of fish**
  - **Group of thinkers**
  - **Educational institution**

- **previous**
- **final**
- **night before**
She drove everyone in her class at school up the wall last night

drove ➔ drive
- through Kam4D “costumes”
- NLP toolkits not available for most languages
- “drive” triggers search for downstream party terms

- annoy

- Social status
- School room
- Style

- Group of fish
- Group of thinkers
- Educational institution

- previous
- final
- night before
**SlowBrew Assisted Translation**

- User selects their meaning on the source side (predisambiguation)
  - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
  - Party Terms are treated as Smurfs in Kam4D
  - Separated expressions easily rejoined (unlike NMT)
- DUCKS finds equivalent term in Language B
SlowBrew Assisted Translation

- User selects their meaning on the source side (predisambiguation)
  - Users can suggest missing senses
- SlowBrew suggests Party Terms (MWEs), or users can mark their own
  - Party Terms are treated as Smurfs in Kam4D
  - Separated expressions easily rejoined (unlike NMT)
- DUCKS finds equivalent term in Language B
- Machine learns from context-specific user selections
  - Crowdsourced dataset of spelling/meaning annotations
  - AI builds from human intelligence on the source-side
Unanswered Questions:

- Will users take the time to predisambiguate?
  - People take time to choose images
  - People take time to spellcheck
- Syntax on the target side?
  - Outside Kamusi wheelhouse – partners needed
- How to pay for it?
1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects…
Kam4D will serve as the linguistic data core for the ACALAN-AU platform
Integrated with Kamusi systems for gathering and disseminating data
Initial focus on 20 VCBLs (Vehicular Cross-Border Languages)
Kamusi Labs projects

1. Gathering data for African languages
2. SlowBrew assisted translation
3. PALE: Platform for African Language Empowerment
4. Many more projects…
Many more Kamusi Labs projects...

- KamuSee 👓 visual dictionary
- Sign languages 👌 gesture video dictionary
- Logikamusi 🙆 ontological dictionary
- Kamedicine ✌️ medical terminology translator
- Kamigrate 👣 refugee and immigrant services translator
- Kamergency 😞👩🚒 phrasebook for accident and disaster first responders
- Kamuseum 🏛️ guides for public spaces
- Box-o-Lex 📚 field lexicography toolkit
- Talkamusi 📣💬📖 talking dictionary
- KamHoosi 🦉 named entities
- EdTech Trio 🎵🎓 for learning IN African languages, learning FROM African languages, and learning OF African languages
- **And Even More:** [http://kamu.si/big-picture-playbook](http://kamu.si/big-picture-playbook)
The Kam4D Linguistic Knowledge Graph:
Putting Smurfs, Ducks, Lemurs, and Party Terms to the Service of African Languages

martin@kamusi.org

recommended reading:
- kamu.si/kam4d
- kamu.si/big-picture-playbook
- teachyoubackwards.com

SADiLaR: DH Colloquium – 17 November, 2021
South African Centre for Digital Language Resources