

Wikimedia Lexicography: Lexemes and Beyond

By:

- Hamidu Rukaya ([Ruky Wunpini](#)) DWUG
- Mohammed Sadat Abdulai ([Masssly](#)) WMDE



wikiindaba



wikimediamorocco

**SCAN
ME!**



<https://bit.ly/WikiIndaba2023Lexemes>



Here's what's in store

- **Brief Introduction to Lexemes on Wikidata**
- **Documenting Pronunciations using Spell4Wiki**
- **Hands-on Editing Session**

Why do languages matter?



"Every language reflects a unique worldview with its own value systems, philosophy and particular cultural features.

The extinction of a language results in the irrecoverable loss of unique cultural knowledge embodied in it for centuries, including historical, spiritual and ecological knowledge that may be essential for the survival of not only its speakers, but also countless others."

- UNESCO



“Every language is a world. Without translation, we would inhabit parishes bordering on silence.”

- Ngugi wa Thiong’o, a Kenyan writer and academic.



“A language becomes extinct when its last native speaker dies, and it’s usually the result of its speakers shifting to a lingua franca like English, Arabic or Spanish. This implies choice, but it’s often a history of marginalisation that leads to the change.”

- Lauren Johnson



Factors of language vitality

According to UNESCO:

- Intergenerational language transmission
- Absolute number of speakers
- Proportion of speakers existing within the total (global) population
- Language use within existing contexts and domains
- Response to language use in new domains and media
- Availability of materials for language education and literacy
- Government and institutional language policies
- Community attitudes toward their language
- Amount and quality of documentation



Factors of language vitality

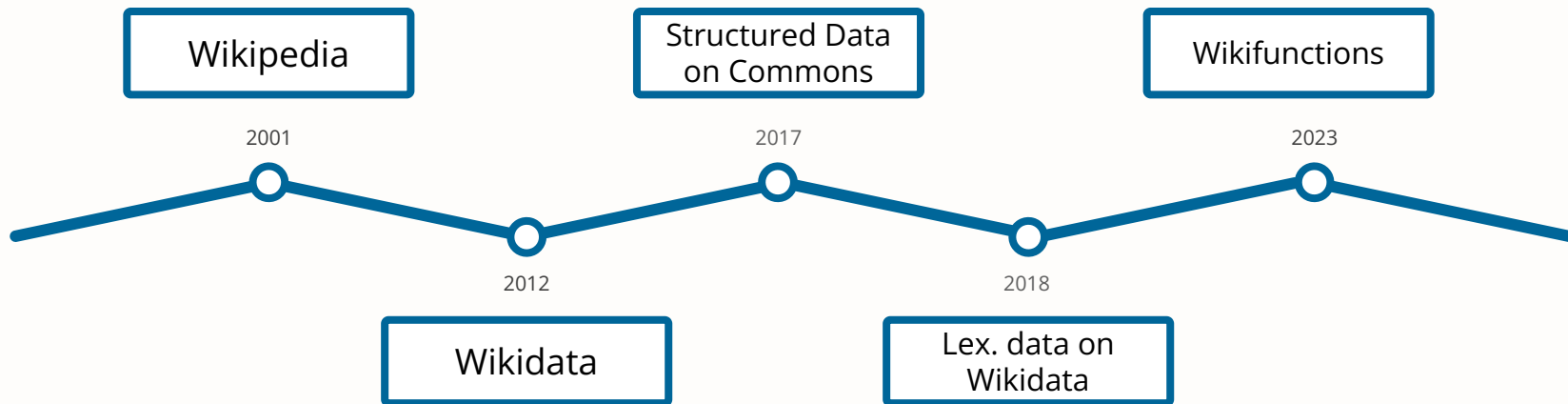
According to UNESCO:

- Intergenerational language transmission
- Absolute number of speakers
- Proportion of speakers existing within the total (global) population
- Language use within existing contexts and domains
- Response to language use in new domains and media
- Availability of materials for language education and literacy
- Government and institutional language policies
- Community attitudes toward their language
- Amount and quality of documentation



Lexicographical data on Wikidata





Until 2018,
Wikidata was
only describing
concepts

Now it also includes

words!



Wait, what's the difference?


Concept "mouse"

- Specie of mammal
- Taxon name
- Average size
- Picture
- Encyclopedia of Life ID



File:House mouse.jpg,
public domain

Lexeme "mouse"

- Language: English
- Lexical category: noun
- Plural form: mice (irregular)
- Etymology: Proto-Germanic *mūs
- Senses: animal, computer device, adjective
- Translations: jəngbariga (dag), beera (ha)
- Audio pronunciation 

/maʊs/

L-id

Lexeme

Lemma - *standard form or dictionary form of the lexeme*

Lexical category

Language

Statements - *e.g. derived-from, homonym, etc.*

Senses

Gloss - *short description*

Statements - *e.g. translations, synonyms, refers-to-concept, etc.*

Forms

Representation

Grammatical features

Statements - *e.g. region, period, pronunciation, etc.*

Glossary:

https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Glossary

More info:

[mw:Extension:Wikibase Lexeme/Data Model](mw:Extension:Wikibase_Lexeme/Data_Model)





- Main page
- Community portal
- Project chat
- Create a new item
- Recent changes
- Random item
- Query Service
- Nearby
- Random Primary Sources item
- Help
- Donate

- Lexicographical data
- Create a new Lexeme
- Recent changes
- Random Lexeme
- Edit as english-noun

- Wikidata Edit
- WEF: Settings
- WEF: AdmEntity
- WEF: Article
- WEF: Award
- WEF: Book / Journal
- WEF: Entity
- WEF: External IDs
- WEF: FRBR Edition
- WEF: FRBR Work**
- WEF: Legal Act
- WEF: Movie
- WEF: Person
- WEF: Software
- WEF: Taxon
- WEF: Trans. Infra

- Tools
- What links here

Enter a schema to check against e.g. E234 Automatically check schema

59 revisions since 2018-05-25 (+80 days), 22 editors, 0 pageviews (30 days), created by: Jsamwrits (5,775,308) · See full page statistics

(L1119) **mouse**

en

Language English
Lexical category noun

Statements (expand all references)

instance of

P31

by LydiaPintscher

count noun / Q1520033

Lookup model item 0 references

described by source

P1343

by LydiaPintscher

Merriam-Webster online dictionary / Q52055677

0 references

usage example

P5831

by LydiaPintscher and Guergana

When the mouse laughs at the cat, there's a hole nearby. (English)

0 references

Twas the night before Christmas, when all through the house -

Example:
<https://www.wikidata.org/wiki/Lexeme:L1119>

- WEF: Book / Journal
- WEF: Entity
- WEF: External IDs
- WEF: FRBR Edition
- WEF: FRBR Work
- WEF: Legal Act
- WEF: Movie
- WEF: Person
- WEF: Software
- WEF: Taxon
- WEF: Trans. Infra

Tools

- What links here
- only lexemes
- Related changes
- Special pages
- Permanent link
- Page information
- Concept URI
- Cite this page
- Get shortened URL
- Automatic addition
- Check sitelink
- Ordia
- Import statements
- Search for lemma in OpenAlex abstracts
- Search for lemma in OpenAlex titles
- Show dependency graphs on lexeme page
- VIP's labels

Browse Primary Sources

▲ back to top ▲

Senses

L578335-S1

Dagbani
Hausa

nyɔm
kamshi koo waari

edit

Statements about L578335-S1

image P18



A curious child, smelling flower, India.jpg
2,461 × 2,000; 580 KB

▼ 0 references

+ add reference
+ add value

item for this sense P5137

sense of smell / Q1541064

▼ 0 references

+ add reference
+ add value

that depicts this sense

translation P5972

L641548-S1

▼ 0 references

+ add reference
+ add value

(L578335) **ziɛɣu**
dag

Language **Dagbani**
Lexical category **noun**

Example:
<https://www.wikidata.org/wiki/Lexeme:L578335>

L578335-S2

Dagbani
English
Hausa

pohim shɛli din mali yaa
storm
iska

edit

Statements about L578335-S2

image edit

P18



2020aug-derecho-shelf-cloud-Sugar-Grove-Illinois.jpg
1,567 × 588; 64 KB

▼ 0 references

+ add reference

+ add value

item for this sense edit

P5137

storm / Q81054

▼ 0 references

+ add reference

+ add value

that depicts this sense

+ add statement

Translations		
Akkadian: imhullu/𒌦𒍪𒌦𒍪	Greenlandic: anorersuaq	Malayalam: മെഘമുതല
Bangla: ঝঞ্ঝা, ঝাটিকা, ঝড়	Hebrew: סופה/סופו	Russian: буря, ураган
Bokmål: storm	Hindko: تھری	Slovak: búrka
Danish: storm	Hindustani: طوفان/توفان	Spanish: tormenta
English: storm	Italian: tempesta	Sumerian: muru/𒍪𒌦𒍪𒌦𒍪, mir/𒍪𒌦𒍪
French: tempête	Latin: tempestas	imhul/𒌦𒍪𒌦𒍪, tumudal/𒍪𒌦𒍪𒌦
German: Sturm	Malay: ribut/ريوت	Swedish: storm

(L578335)

ziɛɣu

dag

Language **Dagbani**

Lexical category **noun**

Example:

<https://www.wikidata.org/wiki/Lexeme:L578335>


Dagbani	bin din moha	edit
Hausa	jaa	
Twi	კაკო	
Polish	kolor czerwony	
Czech	červená barva	
French	couleur rouge	
Ukrainian	Червоний колір	
Spanish	color rojo	
Hungarian	piros szín	
Swedish	röd färg	
Norwegian Bokmål	rød farge	
Catalan	color vermell	
Esperanto	ruĝa koloro	
Turkish	kırmızı renk	
Danish	farven rød	
Bulgarian	Червен цвят	
Finnish	punainen väri	
Italian	colore rosso	
Slovak	červená farba	
Slovenian	rdeča barva	
Latvian	sarkanā krāsa	
Estonian	punane värv	

L578335-S3

Statements about L578335-S3

[image](#) [edit](#)

P18




Red saturations.svg
1,500 × 250; 674 bytes

[0 references](#)

[+ add reference](#)

[edit](#)



(L578335) **ᲗᲠᲔᲘᲚ**
dag

Language **Dagbani**
Lexical category **noun**

Example:
<https://www.wikidata.org/wiki/Lexeme:L578335>

Forms [ə]

› L278-F1 | walk
en

 edit

Grammatical features [simple present](#)

Statements about L278-F1

+ [add statement](#)

› L278-F2 | walks
en

 edit

Grammatical features [third person, singular, simple present](#)

Statements about L278-F2

+ [add statement](#)

› L278-F3 | walked
en

 edit

Grammatical features [simple past](#)

Statements about L278-F3

+ [add statement](#)

› L278-F4 | walking
en

 edit

Grammatical features [present participle](#)

(L278) | **walk**
en

Language [English](#)
Lexical category [verb](#)

+ [add statement](#)

Example:
<https://www.wikidata.org/wiki/Lexeme:L278>

Why is it interesting?

- Structured data = machine readable
- Can be reused by tools, research, dictionaries, translation services
- CC0 = open knowledge, can be reused by all
- Huge variety of languages, including undeserved ones
- International community = more people to help

What's the difference to Wiktionary?

- Wiktionary = plain text + templates, Wikidata = structured data
- Wikidata can be easily parsed and reused
- Wikidata works with Lexemes, Wiktionary combines Lexemes
- Wiktionary may have extra info (examples, quotes...)
- Wikidata = CC0, Wiktionary = CC-BY SA
- Wikidata aims to support Wiktionaries (if they want to)

What's the difference to other services?

- We're providing the background data to build anything on top of it
- We're doing much more than translation: we help machines understand languages
- We give access to the data in CC0
- We include all languages, not only the most profitable ones
- We empower people to contribute to the data

What can we do with it?



Support Wiktionary & other Wikimedia projects

- Provide structured data to be reused on pages
- Working together on the same data
- New tools to make contributing easier and open it up to new contributor groups

Potential users: Wiktionary,
Wikisource, Wikidata Games...



Dictionary applications and more

- Looking up definitions and translations
- Special purpose dictionaries (rhyme, specific topics)
- Thesauri and synonym dictionaries
- Build translation tools (especially for underserved languages that don't have any yet)

Potential users: Leo, Apertium

Language learning tools

- Creating word lists and lessons
- Illustrating words
- Creating games and exercises

Potential users: Parley, Duolingo

Research

- How do languages evolve over time, social class and more?
- Do classes of words change their meaning over time?
- Localizing words on maps

Potential users: The Rosetta Project

Text analysis

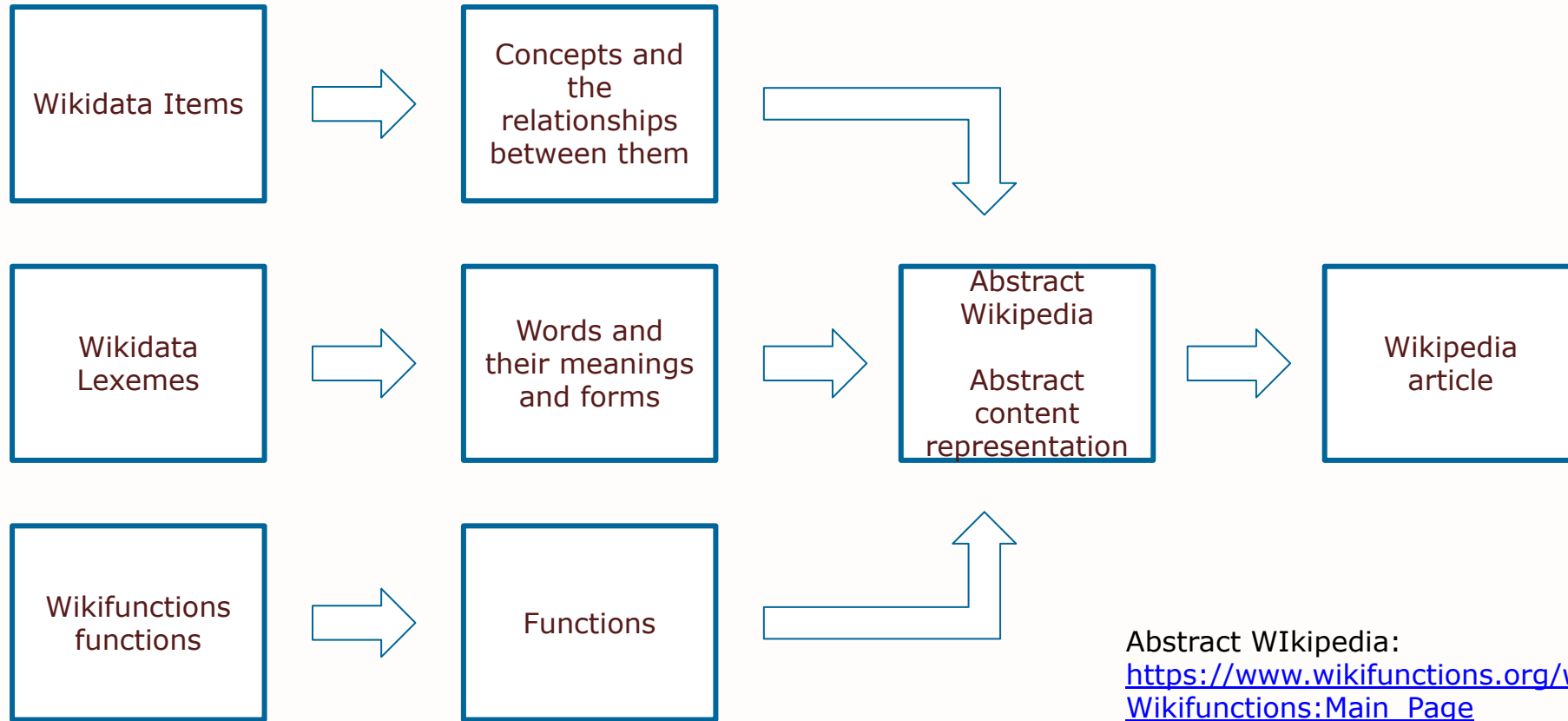
- Sentiment analysis
- Part of speech tagging
- Named entity recognition

Potential users: TextRazor, Wikisource

Text generation

- Generate human-readable text in different languages based on available data
- Text adventures for language learning

Potential users: Abstract
Wikipedia, sports journalism



Abstract Wikipedia:
https://www.wikifunctions.org/wiki/Wikifunctions:Main_Page

The case of “Bachinima”



Documenting Dagbanli native
pronunciations on WikiCommons & Wikidata

...and build first ever dag ASR off it

Azunre, P. et al. (2023). Breaking the Low-Resource Barrier for Dagbani ASR: From Data Collection to Modeling. Retrieved from <https://openreview.net/pdf?id=lje9lI9zV8>



Audio recording tools

- [Spell4Wiki](#)
- [Lingua Libre](#)



Spell4Wiki

Lingua Libre

Record Wizard Discussion Statistics Sound library Help Datasets About


1,000,000

🏆 One million recordings
Lingua Libre just reached this incredible milestone, thanks to 1,400 speakers who contributed in
You can also help documenting your language by [recording](#) you

Welcome to Lingua Libre, the participative linguistic media library of Wikimedia France.

Latest recordings

▶ **Première chapelle Sainte-Barbe de Somain**
French - Jérémy-Günther-Heinz Jähnick



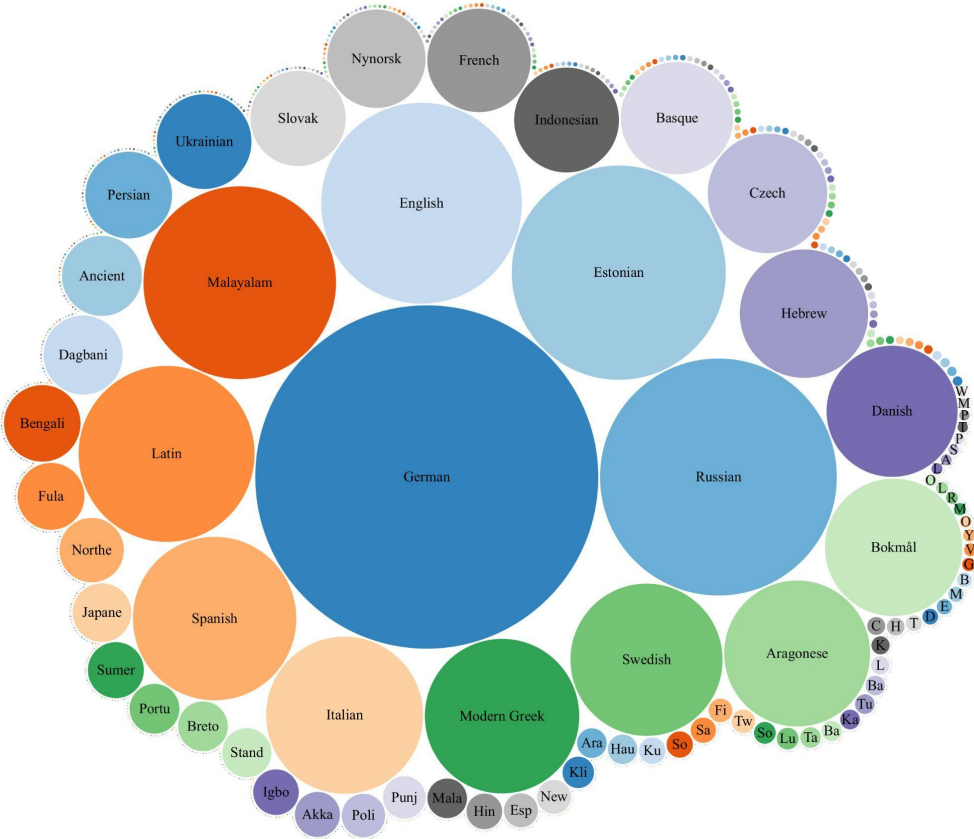
SO...

To enable truly meaningful applications
we need

more data (depth and breadth)

and **more people**

to take care of it.



Abstract Wikipedia Focus languages:

- **Hausa**
- **Igbo**
- **Dagbani**
- Malayalam
- Bengali

Distinct languages of Wikidata Lexemes:
<https://w.wiki/6RiP> (query)

Handson session and Next steps...



Thank you



WikiIndaba



Wikiindaba



wikiindaba



wikimediamorocco

