
WIKIMEDIA
FOUNDATION
Documenting our languages with Lexeme
Asaf Bartov
CEE Meeting 2022
CC by SA 4.0, Lyokoï

WIKIMEDIA FOUNDATION

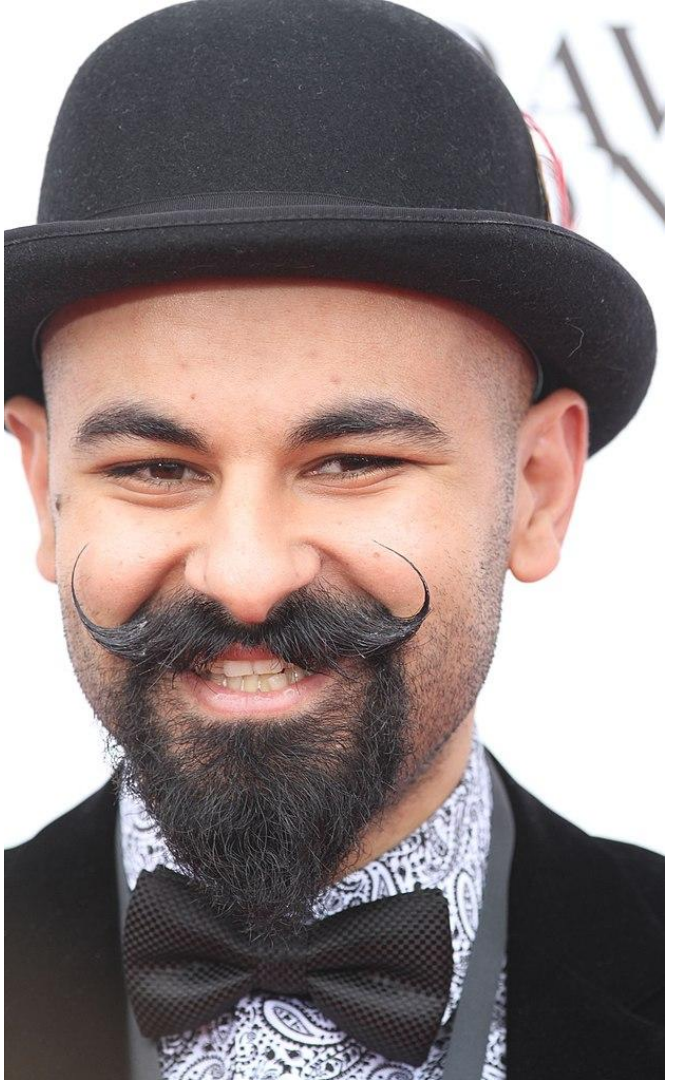
01 What is Lexeme?

And why do I need it in my life?

You are lucky!

You still have time to become **a Lexeme hipster**!

When everybody knows and uses Lexeme, you'll be able to say: "oh yeah, I contributed to Lexeme before it was cool!"

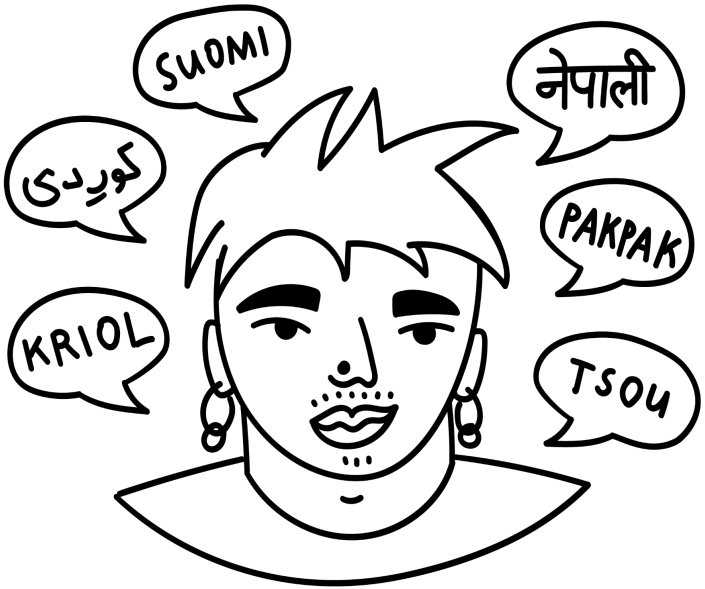


Okay, but why?

Because computers can provide a lot of value for human language **acquisition**, **practice**, **analysis**, **improvement**, and **translation**...

...but for that, they need **structured data** about human languages...

...and human languages are **really complex**!



WIKIMEDIA
FOUNDATION

Is language so complex?

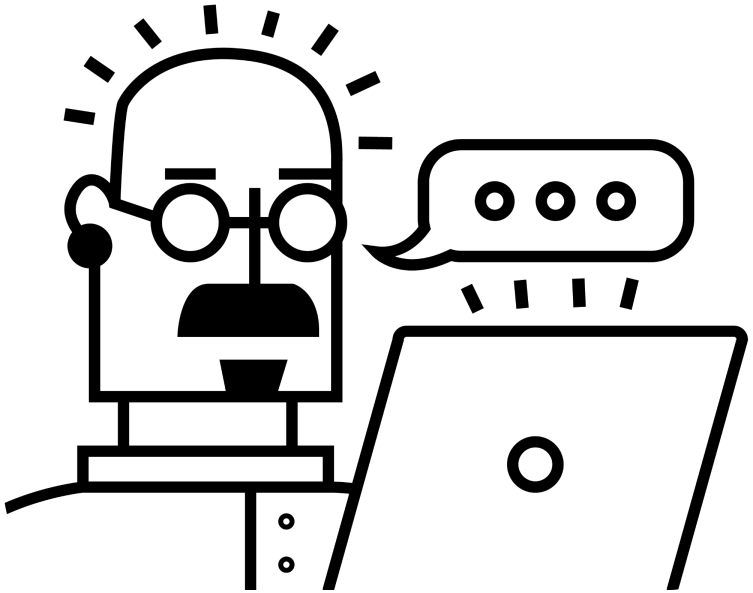
What does 'dog' mean?
-> "guilt began to dog the thief day and night"
What does 'bat' mean?
-> "the owner hit the burglar with a baseball bat"
What does 'mean' mean?
-> he sure was a mean old man
-> the mean income in her country is lower
What translation would be correct and appropriate for 'dog' or 'mean'? It depends on the *specific sense* and *context* in the source text.

WIKIMEDIA
FOUNDATION

Wait, but machine translation exists!

We do have **machine translation** already, and it has gotten a lot better in recent years. But it's still **barely tolerable and generally unreliable** in most languages, including most CEE languages.

The **statistical approach** used by MT *barely understands context*, and therefore flattens nuances, registers, dialects, *even language barriers!* While we all use MT for what it's good for -- getting the gist of a text we cannot read on our own -- there are lots of uses **it is unsuitable for**.



WIKIMEDIA
FOUNDATION

So, language is complex!

- Words have many forms; some irregular (go vs. went) / archaic (thou, dost)
- Words have many senses; some defunct (nice) / regional (lagniappe)
- Senses have many words -- synonyms (peak/summit, clever/smart)
- Homophones (steal vs. steel), homographs (read vs. read)
- Dialectal grammar ("I done saw")
- Register and period ("Hark!"; "Listen!"; "Hey!")
- Lexical overlap and confusion (what soda/pop, doubt, or fanny mean depends on where you live and whom you are speaking to)
- ...and all this is just at the lexeme level, leaving out the world of complexity that is **syntax**!

WIKIMEDIA
FOUNDATION

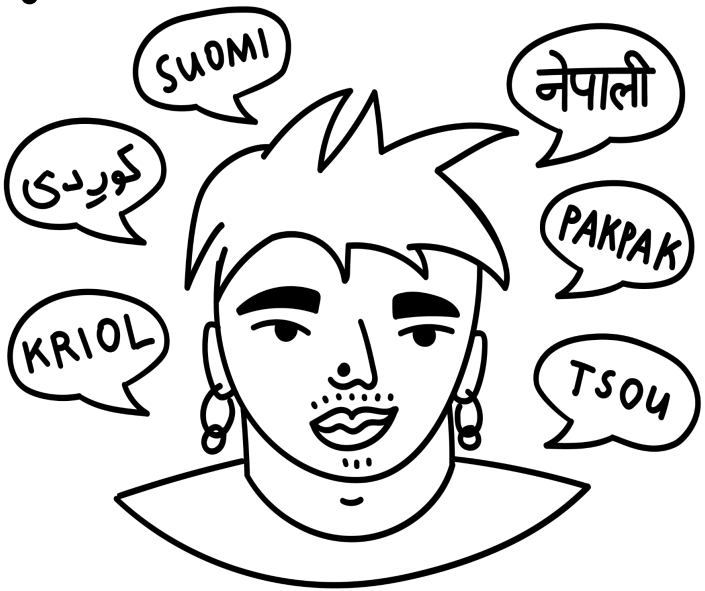
So... it's going to be complex to model as structured data, no?

Yes. :)

But it is really worth it! Because a vast amount of uses will be made possible once we have richly-modeled and linked data about our languages.

Here are just a few prospects. There are others I can think of, and, even more excitingly, still others I *can't even* think of!

And there are cool tools!

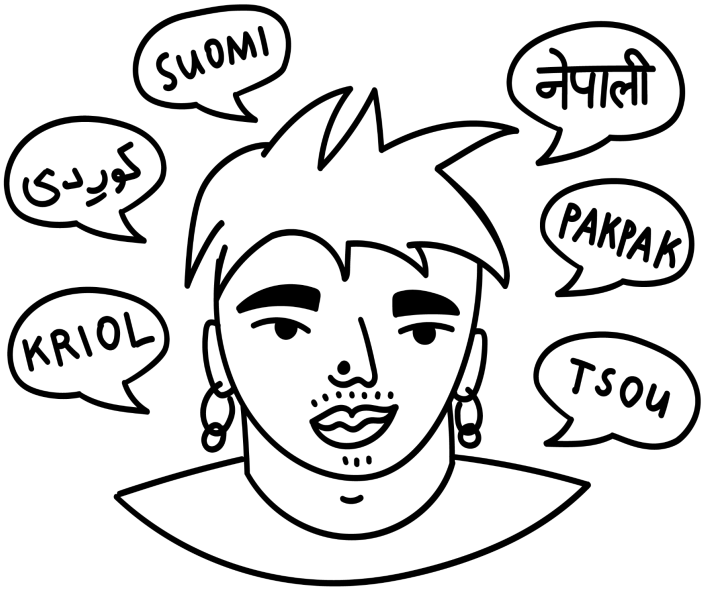


WIKIMEDIA
FOUNDATION

Language acquisition

Structured data about language allows the creation of **language acquisition** software, including:

- flashcard apps
- grammar practice (noun/adjective declension, verb conjugation)
- educational games
- pronunciation practice
- text-reading software with hyper-annotated text (for each word, form and sense analyzed)
- ...and more!

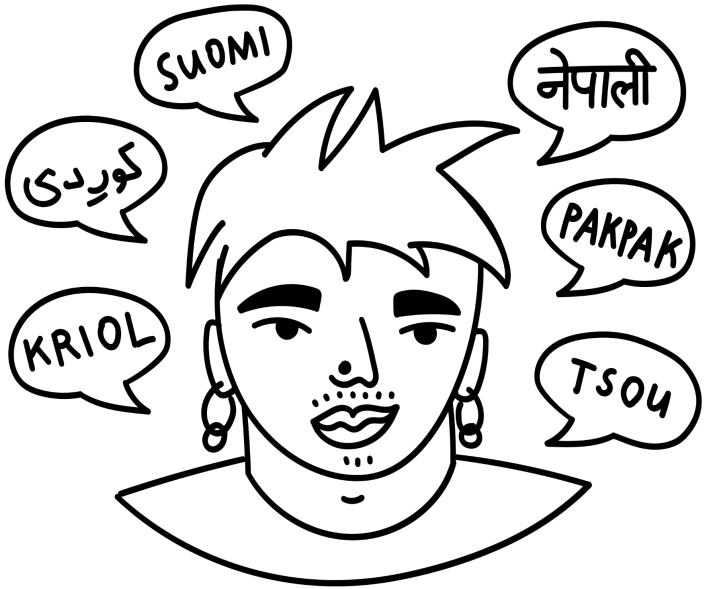


WIKIMEDIA FOUNDATION

Language analysis

Structured data about language allows the creation of **language analysis and improvement** software, including:

- sophisticated spelling/grammar checkers
- crossword solvers
- etymological exploration/research
- stylometry
- stemmatology and phylometry
- ...and more!

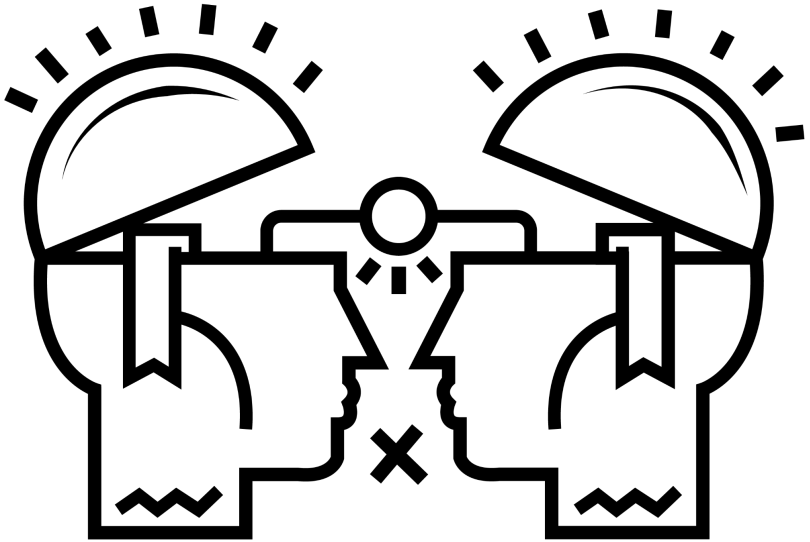


WIKIMEDIA
FOUNDATION

Translation

Correct and adequate translation depends on many factors:

- distinguishing the particular **sense** of the original word/phrase
- contextualizing (genre, register, voice, audience)
- selecting adequate word/phrase in target language, given and preserving the context
- …which is often quite far from a literal word-for-word substitution.

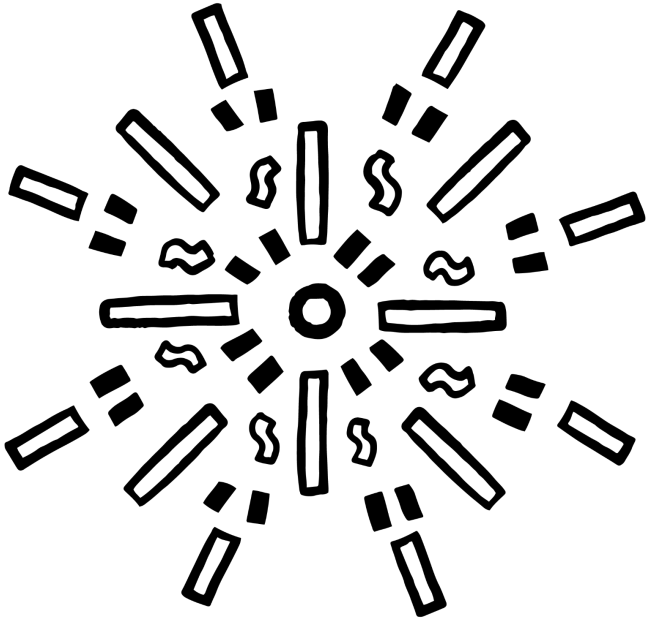


WIKIMEDIA
FOUNDATION

Let's wish upon a star…

What if we had a way to describe lexemes *very precisely*, down to specific **forms** and **senses**?

- To note that *this* form is *nominative* and *this* one *genitive*; this one *imperfect* and this one *pluperfect*?
- To note that a particular form is *regional*, or *archaic*, or *slang*?
- That *one* sense of this lexeme translates into *this* word in German, but *another* sense of this same lexeme translates into *this other* word in German?



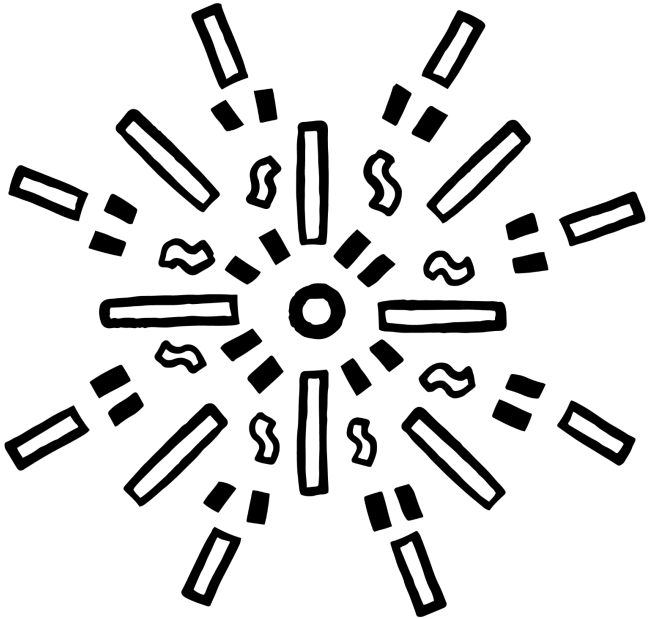
WIKIMEDIA
FOUNDATION

Let's wish upon a star…

That this lexeme **combines** three other lexemes?
That it is **derived** from another lexeme? That it is **borrowed** from another language?
That it denotes *this concept*, which has a (language-independent) Wikidata **item**?

What if we could provide real **example sentences** demonstrating the usage of *each sense* of the lexeme in real texts?

What if we could attach **audio** to each form showing how native speakers **pronounce** it? (Perhaps in more than one way!)

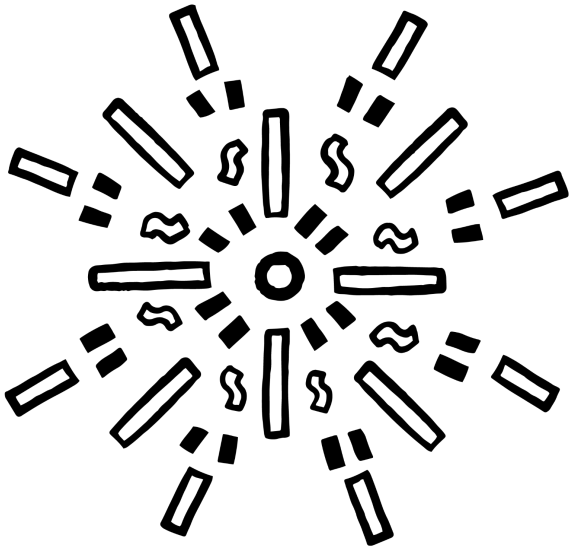


WIKIMEDIA
FOUNDATION

Let's wish upon a star...

What if we could **query** all this, and ask questions like:

- What are some nouns that are masculine in Ukrainian but feminine in German?
- What is the etymological graph of Slavic words for 'horse'?
- What is the longest word in our language without repeating letters?
- What percentage of our language's lexemes have we borrowed from which languages?
- What are some 'false friends' between our language and another? (e.g. Gift in English vs. German)
- How has this lexeme changed in usage over the years, based on actual texts?



WIKIMEDIA
FOUNDATION

WIKIMEDIA FOUNDATION

Guess what?

WIKIMEDIA FOUNDATION

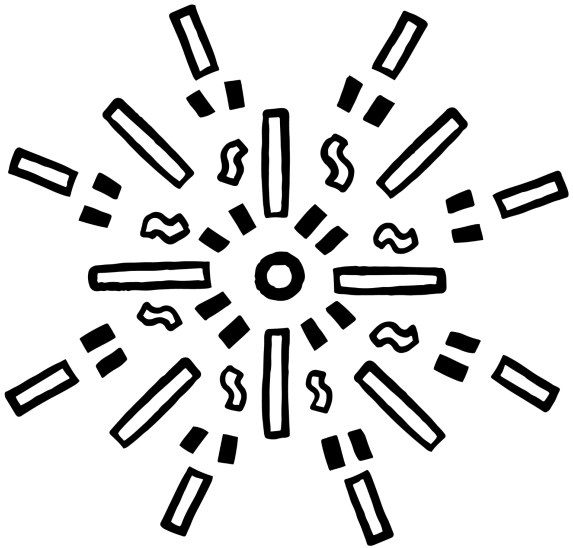
Lexeme can do this right now!

In fact...

Wouldn't it be nice if everyone could speak *your language*?

Until that happens, wouldn't it be nice if we could benefit *in our language* from content written by people who don't speak our language at all, *automagically*?

(o_O)



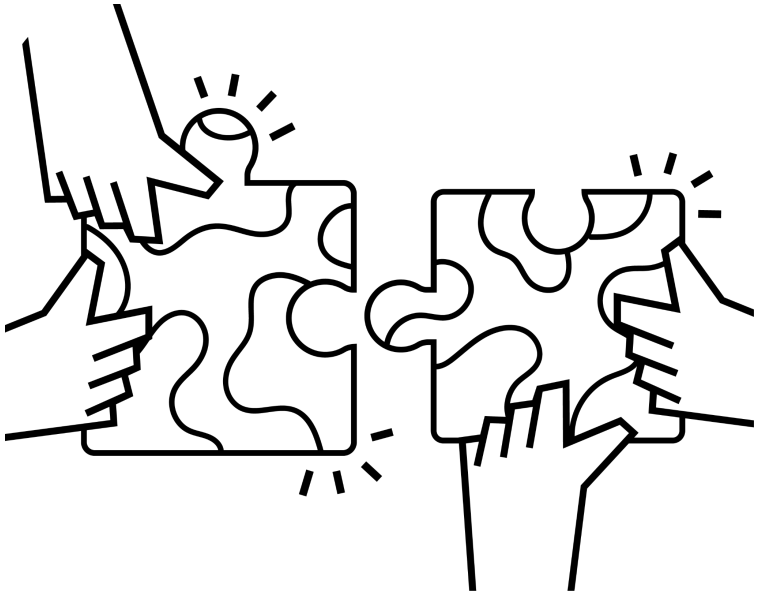
WIKIMEDIA
FOUNDATION

In fact...

Have you heard of **Abstract Wikipedia**? It's going to allow creating "abstract" articles *using code* (programming), from which we could then *generate* human-readable, grammatical, *and accurate* articles *in any language*!

Any language? Well, any language that is *well-described in structured data*!

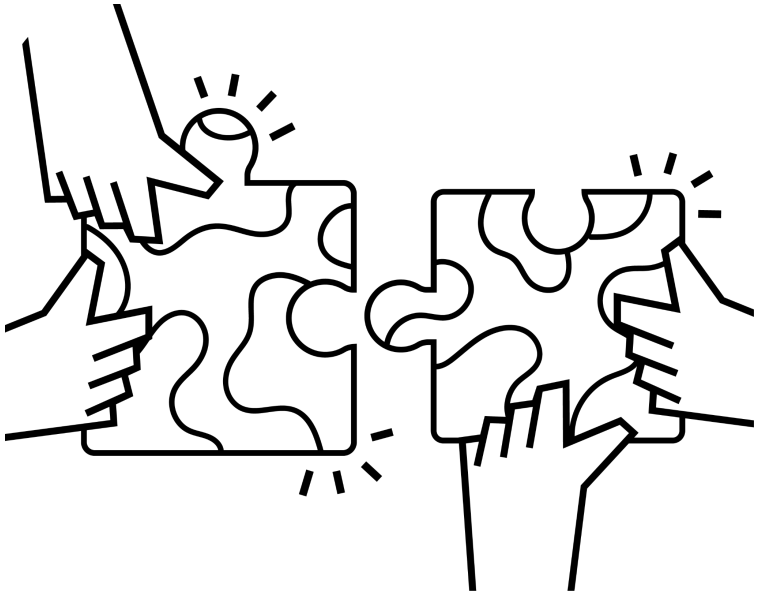
Lexeme is fundamental to Abstract Wikipedia and to **vast** enrichment of content available in your language!



WIKIMEDIA
FOUNDATION

But what *is* Lexeme, exactly?

- It's a **lexicographical layer** on top of the **Wikibase** software running in the **Wikidata** project. "Lexeme" is shorter. :)
- Lexemes are Wikidata entities that exist in parallel to items. **Items ≠ Lexemes**. Items look like **Q212**; Lexemes look like **L34336**.
- We get all the benefits of the wiki; we link into Commons, Wikidata.
- We can use the Wikidata Query Service to query Lexemes (and even Lexemes and items).
- It is a (still) small, friendly, welcoming community



WIKIMEDIA FOUNDATION

Wiktionary?

WIKIMEDIA FOUNDATION

In short:

WIKIMEDIA FOUNDATION

Lexeme is

WIKIMEDIA FOUNDATION

02 A tour of Lexeme

Anatomy and sociology of a lexeme

WIKIMEDIA FOUNDATION

Let's take a look at a lexeme:

https://www.wikidata.org/wiki/Lexeme:L4177

WIKIMEDIA FOUNDATION

Okay, okay, Lexeme is worthwhile!

...

WIKIMEDIA FOUNDATION

But we have lots of questions!

...

WIKIMEDIA FOUNDATION

Such as: how do I know what already exists in my language?

...

WIKIMEDIA FOUNDATION

03 Browsing Lexeme

WIKIMEDIA FOUNDATION

- Ordia
- Hangor
- Lexical coverage report
- …?

WIKIMEDIA FOUNDATION

04 Contributing to Lexeme

Creating a Lexeme

WIKIMEDIA FOUNDATION

1. Lexeme Forms tool (+your lang? +gadget)
2. Orthohin tool (+your lang? +gadget)
3. Entity-suggester script (e.g. L475401)
4. MachtSinn tool -- connect lexemes to items
5. LinguaLibre record pronunciations! (query)
6. Lexeme Party improve by topic (+weekly)
7. Bodh tool -- tabular editing of lexemes.
8. Lexicator tool -- careful mass import from Wiktionary

WIKIMEDIA FOUNDATION

05 Querying Lexeme

WIKIMEDIA FOUNDATION

- Learn SPARQL
- ~~Steal~~Adapt queries

WIKIMEDIA FOUNDATION

06 Fun with Lexeme

WIKIMEDIA FOUNDATION

- Der, Die, Das
- Він, вона, воно
- ...?

Many more are yet to be invented! :)

WIKIMEDIA FOUNDATION

07 Next steps

How do we put our Lexeme hipsterism into action?

WIKIMEDIA FOUNDATION

These are early days!

1. We're figuring things out
2. Your input matters!
3. Take initiative; don't wait
4. Ask, discuss, invite

WIKIMEDIA FOUNDATION

What can you do now?

1. Explore Lexeme on your own
2. Add new lexemes
3. Add new forms, senses, examples to existing lexemes
4. Show Lexeme to your peers

WIKIMEDIA FOUNDATION

Ideally, lead lexeme adoption in your language!

WIKIMEDIA FOUNDATION

1. Check coverage of your lang
2. Start a WikiProject!
   - tutorials
   - to-do lists / queries
   - off-wiki channels

WIKIMEDIA FOUNDATION

Thank you for your attention!