
WIKIMEDIA
FOUNDATION
Documentando nuestros idiomas con Lexeme
Asaf Bartov
CEE Meeting 2022
CC by SA 4.0, Lyokoï

WIKIMEDIA FOUNDATION

01

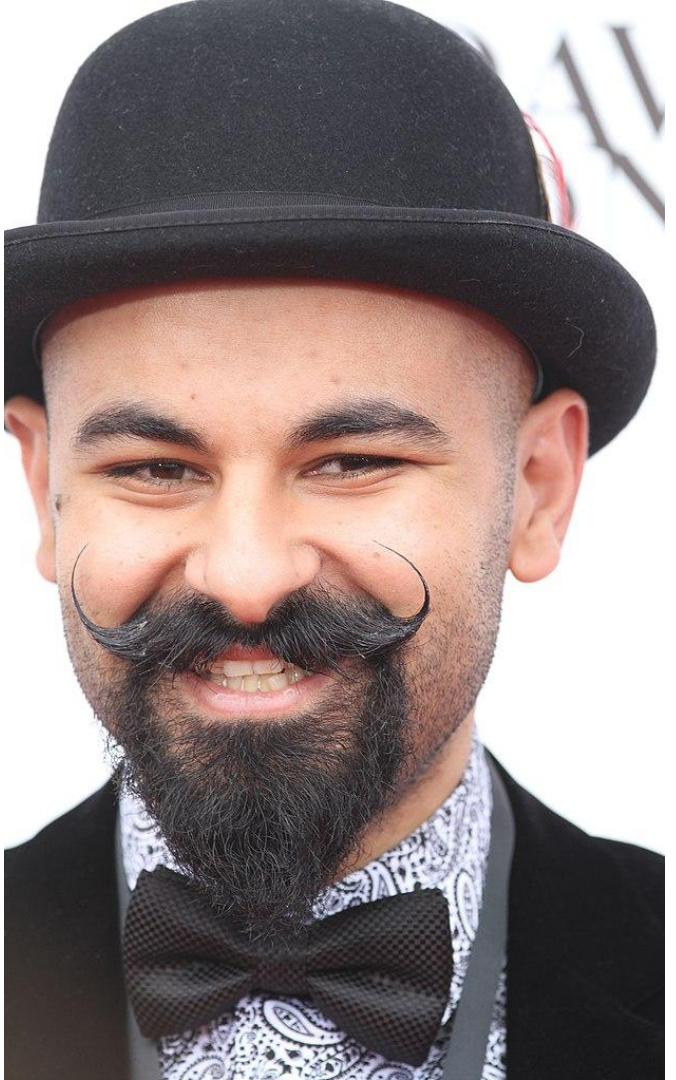
¿Qué es Lexeme?

¿Y porqué lo necesito en mi vida?

¡Estás de suerte!

Aún estás a tiempo de convertirte **en un hipster de Lexeme**!

Cuando todos conozcan y usen Lexeme, podrás decir: "¡Oh, sí, yo contribuía a Lexeme antes de que se pusiera de moda!"

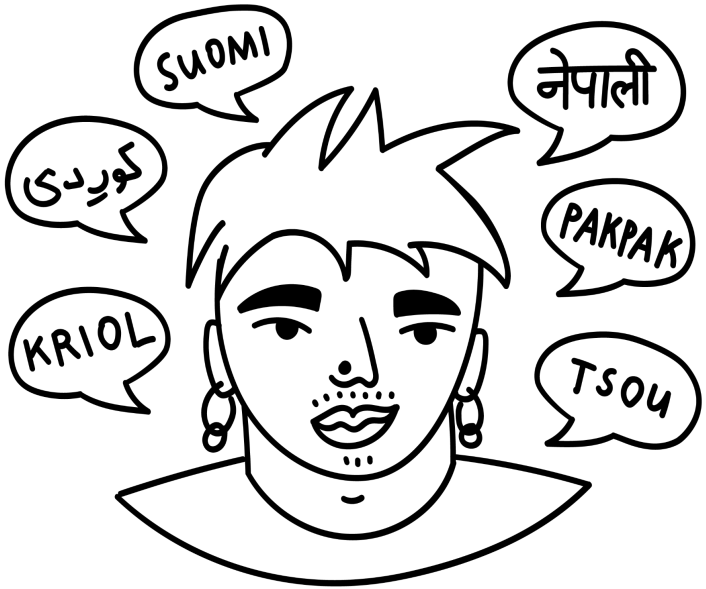


Vale, ¿pero para qué?

Porque los ordenadores pueden proporcionar mucho valor para el lenguaje humano **adquisición, práctica, análisis, mejora y traducción...**

...pero para ello necesitan **datos estructurados** sobre lenguas humanas...

...y las lenguas humanas son **realmente complejas**!



WIKIMEDIA
FOUNDATION

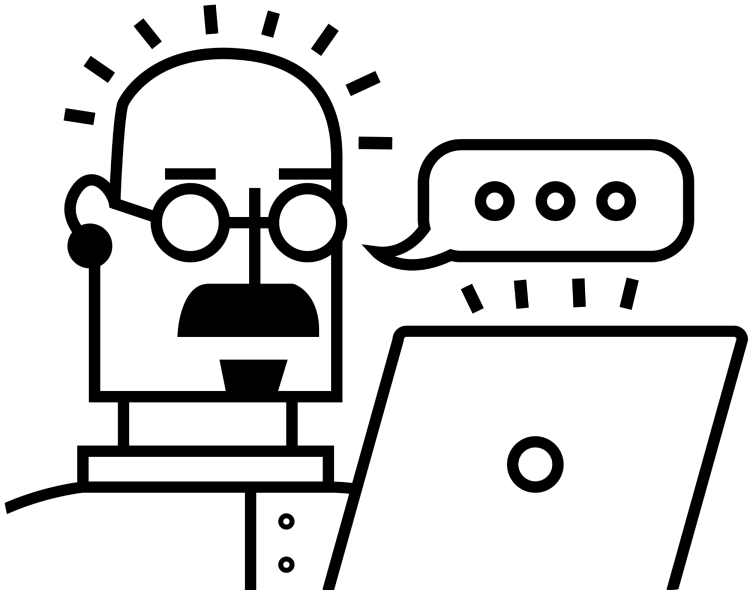
¿Es el lenguaje humano tan complicado?

Qué significa dog?
-> "guilt began to dog the thief day and night"
Qué significa 'bat'?
-> "the owner hit the burglar with a baseball bat"
Qué significa 'mean'?
-> he sure was a mean old man
-> the mean income in her country is lower
¿Cual es traducción acertada para 'dog' o 'mean'? Depende del *sentido específico* y el *contexto* en el texto de origen.

WIKIMEDIA
FOUNDATION

Eh, pero existe traducción automática

Ya tenemos **traducción automática**, y ha mejorado mucho en los últimos años. Pero aún es *apenas tolerable y generalmente poco confiable* en la mayoría de los idiomas, incluidos la mayoría de los idiomas de CEE.



El **enfoque estadístico** utilizado por MT *apenas comprende el contexto* y, por lo tanto, simplifica los matices, los registros, los dialectos, ¡*incluso las barreras del idioma*! Si bien todos usamos MT para lo que es bueno (obtener la esencia de un texto que no podemos leer por nosotros mismos), hay **muchos usos para los que**

WIKIMEDIA FOUNDATION

Así que los idiomas son complejos!

- Las palabras tienen formas; algunas irregulares (ir/fue) / arcaicas (aqueste)
- Las palabras tienen significados; algunos difuntos (orage) / regionales (coger)
- Las palabras tienen sentidos - sinónimos (famoso/célebre)
- Homófonos (a ver y haber), homógrafos (banco)
- Gramática dialectal("La vi/Le vi")
- Registro y periodo (Oiga! Escucha! Ey!)
- Superposición y confusión léxica (estampita significa una cosa u otra dependiendo dónde esté)
- ...y todo esto es solo a nivel de lexemas, dejando fuera el mundo de complejidad que es la **sintaxis**!

WIKIMEDIA
FOUNDATION

Entonces... va a ser complejo modelar datos estructurados, ¿no?

Sí. :)

¡Pero realmente merece la pena! Porque una gran cantidad de usos serán viables una vez que tengamos datos ricamente modelados y vinculados sobre nuestros idiomas.

Aquí hay algunas perspectivas. Hay otras en los que puedo pensar y, aún más emocionante, ¡otros usos de los que ni siquiera puedo imaginar!

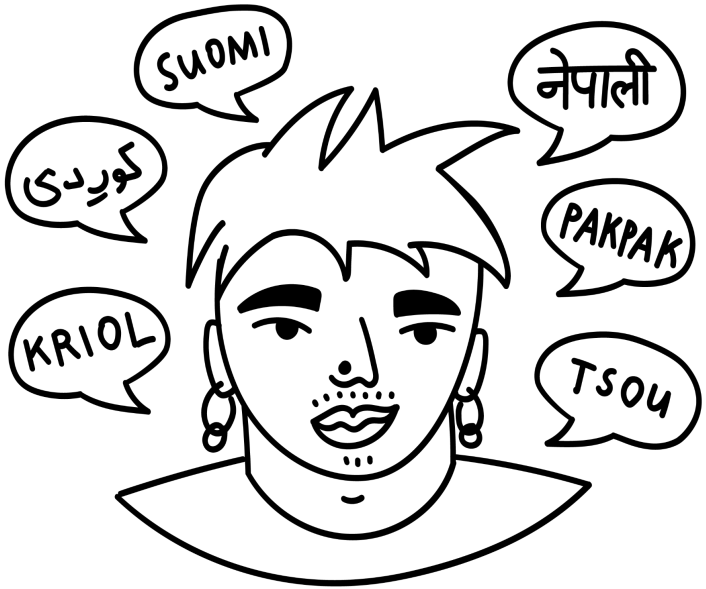
¡Y hay herramientas chulas!

WIKIMEDIA
FOUNDATION

Adquisición de lenguajes

Los datos estructurados sobre el lenguaje permiten la creación de software de **Adquisición de lenguajes**, que incluyen:

- Apps con flashcard
- Prácticas de gramática (declinación sustantivo/adjetivo, conjugación verbal)
- Juegos educacionales
- Prácticas de pronunciación por
- software de lectura de texto con texto hiperanotado (para cada palabra, forma y sentido analizado)
- ...¡y más!

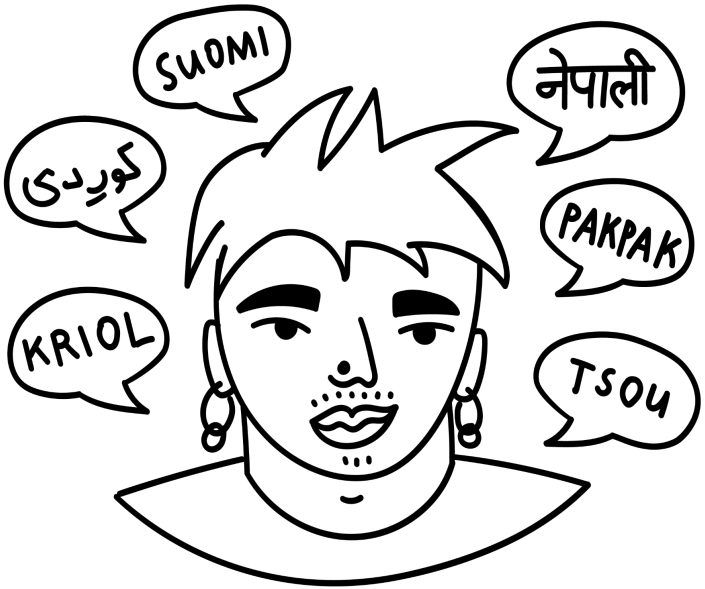


WIKIMEDIA FOUNDATION

Análisis de lenguaje

Los datos estructurados sobre el lenguaje permiten la creación de software de **análisis y mejora del lenguaje**, que incluye:

- sofisticados correctores ortográficos/gramaticales
- solucionadores de crucigramas
- exploración/investigación etimológica
- estilometría
- Talmatologia y filometria
- ...¡y más!

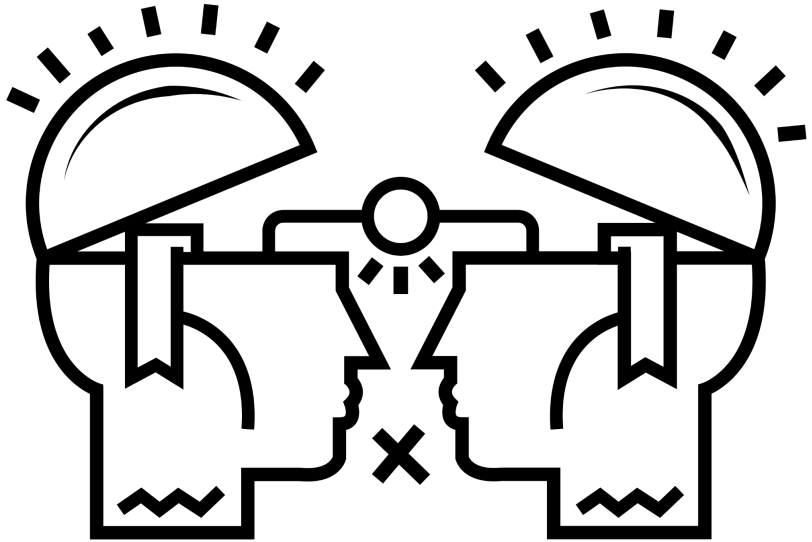


WIKIMEDIA
FOUNDATION

Traducción

Una traducción correcta y adecuada depende de muchos factores:

- distinguir el **sentido** particular de la palabra/frase original
- contextualizar (género, registro, voz, audiencia)
- seleccionar la palabra/frase adecuada en el idioma de destino, y preservar el contexto
- … que a menudo está bastante lejos de una sustitución literal palabra por palabra.

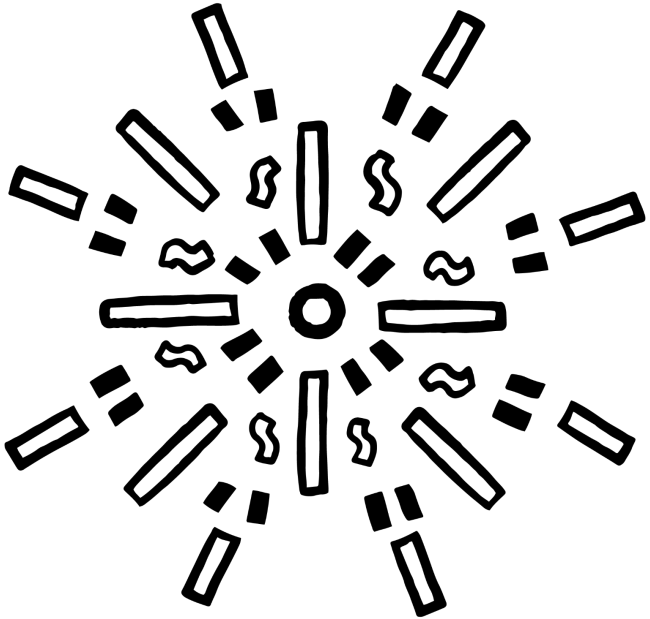


WIKIMEDIA
FOUNDATION

Pidamos deseos a un pozo

¿Qué pasaría si tuviéramos una manera de describir los lexemas con *mucha precisión*, hasta **formas** y **sentidos** específicos?

- Notar que *esta* forma es nominativa y *ésta* genitiva; *este* imperfecto y *este* pluscuamperfecto?
- ¿Notar que una forma particular es *regional*, o *arcaica*, o *jerga*?
- ¿Que un sentido de este lexema se traduce en *esta* palabra en alemán, pero *otro* sentido de este mismo lexema se traduce en *esta* otra palabra en alemán?



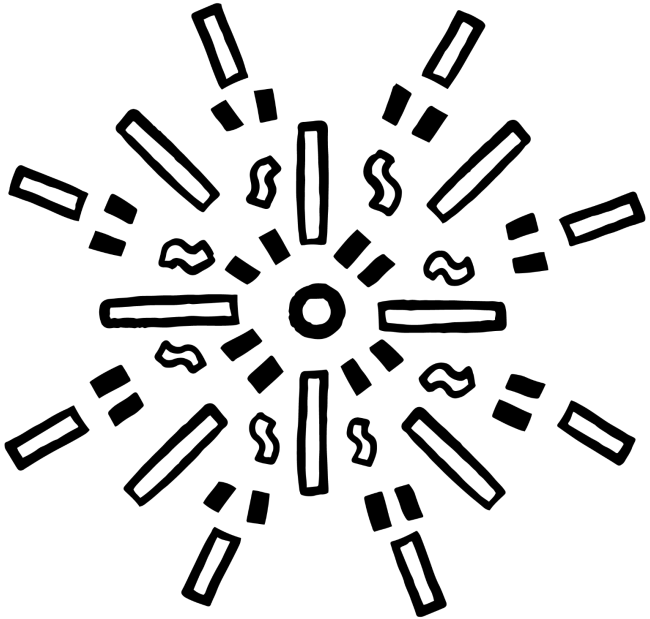
WIKIMEDIA
FOUNDATION

Pidamos deseos a un pozo

¿Que este lexema **combina** otros tres lexemas?
¿Qué se **deriva** de otro lexema? ¿Qué es **prestado** de otro idioma?
¿Que denota *este concepto*, que tiene un elemento de Wikidata (**independientemente** del idioma)?

¿Qué pasaría si pudiéramos proporcionar oraciones con **ejemplos reales** que demuestren el uso de *cada sentido* del lexema en textos reales?

¿Qué pasaría si pudiéramos adjuntar **audio** a cada formulario que muestre cómo lo **pronuncian** los hablantes nativos? (¡Quizás en más de una forma!)

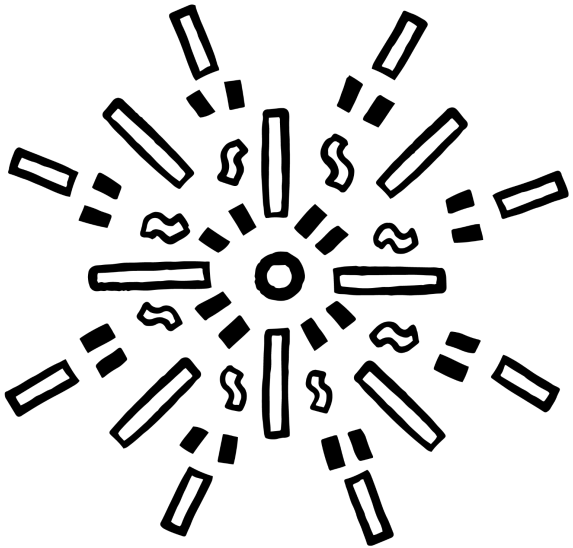


WIKIMEDIA
FOUNDATION

Pidamos deseos a un pozo

¿Qué pasaría si pudiéramos **consultar** todo esto y hacer preguntas como:

- ¿Cuáles son algunos sustantivos que son masculinos en ucraniano pero femeninos en alemán?
- ¿Cuál es el gráfico etimológico de las palabras eslavas para 'caballo'?
- ¿Cuál es la palabra más larga en nuestro idioma sin repetir letras?
- ¿Qué porcentaje de los lexemas de nuestro idioma hemos tomado prestados de qué idiomas?
- ¿Cuáles son algunos 'falsos amigos' entre nuestra lengua y otra? (*sensible* en inglés o español)
- ¿Cómo ha cambiado el uso de este lexema a lo largo de los años, según los textos reales?



WIKIMEDIA FOUNDATION

WIKIMEDIA FOUNDATION

¿A que no sabéis...?

WIKIMEDIA FOUNDATION

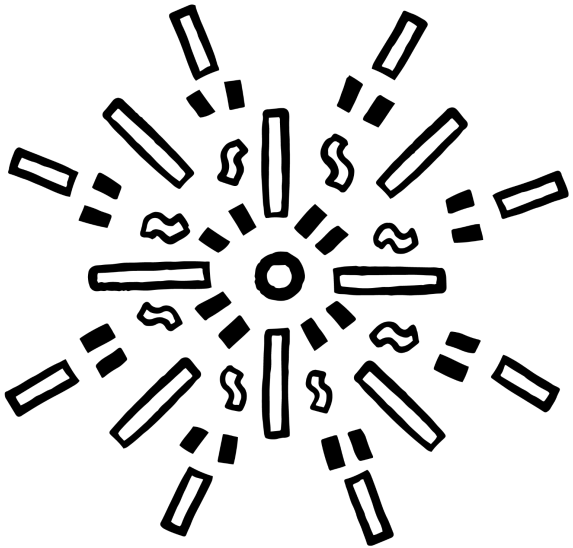
Lexeme ya puede hacer todo esto!

De hecho...

¿No sería bueno si todos pudieran hablar *tu idioma*?

Hasta que eso suceda, ¿no sería bueno si pudiéramos beneficiarnos en *nuestro idioma* del contenido escrito por personas que no hablan nuestro idioma en absoluto, de forma automática?

(o_O)



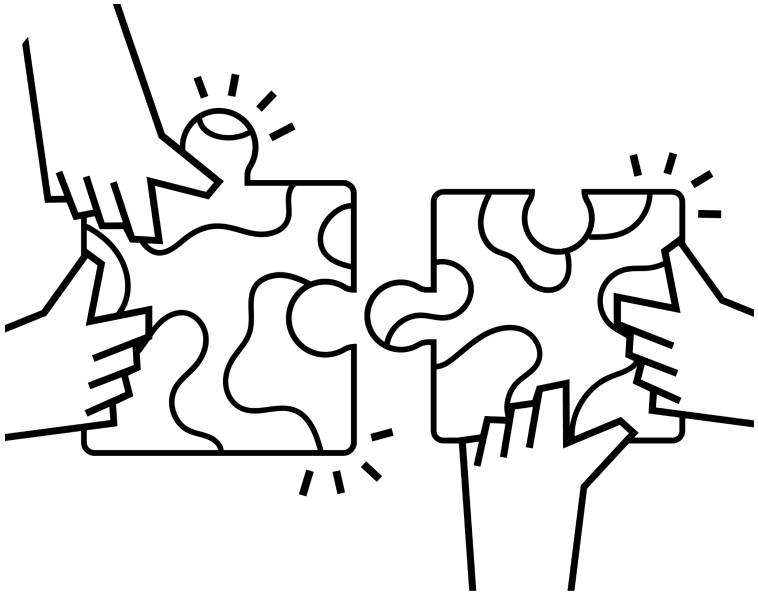
WIKIMEDIA
FOUNDATION

In fact...

¿Has oído hablar de la **Wikipedia abstracta**? ¡Va a permitir crear artículos "abstractos" *usando código* (programación), a partir de los cuales podríamos generar artículos legibles por humanos, *gramaticalmente precisos* en cualquier idioma!

¿Cualquier idioma? Bueno, *¡cualquier idioma que esté bien descrito en datos estructurados!*

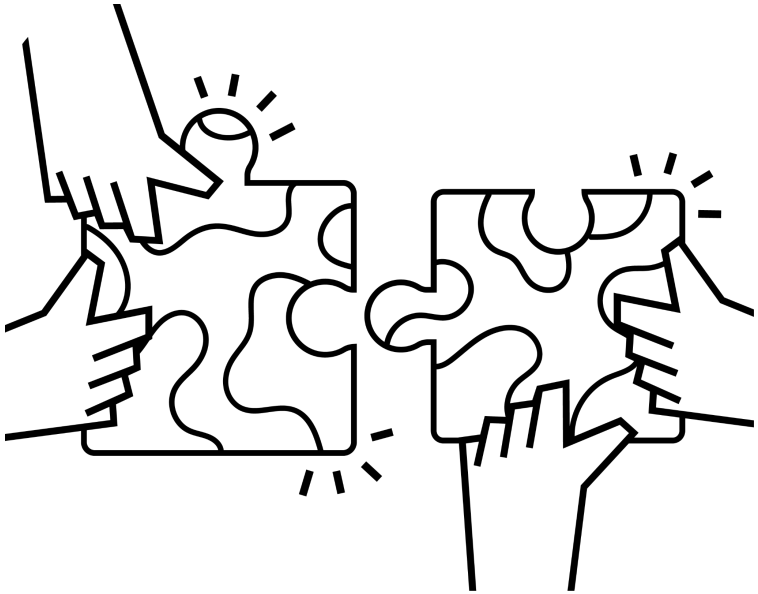
¡**Lexeme** es fundamental para la Wikipedia abstracta y para el gran enriquecimiento del contenido disponible en cualquier idioma!



WIKIMEDIA
FOUNDATION

¿Pero qué es es Lexeme, exactamente?

- Es una **capa lexicográfica** sobre el software **Wikibase** que se ejecuta en el proyecto **Wikidata**. "Lexeme" es más corto. :)
- Lexemes son entidades de Wikidata que existen en paralelo a los elementos. **Items ≠ Lexemes**. Los item tienen formas como **Q212**; Lexemes tienen formas como **L34336**.
- Obtenemos todos los beneficios de los proyectos wiki; vinculamos a Commons, Wikidata.
- Se puede usar Wikidata Query Service para hacer query de Lexemes (e incluso Lexemes y items).
- Es una comunidad (todavía) pequeña, amigable y acogedora.



WIKIMEDIA FOUNDATION

Wiktionary?

WIKIMEDIA FOUNDATION

En resumen:

WIKIMEDIA FOUNDATION

Lexemes

WIKIMEDIA FOUNDATION

02 Un tour de Lexeme

Anatomía y sociología de Lexeme

WIKIMEDIA FOUNDATION

Veamos un lexeme

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

WIKIMEDIA FOUNDATION

Vale, vale, ¡nos has convencido de que Lexeme merece la pena!

...

WIKIMEDIA FOUNDATION

¡Pero aún tenemos muchos interrogantes!

...

WIKIMEDIA FOUNDATION

En plan: ¿cómo sé lo que ya existe en mi idioma?

...

WIKIMEDIA FOUNDATION

03 Navegando entre Lexemes

WIKIMEDIA FOUNDATION

- Ordia
- Hangor
- Lexical coverage report
- …?

WIKIMEDIA FOUNDATION

04 Contribuyendo a Lexeme

Creando un 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 Haciendo Querys de Lexeme

WIKIMEDIA FOUNDATION

- Aprende SPARQL
- ~~Roba~~Adapta query

WIKIMEDIA FOUNDATION

06 Diviértete con Lexeme

WIKIMEDIA FOUNDATION

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

¡Y más que se inventarán! :)

WIKIMEDIA FOUNDATION

07 Siguientes pasos

¿Cómo poner el hipsterismo de Lexeme en acción?

WIKIMEDIA FOUNDATION

¡Estamos empezando!

1. Figurando cosas
2. ¡Tu opinión importa!
3. Lánzate, no esperes
4. Pregunta, conversa, invita

WIKIMEDIA FOUNDATION

¿Qué podemos hacer?

1. Explora Lexeme por ti misma
2. Añade nuevos lexemes
3. Añade nuevas formas y sentidos a los lexemes creados
4. Enseña Lexeme

WIKIMEDIA FOUNDATION

Lo ideal: ¡abandera la adopción de Lexeme en tu idioma!

WIKIMEDIA FOUNDATION

1. Checa cobertura de tu lang
2. Empieza un WikiProyecto!
   - tutoriales
   - Listas / queries de cosas a hacer
   - Canales off-wiki

WIKIMEDIA FOUNDATION

¡Gracias por su atención!