WIKIMEDIA FOUNDATION

Що таке «лексема» і навіщо вона потрібна у вашому житті?

Асаф Бартóв

Українська Вікіконференція 2022

WIKIMEDIA FOUNDATION

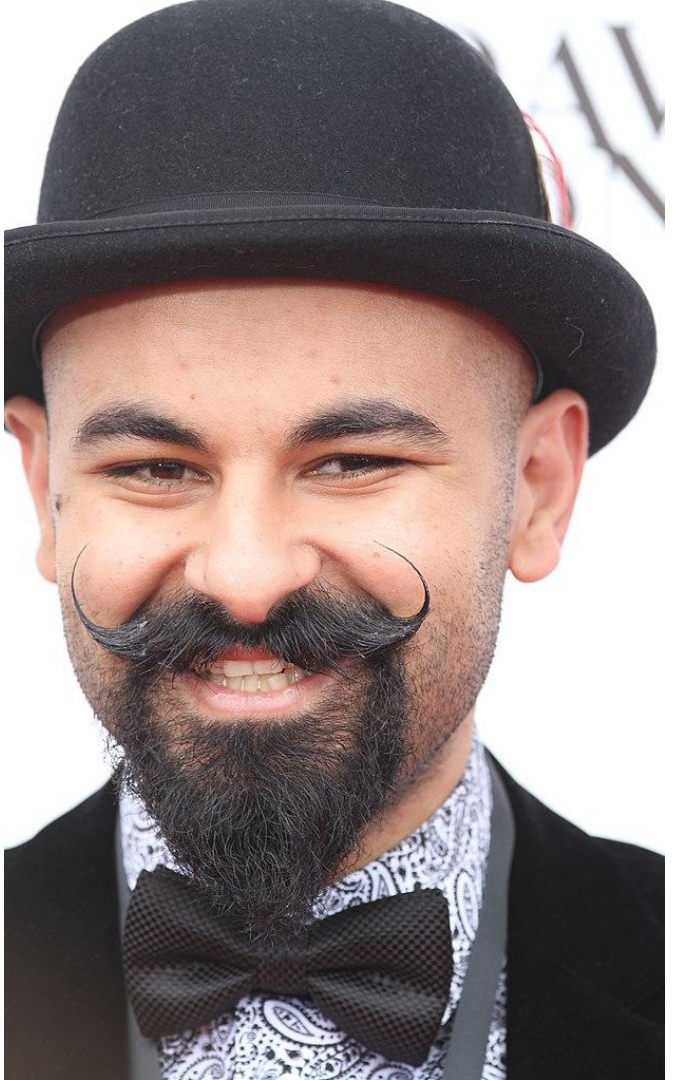
Що таке «лексема»?

І навіщо вона потрібна у моєму житті?

Вам пощастило!

У вас все ще є час стати **лексемним хіпстером**!

Коли всі знатимуть і використовуватимуть лексеми, ви могтимете сказати: «Ага, я роблю внесок до лексем ще з часів, коли це не було так круто».

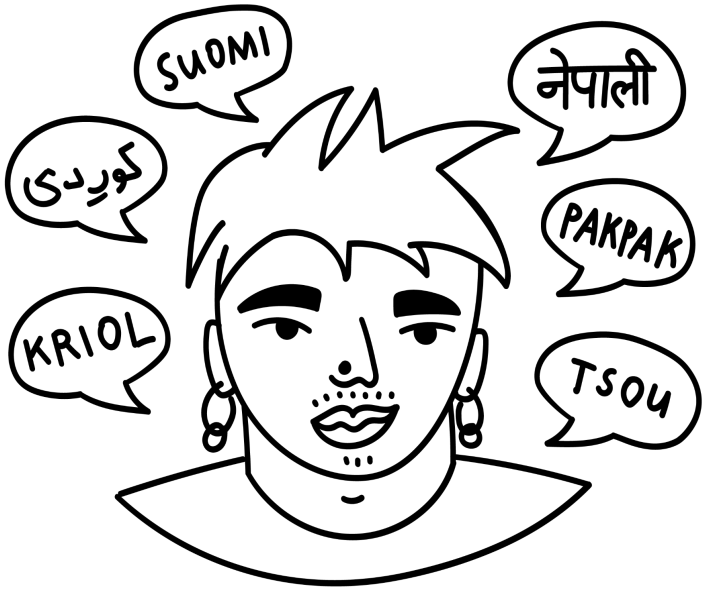


Гаразд, але чому?

Тому що комп'ютери можуть забезпечити дуже багато для **засвоєння** мови людиною, **практики**, **аналізу**, **покращення** і **перекладу**...

...але для цього їм потрібні **структуровані дані** про людські мови...

...а людські мови **дуже складні**!



WIKIMEDIA
FOUNDATION

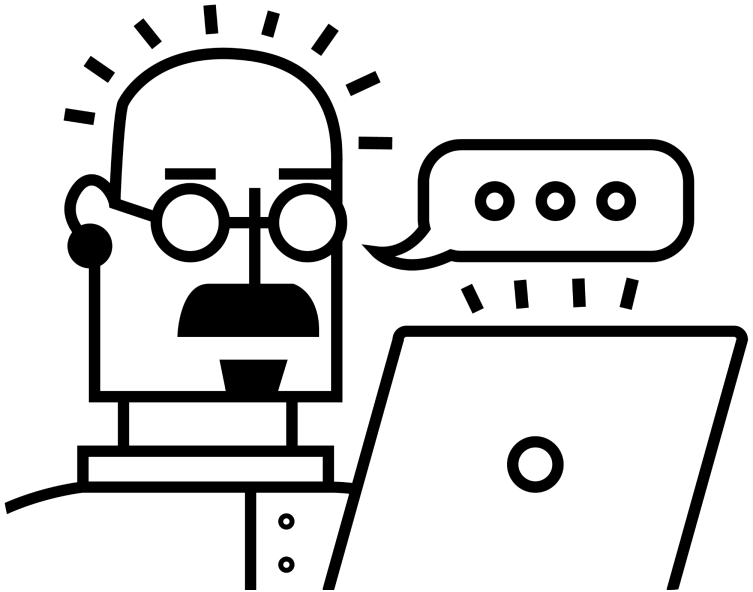
Чи справді мова складна?

dog = собака; effect = вплив; обід = lunch тощо.
правильно?
не… завжди!
Англійською «dog» може бути дієсловом зі значенням переслідувати когось весь час, як мисливський пес, напр. "guilt dogged him day and night".
Слово «effect» також може бути дієсловом, зі значенням спричиняти(!), призводити, напр. "she wants to effect change".
А «обід» може також означати кільце чи коло (обруч).
Який переклад буде правильним і доцільним для слів dog та effect? Це залежить від *конкретного випадку* і *контексту* вихідного тексту.

Чекайте, але ж існує машинний переклад!

Ми вже маємо **машинний переклад**, і за останні роки він значно покращав. Але все ж він **заледве прийнятний і загалом ненадійний** для більшості мов, включно з українською.

Статистичний підхід, який використовує МП, *заледве розуміє контекст*, і тому опускає нюанси, регістри, діалекти, *навіть мовні бар'єри*! Хоча усі ми використовуємо МП там, де він годиться — щоб зрозуміти, про що текст невідомою мовою, — є купа випадків, де **він не підходить**.



WIKIMEDIA FOUNDATION

Тож мова справді складна!

- Слова мають багато форм; деякі неправильні (їм vs. їсти) / архаїчні (дві відрі)
- Слова мають багато значень (обід vs. обід); деякі невживані (благий) / регіональні (вуйко; читать; ш(ь)епка)
- Омоніми, синоніми, омографи (за́мок vs. замо́к)
- Діалектна граматика (я був робив)
- Регістр і період (це vs. се)
- Лексичне накладання і плутанина (що означає «гарбуз», «склеп» чи «пустий», залежить від того, де ви живете і з ким говорите)
- …і все це лише на рівні лексем, навіть не торкаючись цілого світу складності під назвою **синтаксис**!

То… складно буде змоделювати це у вигляді структурованих даних?

Так. :)

Але це справді того варте! Тому що коли ми матимемо детально змодельовані і пов’язані дані про наші мови, для них знайдеться купа використань.

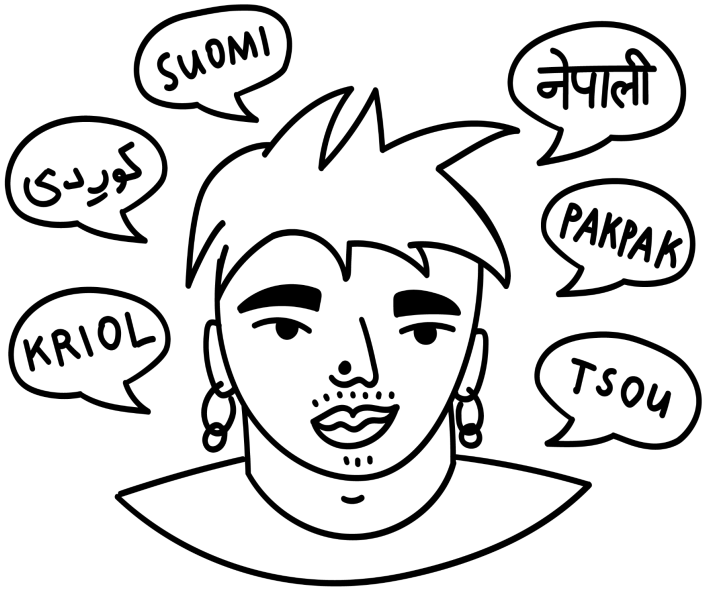
Ось лише кілька перспектив. Є інші, які спадають мені на думку, і що ще більш захопливо, інші, які я *навіть не уявляю*!

WIKIMEDIA FOUNDATION

Засвоєння мови

Структуровані дані про мову дозволяють створювати програмне забезпечення для **засвоєння мови**, зокрема:

- картки
- вправи на граматику (відмінки іменників/прикметників, форми дієслів)
- освітні ігри
- вправи на вимову
- програмне забезпечення для читання тексту (де аналізується кожне слово, форма і значення)
- ...та багато іншого!

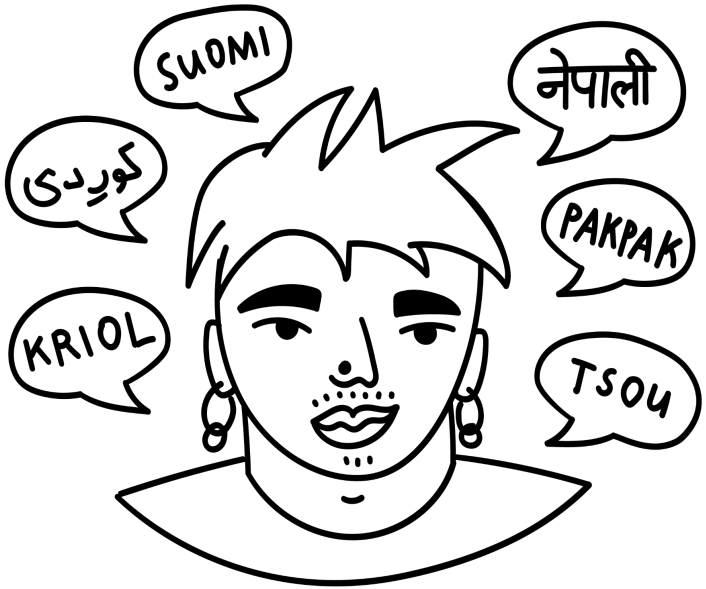


WIKIMEDIA FOUNDATION

Мовний аналіз

Структуровані дані про мову дозволяють створювати програмне забезпечення з **аналізу мови й покращення мовлення** для такого:

- перевірка написання
- розв’язування кросвордів
- етимологічні дослідження
- стилеметрія
- стемматологія та філометрія
- ...та інше!

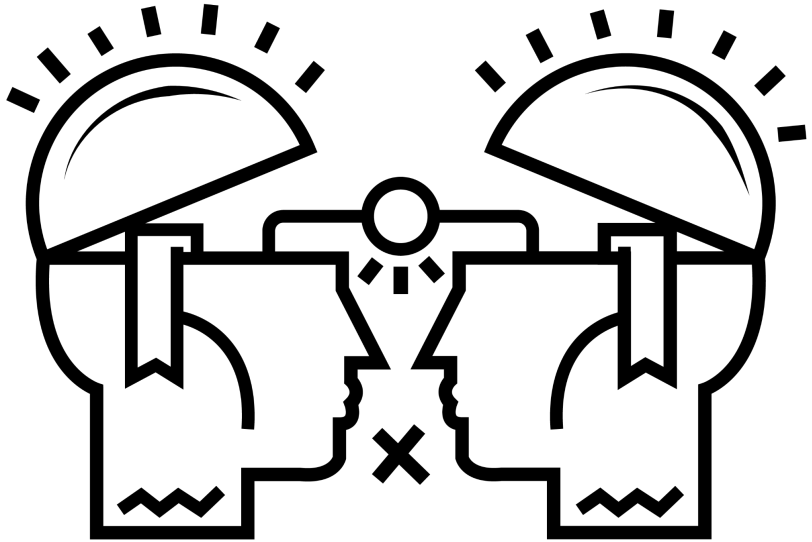


WIKIMEDIA
FOUNDATION

Переклад

Правильний та адекватний переклад залежить від багатьох факторів:

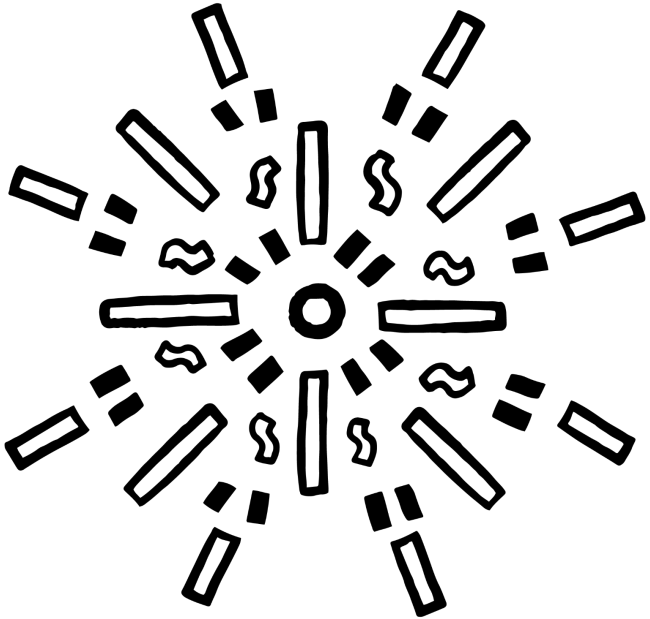
- розрізнення конкретного значення оригінального слова/фрази
- контекстуалізація (жанр, регістр, голос, аудиторія)
- вибір відповідного слова/фрази цільовою мовою, зважаючи на і зберігаючи контекст
- ...що зазвичай доволі далеко від простої дослівної підстановки.



WIKIMEDIA
FOUNDATION

Загадаймо бажання...

А що якби був спосіб описувати лексеми дуже точно, аж до окремих форм і значень?
Зазначати, що ось ця форма є називним відмінком, а ота родовим; це імперфект, а ось те плюсквамперфект?
Зазначати, що ось ця конкретна форма є регіональною, архаїчною чи сленгом?
Що *одне* значення цієї лексеми перекладається *цим* словом німецькою, а *інше* значення цієї самої лексеми перекладається ось *цим іншим* словом німецькою?



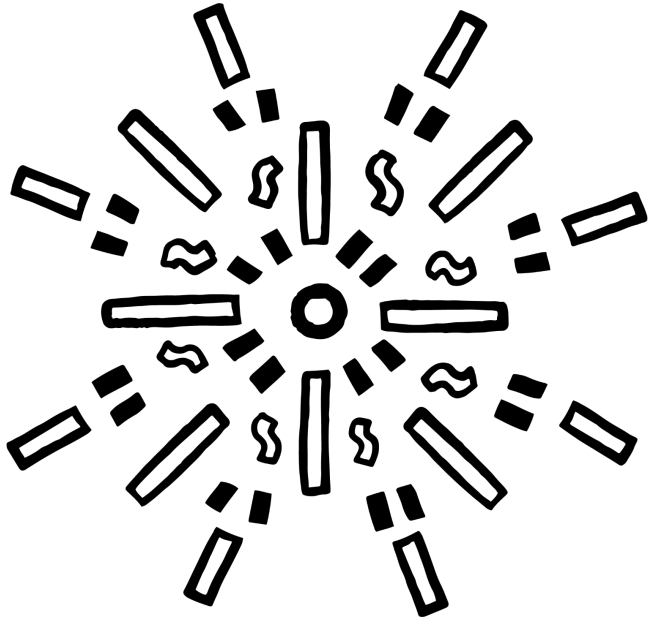
WIKIMEDIA
FOUNDATION

Загадаймо бажання...

Що ця лексема пов’язується з трьома іншими лексемами? Що вона походить від іншої лексеми? Що вона запозичена з іншої мови? Що вона позначає *це явище*, яке має (не залежний від мови) елемент Вікіданих?

Що якби ми могли забезпечити приклади реальних речень, що використовують *кожне значення* лексеми у реальних текстах?

Що якби ми могли прикріпити аудіо до кожної форми, з тим, як мовці її вимовляють? (Може навіть різними способами!)

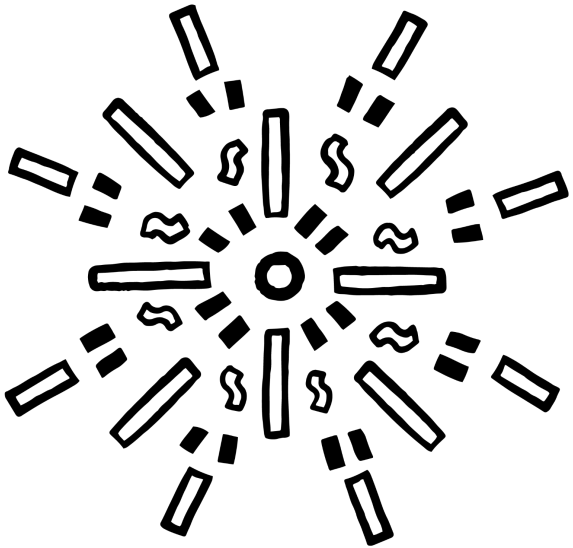


WIKIMEDIA
FOUNDATION

Загадаймо бажання...

Що якби ми могли **робити запити** по всіх цих даних і ставити, наприклад, такі запитання:

- Які іменники мають чоловічий рід українською, але жіночий німецькою?
- Як виглядає етимологічне дерево слов'янських слів зі значенням «кінь»?
- Яке в нашій мові найдовше слово без повторюваних літер?
- Який відсоток лексем нашої мови запозичено з яких мов?
- Які є «хибні друзі перекладача» у нашій і ще якійсь мові? (напр. Gift англійською і німецькою)
- Як змінювалося використання цієї лексеми з часом, на основі реальних текстів?



WIKIMEDIA FOUNDATION

WIKIMEDIA FOUNDATION

Вгадайте що?

WIKIMEDIA FOUNDATION

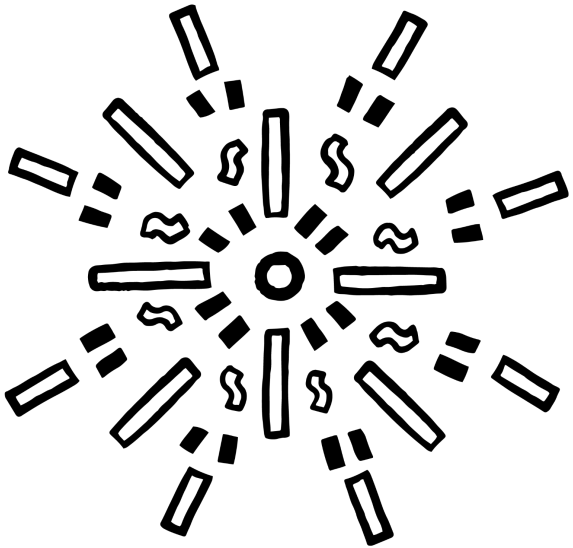
Lexeme може це робити вже зараз!

Якщо подумати...

Хіба не було б чудово, якби всі могли розмовляти українською?

А до того часу, хіба не було б чудово, якби ми могли отримувати користь *українською* від контенту, створеного людьми, які не володіють українською, *автомагічно*?

(o_O)



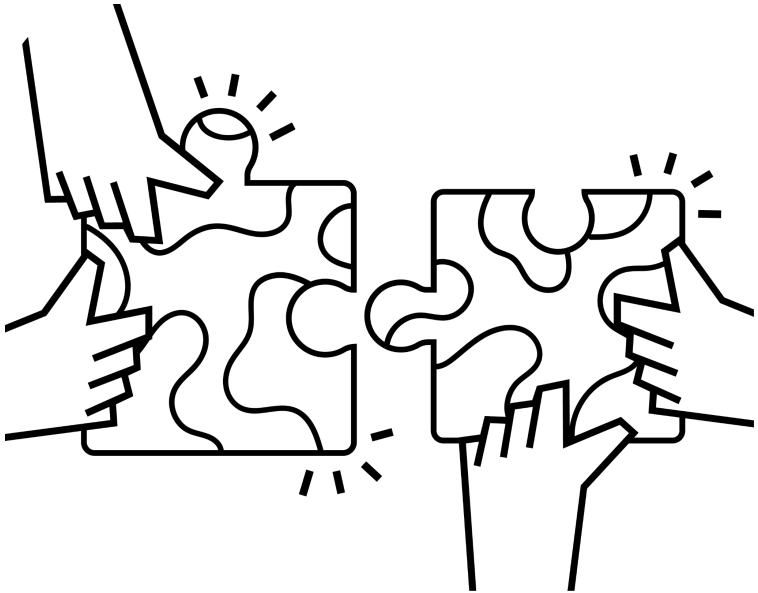
WIKIMEDIA
FOUNDATION

Насправді...

Чи чули ви про **Абстрактну Вікіпедію**? Вона дозволить створювати «абстрактні» статті, *використовуючи код* (програмування), з якого ми зможемо потім *генерувати* читабельні, граматично правильні *та точні* статті *будь-якою мовою*!

Будь-якою мовою? Гаразд, будь-якою, яка *гарно описана структурованими даними*!

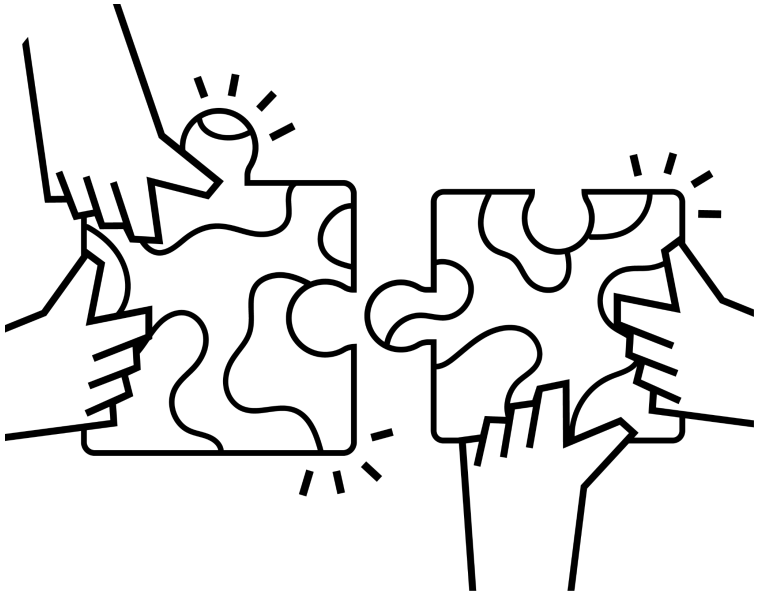
Lexeme фундаментально важлива для Абстрактної Вікіпедії і для **значного** розширення вмісту, доступного українською!



WIKIMEDIA
FOUNDATION

Але *що таке* Lexeme, конкретно?

- Це лексикографічний шар, яким покрите програмне забезпечення **Wikibase**, на якому працює проєкт **Вікідані**.
  «Lexeme» коротше. :)
- Лексеми — це сутності Вікіданих, які існують паралельно з елементами.
  Елементи ≠ Лексеми. Елементи виглядають отак: **Q212**; лексеми виглядають отак: **L34336**.
- Ми користуємося всіма перевагами вікі.
- Ми можемо використовувати Wikidata Query Service, щоб робити запити по лексемах (і навіть лексемах і елементах разом)
- Це (все ще) невелика, дружня, привітна спільнота



WIKIMEDIA FOUNDATION

WIKIMEDIA FOUNDATION

Двома словами:

WIKIMEDIA FOUNDATION

Lexeme це <3

WIKIMEDIA FOUNDATION

Я приніс вам щось від зайчика:

https://w.wiki/5m5B

WIKIMEDIA FOUNDATION

Гаразд, гаразд, лексеми корисні!

...

WIKIMEDIA FOUNDATION

Але у нас є багато запитань!

...

WIKIMEDIA FOUNDATION

Знаю, але маю лише ці 45 хвилин...

...

WIKIMEDIA FOUNDATION

Але не впадайте у відчай:

WIKIMEDIA FOUNDATION

Через два тижні я ґрунтовно вчитиму працювати з лексемами на CEE Meeting (Охрид)

...

WIKIMEDIA FOUNDATION

А за тиждень після того — у Стамбулі.

...

WIKIMEDIA FOUNDATION

Якщо ні там, ні там не буде запису, я запропоную семінар у Meet/Zoom, який буде записано. ...

WIKIMEDIA FOUNDATION

Це навчання включатиме редагування лексем, прості запити, деякі інструменти й додатки...

...

WIKIMEDIA FOUNDATION

Тим часом, ви можете:

1. Досліджувати Lexeme самі
2. Додавати нові лексеми
3. Додавати нові значення до наявних українських лексем
4. Чекати на запис семінару

WIKIMEDIA FOUNDATION

все буде Україна!

asaf@wikimedia.org