
WIKIMEDIA
FOUNDATION
Documenter nos langues avec Lexeme
Asaf Bartov
CEE Meeting 2022
CC by SA 4.0, Lyokoï

WIKIMEDIA FOUNDATION

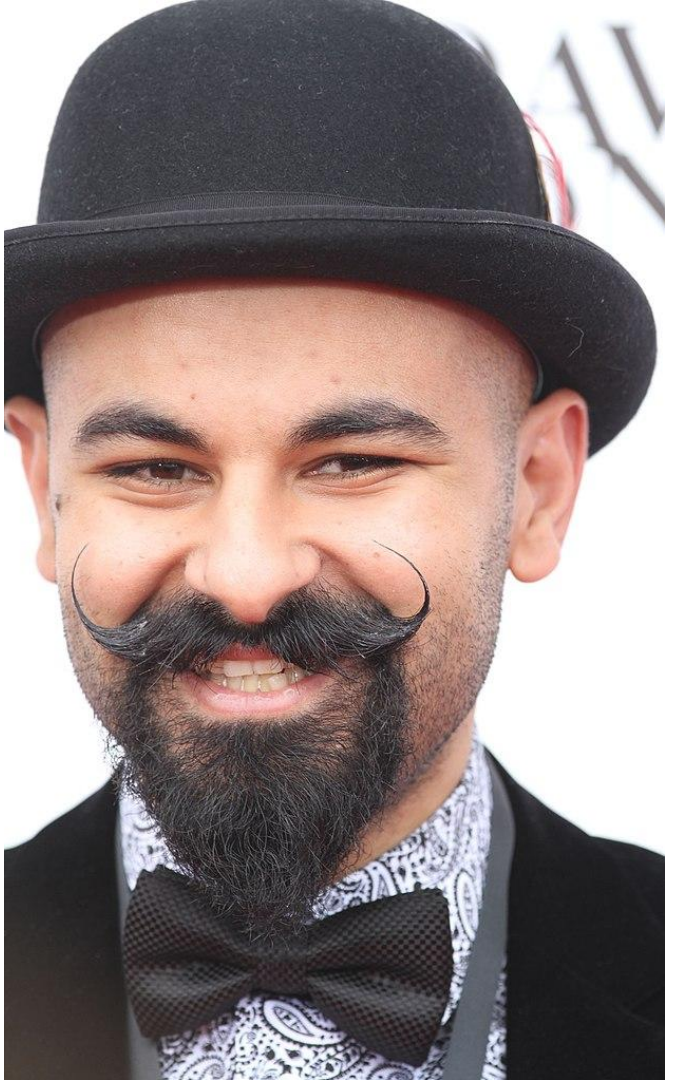
01 Qu'est-ce que Lexeme?

Et pourquoi en ai-je besoin dans ma vie?

Vous avez de la chance!

Vous avez encore le temps pour devenir **un hipster de Lexeme**!

Quand tout le monde connaîtra et utilisera Lexeme, vous pourrez dire : "ah ouais, j'ai contribué à Lexeme avant que ce soit cool!"

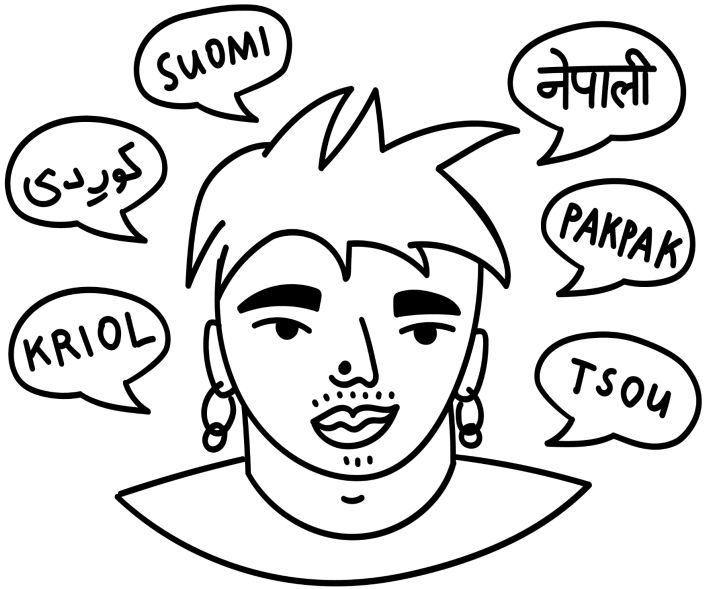


Oui, mais pourquoi?

Parce que les ordinateurs peuvent fournir beaucoup de valeur pour **l'acquisition, la pratique, l'analyse, l'amélioration et la traduction** du langage humain...

...mais pour cela, ils ont besoin de **données structurées** sur les langues humaines...

...et les langues humaines sont **vraiment complexes**!



WIKIMEDIA
FOUNDATION

La langue est-elle si complexe ?

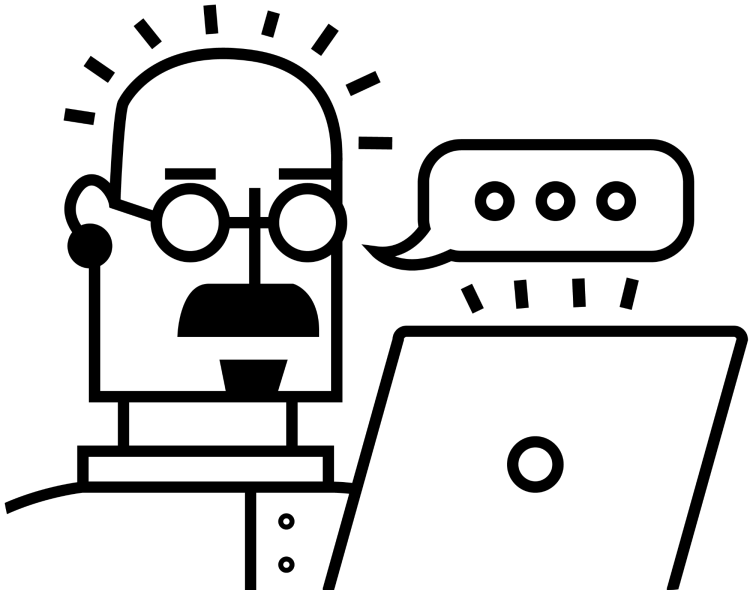
Que veut dire 'livre'?
-> "Combien pèses-tu de livres"
Que veut dire ‘mémoire’?
-> "La chanteuse vient de publier ses mémoires."
Que veut dire ‘somme’?
-> Je te dois une somme de 100 Dollars
-> Je vais faire un petit somme
Quelle traduction serait correcte et appropriée pour ‘mémoire’ ou ‘somme’? Cela dépend du *sens spécifique* et du *contexte* dans le texte d’origine.

WIKIMEDIA
FOUNDATION

Attendez, mais la traduction automatique existe!

Nous avons déjà la **traduction automatique**, et elle s'est beaucoup améliorée ces dernières années. Mais elle est encore **à peine tolérable et généralement peu fiable** dans la plupart des langues, y compris la plupart de celles de l'Europe centrale et orientale.

L'**approche statistique** utilisée par la TA *comprend à peine le contexte*, et aplatit donc les nuances, les registres, les dialectes, v*oire les barrières linguistiques*! Bien que nous utilisions tous la TA pour ce qu'elle fait - obtenir l'essentiel d'un texte que nous ne pouvons pas lire par nous-mêmes - il existe de nombreuses utilisations pour lesquelles **elle ne convient pas.**



WIKIMEDIA
FOUNDATION

Alors, la langue est complexe !

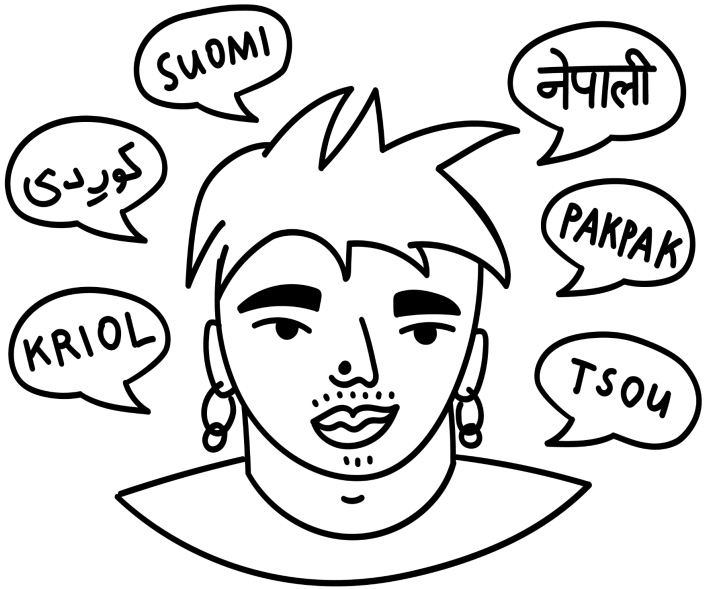
- Les mots ont de nombreuses formes; certains irréguliers (vais vs. ira) / archaïques (hospital, francofyle)
- Les mots ont plusieurs sens; certains défunts (linge) / régionaux (pochon)
- Les sens ont beaucoup de mots -- synonymes (pic/sommet, clever/smart)
- Homophones (ver vs vert vs vers), homographes (son vs son)
- Grammaire dialectale ("J'ai parti")
- Registre et période ("Eh !" ; "Écoute!"; "Oyez!")
- Chevauchement lexical et confusion (ce que signifie soda/pop, doute ou fanny dépend de l'endroit où vous vivez et de la personne à qui vous parlez)
- ...et tout cela n'est qu'au niveau des lexèmes, en laissant de côté le monde de complexité qu'est la **syntaxe**!

WIKIMEDIA
FOUNDATION

Alors... ça sera compliqué à modéliser comme des données structurées, non?

Oui. :)

Mais ça vaut vraiment le coup! Parce qu'une grande quantité d'utilisations sera rendue possible une fois que nous aurons des données richement modélisées et liées sur nos langues.
Voici quelques perspectives. Il y en a d'autres auxquelles je peux penser, et, encore plus intéressant, d'autres encore auxquels je *ne peux même pas* y penser!



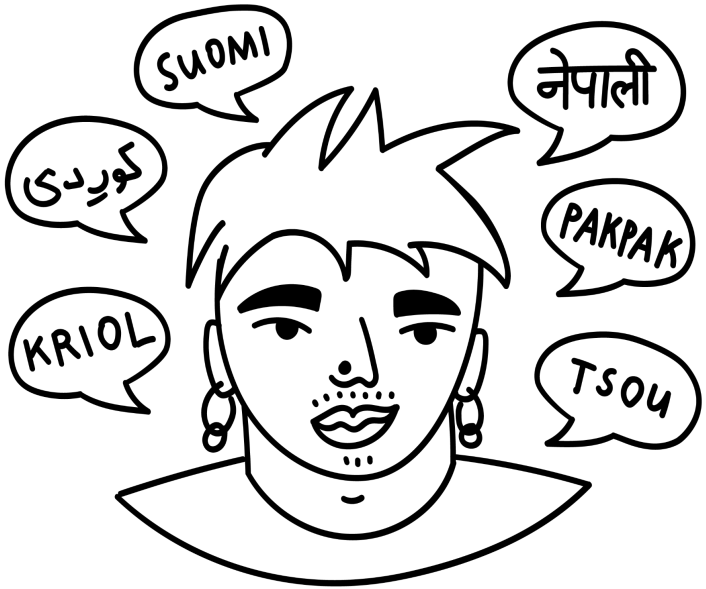
Et il y a des outils sympas!

WIKIMEDIA
FOUNDATION

Acquisition du language

Les données structurées sur le langage permettent la création de logiciels d'**acquisition du langage**, notamment:

- Applications flashcard
- Pratique de la grammaire (déclinaison nom/adjectif, conjugaison des verbes)
- Jeux éducatifs
- Pratique de la prononciation
- Logiciel de lecture de texte avec texte hyper-annoté (pour chaque mot, forme et sens analysés)
- ...et plus!

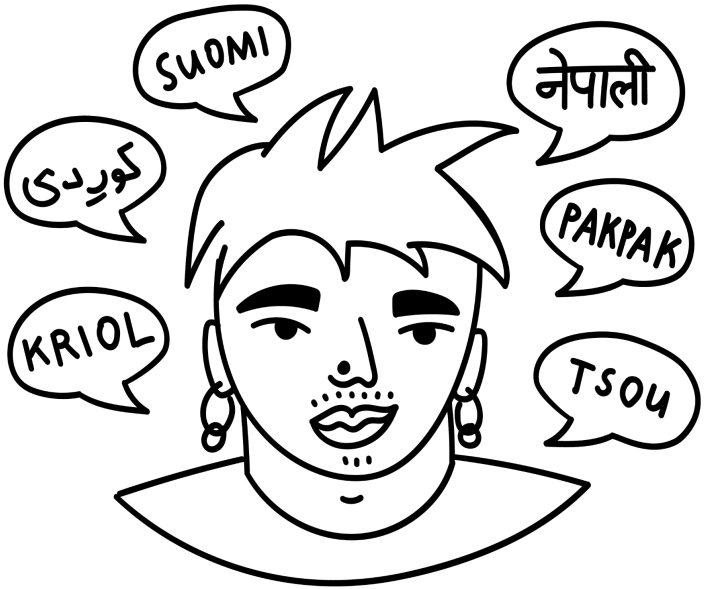


WIKIMEDIA
FOUNDATION

Analyse linguistique

Les données structurées sur la langue permettent la création de logiciels **d'analyse et d'amélioration de la langue**, notamment :

- Vérificateurs d'orthographe/grammaire sophistiqués
- Résolveurs de mots croisés
- Exploration/recherche étymologique
- Stylométrie
- Stemmatologie et phylométrie
- ...et plus!

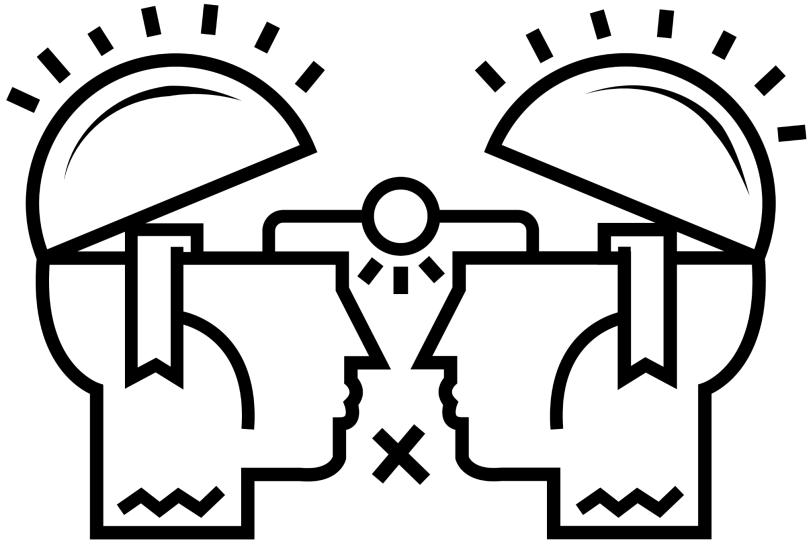


WIKIMEDIA FOUNDATION

Traduction

Une traduction correcte et adéquate dépend de nombreux paramètres:

- Distinguer le **sens** particulier du mot / de la phrase d'origine
- Contextualiser (genre, registre, voix, audience)
- Sélectionner le mot/phrase adéquat dans la langue cible, en tenant compte et en préservant le contexte
- ...ce qui est souvent assez loin d'une substitution mot à mot littérale.

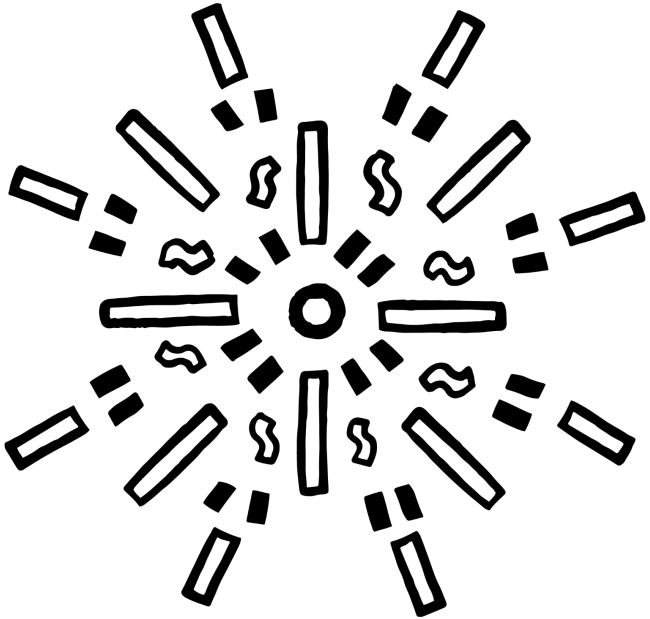


WIKIMEDIA
FOUNDATION

Souhaitons à la vue d'une étoile…

Et si nous avions un moyen de décrire *très précisément* les lexèmes, jusqu'à des **formes** et des **sens** spécifiques?

- A noter que *cette* forme est *nominative* et *celle*-ci *génitive*; celui-ci *imparfait* et celui-ci *plus-que-parfait*?
- A noter qu'une forme particulière est *régionale*, ou *archaïque*, ou *argotique*?
- Qu'*un* sens de ce lexème se traduit par *ce* mot en allemand, mais qu'un *autre* sens de ce même lexème se traduit par *cet* autre mot en allemand?



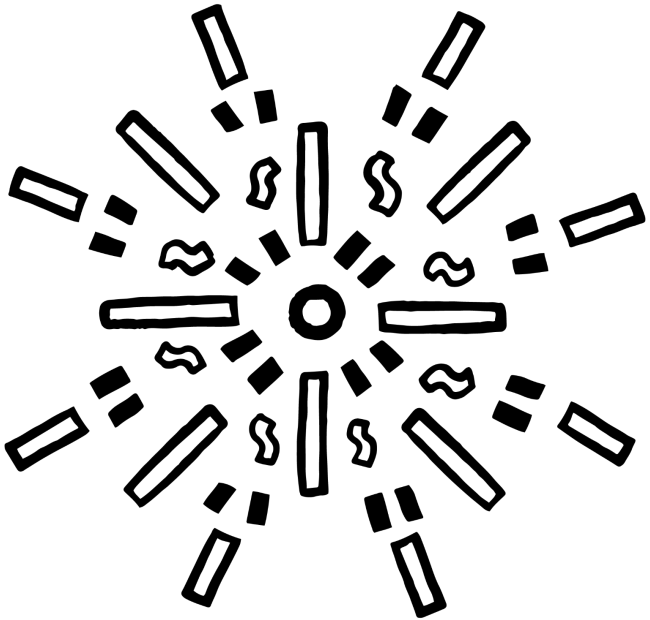
WIKIMEDIA
FOUNDATION

Souhaitons à la vue d’une étoile…

Que ce lexème **combine** trois autres lexèmes? Qu'il **dérive** d'un autre lexème? Qu'il est **emprunté** à une autre langue?
Qu'il désigne *ce concept*, qui a un **élément** Wikidata (indépendant de la langue) ?

Et si nous pouvions fournir de vrais **exemples de phrases** démontrant l'utilisation de *chaque sens* du lexème dans de vrais textes ?

Et si nous pouvions joindre un **son** à chaque forme montrant comment les locuteurs natifs la **prononcent**? (Peut-être à plus d'une manière!)

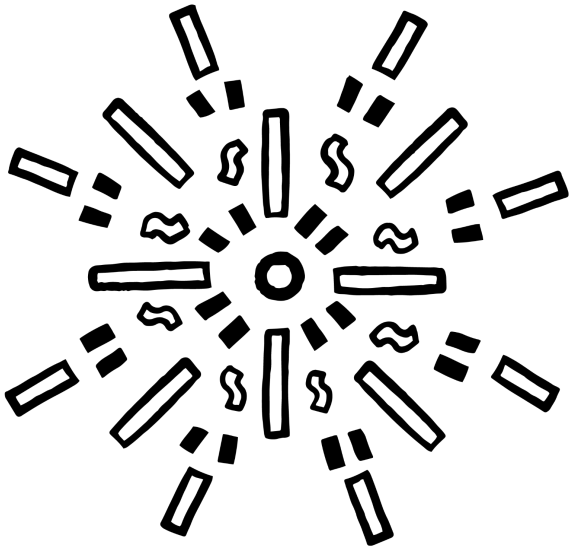


WIKIMEDIA
FOUNDATION

Souhaitons à la vue d’une étoile…

Et si nous pouvions avoir une **requête** pour tout cela, et poser des questions telles que:

- Quels sont les noms qui sont masculins en ukrainien mais féminins en allemand?
- Quel est le graphe étymologique des mots slaves pour ‘cheval’?
- Quel est le mot le plus long de notre langue sans répétition de lettres?
- Quel pourcentage des lexèmes de notre langue avons-nous emprunté à quelles langues?
- Quels sont les ‘faux amis’ entre notre langue et une autre? (par exemple, *Gift* en anglais par rapport à l'allemand)
- Comment ce lexème a-t-il changé d'usage au fil des ans, sur la base de textes réels?



WIKIMEDIA
FOUNDATION

WIKIMEDIA FOUNDATION

Devinez quoi?

WIKIMEDIA FOUNDATION

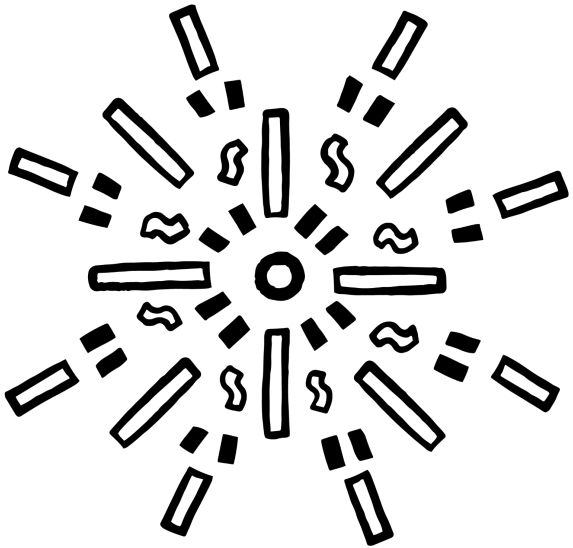
Lexeme peut faire cela dès maintenant !

En effet…

Ne serait-ce pas bien si tout le monde pouvait parler *votre langue*?

En attendant, ne serait-il pas agréable de pouvoir bénéficier *dans notre langue* de contenus écrits par des personnes qui ne parlent pas du tout notre langue, *automagiquement*?

(o_O)



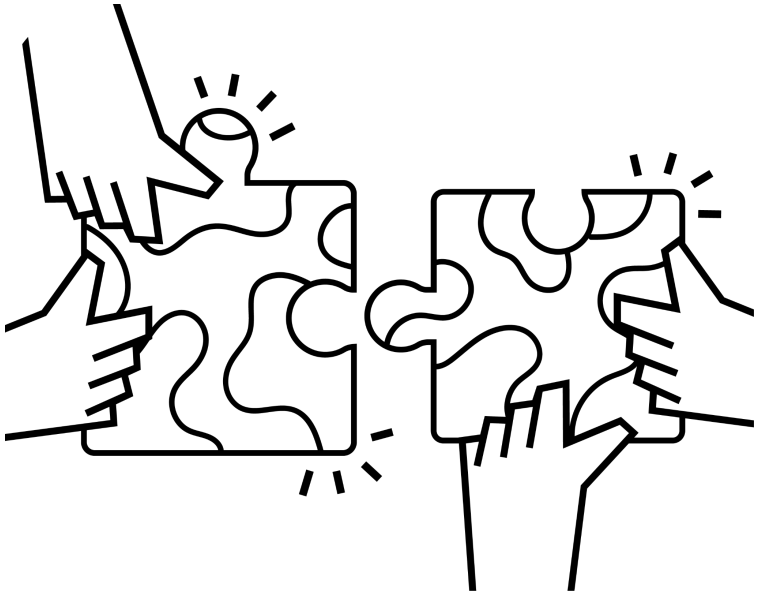
WIKIMEDIA
FOUNDATION

En effet...

Avez-vous déjà entendu parler d'**Abstract Wikipédia**? Cela va permettre de créer des articles 'abstraits' *à l'aide de code* (programmation), à partir desquels nous pourrions ensuite *générer* des articles lisibles par l'homme, grammaticaux et *précis* dans *n'importe quelle langue*!

N'importe quelle langue? Eh bien, n'importe quel langage qui est *bien décrit dans des données structurées*!

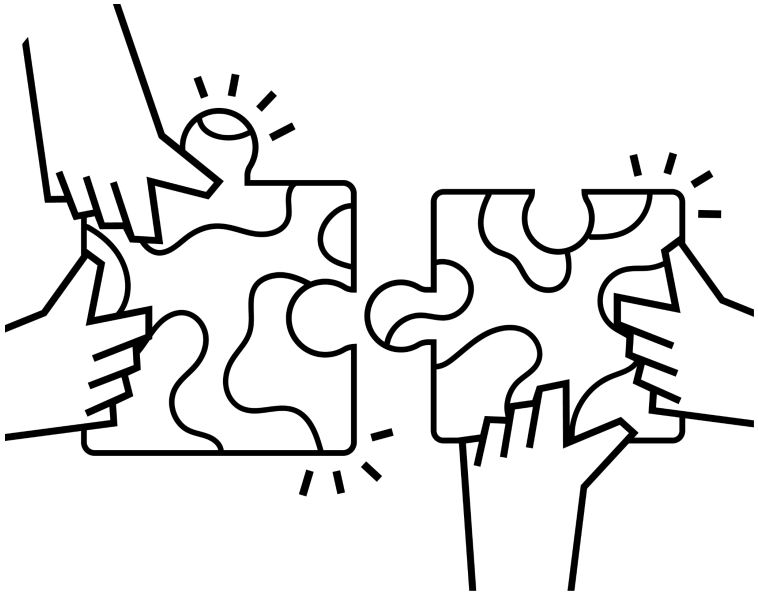
Lexeme est fondamental pour Abstract Wikipedia et pour un **vaste** enrichissement du contenu disponible dans votre langue !



WIKIMEDIA
FOUNDATION

Mais qu'*est*-ce que Lexeme, exactement?

- C'est une **couche lexicographique** au-dessus du logiciel **Wikibase** exécuté dans le projet **Wikidata**. "Lexeme" est plus court. :)
- Les lexèmes sont des entités Wikidata qui existent parallèlement aux éléments. **Éléments ≠ Lexèmes**. Les éléments ressemblent à **Q212**; Les lexèmes ressemblent à **L34336**.
- Nous bénéficions de tous les avantages de wiki ; nous lions avec Commons, Wikidata.
- Nous pouvons utiliser le Wikidata Query Service pour écrire des requêtes sur les lexèmes (et même les lexèmes et les éléments).
- C'est une communauté (encore) petite, amicale et accueillante.



WIKIMEDIA FOUNDATION

Wiktionary?

WIKIMEDIA FOUNDATION

En bref:

WIKIMEDIA FOUNDATION

Lexeme est

WIKIMEDIA FOUNDATION

02 Un tour de Lexeme

Anatomie et sociologie d'un lexème

WIKIMEDIA FOUNDATION

Examinons un lexème:

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

WIKIMEDIA FOUNDATION

D'accord, d'accord, Lexeme vaut la peine!

...

WIKIMEDIA FOUNDATION

Mais nous avons beaucoup de questions!

...

WIKIMEDIA FOUNDATION

Par exemple: comment savoir ce qui existe déjà dans ma langue?

...

WIKIMEDIA FOUNDATION

03 Parcourir Lexeme

WIKIMEDIA FOUNDATION

- Ordia
- Hangor
- Rapport de couverture lexicale
- …?

WIKIMEDIA FOUNDATION

04 Contribuer à Lexeme

Créer un Lexème

WIKIMEDIA FOUNDATION

1. Lexeme Forms Outil (+votre langue? +gadget)
2. Orthohin Outil (+votre langue? +gadget)
3. Entity-suggester script (ex. L475401)
4. MachtSinn Outil -- Lie les lexèmes au éléments
5. LinguaLibre Enregistre les prononciations! (requête)
6. Lexeme Party Améliore par sujet (+hebdomadaire)
7. Bodh Outil - Édition tabulaire des lexèmes.
8. Lexicator Outil -- importation en masse prudente depuis le Wiktionnaire.

WIKIMEDIA FOUNDATION

05 Requêtes sur Lexeme

WIKIMEDIA FOUNDATION

- Apprendre SPARQL
- ~~Vole~~ Adapter les requêtes

WIKIMEDIA FOUNDATION

06 S’amuser avec Lexeme

WIKIMEDIA FOUNDATION

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

Beaucoup d'autres sont encore à inventer ! :)

WIKIMEDIA FOUNDATION

07 Prochaines étapes

Comment mettons-nous en action notre hipstérisme de Lexeme?

WIKIMEDIA FOUNDATION

Ce sont les premiers jours!

1. Nous cherchons à comprendre
2. Votre contribution compte!
3. Prenez des initiatives; n'attendez pas
4. Demandez, discutez, invitez

WIKIMEDIA FOUNDATION

Que pouvez-vous faire maintenant?

1. Explorez Lexeme par vous-même
2. Ajoutez de nouveaux lexèmes
3. Ajoutez de nouvelles formes, sens, exemples aux lexèmes existants
4. Montrez Lexeme à vos pairs

WIKIMEDIA FOUNDATION

Idéalement, menez l'adoption de Lexeme dans votre langue!

WIKIMEDIA FOUNDATION

1. Vérifiez la couverture de votre langue
2. Démarrez un projet Wiki !
   - Tutoriels
   - Listes de tâches / requêtes
   - Canaux hors wiki

WIKIMEDIA FOUNDATION

Merci pour votre attention!