



# Dictionary LODification using Wikibase: Quechua language

Valeria Caruso, Ibai Guillén, Elwin Huaman  
Tutored by David Lindemann

SD-LLOD-22, Cercedilla

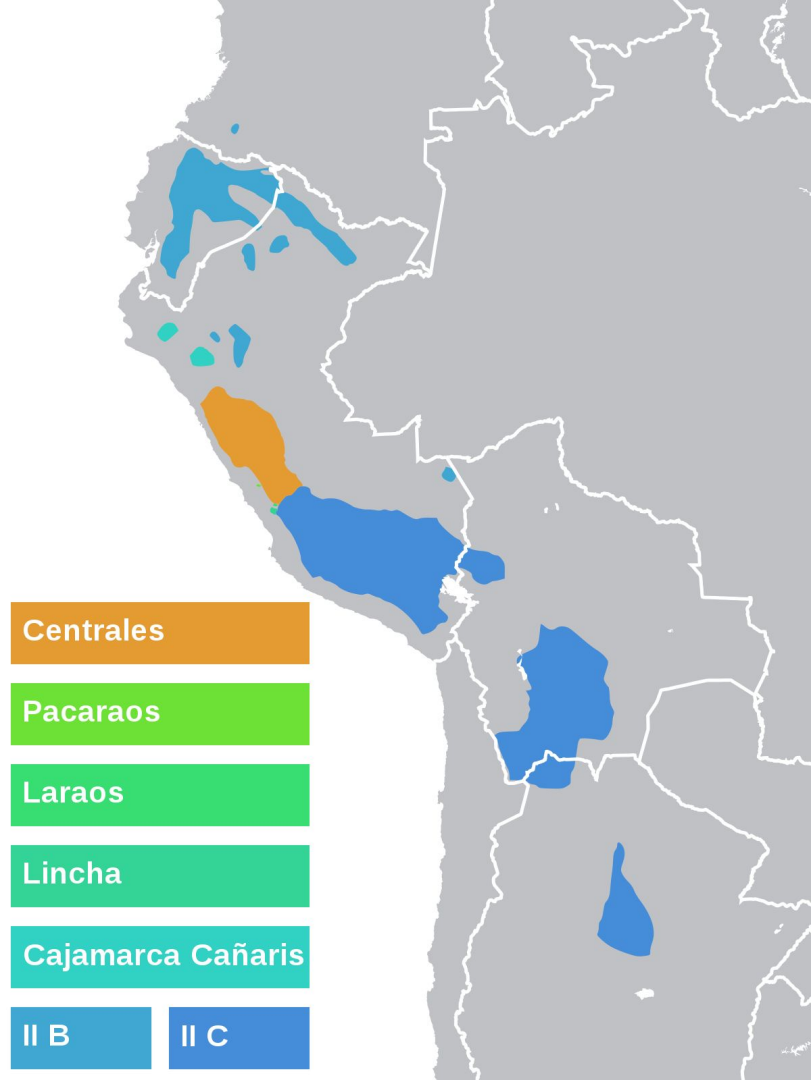


# Outline

1. Why  
Motivation - Problem
2. How  
Approach - Solution
3. What  
Results - Next steps

# Why

- Quechua Language
  - Endangered language
    - 10 million speakers
    - 6 countries
  - Few resources
  - Not present as LLOD
  - ...



# How

1. Identification of adequate sources
2. Preprocessing and setup
3. Modelling and population of the model
4. Exploitation

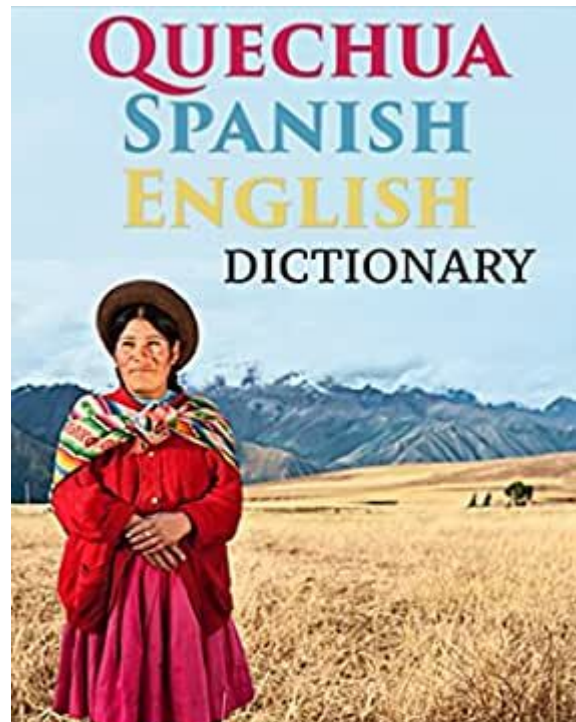
# How

## 1. Identification of adequate resources

- **Data:** Corporuses, dictionaries, bilingual texts
- **Database:** Wikibase, Wikidata, etc.



- LOD knowledge graph
- Community-driven
- Open source
- ...



# How

## 2. Preprocessing and setup

- **Database:** Setting up QICHWABASE and programming a bot
- **Data:** Normalizing, cleaning, and refining

abansay (*to advance*) v.intr.(esp) → v.intr. *POS after cleaning*  
as(ta)wan allin, hatun → astawan allin *lemma*  
(aswan allin) *lemma variant*  
hatun *synonym*



- Create item
- Create lexeme
- Adding sense descriptions
- ...

# How

## 3. Modelling and population of the model

- **Model:** describing ontolex:LexicalEntry, LexicalSense, LexicalForm using [pre-set ontolex application profile](#) (default in Wikibase)
  - Values of wikibase:lexicalCategory (describe lexemes)
  - Values of wikibase:grammaticalFeatures (describe forms)
- **Population:** Ingesting lexemes, POS, sense descriptions to Qichwabase
  - Customise “[Wikibaseintegrator](#)” python bot to Qichwabase

3 June 2022

**N** **b** 07:15 [piñaschay \(L13652\)](#) (diff | hist) .. **(+1,003)** .. [DavidLbot](#) (talk | contribs) (Created a new Lexeme)

**N** **b** 07:15 [piñas \(L13651\)](#) (diff | hist) .. **(+998)** .. [DavidLbot](#) (talk | contribs) (Created a new Lexeme)

**N** **b** 07:15 [piñas \(L13650\)](#) (diff | hist) .. **(+997)** .. [DavidLbot](#) (talk | contribs) (Created a new Lexeme)

**N** **b** 07:15 [piñaqrusqa \(L13649\)](#) (diff | hist) .. **(+1,004)** .. [DavidLbot](#) (talk | contribs) (Created a new Lexeme)

**N** **b** 07:15 [piña \(L13648\)](#) (diff | hist) .. **(+996)** .. [DavidLbot](#) (talk | contribs) (Created a new Lexeme)

**N** **b** 07:14 [pintuq raya \(L13647\)](#) (diff | hist) .. **(+1,003)** .. [DavidLbot](#) (talk | contribs) (Created a new Lexeme)

**List of abbreviations:** [\[Collapse\]](#)

**D** Qichwabase edit

**N** This edit created a new page (also see [list of new pages](#))

**m** This is a minor edit

**b** This edit was performed by a bot

**(±123)** The page size changed by this number of bytes

```
komandoaren gonbita - python -m create_lexeme

Will now upload 13630 piñas Q74
Writing to qichwabase...
Successfully written to item: L13651
Will now upload 13631 piñaschay Q99
Writing to qichwabase...
Successfully written to item: L13652
Will now upload 13632 piñaw Q5
Writing to qichwabase...
```



# How

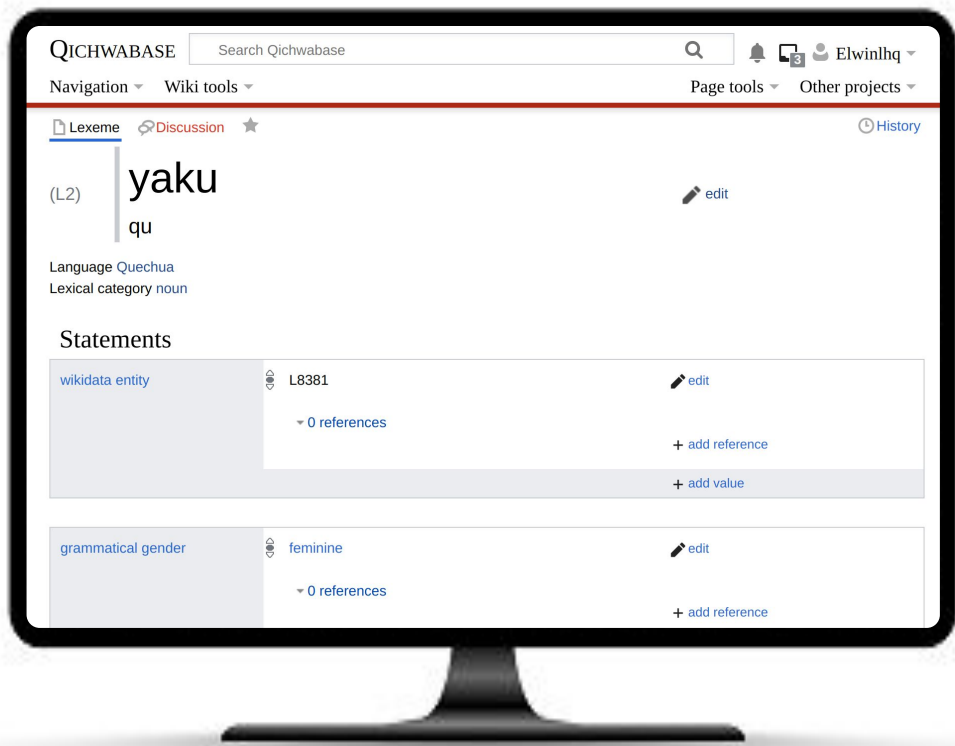
## 4. Exploitation

- <https://qichwa.wikibase.cloud>
- SPARQL query interface / endpoint
- Possible use cases:
  - Language learning resources
  - Dialogue systems
  - NLP tasks

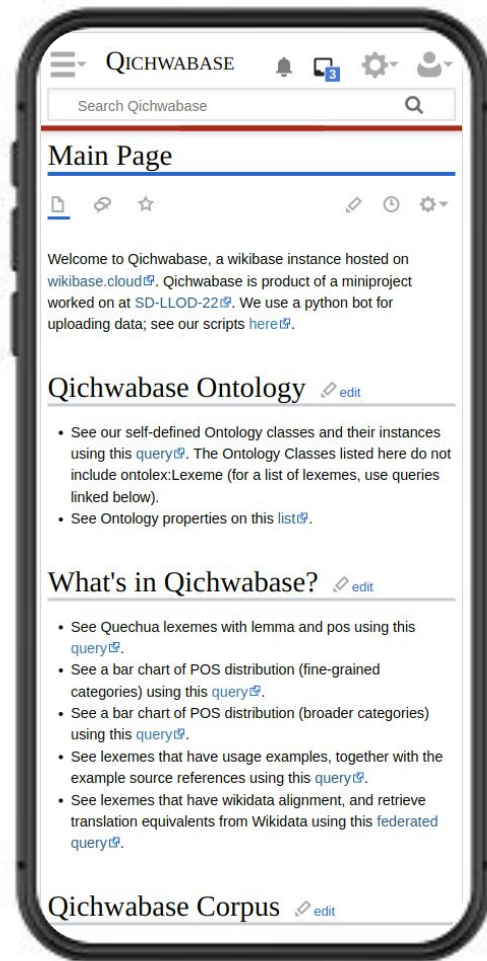


- LOD knowledge graph
- Community-driven
- Open source
- ...

# What: Qichwabase

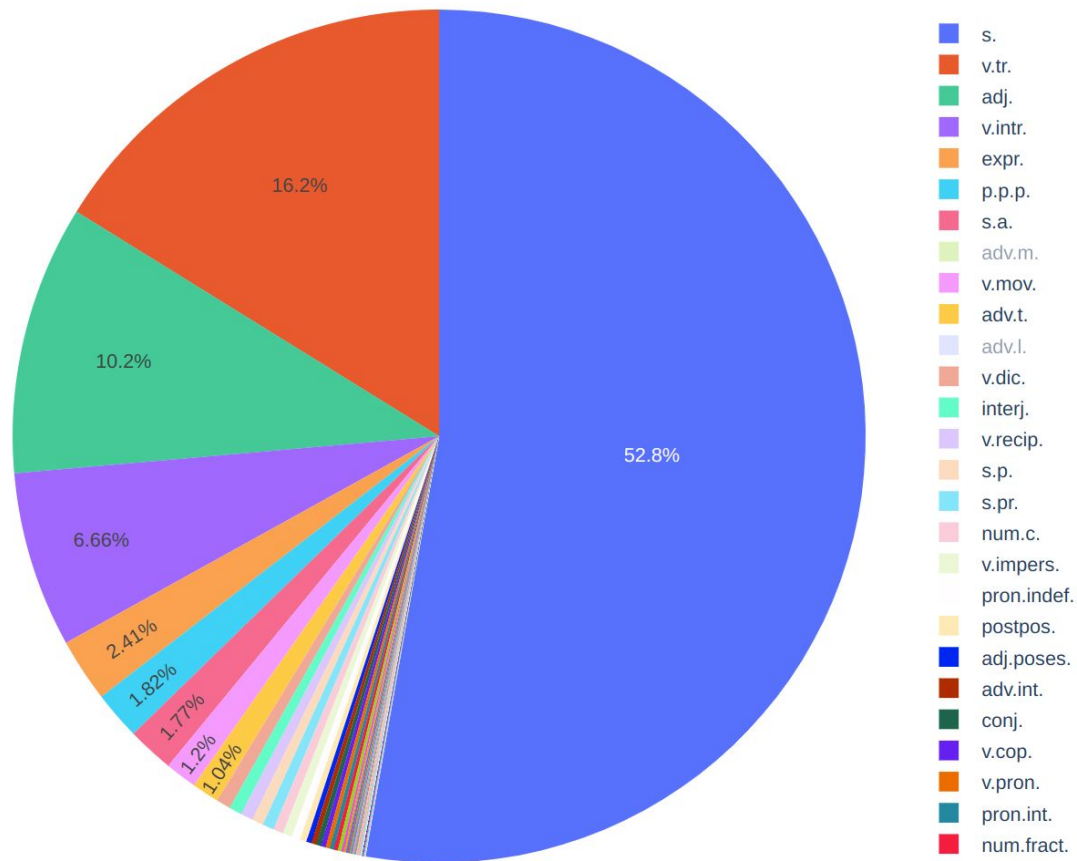


<https://qichwa.wikibase.cloud/>



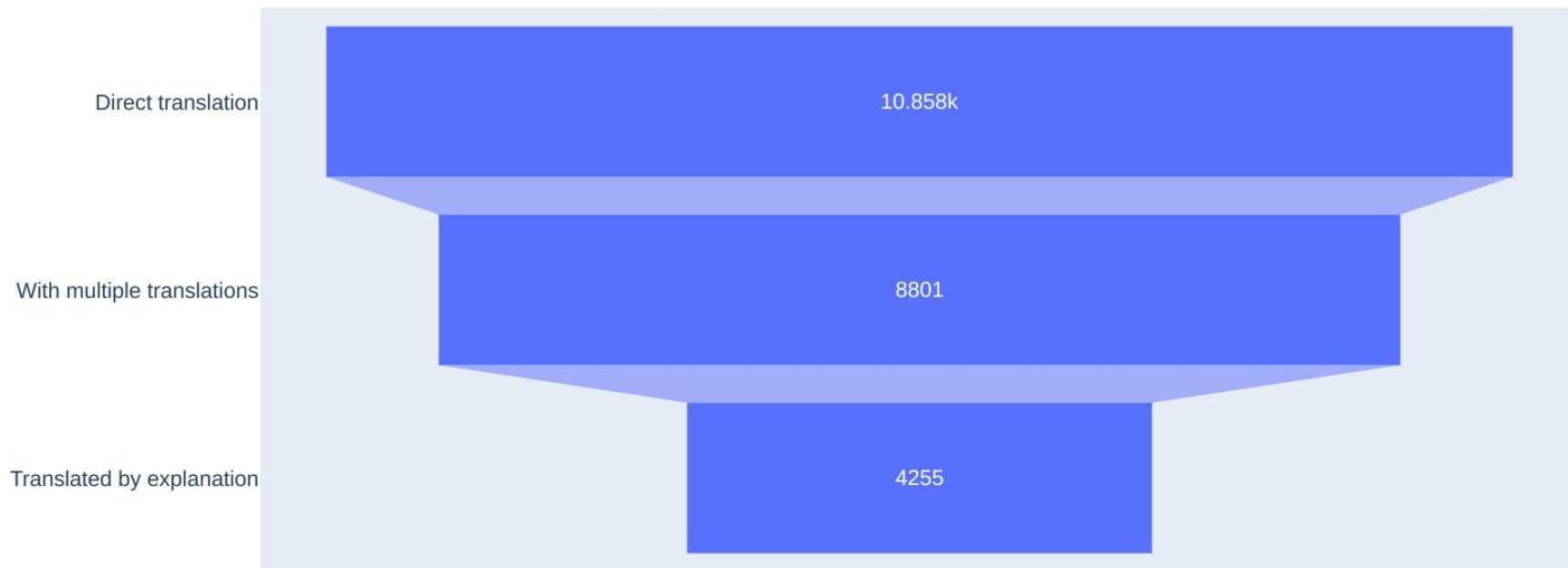
# What: Qichwabase

Lexeme	Count
s.	12981
v.tr	3981
adj.	1512
v.intr.	1638
...	...
<b>TOTAL</b>	<b>25154</b>



[https://nexuslinguarum.github.io/SD-LLOD-22\\_QUECHUA/](https://nexuslinguarum.github.io/SD-LLOD-22_QUECHUA/)

# What: Qichwabase



[https://nexuslinguarum.github.io/SD-LLOD-22\\_QUECHUA/](https://nexuslinguarum.github.io/SD-LLOD-22_QUECHUA/)




# What: Qichwabase

SPARQL queries for retrieving:


- Quechua lexemes with lemma and POS >> [query](#)
- A bar chart of POS distribution (fine-grained categories) >> [query](#)
- A bar chart of POS distribution (broader categories) >> [query](#)
- Lexemes that have usage examples, and their source references >> [query](#).
- Lexemes that have Wikidata alignment, and retrieve translation equivalents from Wikidata >> [query](#).
- Lexemes that have lexical forms >> [query](#)


# What: Qichwabase


Lexemes that have multilingual sense descriptions, and usage examples with their source references >> [query](#)






3 results in 2686 ms

 Code

 Download

 Link

lexeme	lemma	senseglosses	example	source
 <a href="https://qichwa.wikibase.cloud/entity/L106">https://qichwa.wikibase.cloud/entity/L106</a>	achka	viel; viele; a lot; much; many; much; mucho; mucha; muchos; muchas; harto; bastante; abundante; harto; mucho; varios; demasiado; bastante; mucho; bastante; gran cantidad; cuantioso; numeroso; gran porción; profusamente; tanto	11 ñiqin tarpuy killapi 2001 p'unchawpiqa terroristakuna New York llaqtapi pitu hatun wasintintam (World Trade Center, New York Twin Towers nisqata), Washington DC llaqtapi maqanakuy ministiryup (Pentagon nisqap) huk rakinta suwakusqa antanakunawan thuñichirqan, achka waranqa runakunatam wañuchispa. Chaymantam George Walker Bush terrorismu hayu maqanakuyta rimarirqan.	<a href="https://qu.wikipedia.org/wiki/Awacha_turrikuna">https://qu.wikipedia.org/wiki/Awacha_turrikuna</a>
 <a href="https://qichwa.wikibase.cloud/entity/L251">https://qichwa.wikibase.cloud/entity/L251</a>	akllay	Wahl; Auswahl; election; choice; elección; Lotterie; lottery; lotería	Huwan hina Hisuwpa akllay qatinqinkunapas unuchachirqanku	<a href="https://qu.wikipedia.org/wiki/Unuchay">https://qu.wikipedia.org/wiki/Unuchay</a>
 <a href="https://qichwa.wikibase.cloud/entity/L7724">https://qichwa.wikibase.cloud/entity/L7724</a>	kunanlla kasqa	neulich; recently done; reciente	Shuk wawa kunanlla wacharishkata shitashkami shuk llakta ñanpi kay Santa Clara del Tablón llakta	<a href="https://www.policia.gob.ec/dinapen-llankaymanta-chapakkuna-kishpichishkami-kunanlla-wacharishkata/">https://www.policia.gob.ec/dinapen-llankaymanta-chapakkuna-kishpichishkami-kunanlla-wacharishkata/</a>

# What: Qichwabase next steps

- Adding new lexemes and more examples of usage from corpora
  - e.g., synonyms, duplicates,
  - Meronyms in terms of metaphors
    - e.g., anthropomorphic
- Lemmatization of multiword lexical units, e.g., phraseological units
  - e.g., saphi (kunka saphi, sunqu saphi, qusqu saphi, siki saphi)
- Enriching
  - Link Qichwabase ontology properties & classes to Wikidata & Ontolex
  - Link Qichwabase lexemes to Wikidata lexemes
  - Provide more SPARQL queries and visualizations
  - Call in action to the community
- Quality
  - Profiling and Generating SHEX expression to verify the qichwabase knowledge



Project documentation: [at Github](#), and [at Qichwabase](#)  
Datasources and Code: [at Github](#)