Wikidata Lexemes for All

Elwin Huaman
University of Innsbruck, Austria

Abstract

This research aims to analyze the current lifecycle of Wikidata Lexemes, employ Wikidata Lexeme methods and tools for populating and enriching lexemes representation of the Puno Quechua language up to 20 000 lexemes, and propose improvements and solutions for Wikidata Lexemes to engage with Under-Resourced Language communities. We argue that semi-automated approaches play a key role in expanding the representation of lexical data, especially from people historically excluded from technology.

Introduction

Wikidata Lexemes[1] is a specialized subset of the Wikidata Knowledge Graph focused on lexicographical data (i.e., words, phrases, and their linguistic properties). It allows to integrate resources such as data, text, and multimedia that are essential for the long-term preservation of languages and for their communities to use. Not only that, Wikidata Lexemes allows the integration of these resources from other sister projects within the Wikimedia ecosystem. **Figure 1** shows an excerpt of an English lexeme representation, providing information about senses, translations, synonyms, hyperonyms, and usage examples.

In addition, Wikidata Lexemes can integrate usage examples from Wikisource projects, pronunciation audios for lexeme forms from Wikimedia Commons, and images for senses from Wikimedia Commons.

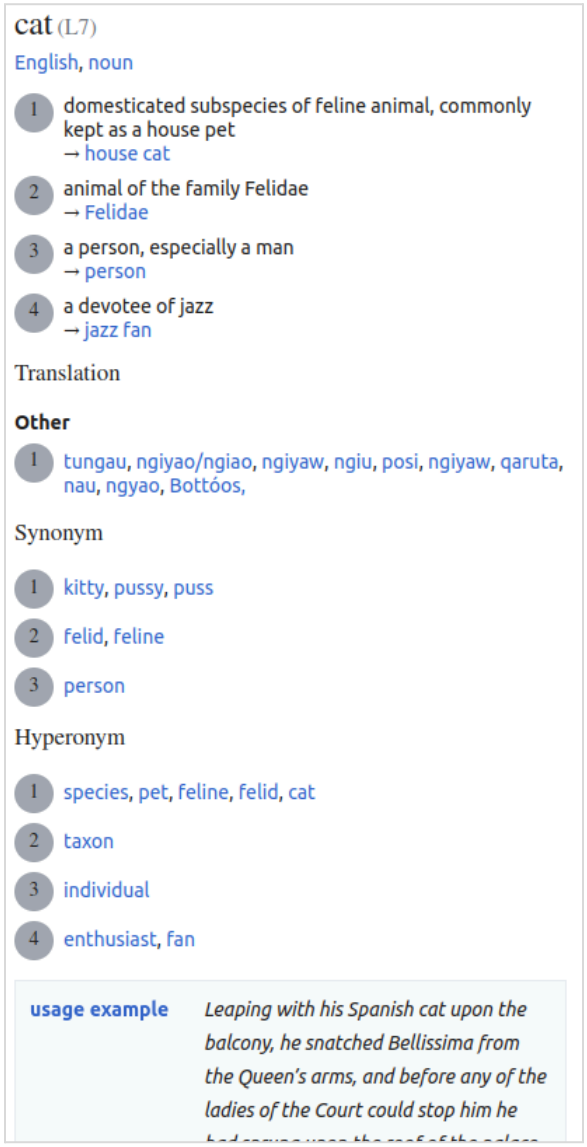

Figure 1: An excerpt view of the representation of the cat noun lexeme on Wikidata Lexemes.

There are more possibilities to extend lexeme descriptions on Wikidata Lexemes, but how simple or complex is it? How does it work to integrate a new language?

[1] https://w.wiki/7v$r (Wikidata Lexemes)

In the world, there are about 7159 languages in use today[2], and only a few of them have the necessary resources to be further implemented on language technologies (Besacier et al., 2014), and about 45% of the languages are endangered, see **Figure 2**.

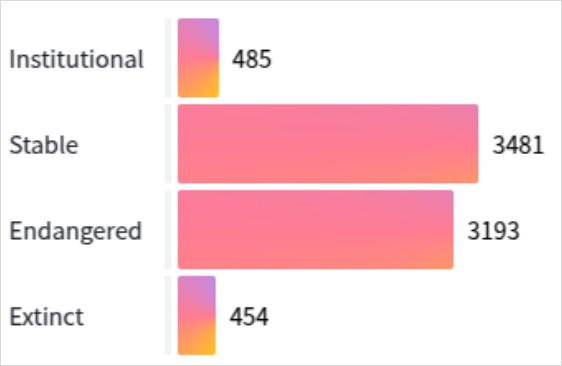

Figure 2: Languages in the world with respect to their level of language vitality by Ethnologue.

If we take a closer look at the Wikimedia ecosystem, and more specifically at Wikidata Lexemes and its representation of languages, we can see that Wikidata Lexemes represents around 1368 languages[3], of which 1240 languages (91%) have less than 100 lexemes, while German, English, and Spanish have 239399, 68498, and 63105 lexemes respectively. **Figure 3,** based on a query to Wikidata[4], shows a range distribution of languages based on the number of their lexemes on Wikidata Lexemes. It shows that 74.9% of the languages on Wikidata Lexemes have less than or equal to 10 lexemes. It also shows that the languages considered on the list are somehow recognized by the Wikidata Lexemes system, such as Middle English, Old English, etc., but the Puno Quechua language is not listed, even though it has lexemes[5]. How can one get their language recognized by the Wikidata Lexemes system? How can one take advantage of the tools developed within the Wikidata Lexemes ecosystem?

Over the last few years, we have collected and modelled ~4000 lexemes of the Puno Quechua language on Qichwabase[6], a Wikibase instance that integrates lexicographical data of Quechua languages based on the Wikidata Lexeme Data Model[7] (**Huaman et al., 2022**; **Huaman et al., 2023**). Then, we started to ingest the Puno Quechua lexemes into Wikidata Lexemes and we realize that there is no clear path for languages to be ingested into Wikidata Lexemes. For example, when we wanted to import ~1000 lexemes into Wikidata, one would need to know programing languages to use tools, e.g., Python for Wikibase Integrator[8], or have a sister project to enrich the lexemes, e.g. Luthor[9] tool links usage examples from Wikisource, but the Quechua languages, we do not have none.

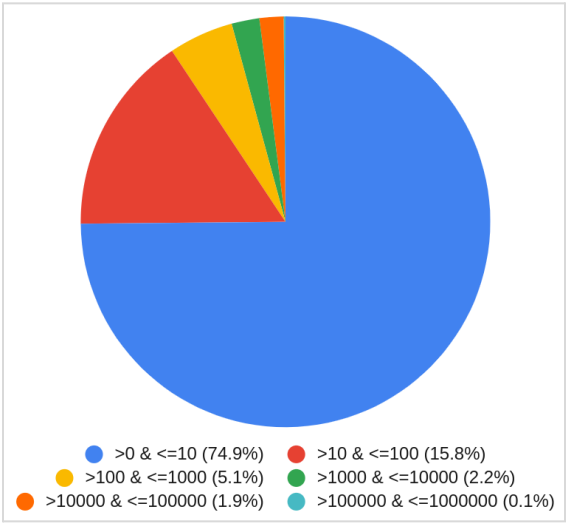

Figure 3: Distribution of languages according to their number of lexemes on Wikidata Lexemes.

[2] https://www.ethnologue.com
[3] https://ordia.toolforge.org/language/
[4] https://w.wiki/DptT
[5] https://w.wiki/DpuY (Lexeme:L1322560)
[6] https://qichwa.wikibase.cloud/
[7] https://w.wiki/7v$r (Wikidata Lexemes)
[8] https://github.com/LeMyst/WikibaseIntegrator
[9] https://luthor.toolforge.org/

As shown in **Figure 3**, only a few languages have succeeded to populate their lexemes on Wikidata Lexemes. As a result, we may all be facing an ecosystem that perpetuates biases of over- or under-representation of languages, resulting in an inaccurate and incomplete representation of the world's knowledge and languages. Recalling the "Innovate in Free Knowledge" recommendation from the Wikimedia 2030 Movement Strategy[10], it calls for attention to innovative approaches to better integrate various tools to include more diverse domains of knowledge, and to support more diverse modes of consumption and contribution to Wikimedia projects.

In order to address a better support of Wikidata Lexemes for all languages, it is necessary to analyse the holistic process of collecting, populating, storing, curating, and deploying lexical resources on Wikidata Lexemes (**Huaman et al., 2025**). The main research question of this research proposal is: **How can Wikidata Lexemes be improved to better support the population of lexicographical data of the Puno Quechua language?** In order to answer this question, we will break it down into specific sub-questions:

RQ1. Which methods and tools can be used within the Wikidata Lexeme ecosystem to **collect** lexical resources?
RQ2. Which approaches can be followed to **populate** Wikidata Lexemes with lexicographical resources?
RQ3. What tools within the Wikidata Lexemes ecosystem can support the **curation** of lexemes?
RQ4. How can Wikidata Lexemes be **deployed** and leveraged?

Through our research and by addressing these questions, we will advance the Wikimedia movement towards the 2030 strategic direction, by supporting contributors and communities (e.g. Puno Quechua) to be represented in the free knowledge ecosystem, and by supporting the building of services and structures (e.g. Wikidata Lexemes) that empower all languages.

Date: July 1, 2025 – June 30, 2026.

Related work

Knowledge Graphs (KGs), like Wikidata, are very large semantic nets that integrate and represent knowledge from various domains (Fensel, Simsek, et al., 2020; Simsek et al., 2023). They allow representing knowledge in structured format and can be understood by humans and machines. Unlike KGs that focus on factual assertions (e.g., "Lima is the capital of Peru"), Wikidata Lexemes encodes linguistic knowledge (e.g., "Rantiy is a verb"), making it a specialized subgraph within Wikidata. However, recent studies highlighted the presence of social and technical biases throughout the KGs lifecycle (Alexandris, 2022; Groth et al., 2022), e.g., biases can emerge during KG creation (Janowicz et al., 2018), hosting (Voit & Paulheim, 2021), curation (Demartini, 2019; **Huaman & Fensel, 2021**), and deployment (Radstok et al., 2021). Looking at **Figure 3**, Wikidata Lexemes may be facing these biases.

Under-Resourced Languages (URLs), are languages that have a limited amount of resources (Del Gratta et al., 2014), struggle to be supported by the pace of technology[11], are marginalized by society, minoritized by dominant languages, and excluded from decision-making processes. Contrary to highly resourced languages, that have sufficient resources such as data, text, and multimedia, and they can be supported by technology. For

[10] https://w.wiki/6$dF

[11] Only 17 languages are supported by Google Assistant. (accessed 10 April 2025) https://support.google.com/googlenest/answer/7550584

example, the top 15 languages in Wikidata Lexemes[12] represent 75.5% of all lexemes.

Puno Quechua Language Use Case, Quechua is a family of languages, 42 identified on Ethnologue[13] and 43 identified on Glottolog[14]. Its vital status has been affected by historical situations, heterogeneous linguistic diversity, governmental language policies, and sociocultural contexts (Hornberger & Coronel-Molina, 2004). In this research proposal, we focus on the Puno Quechua language, identified with the ISO 639-3: qxp[15]. It is mainly spoken in the Puno region of Peru, where the number of Quechua speaker has been decreasing (Andrade Ciudad, 2019; Carbajal Solis et al., 2019; Lópes, 1993): 87,23% in 1940, 50% in 1993, and 38,01% in 2007.

This research will determine the extent to which the Wikidata Lexemes ecosystem supports the Puno Quechua language (a URL) to be represented.

Methods

We propose to explore the holistic lifecycle of Wikidata Lexemes to gain insights into current practices of integrating new languages. We then use the Puno Quechua language to demonstrate the complexity, feasibility and potential impact of Wikidata Lexemes. Finally, we will develop recommendations for supporting Wikidata Lexemes to support all languages, and a toolkit for communities to integrate their languages.

To do this, and to move from data to results, we will carry out a two-phase research study. This will involve qualitative and quantitative analysis.

Phase 1: Qualitative Analysis

Our aim is to understand how other language communities populate the Wikidata lexemes. The idea is to survey Wikidata Lexemes contributors and tool developers[16] about their approach to **collecting**, **populating**, **curating** and **deploying** lexemes.

- (initially) Survey 10 contributors[17], one per language e.g., German, English.
- (initially) The survey will be structured on 4 main topics: collecting, populating, curating and deploying lexemes.
- (initially) Survey 8 tool developers, two per each topic e.g., collecting.

In this phase, the analysis will be used to make qualified strategic decisions: what methods and tools are working in other language communities that can be followed to a) collect, b) populate, c) curate, and d) deploy lexemes on/from Wikidata Lexemes.

Phase 2: Quantitative Analysis

We aim to populate Wikidata Lexemes with Puno Quechua lexicographical data. The main approach is to use the Puno Quechua language and test the tools and approaches mastered by other language communities, e.g., the Wikidata Lexeme Forms[18] tool. It will involve:

- Localization of 4 tools.
- **Collect** Puno Quechua lexemes from Qichwabase and other sources to reach 20000 lexemes.
- Ingest the lexemes into Wikidata Lexemes (i.e. **population**).
- Improve the quality of the lexemes (i.e., **curation**).
- **Deploy** Puno Quechua lexemes from Wikidata Lexemes.

[12] https://w.wiki/DptT (Wikidata Lexemes Count)
[13] https://www.ethnologue.com/language/que/
[14] https://glottolog.org/resource/languoid/id/quec1387
[15] https://iso639-3.sil.org/code/qxp
[16] https://w.wiki/6hML (Wikidata Lexeme Tools)
[17] Wikidata Lexicographical data telegram group https://t.me/+TwVx84TOZfJ0-JvK (237 members)
[18] https://lexeme-forms.toolforge.org/

We will partner, and be assisted, with native Puno Quechua speakers to gather lexical data, and to evaluate the localization of the tools. For example, the Wikidata Lexeme Forms tool helps to generate verb conjugations for lexemes based on Wikifunctions, which needs to be coded to return correct conjugations (see **Figure 4**).

As we approach this phase, we will identify the methods and tools that are (best) suited to the Puno Quechua language or that do not work, and analyse the usefulness of the tools to carry out the tasks when faced with Under-Resourced Languages.

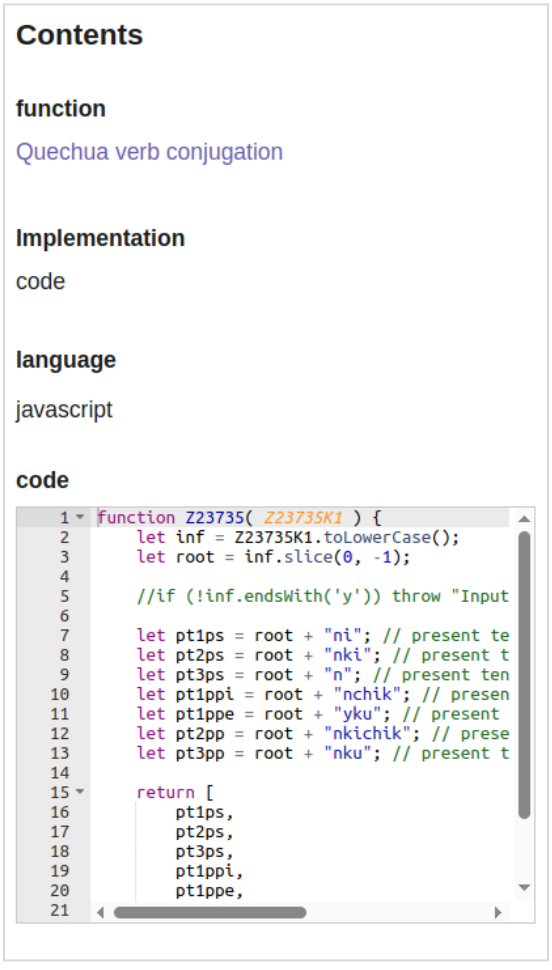

Figure 4: An excerpt view of the "quechua verb form list Javascript" implementation[19] on Wikifunctions.

Expected output

Scientific Publications

Our goal is to publish a journal paper detailing the results and findings of our research, and to present our findings at conferences. The target audiences are the Wikimedia community, researchers interested in Lexicography and Under-Resourced Languages, and Quechua communities.

We are targeting high impact journals such as The Semantic Web[20], but we will also consider other journals where bibliometrics are valued.

Dataset

Approximately 20000 lexemes of the Puno Quechua language will be added to Wikidata Lexemes, describing senses, forms, pronunciation audio, and more, making the dataset available to everyone. The target audience is Quechua communities within the Wikimedia movement, Quechua speakers, researchers, and potential contributors.

Insights to inform decision-making

We will draft initial recommendations for building more equitable and inclusive Wikidata Lexemes that can support all languages. The target audience will be the Wikimedia movement, researchers, and developers. Later, we will ask them for feedback to refine the recommendations. Finally, we will release the recommendations under a CC0 licence on Wikimedia Commons.

Events

We aim to present our findings on venues such as the Wiki Workshop, Wikimania 2026, as well as at meetings within the Wikimedia movement.

[19] https://w.wiki/Dq6v (Quechua Verb Forms)

[20] https://www.semantic-web-journal.net

Risks

There may be factors that compromise the success of our research project:

Wikidata Lexeme tools developers cannot be found or we are not able to contact them, because tools are no longer maintained. To mitigate this risk, we will prioritize the tools that have been updated or were active in the last 6 months.

Wikidata Lexeme tools are not localizable to other languages. This may be due to some project dependencies, or because the tools are simply too complex to localize. For example, when a requirement is to have a Wikisource project in a target language. To mitigate this risk, we will look at other Wikidata Lexeme tools that tackle the same task.

We are not able to recruit lexeme contributors from specific languages who are willing to share their expertise. This can be mitigated by considering contributors from other language communities.

Without a Quechua community, the initiative could certainly have less impact. We plan to mitigate this risk by partnering with institutions in Puno (Peru) and organizing workshops so that a collaborative mentality can be achieved throughout the Quechua community.

The time commitment might represent a challenge. To mitigate this, we have included a salary for the researchers and a cash reward for survey contributors, so that we (all participants) can commit to the success of this project.

Community impact plan

We want to achieve a systemic change, which requires the collaboration of diverse stakeholders. We want to ensure that the results of this research have a real-world impact, where Wikimedia volunteer developers, Wikimedia researchers, and language communities collaborate together. The findings can inform them to build more inclusive tools, promote responsible language data collection practices, and foster knowledge ownership rights of language communities.

This research aims to empower language communities within the Wikimedia ecosystem to contribute their language data to Wikidata Lexemes, ensuring that their voices are heard and represented in the free knowledge ecosystem and beyond.

The long-term goal is to strengthen the position of Wikidata Lexemes to support all languages in the digital space and beyond. For example, the published lexemes can be used to generate natural language text and enrich other Wikimedia projects, and to create open educational resources.

Evaluation

The project can be evaluated as successful if we are able to:

- Interview at least 10 lexeme contributors.
- Interview at least 8 Wikidata Lexeme tool developers.
- Localize 4 Wikidata Lexeme tools.
- Integrate at least 75% of all collected Puno Quechua lexemes into Wikidata Lexemes.
- Present the research outcomes at least twice.
- Submit a paper to a journal (or at least to a conference).

Budget

Details of the budget supporting our research project can be found here.

References

Alexandris, C. (2022). Sense and sensitivity: Knowledge graphs as training data for processing cognitive bias, context and information not uttered in spoken interaction. In T. Kido & K. Takadama (Eds.), Proceedings of the symposium how fair is fair? achieving wellbeing AI co-located with association for the advancement of artificial intelligence 2022 spring symposium (aaai-spring symposium 2022), stanford, ca, march 21-23, 2022 (pp. 47–54, Vol. 3276). CEUR-WS.org.

Andrade Ciudad, L. (2019). Ten News About Quechua in the last Peruvian census. Letras (Lima), 90, 41–70.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Commun., 56, 85–100.

Carbajal Solis, V., Garcia Rivera, F. A., Huamancayo Curi, E. Y., Mori Clement, M., Rodriguez Aguero, M., Gutierrez, C., & Verastegui Walqui, N. (2019). Lenguas originarias del Perú. Ministerio de Educación del Perú.

Demartini, G. (2019). Implicit bias in crowdsourced knowledge graphs. In S. Amer-Yahia, M. Mahdian, A. Goel, G. Houben, K. Lerman, J. J. McAuley, R. Baeza-Yates, & L. Zia (Eds.), Companion of the 2019 world wide web conference, WWW 2019, san francisco, ca, usa, may 13-17, 2019 (pp. 624–630). ACM.

Fensel, D., Simsek, U., Angele, K., **Huaman, E.**, Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., & Wahler, A. (2020). Knowledge graphs - methodology, tools and selected use cases. Springer.

Groth, P., Simperl, E., van Erp, M., & Vrandecic, D. (2022). Knowledge graphs and their role in the knowledge engineering of the 21st century (dagstuhl seminar 22372). Dagstuhl Reports, 12(9), 60–120.

Hornberger, N. H., & Coronel-Molina, S. M. (2004). Quechua language shift, maintenance, and revitalization in the andes: The case for language planning. International Journal of the Sociology of Language, 2004(167), 9–67.

Huaman, E., & Fensel, D. (2021). Knowledge graph curation: A practical framework. IJCKG'21: The 10th International Joint Conference on Knowledge Graphs, Virtual Event, Thailand, December 6 - 8, 2021, 166–171.

Huaman, E., Huaman, J. L., & Huaman, W. (2022). Getting quechua closer to final users through knowledge graphs. In J. A. Lossio-Ventura, J. Valverde-Rebaza, E. Díaz, & H. Alatrista- Salas (Eds.), Information management and big data - 9th annual international conference, simbig 2022, lima, peru, november 16-18, 2022, proceedings (pp. 61–69, Vol. 1837). Springer.

Huaman, E., Huaman, W., & Huaman, J. (2025). Making an Under-Resourced Language available on the Wikidata Knowledge Graph: Quechua Language. Information management and big data - 9th annual international conference, SIMBig 2024, Ilo, Peru, november 20–22, 2024, to be published on Springer.

Huaman, E., Lindemann, D., Caruso, V., & Huaman, J. L. (2023). QICHWABASE: A Quechua Language and Knowledge Base for Quechua Communities. arXiv preprint arXiv:2305.06173.

Janowicz, K., Yan, B., Regalia, B., Zhu, R., & Mai, G. (2018). Debiasing knowledge graphs: Why female presidents are not like female popes. In M. van Erp, M. Atre, V. López, K. Srinivas, & C.

Fortuna (Eds.), Proceedings of the ISWC 2018 posters & demonstrations, industry and blue sky ideas tracks co-located with 17th international semantic web conference (ISWC 2018), monterey, usa, october 8th - to - 12th, 2018 (Vol. 2180). CEUR-WS.org.

Lópes, L. E. (1993). Educación bilingüe en puno-perú: Activos y pasivos de un programa de edu- catión rural en los andes. Lingüística indígena e educação na América Latina. Campinas: UNICAMP, 13–70.

Radstok, W., Chekol, M. W., & Schäfer, M. T. (2021). Are knowledge graph embedding models biased, or is it the data that they are trained on? In L. Kaffee, S. Razniewski, & A. Hogan (Eds.), Proceedings of the 2nd wikidata workshop (wikidata 2021) co-located with the 20th international semantic web conference (ISWC 2021), virtual conference, october 24, 2021 (Vol. 2982). CEUR-WS.org.

Simsek, U., Kärle, E., Angele, K., **Huaman, E.**, Opdenplatz, J., Sommer, D., Umbrich, J., & Fensel, D. (2023). A knowledge graph perspective on knowledge engineering. SN Comput. Sci., 4(1), 16.

Voit, M. M., & Paulheim, H. (2021). Bias in knowledge graphs - an empirical study with movie recommendation and different language editions of dbpedia. In D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, & B. Heinisch (Eds.), 3rd conference on language, data and knowledge, LDK 2021, september 1-3, 2021, zaragoza, spain (14:1–14:13, Vol. 93). Schloss Dagstuhl - Leibniz-Zentrum für Informatik.