netidee STIPENDIEN

Building Knowledge Graphs for Under-Resourced Languages

Endbericht | Call 19 | stip7335

Contents

1 Introduction

This final report provides insights on my doctoral dissertation on Building Knowledge Graphs (KGs) for Under-Resourced Languages (URLs). This work sits at the intersection of natural language processing, semantic web technologies, and AI. The structure of this report is as follows: Section 2 introduces the topic, motivation, and research questions. In Section 3, we explain the existing results and insights. Afterwards, Section 4 provides details on what needs to be accomplished in order to complete the answer to all research questions. Finally, Section 5 outlines some future directions.

2 General information

Summary of the thesis, research question, aim of the thesis

In the last decades, the development of KGs by the major tech companies[1][2] have got a lot of attention [FenselSAHKPTUW20, 2021Hogan,KaffeeBKVM23] because KGs provide structured information that AI applications can use. KGs encode human knowledge explicitly, facilitating AI reasoning and decision-making. However, current KG construction methods lack transparency and equity regarding the knowledge and data of URLs. For instance, KGs are mainly built based on data from the web, which may over-represent certain languages.

In this research project, I explored the problems of building KGs with respect to URLs, which represents significant challenges due to the limited availability of data, linguistic resources, and computational tools. Furthermore, URLs are languages that are affected by different aspects, such as language vitality, language policies, and dominant languages. Enabling URLs and their communities to represent their knowledge and language as KGs, will act as a common sense layer for AI applications, facilitating AI reasoning to operate more responsibly and transparently. Therefore, I researched the holistic process of KG generation with respect to URLs to answer the question: **How can knowledge graphs be built to support lexicographical data of under-resourced languages?**

To address this question, I focused on the Puno Quechua language as a use case. First, I researched the current practices for building KGs, then I collected Puno Quechua lexical resources, such as dictionaries, vocabularies, and pronunciation speech audios to model their representation in a KG. Then, based on the KG, I collected missing information, such as part-of-speech tags, forms, senses, usage examples, and pronunciation audio. Finally, in order to build a community-centred language-aligned KG for the Puno Quechua language,

[1] https://developer.amazon.com/en-US/docs/alexa/custom-skills/develop-skills-in-multiple-languages.html
[2] https://support.google.com/googlenest/answer/7550584

the KG will be made accessible online and can be curated by the Quechua communities. In addition, I studied to what extent the KG improves the learning experience of the Puno Quechua language. Last but not least, I aim to raise awareness about the bias on KGs, improve data governance over which communities of URLs like Quechua have no control, and empower these communities to reduce linguistic inequality, and promote linguistic diversity.

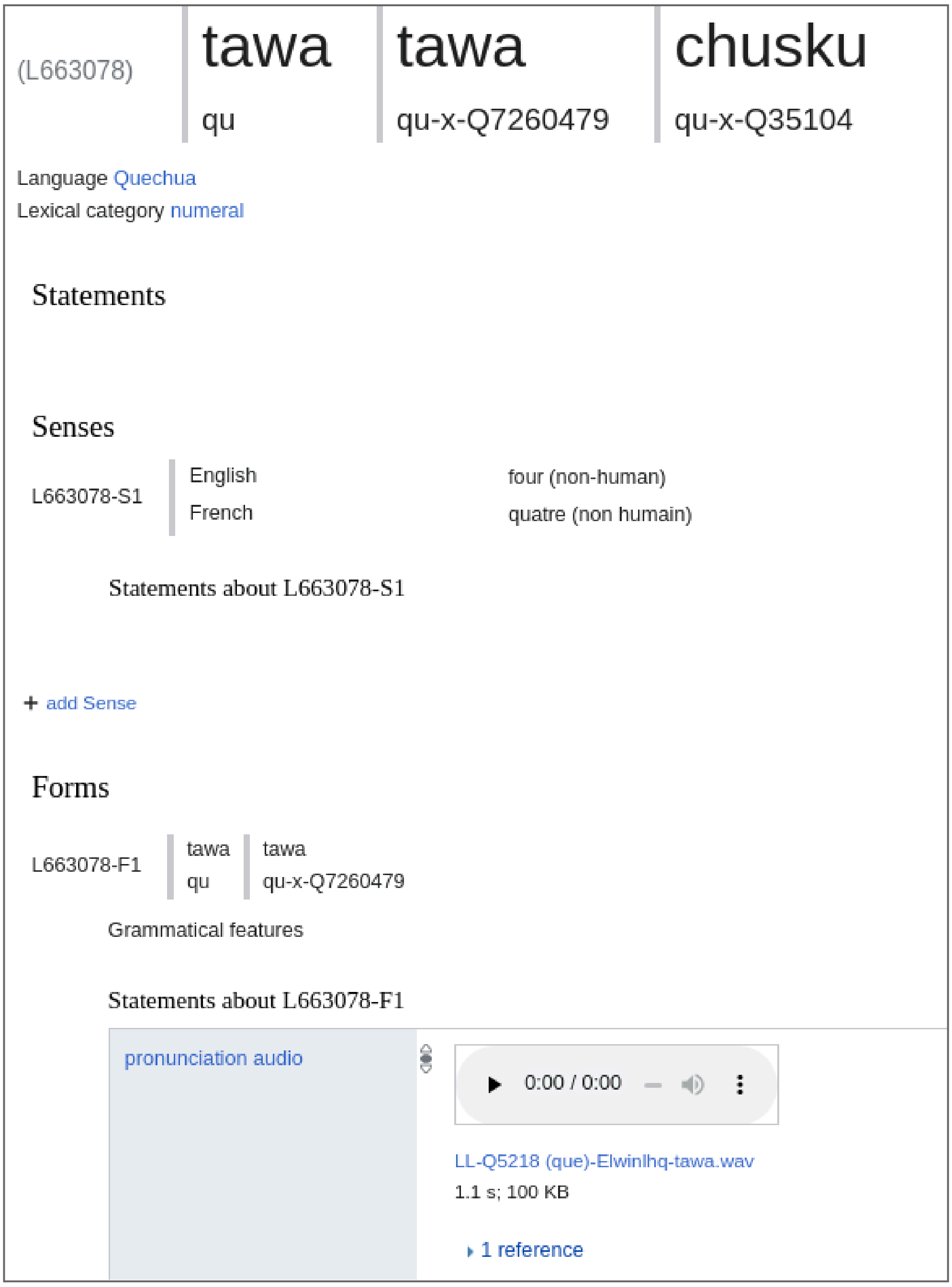

Figure 1. A lexeme representation of a Quechua word

Figure 1 represents a lexeme with its seven main parts. The represented lemma is “Tawa” (en: four) and it can be described as follows:

1. **Lexeme ID**: This is the word's unique ID number. It always starts with an “L”. For our example, it is L663078.
2. **Lemmas** (The Headword): This is the dictionary form of the word, the label in Quechua language.
    a. For “tawa”, the lemma is the string "tawa".
    b. For “tawa” the Quechua language tag is “qu”. But what if you want to represent a word in a specific regional variety, like Puno Quechua?
    c. For “tawa” the Puno Quechua language tag is “qxp” or “qu-x-Q7260479”. This allows us to represent the language's diversity accurately.
3. **Language**: This simply states which language the word belongs to. For "tawa", this is Quechua.
4. **Lexical Category**: This describes the part of speech of the word. "tawa" is a “numeral”.
5. **Statements**: This describes the facts about the word itself, not its meanings.
    a. For example, we could have a statement that says "tawa" is derived from the root "ta-".
    b. For example, it can describe the provenance of the word, where was taken from, like a citation of the original work.
6. **Senses**: These are the different meanings of the word. I.e, It shows equivalences of the word in other languages. For example, the word “tawa” has equivalents in English and French.
7. **Forms**: This lists all the different ways the word can change. For example, singular, plural, etc. Additionally, it can describe properties such as “pronunciation audio”, which can link to audio files for that word.

This dissertation focused on the creation of Knowledge Graphs (KGs) for Under-Resourced Languages (URLs). Our core contributions are as follows:

- Addressing the challenges of collecting resources of languages that have low amounts of resources, for instance, Quechua languages.
- Design a process model for modelling the integration of linguistic data of under-resourced languages into knowledge graphs.
- Tackle the hosting and storing of knowledge graphs that support under-resourced languages based on FAIR and LOD principles, as well as their licensing and distribution.
- Provide guidelines on a community-driven approach to the curation of knowledge graphs that are focused on under-resourced languages.

- Develop applications that can enhance the learning experience of language learners.
- Disseminate research findings through reports, presentations, and publications.

3 Results

Description of the results achieved

- The literature review confirms a significant gap on knowledge graphs for under-resourced languages. While large multilingual knowledge graphs exist, their coverage of under-resourced languages is often shallow, incomplete, and English-centric, which introduces errors and cultural biases.
- The schema alignment and knowledge graph creation has been completed, and will continue. It is an iterative process and more knowledge can be ingested as they come. The advantage of Knowledge graphs is that even if you do not have a Graph database, it can still be maintained in structured format and files, e.g., JSON-LD format.
- This research addresses the challenges of collecting resources of languages that have low amounts of resources and present them in a structured format. The resources we have collected for the Puno Quechua language are stored on Mozilla Data Collective and can be retrieved from https://mozilladatacollective.com. The current results show the collection of more than 30 hours of Spontaneous Speech and more than 34 hours of Scripted speech for the Puno Quechua language.
- A process model for modelling the integration of data and knowledge from different points of view and that are particular to under-resourced languages (see Figure 1).
- The hosting and storing of under-resourced language knowledge graphs based on FAIR and LOD principles, as well as their licensing and distribution. It uses an Open Source and Community-drive approach, i.e., Wikibase: http://wikibase.cloud/

4 Planned follow-up activities

Specification of remaining activities (timetable)

February to May 2026:

- Linking Lexemes with Pronunciation Audio files based on the Lexeme model
- User testing of the final platform that provides access to the knowledge graph

June 2026:

- Proof-reading of the thesis and minor corrections

5 Suggestions for continuation by third parties

What continuations for third parties result from your work?

Our work provides a lexeme representation model that can be applied to any language. For instance, in order to create a knowledge graph (KGs) for an under-resourced language (URL) the following tasks need to be followed:

- **Knowledge Acquisition**, is the task of collecting or retrieving lexicographical data from heterogeneous sources, and then semantically annotating them based on a bottom-up or top-down approach.
- **Knowledge Hosting**, refers to the approach of storing the generated knowledge in a database, which must take into account the size and application of KGs.
- **Knowledge Curation**, is the task of ensuring the highest possible correctness and completeness of the KG, the overall goal of knowledge curation is to provide cost-sensitive methods to improve KG's quality and usefulness for applications.
- **Knowledge Deployment**, is the task of building applications on top of KGs that can serve the URL community, e.g., web search, personal assistants, learning management systems, and more.

We encourage third parties, such as public institutions, linguistic departments that are working with under-resourced languages across Europe and the world, to create a knowledge graph for their languages of interest. This will allow those languages to be part of a bigger knowledge ecosystem where all words are interlinked between knowledge graphs. I.e. Word equivalents between languages can be linked.

6 Bibliography

- Aghaei, S., Angele, K., Huaman, E., Bushati, G., Schiestl, M., & Fensel, A. (2022). Interactive search on the web: The story so far. Inf., 13(7), 324. https://doi.org/10.3390/INFO13070324

- 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. http://ceur-ws.org/Vol-3276/SSS-22%5C_FinalPaper%5C_14.pdf

- Andrade Ciudad, L. (2019). Ten news about quechua in the last peruvian census. Letras (Lima), 90, 41–70. http://dx.doi.org/10.30920/letras.90.132.2

- Angele, K., Fensel, D., Huaman, E., Kärle, E., Panasiuk, O., Şimşek, U., Toma, I., & Wahler, A. (2020). Semantic web empowered e-tourism. In Z. Xiang, M. Fuchs, U. Gretzel, & W. Höpken (Eds.), Handbook of e-tourism (pp. 1–46). Springer International Publishing. https://doi.org/10.1007/978-3-030-05324-6_22-1

- Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., Breiner, T., Axelrod, V., Riesa, J., Cao, Y., Chen, M. X., Macherey, K., Krikun, M., Wang, P., Gutkin, A., Hughes, M. (2022). Building machine translation systems for the next thousand languages. CoRR, abs/2205.03983. https://doi.org/10. 48550/ARXIV.2205.03983

- Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Commun*., 56, 85–100. https://doi.org/10.1016/J.SPECOM.2013.07.008

- Butryna, A., Chu, S. C., Demirsahin, I., Gutkin, A., Ha, L., He, F., Jansche, M., Johny, C., Katanova, A., Kjartansson, O., Li, C., Merkulova, T., Oo, Y. M., Pipatsrisawat, K., Rivera, C., Sarin, S., Silva, P. D., Sodimana, K., Sproat, R., . . . Wibawa, J. A. E. (2020). Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: An overview. CoRR, abs/2010.06778. https://arxiv.org/abs/2010.06778

- 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 peru. Ministerio de Educacion del Peru. https://doi.org/https://hdl.handle.net/20.500.12799/6261

- Casanova, P., & Jones, M. (2013). What is a dominant language?: Giacomo leopardi: Theoretician of linguistic inequality. New Literary History, 44(3), 379–399. https://muse.jhu.edu/article/525064

- Caswell, I., Breiner, T., van Esch, D., & Bapna, A. (2020). Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In D. Scott, N. Bel, & C. Zong (Eds.), Proceedings of the 28th international conference on computational linguistics, COLING 2020, barcelona, spain (online), december 8-13, 2020 (pp. 6588–6608). International Committee on Computational Linguistics. https://doi.org/10.18653/V1/2020.COLING-MAIN.579

- Del Gratta, R., Frontini, F., Khan, A. F., Mariani, J., & Soria, C. (2014). The lre map for under-resourced languages. Workshop Collaboration and Computing for Under-Resourced Lan-guages in the Linked Open Data Era, Satellite Workshop of LREC, 14.

- 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. https://doi.org/10.1145/3308560.3317307

- Dwyer, A. M. (2012). Tools and techniques for endangered-language assessment and revital-ization. GYA, K., SNAVELY, A., & SPERLING, E.(Édit.), Minority Language inToday's Global Society, 201–251. https://kuscholarworks.ku.edu/bitstream/handle/1808/7109/Dwyer2011%5C_AssessRevitalize.pdf

- 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. https://doi.org/10.1007/978-3-030-37439-6

- Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., & Wahler, A. (2020). Why we need knowledge graphs: Applications. In Knowledge graphs: Methodology, tools and selected use cases (pp. 95–112). Springer International Publishing. https://doi.org/10.1007/978-3-030-37439-6_4

- 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. https://doi.org/10.4230/DAGREP.12.9.60

- Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). *Knowledge graphs*. Morgan & Claypool Publishers. https://doi.org/10.2200/S01125ED1V01Y202109DSK022

- 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. https://doi.org/10.1515/IJSL.2004.025

- 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. https://doi.org/10.1145/3502223.3502247

- 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. https://doi.org/10.1007/978-3-031-35445-8_5

- Jacobsen, A., de Miranda Azevedo, R., Juty, N. S., Batista, D., Coles, S. J., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T. A., Goble, C. A., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K. M., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., Schultes, E. (2020). FAIR principles: Interpretations and implementation considerations. Data Intell., 2(1-2), 10–29. https://doi.org/10.1162/DINT_R_00024

- 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. https://ceur-ws.org/Vol-2180/ISWC%5C_2018%5C_Outrageous%5C_Ideas%5C_paper%5C_17.pdf

- Kaffee, L., Biswas, R., Keet, C. M., Vakaj, E., & de Melo, G. (2023). Multilingual knowledge graphs and low-resource languages: A review. TGDK, 1(1), 10:1–10:19. https://doi.org/10.4230/TGDK.1.1.10

- Körner, E., Helfer, F., Schröder, C., Eckart, T., & Goldhahn, D. (2022). Crawling under-resourced languages-a portal for community-contributed corpus collection. In J. Sälevä & C. Lignos (Eds.), Proceedings of the workshop on dataset creation for lower-resourced languages within the 13th language resources and evaluation conference (pp. 36–43). European Language Resources Association. https://aclanthology.org/2022.dclrl-1.5

- Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suárez, P. J. O., . . . Adeyemi, M. (2022).

Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Comput. Linguistics, 10, 50–72. https://doi.org/10.1162/TACL_A_00447

- Lópes, L. E. (1993). Educación bilingüe en puno-perú: Activos y pasivos de un programa de educación rural en los andes. Lingüística indígena e educação na América Latina. Campinas: UNICAMP, 13–70. https://etnolinguistica.wdfiles.com/local--files/biblio%3Alopez-1993-educacion/Lopez_1993_EducacionBilinguePunoPeru.pdf
- 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. https://ceur-ws.org/Vol-2982/paper-5.pdf
- Sallabank, J., & Austin, P. K. (2022). Endangered languages. In The routledge handbook of applied linguistics (pp. 362–373). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003082637-31/endangered-languages-julia-sallabank-peter-austin
- 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. https://doi.org/10.1007/S42979-022-01429-X
- 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. https://doi.org/10.4230/OASICS.LDK.2021.14