WordNet-QU: Development of a Lexical Database for Quechua Varieties

Nelsi Melgarejo
Pontifical Catholic University of Peru
nelsi.melgarejo@pucp.edu.pe

Rodolfo Zevallos
Pompeu Fabra University
rodolfojoel.zevallos@upf.edu

Héctor Gómez
Pontifical Catholic University of Peru
hector.gomez@pucp.edu.pe

John E. Ortega
Northeastern University
j.ortega@northeastern.edu

Abstract

In the effort to minimize the risk of extinction of a language, linguistic resources are fundamental. Quechua, a low-resource language from South America, is a language spoken by millions but, despite several efforts in the past, still lacks the resources necessary to build high-performance computational systems. In this article, we present **WordNet-QU** which signifies the inclusion of Quechua in a well-known lexical database called *wordnet*. We propose **WordNet-QU** to be included as an extension to wordnet after demonstrating a manually-curated collection of multiple digital resources for lexical use in Quechua. Our work uses the *synset* alignment algorithm to compare Quechua to its geographically nearest high-resource language, Spanish. Altogether, we propose a total of 28,582 unique synset IDs divided according to region like so: 20510 for Southern Quechua, 5993 for Central Quechua, 1121 for Northern Quechua, and 958 for Amazonian Quechua.

1 Introduction and related work

Lexical databases and resources have been used in the past for various natural language processing (NLP) tasks ranging from information retrieval (IR) to machine translation (MT). While many recent NLP approaches rely on deep learning techniques like transformers, namely BERT (Devlin et al., 2018), where attention (Vaswani et al., 2017) is used to create a semantic representation of text, more traditional approaches relied on purely linguistic and syntactic features. More often than not, recent deep-learning approaches require a large amount of data to perform better than traditional ones (e.g. on the order of millions of words for machine translation (Koehn and Knowles, 2017; Bahdanau et al., 2014)). This makes NLP approaches with low-resource languages, languages that are measured in the thousands typically, much more difficult to solve with recent approaches thus forcing the use of traditional approaches to solve problems.

One low-resource language from South America, called Quechua, is spoken by nearly 8 million people[1] yet still does not have enough resources to effectively compete with other high-resource languages as has been shown in previous research (Ebrahimi et al., 2021; Ortega et al., 2021, 2020). Oftentimes, due to insufficient resources, scores such as BLEU (Papineni et al., 2002) and accuracy are more than three times lower. This lack of resources thus drives the need for traditional techniques such as the use of lexical databases, grammars, and other linguistic cues such as tree banks and more. One such resource that has been commonly used for traditional approaches is called *wordnet* (Fellbaum, 1998) which was originally created in the 1990s yet is still used today, especially for low-resource languages like Quechua.

The need to build digital resources is greater for endangered languages like Quechua and others since there is a clear desire to save the language from extinction. However, the desire is typically not supported by those agencies that are responsible for its survival. Berment (Berment, 2002) and others have expressed the need for further analysis and research stating that the current effort "may be insufficient to aid preservation efforts". In this work, we provide several lexical resources for Quechua to increase its inclusion in wordnet (Fellbaum, 1998). We call the collection of resources **WordNet-QU** which corresponds to its commonly-used language-pair symbol (QU) found in most corpora for NLP in Quechua. To elaborate on its inclusion, in Section 2 we provide details on how the corpus was compiled and the annotations done. Then, in Section 3, we cover the wordnet implementation of the corpus. Finally, in Section 4 we provide insight into our future downstream tasks.

[1] https://en.wikipedia.org/wiki/Quechuan_languages

Proceedings of the 29th International Conference on Computational Linguistics, pages 4429–4433
October 12–17, 2022.

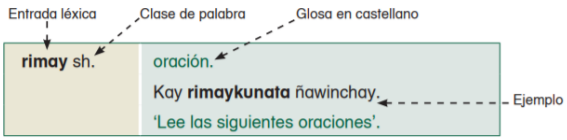

Figure 1: Example of the structure of the some dictionaries.

2 Corpus details

The corpus presented here made available publicly[2] has been created using a manually curated collection of dictionaries. The dictionaries were mostly gotten from Ministry of Education in Peru (MINEDU in Table 4, see Appendix) and consist of five regional varieties of Quechua (Southern Quechua (Collao), Southern Quechua (Chanka), Central Quechua, Northern Quechua, and Amazonian Quechua) ranging from 1976 to 2005 in the years they were collected.

In order to organize the dictionaries into a format that can be used by wordnet (Fellbaum, 1998), the corpus is structured in a format that consists of the following labels: (i) branch, (ii) variety, (iii) region, (iv) author, (v) dictionary, (vi) year, (vii) lexical entry, (viii) grammatical category, (ix) glossary entry, (x) Quechua definition, (xi) Quechua synonym, (xii) Spanish synonym and (xiii) notes or clarifications. An example of the original dictionary as found from the Ministry of Education is seen in Figure 1.

Since there are several dialects of Quechua spoken in Peru (Cerrón-Palomino, 2021), it was important to compile the corpus by variety or region. In order to better illustrate the differences in parts of speech for dialects, we break each region's dialect into the following categories: noun, adjective, adverb and verb as shown in Table 1. The variety with the highest lexical entries are Southern Quechua (South) followed by Central, Northern (North), and Amazonian (Amaz). For each of the parts of speech and varieties of Quechua, there is a corresponding Spanish glossary entry. Additionally, for the southern and central varieties, apart from the part of speech and glossary, there is a definition in Quechua and translation in Spanish. In some cases the translation is gotten from MINEEDU and in other cases native speakers translated for us.

[2] https://github.com/Llamacha/wordnet-qu

POS	Quechua variety			
	South	Central	North	Amaz
Noun	16 717	3 241	579	537
Verb	8 000	3 145	506	423
Adjective	4 116	904	157	160
Adverb	985	384	126	48
Total	**29818**	**7677**	**1368**	**1204**

Table 1: Number of words per part of speech (POS) for each Peruvian region.

3 Methodology

In order to use and distribute **WordNet-QU** we had to make it compatible with wordnet (Fellbaum, 1998). Constructing a wordnet, whether from scratch or by expanding a previous one, is a labor intensive process that requires several steps and extensive use of both human labor and automated systems. Since the creation of the first wordnet (Princeton WordNet (PWN)) in 1995 (Miller, 1995), many other wordnets have been created for several languages. For example EuroWordNet (EWN) is a multilingual wordnet project that links wordnets of multiple European languages (English, Dutch, Italian, Spanish, German, French, Czech and Estonian) (Vossen, 1997). In EWN, wordnets were created for each language separately and then linked through an index based on PWN. In the same way, BalkaNet is a multilingual wordnet project consisting of six Balkan languages (Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish). (Tufis et al., 2004)

Two of the most-commonly used approaches for creating a wordnet are based on what are known as the *expand* and *merge* approaches. Both approaches use *synsets* – groups of synonyms that express the same concept in wordnet. One synset can have multiple words and one word can have multiple synsets. In the *expand* approach, a set of synsets from PWN, including their semantic database, are first translated into the target language and then relations are transferred from English and checked in a manual fashion as is done for Scottish Gaelic (Bella et al., 2020) and the French (Sagot and Fišer, 2008). The *merge* approach builds bilingual relations from scratch, without any links to English, the main language for wordnet. Both the Polish wordnet (Derwojedowa et al., 2008) and Norwegian wordnet (Fjeld and Nygaard, 2009) use the merge approach.

Model	Size	Spearman
Pre-trained Model (Wiki)	29k	0.35
WordNet-QU (Wiki + WordNet Corpus)	31k	**0.61**

Table 2: A comparison of Spearman correlation coefficients (Wissler, 1905) between human judgement and similarity scores for pre-trained model on tokens of Wikipedia alone and Wikipedia with the WordNet-QU corpus.

Our implementation is based on a few steps. The first step is to construct a wordnet for Spanish because translations for Quechua are more available in the high-resource language (Spanish) in Peru. Using the main wordnet in English, we create a Spanish wordnet using the expansion technique described above based on similarity alone. The abundance of on-line Spanish glossaries and other relationships helped when creating the Spanish wordnet. Once translated, the Spanish wordnet became what is known as our multi-lingual central repository (MCR) for Quechua. This, in turn, facilitates the next steps which are to create and align synsets to with their corresponding concept which is validated manually by a human.

3.1 Synset alignment

The most important part of creating a wordnet is the alignment of synsets to their main concept. Our algorithm focuses on a straightforward process. First, the algorithm iterates through the entire wordnet MCR in Spanish for each word from the Quechua corpus.[3] When an exact Quechua–Spanish match is found and verified (manually), all of the related words from the Quechua vocabulary are mapped to their corresponding Spanish concept. This process constitutes the creation of a Quechua synset for one or more words that exist in their Spanish counterpart. After the synset creation, part-of-speech tags are created according to their grammatical category.

3.2 Wordnet validation

In order to validate the feasibility of **WordNet-QU**, we measure the cosine similarity distance between two FastText (Grave et al., 2018) models: (1) a baseline model[4] based on Wikipedia[5] which contains Quechua text and (2) a model based on Wikipedia with the addition of the **WordNet-QU** corpus. Our FastText (Grave et al., 2018) model is trained using 31 thousand tokens and identical hyper-parameters and algorithm as the baseline (skipgram algorithm, an embedding size of 300 dimensions, a context window size of 5, and n-grams ranging from 3 to 6 characters). The cosine similarity is measured for a 1000 randomly collected synsets. The distance results are then compared to the annotator's yes/no decision of whether or not each synset corresponds to the words from **WordNet-QU**. Human judgement is found to correspond much higher with the WordNet-QU model than the pre-trained model as shown in Table 2. We leave further improvement for future work.

4 Results and future work

Variety	**Synsets**	**Def.**	**Sent.**
Southern	20 510	1 873	1 827
Central	5 993	1 191	1 191
Northern	1 121	-	-
Amazonian	958	-	-
Total	**28 582**	**3 064**	**3 018**

Table 3: A count of synsets, definitions, and sentences per variety.

We have presented the process and resources used to create a wordnet-based resource for Quechua called **WordNet-QU**. We use fastText embeddings as a manner of measuring the similarity between Quechua words and Spanish concepts which provides nearly the 29k synsets illustrated in Table 3. We make the synsets and various lexicons created available publicly. For more details on specific dialects and other information related to our processing, please consult the Appendix.

This research was focused on the development of a Quechua wordnet using synonyms between different varieties of Quechua. The dictionaries used from different sources had to be identified for there region and dialect which became an after-the-fact asset to our work.

Future lines of investigations are based on work that is planed with several renown authors in

[3] Translations from Quechua to Spanish are performed beforehand.

[4] https://fasttext.cc/docs/en/crawl-vectors.html

[5] https://www.wikipedia.org/

the field of NLP processing of Quechua to use **WordNet-QU** in downstream tasks. Some of the NLP approaches that are currently in discussion are **WordNet-QU** for Quechua–Spanish translation and **WordNet-QU** for POS tagging in treebanks.

Acknowledgements

This work has been partially supported by the Project PID2019-104512GB-I00, Ministerio de Ciencia e Innovación and Agencia Estatal de Investigación (Spain).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Gábor Bella, Fiona McNeill, Rody Gorman, Caoimhín Ó Donnaíle, Kirsty MacDonald, Yamini Chandrashekar, Abed Alhakim Freihat, and Fausto Giunchiglia. 2020. A major wordnet for a minority language: Scottish gaelic. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2812–2818.

Vincent Berment. 2002. Several directions for minority languages computerization. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

Rodolfo Cerrón-Palomino. 2021. The languages of the inkas. In *The Inka Empire*, pages 39–54. University of Texas Press.

Magdalena Derwojedowa, Maciej Piasecki, Stanis[a]aw Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of polish wordnet. *Proceedings of GWC 2008*, pages 162–177.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, et al. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

C Fellbaum. 1998. Wordnet: Wiley online library. *The Encyclopedia of Applied Linguistics*, 7.

Ruth Vatvedt Fjeld and Lars Nygaard. 2009. Nornet-a monolingual wordnet of modern norwegian. In *NODALIDA 2009 workshop: WordNets and other Lexical Semantic Resources-between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7, pages 13–16.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2021. Love thy neighbor: Combining two neighboring low-resource languages for translation. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 44–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *OntoLex*.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Piek J.T.M. Vossen. 1997. Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997 Zurich*. Vrije Universiteit.

Clark Wissler. 1905. The spearman correlation formula. *Science*, 22(558):309–311.

Variety of Quechua	Dictionary	Author	Year
Southern (Collao)	Yachakuqkunapa Simi Qullqa	MINEDU	2005
	Diccionario quechua: Cuzco– Collao.		2005
Southern (Chanka)	Yachakuqkunapa Simi Qullqa	MINEDU	2005
	Diccionario quechua: Cuzco– Chanka		2005
Central	Chawpi Qichwapa Chimi Qullqan	MINEDU	2017
	Yachachinapaq shimikunachawpin qichwa		2005
Northern	Diccionario quechua: Cajamarca – Cañaris	MINEDU	1976
Amazonian	Shimikunata asirtachik killka Inka Castellanu	Inst. ling. de verano	2002

Table 4: Dictionaries used for the construction of the corpus.

N°	Grammatical category			
	Noun	**Verb**	**Adjective**	**Adverb**
1	9 190	4 682	2 924	474
2	1 186	578	382	32
3	374	211	112	11
4	134	88	26	8
5	37	32	9	2
6	10	3	4	1
Total	**10 931**	**5 594**	**3 457**	**528**

Table 5: Number of words per sense for each grammatical category of Southern Quechua wordnet.

N°	Grammatical category			
	Noun	**Verb**	**Adjective**	**Adverb**
1	1 702	1 974	603	210
2	362	292	136	30
3	95	67	33	15
4	19	14	8	3
5	4	2	1	3
Total	**2 182**	**2 349**	**781**	**261**

Table 6: Number of words per sense for each grammatical category of Central Quechua wordnet.

N°	Grammatical category			
	Noun	**Verb**	**Adjective**	**Adverb**
1	382	372	123	31
2	38	12	26	16
Total	**424**	**384**	**149**	**47**

Table 7: Number of words per sense for each grammatical category of Amazonian Quechua wordnet.

N°	Grammatical category			
	Noun	**Verb**	**Adjective**	**Adverb**
1	439	392	141	71
2	21	40	5	10
Total	**460**	**433**	**146**	**82**

Table 8: Number of words per sense for each grammatical category of Northern Quechua wordnet.