PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Desarrollo de recursos léxicos multi-dialécticos para el quechua

Tesis para optar el grado académico de Magíster en Informática con mención en Ciencias de la Computación que presenta:

Nelsi Belly Melgarejo Vergara

Asesor:

Héctor Erasmo Gómez Montoya

Lima, 2022

Informe de Similitud

Yo, Héctor Erasmo Gómez Montoya, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado Desarrollo de recursos léxicos multi-dialécticos para el quechua, de la autora Nelsi Belly Melgarejo Vergara dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 19 %. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 20/01/2023.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 20 de enero del 2023.

Apellidos y nombres del asesor / de la asesora: Gómez, Héctor Erasmo	
DNI: 70599170	Firma
ORCID: 0000-0002-1338-3392	

Resumen

Las lenguas de bajos recursos como el quechua no cuentan con recursos léxicos a pesar de ser importantes para contribuir en las investigaciones y en el desarrollo de muchas herramientas de Procesamiento de Lenguaje Natural (NLP) que se benefician o requieren de recursos de este tipo, de esa forma poder contribuir en la preservación de la lengua. El objetivo de esta investigación es construir una WordNet (base de datos léxica) para las variedades quechua sureño, central, amazónico y norteño, y un un etiquetado gramatical de secuencias de palabras (POS tagging) para la variedad del quechua sureño. Para el desarrollo de esta investigación se recopiló información de los diccionarios y se creó corpus paralelo quechua - español, se implementó un algoritmo de clasificación para alinear el sentido de las palabras con el synset del significado en español para cada variedad de la lengua quechua y finalmente se creó un modelo de etiquetación gramatical basado en el modelo BERT. El score obtenido para el POS tagging de la variedad quechua sureño fue 0.85 % y para el quechua central 0.8 %.

Palabras claves: WordNet, POS tagging, Quechua, Lenguaje de bajos recursos

Agradecimientos

A mis padres por ser el pilar fundamental y apoyarme incondicionalmente en todo momento. También me gustaría agradecer a mi asesor Erasmo Gómez por guiarme durante todo el proceso del proyecto, tambien a Rodolfo Zevallos por su constante apoyo y aporte en la investigación. Finalmente, a los lingüistas por su apoyo en la construcción y validación del corpus para la investigación.

Publicaciones

El presente trabajo de investigación para optar el grado académico de Magíster en Ingeniería Informática con mención en Ciencias de la Computación de la Pontificia Universidad Católica del Perú y como parte de este se ha realizado la siguiente publicación, la cual ha sido incluida como anexo:

- Nelsi Melgarejo, Rodolfo Zevallos, Hector Gomez, and John E. Ortega. 2022. **WordNet-QU: Development of a Lexical Database for Quechua Varieties.** In Proceedings of the 29th International Conference on Computational Linguistics, pages 4429–4433, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Índice general

Índice de figuras

Índice de tablas

Capítulo I

Planteamiento de la investigación

1.1. Problemática

Una cultura se transmite a través de la lengua, es así que la cultura y la lengua van de la mano, una la lengua no solo sirve para comunicarnos sino también para poder recepcionar y seguir transmitiendo información de nuestros antepasados a las generaciones siguientes [Fishman, 1991].

Una lengua está en peligro cuando se encuentra en vías de extinción, cuando sus hablantes dejan de utilizarla, cuando se usan en un número cada vez más reducido en ámbitos de comunicación y cuando dejan de transmitirse de una generación a la siguiente [Drude et al., 2003].

El Perú cuenta con 48 lenguas vigentes, entre las cuales se encuentra el Quechua. El Quechua es una entidad lingüística y sus variantes se hablan en siete países de América del Sur: Argentina, Bolivia, Brasil, Chile, Colombia, Ecuador y Perú, siendo el Perú el país con el mayor número de quechua hablantes. Según el censo realizado por el Instituto Nacional de Estadísticas e Informática en el año 2017 [INEI, 2017], 3 millones 261 mil 750 son Quechua hablantes.

El quechua tiene cuatros ramas: Quechua Sureño, Quechua Central, Quechua Norteño, Quechua Amazónico, y estas se agrupan en 12 variedades según el Ministerio de Educación

[MINEDU, 2013]. En el año 2018 el Ministerio de Educación presentó las variedades con su respectivo estado de vitalidad, donde cinco de las variedades se encuentran seriamente en peligro, dos en peligro y cinco vital [MINEDU, 2018].

El Quechua es una lengua netamente hablada razón por la cual se cuenta con pocos recursos digitalizados, además de la presencia de variedades lo cual hace que los pocos recursos digitalizados no se encuentren estandarizados gramaticalmente, lo que dificulta el desarrollo de herramientas que apoyen en la revitalización de la lengua.

La construcción de recursos lingüísticos digitales es un gran apoyo para las lenguas en peligro, ya que ayudan a preservar la información relevante y el conocimiento relacionado, no solo para el idioma en sí, sino también para la comunidad que lo habla. Sin embargo, si esos recursos no se desarrollan para poder realizar más análisis e investigaciones, pueden ser insuficientes para ayudar en los esfuerzos de preservación [Berment, 2002].

Fellbaum menciona que lingüística computacional es un área de investigación que tiene como objetivo comprender los fenómenos lingüísticos, de forma automática, a través del procesamiento y explotación de los corpus lingüísticos. Para lograr ese objetivo, el corpus debe estar en un formato legible por máquina y puede incluir información estructurada y metadatos lingüísticos que ayuden a comprender automáticamente los patrones del idioma. Entre los recursos léxicos y estructurados más importantes,se incluye la WordNet [Fellbaum, C, 1988].

El último punto a considerar es que no existen recursos léxicos, los cuales son importantes para contribuir en las investigaciones y en el desarrollo de muchas herramientas de Procesamiento de Lenguaje Natural (NLP) que se benefician o requieren de recursos de este tipo de esa forma poder contribuir en la preservación y expansión de la lengua.

1.2. Objetivo general

Desarrollo de recursos léxicos (WordNet y etiquetado de palabras) multi-dialécticos para Quechua

1.3. Objetivos específicos

1. Creación de corpus paralelo quechua - español
2. Implementar un algoritmo de clasificación para alinear el sentido de las palabras con el synset del significado en español.
3. Crear modelo de etiquetación gramatical basado en BERT para el quechua sureño.
4. Elaborar un prototipo de interfaz web accesible a la base de datos léxica en quechua.

1.4. Resultados esperados

1. Para objetivo específico 1:
    - Algoritmo para la creación del synset en español y quechua.
    - Corpus paralelo quechua - español
2. Para objetivo específico 2:
    - Algoritmo de clasificación del synset de la WordNet en español y su correspondiente relación en quechua
    - Recurso lingüístico con los synset en quechua y sus respectivas alineaciones
3. Para objetivo específico 3:

- Asignar categorías sintácticas a cada palabra del corpus.
- Algoritmo de asignación de etiquetas.

4. Para objetivo específico 4:

   - Diseño de un prototipo de interfaz web para consultar la WordNet.
   - Arquitectura del servicio web para realizar consultas en la WordNet del quechua.

1.5. Recursos, Herramientas, métodos y procedimientos

1.5.1. Recursos y herramientas

1. **Diccionarios bilingüe quechua - español**

   Se empleó diccionarios digitales de diferentes variedades emitidos por el Ministerio de Educación, gobiernos regionales y otros autores. Los diccionarios contienen palabras en quechua con su traducción al español, así cómo la categorías gramatical de cada palabra y en algunos de los diccionarios tienen definición de la palabra en quechua y ejemplos en quechua empleando la palabra con su respectiva traducción al español.

2. **WordNet en español**

   Se empleó la WordNet en español 3.0 del Repositorio Central Plurilingüe (MCR) como base para realizar la alineación. MCR es una base de conocimientos léxicos que integra WordNets de cinco diferentes idiomas: catalán, castellano, euskera, gallego e inglés bajo la misma infraestructura [Gonzalez-Agirre et al., 2012].

   La WordNet en español en su versión 3.0 cuenta con 38,702 synsets, 39,142 sustantivos, 10,824 verbos, 6,967 adjetivos y 1,051 advervios [Gonzalez-Agirre, 2013].

3. **Métricas de evaluación**

La validación de la eficiencia del algoritmo se realizó a través de similitud semántica y la validación por expertos.

- **Similitud semántica**

  Para la medir la similitud entre los synsets se empleará la el vocabulario de la WordNet y similitud de coseno para medir la distancias entre dos synsets. La idea de este método es evaluar la capacidad representativa de los embeddings a través de medir qué tanto pueden encapsular la relación que existe entre pares de palabras. Concretamente, este método busca determinar cuál es el grado de correlación entre la "similitud" de pares de palabras, respecto a la "similitud" de los pares de vectores correspondientes en el embedding.

  Para la realización de este método, se necesita de un conjunto de pares de palabras con los cuales realizar los experimentos, definir qué es la similitud entre pares de palabras y qué es la similitud entre vectores de un embedding.

  La similitud entre pares de vectores se define como una función de distancia. Esta distancia coseno, se define como el coseno del ángulo formado entre un par de vectores.

$$\cos \vec{u}, \vec{v} = \frac{\vec{u}.\vec{v}}{\|\vec{u}\| , \|\vec{v}\|} \tag{1.1}$$

  El resultado final del test se obtiene a partir de la correlación que existe entre los dos valores de similitud, de pares de palabra y pares de vectores. Los coeficientes adoptan valores entre -1 y +1, un valor de -1 indica que la relación es decreciente y cuando un valor sea cercano a +1 indica que existe una relación creciente entre la similitud percibida en las palabras y la similitud presente en el embedding. La

principal diferencia que existe entre coeficientes, es la sensibilidad que tienen para detectar ciertos tipos de relaciones.

- **Evaluación por expertos**

  En base a una muestra de synsets de la WordNet en quechua un grupo de lingüistas validaron las entradas de cada synset y la métrica para verificar la eficiencia del algoritmo fue la precisión, donde se calculó el porcentaje de predicciones positivas correctas. La Precisión nos brinda el porcentaje de las entradas de los synsets alineaciones correctamente.

$$\text{Precisión} = \frac{\textit{Número de entradas de los synsets correctamente alineados}}{\textit{Número total de entradas de los synsets seleccionadas para validación}} \tag{1.2}$$

4. **Python**

   Python es un lenguaje de programación ampliamente utilizado en el desarrollo de software, aplicaciones web, la ciencia de datos y el aprendizaje automático. Una potencia notable de python es que cuenta con una sintaxis muy clara. Se ejecuta en muchas variantes de Unix, incluidos Linux y macOS, y en Windows, por estas razones se hizo el uso de este lenguaje [Python,].

   Para el procesamiento y análisis de texto, se utilizarán librerías del paquete NLTK[1], el cual es un conjunto de herramientas de Python para el procesamiento de lenguaje natural (NLP, por sus siglas en inglés).

   Por otro lado, se utilizará Django como herramienta para desarrollo web, el cual, implementa el patrón de software: Modelo-Vista-Controlador (MVC), y a su vez facilita la

[1] https://www.nltk.org/

creación de sitios complejos debido a que está enfocado en la reusabilidad, conectividad, extensibilidad de componentes, desarrollo rápido y el principio de diseño de software DYR (Don't Repeat Yourself)

1.5.2. Métodos y procedimientos

Construcción del Algoritmo de Alineamiento

Se tomó la definición de cada sentido por cada palabra encontrada en los diccionarios del quechua para ubicar el synset del español que represente el mismo concepto. Una vez que todos los sentidos fueron alineados, los grupos resultantes de las palabras en quechua son los synsets, las variedades fueron alineados por separado.

Construcción del Algoritmo de similitud

Para la medir la similitud entre los synsets se empleó el vocabulario de la WordNet y similitud de coseno para medir la distancias entre dos synsets.

Validación de una muestra de la WordNet del quechua

Para la validación de las WordNet, se solicitó apoyo de un grupo de lingüistas quienes verificaron una muestra de synsets de la Wordnet quechua, donde verificaron las entradas de los synsets. Finalmente se evaluó considerando como métrica la precisión. La precisión nos brinda una medida de calidad de la WordNet de cada varieda del quechua generada.

Construcción del Algoritmo de asignación de etiquetas

Se seleccionó las oraciones encontradas en los diccionarios del quechua para después asignar la categoría gramatical a cada palabra de la oración del conjunto de datos. Se empleó el modelo para después crear el algoritmo de asignación.

Construcción del servicio web

La interfaz construida es similar a la WordNet de Princeton [Fellbaum, 1998]. El servicio mostrará los diferentes sentidos para la palabra, una definición de la palabra en quechua y ejemplos empleando la palabra.

1.6. Alcances

Este trabajo de investigación busca desarrollar una base de datos léxica más conocida como WordNet para las variedades de la lengua quechua, buscando facilitar tareas de computación lingüística.

Para el recurso lingüístico a construir en este proyecto se desarrolló un algoritmo que permita la clasificación y alineamiento entre synsets en español y quechua. Las relaciones semánticas fueron heredadas automáticamente de la WordNet en español al realizar el alineamiento.

Se implementó una interfaz web para realizar consulta de la WordNet en quechua, esta es similar a otras WordNets que se pueden acceder a través de internet.

La implementación de una herramienta que permita obtener el lema de una palabra en quechua en base a los diccionarios de términos y reglas de derivación propias del quechua.

1.7. Limitaciones

- Diccionarios escritos sin respetar la gramática estipulada por el Ministerio de Educación
- Palabras sin categoría gramatical

1.8. Justificación

Las bases de datos léxicas y el etiquetador gramatical son recursos eficaces para el procesamiento del lenguaje natural y la recuperación de información, especialmente para el procesamiento semántico y las tareas relacionadas con el significado. A la fecha se ha llegado a construir recursos de procesamiento de lenguaje natural para muchos idiomas, a pesar de ello, para las lenguas de bajos recursos como el quechua no han sido bien estudiados. La presente investigación busca crear una base de datos léxica de alta calidad y gran escala para las variedades del quechua y crear un modelo de etiquetación gramatical para contribuir en las investigaciones y creación de herramientas para la lengua quechua.

1.9. Organización del documento

El presente documento está dividido por capítulos. El primer capítulo contiene el planteamiento de la investigación. El segundo capítulo de la investigación contiene la base teórica. El tercer capítulo contiene la revisión del estado del arte y el cuarto capítulo presenta el desarrollo de la investigación. El quinto capítulo contiene los resultados, el sexto capítulo las conclusiones y finalmente el sétimo capítulo contiene los trabajos futuros.

Capítulo II

Marco conceptual

En la siguiente sección se muestran los conceptos relacionados directamente con el tema de la investigación para asegurar su entendimiento.

2.1. Procesamiento del Lenguaje Natural (NLP)

NLP es el área que maneja el procesamiento computacional de información, expresada en lenguaje natural, cuya finalidad es que las computadoras cuenten con la capacidad de comprender textos escritos por humanos y producirlos en una lengua familiar para que los humanos puedan entenderlos [Jurafsky and Martin, 2009].

Según Carbonell [Carbonell, 1994] los objetivos del Procesamiento del Lenguaje Natural se basan en tres tipos de aplicaciones que son Interfaces en lenguaje natural, Procesamiento de textos y Traducción automática.

- Interfaces en lenguaje natural: ¿No estaría bien dar las órdenes en el mismo lenguaje a todos los ordenadores, y tanto más aún si ese lenguaje fuera uno que los usuarios que ya conocieran bien, como su propio lenguaje natural nativo?

- Procesamiento de textos: las necesidades de los usuarios van más allá de la recuperación de información e incluyen la extracción de los datos más significativos, la elaboración de resúmenes, etc. Actualmente las investigaciones en el campo del Procesamiento del Lenguaje Natural buscan tratar este tipo de problemas.

- Traducción automática: diversos sistemas plurilingües eficaces de traducción automática ya están siendo explotados industrialmente y continuarán evolucionando de manera rápida en un futuro inmediato.

La Ambigüedad de sentidos y variedad de léxicos son las dificultades que destacan dentro del área de Procesamiento del Lenguaje Natural. El lenguaje natural es ambiguo, y la resolución de ambigüedades es muy necesaria para un procesamiento eficaz. Otra de las dificultades es la extensión y variedad de léxicos. Se debe tener registro de las miles de palabras y sus respectivos sentidos de la lengua. Además, existen palabras de la misma lengua que no se usan o tienen otro sentido según la comunidad o región.

2.2. Lengua de bajos recursos

Los idiomas que han recibido relativamente menos atención de la NLP suelen ser menos populares debido a la falta de recursos disponibles y, a menudo, se denominan idiomas de bajos recursos. Los idiomas que tienen una gran cantidad de recursos y herramientas de PNL generalmente lo hacen por razones sociales, políticas y financieras. Aunque concentrar el dinero y el esfuerzo en los idiomas más hablados tiene sentido desde un punto de vista económico, es difícil para los investigadores producir recursos significativos para los idiomas actuales de bajos recursos sin financiación. Se necesita realizar la investigación de métodos de NLP independientes del lenguaje que sean apropiados en entornos de bajos recursos, ya que estas técnicas se pueden aplicar a muchos idiomas de bajos recursos a la vez. A la urgencia necesaria para la

PNL en los idiomas de bajos recursos se suma la desaparición de los idiomas en peligro de extinción [King, 2015]

La World Wide Web comenzó como un fenómeno predominantemente con el inglés, a medida que ha ido madurando, ha comenzado a diversificarse y ahora contiene texto en miles de idiomas, aunque las presiones políticas, sociales y económicas incluso han provocado la desaparición de algunos idiomas de la Web [Streiter et al., 2006]. La Web podría servir para acelerar el proceso de desaparición de la lengua si los hablantes de lenguas minoritarias y en peligro de extinción descubren que tienen que usar un idioma diferente para comunicarse en la Web, o podría ayudar a preservar las lenguas si los hablantes sienten que pueden comunicarse en la Web en esos idiomas. La disponibilidad de herramientas de PNL, como la traducción automática (MT), en la Web podría ayudar a los hablantes de idiomas de bajos recursos a seguir usando ese idioma en lugar de abandonarlo en favor de un idioma mayoritario.

2.3. Lengua quechua

El quechua en la actualidad es la lengua originaria con mayor predominancia de Sudamérica [Torero, 2002].

Según clasificación lingüística de Torero, el quechua es una familia lingüística que presenta dos grandes grupos: Quechua I o Huáihuash y Quechua II o Huámpuy. El Quechua II, se divide en tres variedades: Quechua II A o quechua norteño (Perú), Quechua II B o quechua amazónico (Perú, Ecuador y Colombia), y Quechua II C o quechua sureño (Perú, Bolivia y Argentina) [Torero, 1964]. La imagen 4.2 podemos observar la clasificación mencionada.

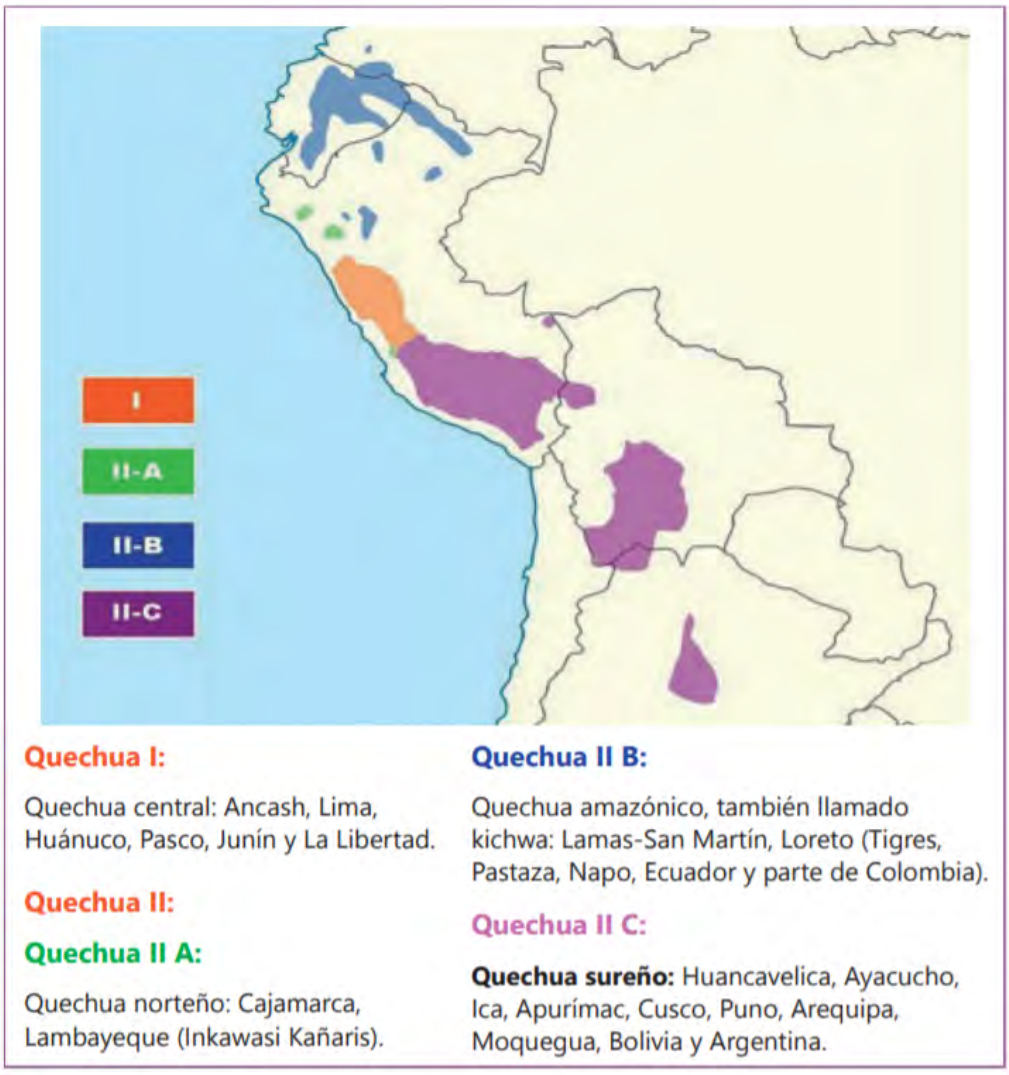


Figura 2.1: Familia lingüística quechua - imagen fue tomada del manual de escritura del quechua sureño del MINEDU

Respecto a la distribución actual, el quechua es una entidad lingüística y sus variantes son habladas en siete países de América del Sur: Argentina, Bolivia, Brasil, Colombia, Chile, Ecuador y Perú, siendo el Perú el país con el mayor número de Quechua hablantes. Según el censo realizado por el Instituto Nacional de Estadísticas e Informática en el año 2017 [INEI, 2017], 3 millones 805 mil 531 hablantes mayores de 5 años son Quechua hablantes.

La lengua Quechua presenta cuatros ramas: Quechua sureño, Quechua, Quechua central

norteño y Quechua amazónico, estas se agrupan en 12 variedades según el Ministerio de Educación(2013) [MINEDU, 2013]. El Ministerio de Educación en el año 2018 presentó las variedades con su respectivo estado de vitalidad, donde cinco de las variedades se encuentran seriamente en peligro, dos en peligro y cinco vital [MINEDU, 2018], como se puede observar en la tabla 2.1.

2.3.1. Características fonológicas

Los principales rasgos fonológicos que caracterizan son los siguientes:

- Se puede observar contraste entre vocales largas y breves, que ha permitido que el sistema trivocálico original (/a/, /i/, /u/) se desdoble en seis vocales (/a/, /i/, /u/,/a:/, /i:/, /u:/).
- Preserva la oposición de las africadas del protoquechua /č/ (palatal) y /ĉ/ (retrofleja). En relación con este rasgo fonológico, comúnmente, se ha usado como un criterio que ayudaría diferenciar entre quechua central y quechua sureño-norteño. No obstante, este criterio no goza de mucha validez, debido a que algunas variedades del sureño-norteño o quechua II, a saber, la variedad de Cajamarca, la variedad de Ferreñafe, el quechua de Yauyos y el quechua de Chachapoyas, preservan la diferencia entre las consonantes africadas /č/ (palatal) y /ĉ/ (retrofleja); así, por ejemplo, čaki 'seco' y ĉaki 'pie' se diferencian, porque se mantiene la distinción original del protoquechua. De igual modo, tampoco en toda la rama central se hace esta distinción, en virtud de que se han transformado dichas africadas en algunas variedades, por ejemplo: čaki 'seco' y čaki 'pie' en el quechua de Alto Huallaga); tsaki 'seco' y čaki 'pie' en el quechua huaylino de Ancash, etc. [MINEDU, 2020]

Tabla 2.1: Estado de vitalidad de las variedades del quechua - MINEDU 2018

Rama	***Variedad***	***Región***	***Estado vital***
Quechua sureño	Quechua Chanka	Huancavelica, Ayacucho y Apurímac (Andahuaylas, Aymaraes y Chincheros)	Vital
	Quechua Collao	Apurímac (Abancay, Grau, Antabamba y Cotabambas), Cusco, Puno, Arequipa y Moquegua	Vital
Quechua central	Quechua Pataz	La Libertad	Vital
	Quechua Cajatambo, Oyón, Huaura	Lima	Seriamente en peligro
	Quechua Yauyas	Lima	Seriamente en peligro
	Quechua Áncash	Áncash	Vital
	Quechua Huánuco	Huánuco	En peligro
	Quechua Pasco	Pasco	Seriamente en peligro
	Quechua Wanka	Junín	Seriamente en peligro
Quechua norteño	Quechua Cajamarca	Cajamarca	Seriamente en peligro
	Quechua Inkawasi Kañaris	Lambayeque y Piura (Comunidad de Chilcapampa, distrito de Huarmaca, provincia de Huancabamba; centro poblado La Pilca, distrito de Buenos Aires, provincia de Morropón)	Vital
Quechua amazónico	Kichwa amazónico: Pastaza, Napo, Putumayo, Tigre, Alto Napo (Santarrosino-Madre de Dios) y Chachapoyas y San Martín	Loreto, Madre de Dios, Chachapoyas y San Martín	En peligro

2.3.2. Características morfológicas

En el libro de manual de escritura del quechua sureño y central emitida el año 2021 por el Ministerio de Educación [MINEDU, 2020] [MINEDU, 2021] menciona las siguientes características:

- En el quechua central la primera persona con función de sujeto (con verbos) y de poseedor (con sustantivos) se indica por medio de alargamiento vocálico (aa, ii, uu), por ejemplo mama-a 'mi madre'. El sufijo -ni en el quechua sureño-norteño se emplea para primera persona sujeto en el presente indicativo y el sufijo -y en el resto del paradigma verbal y como marca nominal de la primera persona poseedora.

- La primera persona con función de objeto se indica por medio del sufijo verbal -ma(a)-, observamos este ejemplo, muna-ma-ki que quiere decir 'tú me quieres'. En el caso del quechua sureño-norteño, el sufijo equivalente que marca prototípicamente la primera persona objeto es -wa-.

- Los sufijos de caso locativo, ablativo y comparativo no coinciden en quechua sureño-norteño y central. El sufijo de caso locativo es -chaw por ejemplo Llatachaw 'en Llata'; el ablativo se expresa con -piq, -piqta o -pita por ejemplo qanyan-piq, qanyan-piqta, qanyan-pita 'desde ayer'; el comparativo es -naw ejemplo qam-naw 'como tú'. El caso locativo en el quechua sureño-norteño se marca con -pi, el ablativo (con excepción de Laraos) se señala con -manta y el sufijo de caso comparativo se marca por lo generalmente con -hina.

2.3.3. Características léxicas

En el quechua a nivel lexical existen algunas diferencias, tales como se muestra en la tabla 2.2.

Tabla 2.2: Diferencia a nivel léxical

Léxico quechua sureño	***Léxico quechua central***	***Léxico quechua norteño***	***Glosa***
quyi	haka	quwi	cuy
punchaw	hunaq	punchaw/ p'unchaw	día
ri-	aywa-	ri-	ir
qishya-	qishya-	unqu-	enfermarse

2.4. WordNet

WordNet es una base de conocimiento léxica para el idioma inglés. Esta inspirada por teorías psico-linguisticas y computacionales sobre la memoria léxica humana. Contiene información codificada manualmente sobre sustantivos, adjetivos, verbos, y adverbios del inglés, y esta organizada en relación a la noción de synset. Un synset es un conjunto de palabras de la misma categoría morfosintáctica que pueden ser intercambiados en un contexto dado. Es una gran base de datos léxica del inglés [Miller, 1998]. Sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos, cada uno expresando un concepto distinto. Los synsets están vinculados entre sí por medio de las relaciones conceptuales, semánticas y léxicos.

La red resultante de las palabras y los conceptos relacionados de manera significativa se puede navegar. WordNet se encuentra públicamente disponible para descarga de forma libre. La estructura de la WordNet hace que sea una herramienta útil para la lingüística computacional y procesamiento del lenguaje natural.

De manera superficial la WordNet se parece a un diccionario de sinónimos, a pesar de ello, hay algunas diferencias, la WordNet divide el lexicón en cinco categorías: sustantivos, verbos, adjetivos, adverbios y elementos funcionales. Por otra parte, este tipo de organización, facilita enormemente el análisis de las diferencias de organización semántica que existen entre esas categorías sintácticas, cabe mencionar que es importante destacar que, al no tener que forzar

las diferentes categorías en un mismo esquema representacional, así mismo, se puede buscar la forma más adecuada para cada una de ellas por separado.

2.5. Etiquetado gramatical (POS tagging)

El etiquetado gramatical es un proceso de los más usados en el Procesamiento del Lenguaje Natural (NLP). El etiquetado POS intenta etiquetar cada palabra con su parte correcta del discurso (también llamada categoría de palabra, clase de palabra o categoría léxica) [Güngör, 2010]. El etiquetado POS se trata de asignar una etiqueta POS a la palabra en un texto (corpus) correspondiente a una parte particular de un discurso, según la definición de la palabra y su contexto. Los etiquetadores gramaticales describen la estructura característica de los términos léxicos dentro de una oración o texto, por lo tanto, pueden ser usados para hacer suposiciones sobre semántica. Algunas aplicaciones del etiquetado gramatical son: Reconocimiento de Entidad Nombrada (ENR: Entity Named Recognition), Resolución de Co-Referencia (Co-reference Resolution) y Reconocimiento de Voz (Speech Recognition).

Cuando se realiza el etiquetado gramatical, a menudo ocurre que el etiquetador encontrará palabras que no estaban dentro del vocabulario que se utilizó, por ello, en consecuencia, aumentar el conjunto de datos para incluir tokens de palabras desconocidas ayudará al etiquetador a seleccionar las etiquetas adecuadas para esas palabras.

Dentro de un texto, para cada palabra es posible identificar su propio término léxico, sin embargo, tener que identificar constantemente estos términos completos cuando se realiza un análisis de texto puede convertirse en un trabajo laborioso, especialmente a medida que aumenta el tamaño del corpus. De ahí, que se utiliza una representación corta denominada etiquetas para representar las categorías de las palabras.

Las etiquetas gramaticales son útiles para crear árboles de análisis, que se utilizan para crear ENR (la mayoría de las entidades nombradas son sustantivos) y extraer relaciones entre

palabras. El etiquetado gramatical también es esencial para crear lematizadores que se utilizan para reducir una palabra a su forma raíz.

Existen diferentes técnicas para el etiquetado gramatical [Ramachandran,]:

- Métodos basados en léxicos: asigna la etiqueta que aparece con más frecuencia con una palabra en el corpus de entrenamiento.
- Métodos basados en reglas: asigna etiquetas según las reglas. Por ejemplo, se puede tener una regla que diga que las palabras que terminan con .ªdo.º .ªndo"deben asignarse a un verbo. Las técnicas basadas en reglas se pueden usar junto con los enfoques basados en léxico para permitir el etiquetado de palabras que no están presentes en el corpus de entrenamiento, pero sí en los datos de prueba.
- Métodos probabilísticos: este método asigna las etiquetas en función de la probabilidad de que ocurra una secuencia de etiquetas en particular. Los Campos Aleatorios Condicionales (CRF: Conditional Random Fields) y los Modelos de Markov Ocultos (HMM: Hidden Markov Models) son enfoques probabilísticos para asignar una etiqueta POS.
- Métodos de aprendizaje profundo: las redes neuronales recurrentes se pueden utilizar para el etiquetado gramaticales.

2.6. Transformers

El Transformer es el primer modelo de transducción que se basada únicamente en mecanismos de atención, prescindiendo por completo de recurrencias y convoluciones[Vaswani et al., 2017]. La idea clave del Transformer es gestionar completamente las dependencias entre la entrada y la salida con atención y recurrencia. Transformers proporciona miles de modelos previamente

entrenados para realizar diferentes modalidades de tareas, como texto, visión y audio. Estos modelos se pueden aplicar en:

- Texto, para tareas como: clasificación de texto, extracción de información, resumen, respuesta a preguntas, traducción, generación de texto y más, compatible con más de 100 idiomas.
- Imágenes, para tareas como clasificación de imágenes, detección de objetos y segmentación.
- Audio, para tareas como reconocimiento de voz y clasificación de audio.

Transformers actualmente proporciona las muchas arquitecturas.

2.6.1. BERT

BERT es el acrónimo para Bidirectional Encoder Representations from Transformers (Representaciones de Codificador Bidireccional de Transformadores). Se trata de una técnica basada en Redes Neuronales Artificiales (RNA) aplicado al campo del NLP (Natural Language Processing), específicamente al subcampo del Natural Language Understanding (NLU).

BERT está diseñado para entrenar previamente representaciones bidireccionales profundas a partir de texto sin etiquetar al condicionar conjuntamente el contexto izquierdo y derecho en todas las capas. El modelo preentrenado BERT se puede ajustar con solo una capa de salida adicional para crear modelos de última generación para una amplia gama de tareas, como la respuesta a preguntas y la inferencia de lenguaje, sin modificaciones sustanciales en la arquitectura específica de la tarea [Devlin et al., 2018].

2.6.2. RoBERTa

RoBERTa es un modelo robustamente optimizado para preentrenamiento de sistemas de procesamiento de lenguaje natural (NLP) que mejora las representaciones de BERT.

RoBERTa se basa en la estrategia de enmascaramiento de idiomas de BERT, en la que el sistema aprende a predecir secciones de texto ocultas intencionalmente dentro de ejemplos de idiomas sin anotaciones. RoBERTa, que se implementó en PyTorch, modifica los hiperparámetros clave en BERT, incluida la eliminación del objetivo de entrenamiento previo de la siguiente oración de BERT y el entrenamiento con mini lotes y tasas de aprendizaje mucho más grandes. Esto permite que RoBERTa mejore el objetivo de modelado del lenguaje enmascarado en comparación con BERT y conduce a un mejor desempeño de las tareas posteriores. [Liu et al., 2019].

2.6.3. QuBERT

QuBERT (A Large Monolingual Corpus and BERT Model for Southern Quechua) es un modelo basado en RoBERTa desarrollado por [Zevallos et al., 2022]. QuBERT es el modelo entrenado con mayor volumen de datos existente a la fecha del quechua. El conjunto de datos consiste de 4,408,953 tokens y 384,184 sentencias del quechua sureño que incluye las variedades del quechua "Chanka" y "Collao". El conjunto de datos fue limpiada a fondo por estudiantes universitarios y hablantes nativos de quechua, emplearon el analizador y normalizador morfológico quechua desarrollado por [Rios, 2015] para convertirlo en quechua sureño estándar.

2.6.4. FastText

FastText[1] es una biblioteca liviana, gratuita y de código abierto que permite a los usuarios aprender representaciones de palabras y clasificadores de oraciones. Funciona en un hardware

[1]https://fasttext.cc/

genérico estándar.

Facebook AI Research [Bojanowski et al., 2017] realizaron investigaciones sobre un método simple para aprender las representaciones de palabras teniendo en cuenta la información de las subpalabras donde demostraron que debido a su simplicidad del modelo, entrena rápido y no requiere ningún procesamiento previo o supervisión.

Entre sus características destacan su velocidad de entrenamiento y la resistencia a palabras desconocidas. fastText, a diferencia de word2vec, genera buenos embeddings incluso para palabras que previamente no había visto durante el entrenamiento debido a que fastText no supone que la unidad más pequeña son las palabras, por el contrario, trata las palabras como la combinación de sus caracteres individuales, lo que le permite interpretar palabras que no reconoció durante el entrenamiento pero se parecen a otras que sí reconoce.

FastText provee vectores de palabras pre entrenado para 157 idiomas entrenados usados datos recopilados de Common Crawl [2] y Wikipedia [3] usando fastText [Grave et al., 2018]. Entre la lista de los modelos pre entrenados se encuentra el quechua.

[2]https://commoncrawl.org/
[3]https://www.wikipedia.org/

Capítulo III

Estado del arte

3.1. Introducción

Este capítulo presenta el estado del arte en el de la revisión bibliográfica de los métodos y técnicas aplicadas. Durante la revisión bibliográfica se analizaron artículos y libros obtenidos de bases de datos como SCOPUS, ACL Anthology, Scholar Google, entre otros. Se buscaron artículos relacionados con el tema de la investigación. Para ello, se buscó empleando términos específicos para que los resultados de la búsqueda sean más detallados y centrados en el tema de estudio. Además, se incluyeron criterios de exclusión e inclusión con el fin de reducir el número de referencias.

3.2. WordNet

Construction of an English-Uyghur WordNet Dataset. [Abiderexiti and Sun, 2019]

La investigación tuvo como objetivo crear un conjunto de datos de evaluación disponibles públicamente para la creación automática de recursos semánticos como WordNet para la lengua Uigur que es una lengua de escasos recursos.

Para la construcción de este recurso inicialmente crearon diccionarios uigur-inglés e inglés-uigur rastreando diccionarios en linea y diccionarios físicos. Emplearon CUDD (Diccionario detallado ouigur contemporáneo) que utiliza principalmente la forma de diccionario de una palabra (lema) como objeto de descripción y explica cada sentido de la palabra donde cada sentido tiene una explicación y la mayoría de los sentidos tienen oraciones de ejemplo y Princeton WordNet (PWN) 3.0 para la construcción de datos.

En este conjunto de datos obtenido contiene más de 73,000 synsets en inglés se mapean en uigur automáticamente, en los que más de 20,000 se anotaron manualmente para obtener 6,642 synsets en inglés.

Automatic wordnet development for low-resource languages using cross-lingual WSD [Taghizadeh and Faili, 2016]

El objetivo de la investigación fue crear una WordNet de alta calidad y de gran escala para lenguas de escasos recursos como el persa empleando un algoritmo EM (Expectation–maximization), los recursos empleados para este algoritmo propuesto incluyen un diccionario bilingüe y un corpus monolingüe. La investigación propuso un método automático para extraer synsets para idiomas de bajos recursos. El método propuesto pertenece al enfoque de expansión y, por lo tanto, crea una WordNet multilingüe para cada palabra en el idioma de destino, se conoce el synset equivalente en WordNet, todas las traducciones de palabras en inglés dentro se extraen del diccionario bilingüe y se establecen vínculos entre las palabras de traducción y los synsets de la WordNet. Dado que los diccionarios traducen palabra a palabra, no de un sentido a otro, las traducciones son ambiguas. Por tanto, la tarea consiste en puntuar enlaces y encontrar los incorrectos. El método usado no emplea ninguna característica específica del idioma de destino, por esta razón se puede emplear en otros idiomas para generar redes de palabras.

El método propuesto se aplicó al idioma persa y la calidad del WordNet resultante se examinó a través de varios experimentos. Su precisión fue del 18 % según FarsNet y del 90 % según el juicio manual. La red de palabras resultante contiene alrededor de 12.000 palabras del idioma persa de solo usando el 13 % del corpus de Bijankhan,

Finalmente, se observó que cuando el tamaño del corpus aumentó, toda las medidas aumentaron excepto la precisión, que no cambió o solo cambió ligeramente. El resultado no superó las expectativas. La precisión de la red de palabras resultante depende de la precisión del procedimiento WSD (Word Sense Disambiguation) y, por lo tanto, no depende del tamaño del corpus. Sin embargo, se descubrieron nuevos sentidos posibles de las palabras aumentando el tamaño del corpus y, por lo tanto la precisión, la cobertura y el tamaño de la WordNet aumentan con el crecimiento del tamaño del corpus.

A major wordnet for a minority language: Scottish gaelic. [Bella et al., 2020]

El objetivo de la presente investigación fue proporcionar un recurso léxico-semántico que sea de un tamaño utilizable para cubrir una parte considerable del vocabulario común (incluso si está lejos de ser exhaustivo en su etapa inicial), que esté alineado con el sentido no solo con Inglés pero también con otros idiomas, que es de alta calidad debido a la supervisión humana, y finalmente que es explotable tanto computacionalmente como por humanos.

Para la construcción de la WordNet adoptaron el enfoque de expansión de fuentes expertas, es decir, un subconjunto de palabras, glosas y ejemplos del inglés Princeton WordNet fueron traducidos y validados por expertos en gaélico. Los pasos de la metodología empleada fueron los siguientes:

1. Generación de tareas de traducción: primero especificaron qué subconjunto de Princeton WordNet (PWN) para traducir.

2. Traducción: el esfuerzo de traducción real realizado por un experto en idioma gaélico.

3. Unir: fusionar los resultados de la traducción con las palabras proporcionadas por la wordnet gaélica existente basada en Wiktionary.

4. Validación: un subconjunto de los términos traducidos fue evaluado y corregido por otro experto en idiomas.

WordNet-SHP: Towards the building of a lexical database for a Peruvian minority language. [Maguino-Valencia et al., 2018]

La investigación se enfocó en la construcción de una base de datos WordNet inicial para la lengua Shipibo-Konib(shp), la cual es una lengua indígena de bajos recursos en el Perú. Para la construcción de la WordNet la primera tarea consistió en la digitalización y preprocesamiento de un diccionario bilingüe shipibo - español. La segunda tarea consistió en la tarea de alineación de synset mediante el uso de una métrica de similitud con las glosas de definición en el diccionario y el WordNet en español del Repositorio Central Multilingüe (MCR). Finalmente, realizaron un proceso de evaluación para los synsets, utilizando un Gold Standard anotado manualmente en Shipibo-Konibo. Los resultados obtenidos de la alineación mostraron una estrecha similitud en la distribución del sentido de las palabras entre shipibo-konibo y el español.

3.3. POS tagging

Ship-LemmaTagger: Building an NLP Toolkit for a Peruvian Native Language

[Pereira-Noriega et al., 2017]

Para la investigación con la ayuda de una herramienta llamada ChAnot, desarrollaron un corpus de 219 oraciones anotadas, donde cada palabra de la oración contiene: una anotación del lema, POS-tag, sub-POS-tag y una lista de todos los afijos que aparece en la palabra. Este cor-

pus emplearon para el entrenamiento del etiquetador POS. Para el POS tagging Utilizaron dos métodos: un modelo SVM y un modelo de árbol de decisión, donde para el modelo SVM obtuvieron 0,848 y para el "decision tree"0.811.

Capítulo IV

Desarrollo

En el presente capítulo se detalla la solución de los objetivos específicos planteados para la construcción de los recursos léxicos.

4.1. Corpus

Uno de los requisitos más importantes para el desarrollo de recursos léxicos como la WordNet y el POS tagging es sin duda el conjunto de datos debidamente etiquetados según la categoría gramatical, en esta sección se detalla el proceso que se siguió para obtener los datos, así cómo la cantidad de la data empleada para la construcción de cada recurso léxico.

4.1.1. Recopilación de datos

La recopilación de los diccionarios fueron a través de diferentes fuentes en linea por variedades, en su mayoría de acceso público. Durante el proceso de recolección se dio prioridad a los diccionarios emitidos por el Ministerio de Educación, las cuales se encuentran escritas según la gramática emitida por el Ministerio de Educación y Ministerio de Cultura. En la tabla 4.1 se lista los diccionarios recopilados.

Los diccionario recopilados se encontraban en diferentes formatos digitales como PDF, Json, Word y Excel. Para extraer la información de los diccionarios en formato PDF se empleó la libreria PyPDF2 de Python para manejar este tipo de archivos lo cual nos permitió extraer la información y poder manipular los datos.

Tabla 4.1: Diccionarios empleado para la recopilación del corpus

Rama	***Variedad***	***Diccionario***	***Diccionario***
Quechua sureño	Quechua Chanka	-Yachakuqkunapa Simi Qullqa -chanka qichwa simi	Ministerio de Educación
	Quechua Collao	Niraq Masi Rimaykuna Taqi -Diccionario quechua: Cuzco – Collao.	Ministerio de Educación
Quechua central	Quechua Pataz/ Quechua Cajatambo, Oyón, Huaura /Quechua Yauyas/- Quechua Áncash/ Quechua Huánuco/ Quechua Pasco/ Quechua Wanka	- Yachachinapaq shimikuna chawpin qichwa - Chawpi qichwapa shimi qullqan - Shimikunata asirtachik killka inka-kastellanu	-Ministerio de Educación -Instituto lingüístico de verano
Quechua norteño	Quechua Cajamarca		Ministerio de Educación
	Quechua Inkawasi Kañaris	Diccionario quechua: Cajamarca – Cañaris	Ministerio de Educación
Quechua amazónico	Kichwa amazónico: Pastaza, Napo, Putumayo, Tigre, Alto Napo (Santarrosino-Madre de Dios) y Chachapoyas y San Martín	Diccionario visual - quechua amazónico DIGITAL	

Alguno de los diccionarios explican detalladamente las reglas generales de la estructura para cualquier palabra, cabe mencionar que no toda las excepciones son mencionadas. A continuación se presenta la estructura general de algunos diccionarios de las ramas quechua sureño y

quechua central:

Entrada léxica: es el neologismo forma básica de la palabra.

Categoría gramatical: informa si la entrada es un sustantivo, verbo, adjetivo, adverbio u otra categoría.

Definición en quechua: menciona significado de la palabra.

Glosa: se llama glosa a la forma equivalente en la lengua castellana.

Ejemplo: es el uso del neologismo en una oración o en oraciones, de acuerdo al contexto, con su traducción o traducciones al castellano.

Sinónimo: en algunos casos se menciona los sinónimos de la palabra tanto en español como en quechua.

Los elementos que fueron mencionados anteriormente se ilustran en la figura 4.1. Es importante mencionar que no todo los diccionarios tienen los elementos mencionados anteriormente y algunos diccionarios son de tipo vocabulario (son aquellos q tienen solo la pabra, la categoria gramaticas y el significados).

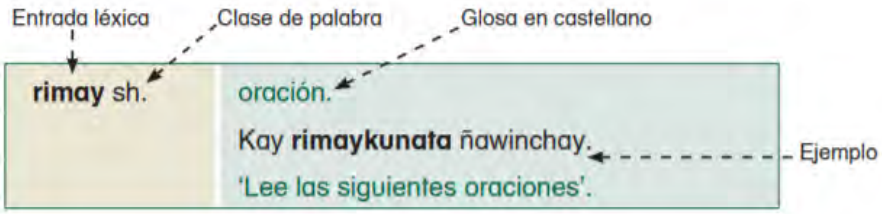


Figura 4.1: Ejemplo de la estructura - imagen fue tomada del diccionario del central del MINEDU

Se integraron los diccionarios en formato JSON donde adicionó 5 columnas a continuación se detalla las columnas:

Rama: quechua amazónico, quechua central, quechua, norteño o quechua sureño

Variedad: menciona la variedad del quechua al que pertenece.

Región: (opcional) región de referencia del diccionario

Autor: (opcional) autor del diccionario.

Año: (opcional) fecha de publicación del diccionario.

En la tabla 4.2 se lista la cantidad entradas léxicas por variedad de la lengua y cantidad porcentual que representan de la data recopilada previo al preprocesamiento.

Tabla 4.2: Cantidad de entradas léxicas por rama de la lengua quechua

Rama	***Entrada léxica***	***Porcentaje (%)***
Quechua amazónico	1812	1,82
Quechua central	15947	15,98
Quechua norteño	3512	3,52
Quechua sureño	78534	78,69
Total	99805	100

Se realizó el conteo de las entradas léxicas por categoría gramatical, sólo considerando sustantivo, adjetivo, adverbio y verbo, así como se observa en la la tabla 4.3, donde el sustantivo junto con el verbo son los que tienen mayor cantidad de entradas.

Tabla 4.3: Cantidad de palabras por categoría gramatical

	Variedad			
Pos	***Quechua sureño***	***Quechua central***	***Quechua norteño***	***Quechua amazónico***
Sustantivo	35282	5578	1119	700
Verbo	16050	4365	905	431
Adjetivo	8360	1464	314	181
Adverbio	2229	646	207	56
Total	**65234**	**12965**	**2777**	**1454**

4.1.2. Preprocesamiento del corpus

Después de la recolección del conjuntos de datos se realizaron las siguientes acciones:

1. Se separaron los sinónimos de la columna glosa y entrada léxica a la columna sinónimos español y sinónimos quechua respectivamente.

2. Se corrigió los errores de caracteres tales como: signos de puntuación o letras que contienen caracteres especiales. Se empleó Python para realizar limpieza de los caracteres especiales masivamente, también se realizó de manera manual.

3. Se eliminaron las entradas duplicadas por variedad.

4. Se corrigieron y/o eliminaron las entradas léxicas que no se encontraban escritas según la gramática del Ministerio de Educación y Ministerio de Cultura.

5. Se revisó manualmente los casos que contenían más de una categoría gramatical.

6. Finalmente se filtró la data de las cuatro categorías gramaticales: Sustantivo, adjetivo, adverbio y verbo.

Después de realizar las acciones mencionadas anteriormente se obtuvo un total de 40 067 entradas léxicas únicas por las cuatro variedades de la lengua, donde el sustantivo y el verbo son las categorías gramaticales con mayor entrada léxica, así mismo, la variedad quechua sureño es quien cuenta con mayor cantidad de entradas seguida por el quechua central como se observa en la tabla 4.4.

4.1.3. Conjunto de datos recopilados para la WordNet

Después del preprocesamiento del corpus, la estructura usada de la salida es la siguiente: Rama; Variedad; Región; Autor; Diccionario; Año; Entrada léxica; Categoría gramatical; glosa;

Definición en quechua; Ejemplo en quechua; Traducción del ejemplo en quechua; Sinónimo quechua; Sinónimo español; Notas (aclaraciones) A continuación se muestra un ejemplo:

Rama	Variedad	Región	Autor	Nombre del Diccionario	Año	Entrada léxica	Categoría gramatical	Definición en quechua	Glosa	Ejemplo - quechua	Ejemplo - español	Sinónimo quechua	Sinónimo español	Notas
Quechua central	Ancash	Ancash	Ministerio de Educación	Chawpi qichwapa shimi qullqan	2017	yarpay	s	Umanchikchaw, piqanchikchaw yarpaachimaqninchik	memoria	Llapan yachakushqanchikmi yuyayninchikman aywan.	Todo lo que aprendemos va a la memoria.			

Figura 4.2: Estructura del corpus recopilado - Elaboración propia

En la tabla 4.4 se observa la cantidad de entradas léxicas por variedad de la lengua y según la categoría gramatical. Solo algunas entradas léxicas de de las las variedades quechua sureño y central son los que cuentan con definición en quechua, ejemplo empleando la entrada léxica en una frase u oración con su respectiva traducción.

Tabla 4.4: Entrada léxica por variedad y categoría gramatical

	Variedad			
POS	***Quechua sureño***	***Quechua central***	***Quechua norteño***	***Quechua amazónico***
Sustantivo	16717	3241	579	537
Verbo	8000	3145	506	423
Adjetivo	4116	904	157	160
Adverbio	985	384	126	48
Total	**29818**	**7674**	**1368**	**1168**

Los información recopilada para el quechua amazónico y quechua norteño fueron vocabularios donde la entrada léxica tiene su respectiva glosa en español y su categoría gramatical.

4.1.4. Conjunto de datos para el Pos Tagging

Debido a que no se cuenta con un corpus etiquetado para el quechua, se desarrollo un corpus POS se creó con el apoyo de lingüistas, quienes son expertos en las variedades del quechua

chanka, collao y central. Fue posible desarrollar un corpus de 5 367 oraciones etiquetadas, donde cada palabra de la oración contiene la categoría gramatical . Además, con la ayuda de los diccionarios recopilados para la construcción de la WordNet cuyas entradas léxicas incluían informacion de etiqueta POS se logró etiquetar el 38 % las cuales solo fueron validadas por los lingüistas.

El marco empleado para la anotación fue Dependencias Universales (UD). El número de etiquetas POS principales utilizadas en el corpus fueron 11 y con subcategorías 13. La Tabla 4.5 presenta la descripción detallada del conjunto de etiquetas y la abreviatura empleada.

Tabla 4.5: Descripción detallada del conjunto de etiquetas

Etiqueta principal	***Sub categorías***	***Representación***
Sustantivo	–	NN
Verbo	–	VB
Adverbio	–	ADV
Adjetivo	–	ADJ
Determinante	–	DET
Interjección	–	INJ
Pronombre	Indefinido	INDP
	Demostrativo	DP
	Interrogativo	INTP
Conjunción	–	CONJ
Preposición	–	PREP
Locución	–	LOC
Numeral	–	NUM

La tabla 4.6 presenta la frecuencia de cada etiqueta POS en el corpus. La etiqueta que se encontró con más frecuencia fue una etiqueta de sustantivo (NN), seguida por verbo (VB), adjetivo (ADJ) y adverbio (ADV). Finalmente se obtuvo un total de 31 356 palabras y 13 650 palabras de entrada únicas (con su lema correspondiente y etiqueta POS) distribuidas por categoría de palabra como se muestra en la Tabla 4.6.

Tabla 4.6: Conjunto de etiquetas POS por variedades del quechua.

Etiqueta principal	***Quechua central***	***Quechua sureño***	***Total***
Sustantivo (NN)	4 406	10 867	15 273
Verbo (VB)	2282	5658	7 944
Adverbio (ADV)	478	2 141	2 619
Adjetivo (ADJ)	936	3324	4 260
Determinante (DET)	4	418	422
Interjección (INJ)	14	26	40
Pronombre (PRON)	115	459	574
Conjunción (CONJ)	–	116	116
Preposición (PREP)	–	2	2
Locución (LOC)	3	13	16
Numeral (NUM)	2	36	38
Total	**8240**	**23060**	**31300**

4.2. WordNet

En esta sección se muestra la implementación del algoritmo para la construcción de la WordNet.

La construcción de la WordNet se realizó por cada variedad de la lengua (quechua sureño, central, amazónico y norteño), esto debido a que cada variedad tiene sus características fonológicas, morfológicas y léxicas diferente.

4.2.1. Implementación del algoritmo de clasificación para alinear el sentido de las palabras con el synset del significado en español.

La siguiente sección aborda el proceso de implementación del algoritmo de clasificación que permita construir la estructura interna de la WordNet quechua.

Seguidamente se explicará cómo se desarrolló el objetivo específico dos: implementar un algoritmo de clasificación que permita alinear cada sentido de cada palabra del quechua con el synset que tiene igual significado en el español.

(Vossen, 1998) sostiene que hay dos formas principales, fundamentalmente diferentes de crear nuevas redes de palabras, a través de lo que él llama la expansión y el unir enfoques. El primero toma como base un wordnet existente, generalmente el Princeton WordNet en inglés, ya que es el más completo, y proporcionando traducciones para un subconjunto cuidadosamente seleccionado de synsets, basados tanto en las palabras fuente como en la glosa.

En el caso de la expansión (traducción), el nuevo WordNet estará completamente alineado con el significado del idioma de origen, que es ideal para usos multilingües: como la mayoría de los WordNets ya están alineados con PWN, obtenemos traducciones bilingües a todos esos idiomas 'gratis'.

Para la construcción de nuestro WordNet se usó el mismo principio de la expansión así como la construcción de la WorNet del español que se construyó empleando la WordNet del Inglés, considerando que tenemos una gran cantidad de glosas en español.

Se empleó la WordNet en español del Multilingual Central Repository (MCR) como base para realizar la clasificación, como se menciona en la parte de recurso y herramientas, este recursos se encuentra disponible para ser descargada de forma libre. Para la alineación se empleó cuatro librerias Nltk[1], Pandas[2], Json[3] y Numpy[4].

[1] https://www.nltk.org/
[2] https://pandas.pydata.org/
[3] https://www.json.org/
[4] https://numpy.org/

```
Para cada término en Quechua
    indice_for = 0;
        Si ( Rama del termino es "Rama del quecha" ) entonces
            Si (El numero de significados el termino es 1 ) entonces
                quechua_work = Palabra en Quechua del Termino
                spanish_work = Significado en español del termino

                result = hallar_synset(quechua_work,hallar_work(Categoría gramatical del Termino))

                synset_word = Unir_texto(quechua_work, hallar_work(Categoría gramatical del Termino), result))

            lista_synsets = []
            Si (Existe synsets encontrados con el spanish_work ) entonces
                Para synset encontrado de la lista de synsets
                    syn = obtener solo el synset sin la estructura
                    agregar_lista_synsets(syn)
                Fin

            Si (Existe sinonimo encontrados con el spanish_work ) entonces
                Para cada synomy de la lista de sinonimo
                    Si (synom == spanish_work) entonces
                        indice_for = synomy.ObtenerIndice(spanish_work)
                        agregar_lista_synsets(synomy[indice_for])
                Fin

            agregar_lista_synomys(spanish_work)
            agregar_lista_synsets(synset_word)

            p = texto formato Json

            agregar_lista_quechua(p)
Fin
```

Figura 4.3: Pseudocódigo del algoritmo

El algoritmo funciona de la siguiente manera: cada definición en español de la palabra quechua encontrada en el corpus del quechua fue comparada con cada synset de la WordNet en español, una vez encontrado su correspondiente a una palabra en quechua se insertó en la base todo lo relacionado a ella siguiendo el estándar usado por Multilingual Central Repository (MCR) su correspondiente en la traducción, se selecciona la información del español y se adiciona al synset del quechua, en el caso de encontrar palabras en el mismo quechua son synsets, lo que hace es recorrer todo y se adiciona a esas mismas palabras ese synset a esas palabras. Se crean las etiquetas de acuerdo a la categoría gramatical, Se almacenó los resultados en un formato Json para cada variedad del quechua, así mismo nos permitirá realizar las consultas en línea.

Una vez que se ejecutó el algoritmo con una palabra como muestra, se obtuvo el synset con sus respectivas entradas, en la Figura (4.4) podemos observar la clasificación. Por ejemplo, hamay (uno de sus significados en el español es respirar) fue clasificado correctamente a uno de

los synsets.


```
{
    "quechua_word": "hamay",
    "synset_word": "hamay.v.01",
    "synsets_word": [
        "aspirar.v.01",
        "respirar.v.02",
        "respirar.v.03",
        "haamay.v.01",
        "hamay.n.01",
        "hamay.v.01",
        "qushyay.v.01",
        "shuutay.v.01",
    ],
    "pos_word": "VERB",
    "definition_word": "Puywanninchikman wayrata yaykuchiy, hitaypis",
    "lemma_word": "hamay",
    "example_word": "Llullu hamaykaqta wiyaaramushaq."
},
```


Figura 4.4: Resultados de ejecución, elaboración propia

4.2.2. Validación de la WordNet

Validación por similitud

Fasttext es un método o algoritmo para vectorizar de una manera un poco más rápido y más precisa que word2vec. Se reentrenó el modelo fasttext del quechua con el conjunto de la WordNet con un total de 31 mil tokens más el conjunto de datos de quechua de fasttext para mejorar el rendimiento. Como paso inicial el algoritmo verifica si existe la entrada léxica en nuestro WordNet y seguidamente empleamos similitud de coseno para calcular la similitud entre 2 synset, finalmente se extrajo el promedio de la similitud por las variedades del quechua.

Validación por expertos

La validación de expertos en trabajos de este tipo es muy importante es así que a son de complementar la validación por similitud se seleccionó un grupo de synset del quechua sureño y

quechua central para que los expertos puedan validar si las entradas de los synsets son correctas. La validación se realizó solo de las dos variedades debido a que no se pudo contactar con expertos en las variedades de quechua amazónico y norteño.

4.3. POS tagging

4.3.1. Modelo de etiquetación gramatical basado en BERT

Para el desarrollo del POS tagging se realizó un fine-tuning al modelo QuBERT.

4.3.2. Preprocesamiento

En la fase de preparación de los datos, se aplicó tratamientos de limpieza de datos y estandarización de las etiquetas. El proceso de etiquetado lo realizaron diferentes expertos y en algunos casos emplearon la representación o abreviatura de las Dependencias Universales (UD) y en otras el nombre completo tanto en español como en inglés de la Categoría Gramatical, también añadieron comentarios aclaratorios o sugerencias de las frases del etiquetado.

Se renombró y eliminó algunas columnas que no se requerían y el conjunto de datos final quedó con las siguientes columnas: sentence-id (número de la sentencia), words(palabras de la sentencia) y labels (Categoría Gramatical de la sentencia).

El pre procesamiento se realizó al conjuntos de datos del quechua sureño y del quechua central, el conjunto de datos de las dos variedades se encuentran por separado.

4.3.3. Experimentación

Para el experimento se dividió el Dataset en dos conjuntos, datos de entrenamiento y pruebas donde los datos de entrenamiento representan el 80 % y los datos prueba 20 %.

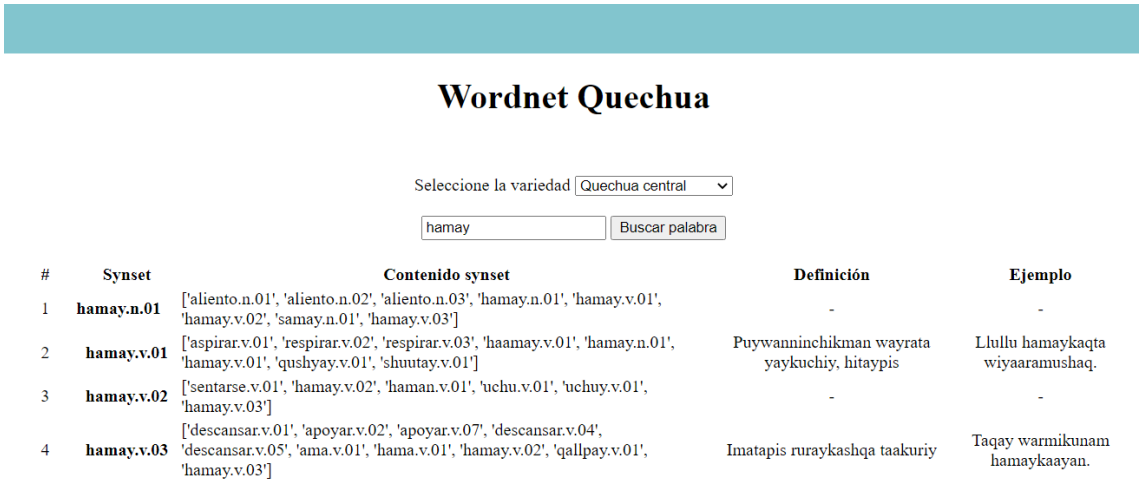
Los hiperparámetros empleados fueron los siguientes: 100 épocas con un learning rate de 1e-4 un batch size de 32. El desarrollo se realizó en una computadora portátil Google Colab. El entrenamiento para el quechua sureño duró 1 horas con 58 minutos y para el quechua central duró 1 horas con 06 minutos.

4.4. Prototipo de interfaz web

Se construyó una interfaz web para realizar consultas de los synsets y sus relaciones a partir de una base de datos no estructurada en formato JSON (WordNet). Se ha utilizado el framework de desarrollo de aplicaciones web Django escrito en lenguaje Python para la construcción del prototipo; un framework es un conjunto de componentes que ayudan a desarrollar sitios web de forma más fácil y rápida.

En la interfaz de consulta, se espera que el usuario seleccione un tipo de variedad de Quechua e ingrese una palabra en Quechua, luego de lo cual, se procede a ejecutar la búsqueda y por último, se mostrará el resultado de toda la información disponible sobre dicha palabra en los sinónimos y su relación semántica encontrada en la WordNet, como se observa en la Figura 4.6.

Cabe mencionar que cada una de las variedades de quechua son publicadas para consulta mediante un servicio web a partir de un archivo .json que contiene la información. El prototipo consume el servicio web correspondiente a la variedad de quechua seleccionada por el usuario a través de la interfaz web y junto con la palabra buscada es filtrada para mostrar la información en la estructura de salida de la búsqueda como se observa en la imagen 4.6.



Wordnet Quechua

Seleccione la variedad Quechua central

hamay Buscar palabra

#	Synset	Contenido synset	Definición	Ejemplo
1	**hamay.n.01**	['aliento.n.01', 'aliento.n.02', 'aliento.n.03', 'hamay.n.01', 'hamay.v.01', 'hamay.v.02', 'samay.n.01', 'hamay.v.03']	-	-
2	**hamay.v.01**	['aspirar.v.01', 'respirar.v.02', 'respirar.v.03', 'haamay.v.01', 'hamay.n.01', 'hamay.v.01', 'qushyay.v.01', 'shuutay.v.01']	Puywanninchikman wayrata yaykuchiy, hitaypis	Llullu hamaykaqta wiyaaramushaq.
3	**hamay.v.02**	['sentarse.v.01', 'hamay.v.02', 'haman.v.01', 'uchu.v.01', 'uchuy.v.01', 'hamay.v.03']	-	-
4	**hamay.v.03**	['descansar.v.01', 'apoyar.v.02', 'apoyar.v.07', 'descansar.v.04', 'descansar.v.05', 'ama.v.01', 'hama.v.01', 'hamay.v.02', 'qallpay.v.01', 'hamay.v.03']	Imatapis ruraykashqa taakuriy	Taqay warmikunam hamaykaayan.

Figura 4.5: Resultado de la búsqueda en la WordNet

Este prototipo utiliza una arquitectura cliente - servidor para la interacción entre el cliente (a través de la cual interactúa el usuario y envía las solicitudes web) y el servidor de aplicaciones (a través de la cual la aplicación web procesa y retorna las respuestas a las solicitudes web con los datos obtenidos). El servidor de aplicaciones a su vez interactúa con un servidor de servicios web, el cual recibe y retorna las solicitudes de información de las variedades de quechua.

Figura 4.6: Arquitectura cliente - servidor, elaboración propia de la imagen

Capítulo V

Resultados

5.1. WordNet

Una vez construida la WordNet para cada variedad del quechua se procedió a obtener las estadísticas por categoría gramatical. La tabla 5.1 presenta la WordNet por variedades el número actual de synsets y por categoría gramatical. La WordNet con mayor cantidad de synsets es el quechua sureño que representa un 76.6 % del total de los synsets del quechua, seguido por el quechua central que representa 16.9 %, el quechua norteño representa el 3.4 %, y mientras el quechua amazónico un 3.1 %. Cabe mencionar que en la WordNets la categoría gramatical que más prevalece es el sustantivo representando un 50.2 %, seguido por el verbo con 30.8 %, el adjetivo un 16 % y el adverbio un 2.9 %.

Se extrajo información del número de sentidos de cada palabra en quechua en la WordNet, por ejemplo tenemos palabra "pisiy"que podría significar estrecho (POS: adjetivo), falta (POS: sustantivo), escasear (POS: verbo) y faltar (POS: verbo), la palabra "pisiy"tiene un total de cuatro synsets. Las tablas 5.2, 5.3, 5.4 y 5.5 muestran por categoría gramatical el número de synsets que tiene una palabra en quechua.

Tabla 5.1: Número de synsets por cada WordNet y por categoría gramatical

WordNets	***Sustantivo***	***Verbo***	***Adjetivo***	***Adverbio***	***Synsets***
WordNet quechua sureño	13 444	6 969	4 187	575	**25 175**
WordNet quechua central	2 182	2 349	781	261	**5 573**
WordNet quechua norteño	460	433	146	82	**1 121**
WordNet quechua amazónico	424	384	149	47	**1 004**
Total	**16 510**	**10 135**	**5 263**	**965**	**32 873**

Tabla 5.2: Número de synsets de cada palabra en quechua por cada categoría gramatical en la WordNet del quechua sureño

N°	**Categoría gramatical**			
	Sustantivo	***Verbo***	***Adjetivo***	***Adverbio***
1	9 208	4 700	2 926	445
2	1 176	565	376	30
3	373	211	112	10
4	130	82	26	6
5	37	32	9	2
6	10	3	4	1
Total	**13 444**	**6 969**	**4 187**	**575**

Tabla 5.3: Número de synsets de cada palabra en quechua por cada categoría gramatical en la WordNet del quechua central

N°	**Categoría gramatical**			
	Sustantivo	***Verbo***	***Adjetivo***	***Adverbio***
1	1 888	1975	623	209
2	423	312	142	30
3	156	96	42	14
4	37	19	11	5
5	3	2		
6	2		1	3
Total	**2182**	**2349**	**781**	**261**

Tabla 5.4: Número de synsets de cada palabra en quechua por cada categoría gramatical en la WordNet del quechua norteño

N°	Categoría gramatical			
	Sustantivo	***Verbo***	***Adjetivo***	***Adverbio***
1	439	392	141	71
2	21	40	5	10
Total	**460**	**433**	**146**	**82**

Tabla 5.5: Número de synsets de cada palabra en quechua por cada categoría gramatical en la WordNet del quechua amazónico

N°	Categoría gramatical			
	Sustantivo	***Verbo***	***Adjetivo***	***Adverbio***
1	382	372	123	31
2	38	12	26	16
Total	**424**	**384**	**149**	**47**

Validación por similitud

En la validación de la WordNet por similitud se realizó de todo el conjunto de datos. La variedad que obtuvo un mejor resultado fue el quechua amazónico seguido por el quechua sureño, quechua central y quechua norteño como se observa en la tabla 5.6. El quechua amazónico el variedad con menos synsets debido a esto la similitud es mayor en esta variedad.

Tabla 5.6: Similitud en la WordNet obtenido por variedades

Variedad	**Similitud**
Quechua sureño	0.855
Quechua central	0.790
Quechua norteño	0.760
Quechua amazónico	0.879

Validación por expertos

Después de la validación de la muestra seleccionada de las entradas de cada synset se procedió a evaluar precisión de la alineación por el algoritmo y se obtuvo un 0.93 % de entradas de los synsets correctamente alineadas para el quechua sureño, mientras que para el quechua central un 0.94 % como se observa en la tabla 5.7.

Tabla 5.7: Precisión de la alineación de los synsets para el qechua sureño y quechua central

Variedades	Precisión
Quechua sureño	0.93
Quechua central	0.94

5.2. POS tagging

Después de realizar el entrenamiento se obtuvo los resultados que se muestran en la tabla 5.8 donde se obtuvo 0.85 para quechua sureño.

Tabla 5.8: Comparación de las puntuaciones F1 utilizando el conjunto de datos del POSTagging del quechua sureño y central en los modelos plurilingües mBERT y XLM-R, MetaXL y MAD-X, y QuBERT.

Modelos	Quechua sureño
mBERT	0.54
XLM-RoBERTa	0.75
MetaXL	0.77
MAD-X-large	0.79
QuBERT	**0.85**

En la Tabla: 5.8 se puede observar que con el modelo de QuBERT se obtiene mejores resultados. El modelo QuBERT fue entrenado con el conjunto de datos del quechua sureño (quechua

chanka y quechua collao) por está razón se realizó el entrenamiento con otros modelos para visualizar las puntuaciones y ver con cual de los modelos se obtiene un mejor resultado. El objetivo fue crear un modelo de etiquetación gramatical basado en BERT para el quechua sureño, sin embargo en proceso de recopilación y construcción de los datos también se obtuvo datos etiquetados para el quechua central y se procedió hacer el entrenamiento a pesar de saber que QuBERT fue entrenado solo con datos del quechua sureño, la puntuación obtenida fue 0.8, el resultado fue más que el esperado.

Capítulo VI

Conclusiones

El presente trabajo de investigación tuvo como objetivo general desarrollar una WordNet basado en sinonimia y un POS tagging para las variedades del quechua. Para el desarrollo de los recursos mencionados inicialmente se tuvo que recopilar diccionarios digitales de diferentes fuentes, posteriormente se alinear a una sola estructura.

Para el desarrollo de la WordNet uno de los objetivos específicos planteados fue crear corpus paralelo quechua - español, para lo cual después de la recopilación del corpus se realizó un preprocesamiento. Seguidamente se construyó el algoritmo usando la WordNet en español para ubicar cada sentido de cada palabra en quechua con su respectivo synset en español. Para la validación se empleó el algoritmo de similitud de coseno mediante el modelo fastText y validación con expertos donde se extrajo una grupos de synsets para ser validados.

Sin embargo para para el desarrollo del POS tagging no se contaba con la data etiquetada, se empleó el corpus de la WordNet para etiquetar masivamente. Se logró etiquetar el 38 % del corpus, así mismo, el resto de la data fueron etiquetados y validados por cinco lingüistas expertos en las variedades chanka, collao y central, cabe mencionar que no se obtuvo oraciones o frases

de las variedades quechua amazónico y norteño, razón por la cual el POS tagging se desarrolló solo para quechua central y sureño. Los resultados obtenidos fueron mejorando a medida que la data fue aumentado.

Finalmente, se implementó una interfaz web para realizar consultas de la WordNet quechua, buscando facilitar y motivar su uso. La consulta se realiza por cada variedad de la lengua quechua.

Capítulo VII

Trabajos futuros

Durante el desarrollo de la investigación se observado que el tamaño y la calidad del corpus utilizado influye directamente a los resultados obtenidos, razón por la cual se espera que se siga recopilando más conjunto de datos y se realice experimentos para obtener mejores resultados. Se sugiere recopilar información de las variedades de quechua amazónico y quechua norteño ya que estas dos variedades son los que cuentan con menos conjunto de datos disponible.

Finalmente, se espera que en el futuro los usuarios puedan seguir adicionando información a la WordNet desarrolla en la presente investigación, para contar con un recurso muy potente y emplear para diferentes investigaciones y creación de herramientas para la lengua quechua. Cabe mencionar que las variedades con menos recursos son el quechua amazónico y el quechua norteño.

Bibliografía

[Abiderexiti and Sun, 2019] Abiderexiti, K. and Sun, M. (2019). Construction of an english-uyghur wordnet dataset. In *China national conference on Chinese computational linguistics*, pages 382–393. Springer.

[Bella et al., 2020] Bella, G., McNeill, F., Gorman, R., Donnaíle, C. Ó., MacDonald, K., Chandrashekar, Y., Freihat, A. A., and Giunchiglia, F. (2020). A major wordnet for a minority language: Scottish gaelic. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2812–2818.

[Berment, 2002] Berment, V. (2002). Several directions for minority languages computerization. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

[Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

[Carbonell, 1994] Carbonell, J. (1994). El procesamiento del lenguaje natural, tecnología en transición. In *Actas del Congreso de la Lengua Española: Sevilla, 7 al 10 octubre, 1992*, pages 247–250. Instituto Cervantes.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Drude et al., 2003] Drude, S. et al. (2003). Language vitality and endangerment.

[Fellbaum, C, 1988] Fellbaum, C (1988). Wordnet: Wiley online library.

[Fishman, 1991] Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*, volume 76. Multilingual matters.

[Gonzalez-Agirre et al., 2012] Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012). Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

[Gonzalez-Agirre, 2013] Gonzalez-Agirre, Aitor y Rigau, G. (2013). Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1):13–28.

[Grave et al., 2018] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

[Güngör, 2010] Güngör, T. (2010). Part-of-speech tagging. *Handbook of natural language processing*, 2:205–235.

[INEI, 2017] INEI (2017). Instituto nacional de estdistica e informática.

[Jurafsky and Martin, 2009] Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.

[King, 2015] King, B. P. (2015). *Practical Natural Language Processing for Low-Resource Languages*. PhD thesis.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Maguino-Valencia et al., 2018] Maguino-Valencia, D., Oncevay, A., and Cabezudo, M. A. S. (2018). Wordnet-shp: Towards the building of a lexical database for a peruvian minority language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[Miller, 1998] Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

[MINEDU, 2013] MINEDU (2013). Documento nacional de lenguas originarias del perú.

[MINEDU, 2018] MINEDU (2018). Documento nacional de lenguas originarias del perú.

[MINEDU, 2020] MINEDU (2020). Chawpi qichwata alli qillqanapaq maytu 2= manual de escritura en lengua originaria quechua central.

[MINEDU, 2021] MINEDU (2021). Urin qichwa qillqay yachana mayt'u= manual de escritura quechua sureño.

[Pereira-Noriega et al., 2017] Pereira-Noriega, J., Mercado-Gonzales, R., Melgar, A., Sobrevilla-Cabezudo, M., and Oncevay-Marcos, A. (2017). Ship-lemmatagger: Building an nlp toolkit for a peruvian native language. In *International Conference on Text, Speech, and Dialogue*, pages 473–481. Springer.

[Python,] Python. The python tutorial. https://docs.python.org/3/tutorial/index.html.

[Ramachandran,] Ramachandran, A. Nlp guide: Identifying part of speech tags using conditional random fields. https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31.

[Rios, 2015] Rios, A. (2015). *A basic language technology toolkit for quechua*. PhD thesis, University of Zurich.

[Streiter et al., 2006] Streiter, O., Scannell, K. P., and Stuflesser, M. (2006). Implementing nlp projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289.

[Taghizadeh and Faili, 2016] Taghizadeh, N. and Faili, H. (2016). Automatic wordnet development for low-resource languages using cross-lingual wsd. *Journal of Artificial Intelligence Research*, 56:61–87.

[Torero, 1964] Torero, A. (1964). *Los dialectos quechuas*. Univ. Agraria.

[Torero, 2002] Torero, A. (2002). *Idiomas de los Andes: lingüística e historia*. Number 162. IFEA, Instituto Francés de Estudios Andinos.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Zevallos et al., 2022] Zevallos, R., Ortega, J., Chen, W., Castro, R., Bel, N., Toshio, C., Venturas, R., Aradiel, H., and Melgarejo, N. (2022). Introducing qubert: A large monolingual corpus and bert model for southern quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13.

Anexos

Anexo 1: WordNet-QU: Development of a Lexical Database for Quechua Varieties

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.

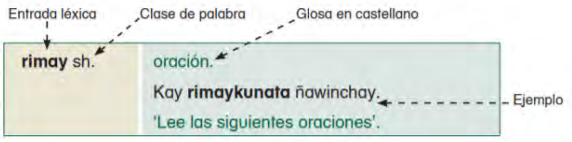


Figure 1: Example of the structure of the some dictionaries.

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.