PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

DESARROLLO DE UNA BASE DE DATOS LÉXICA BASADA EN SINONIMIA PARA SHIPIBO-KONIBO

Tesis para obtener el título profesional de Ingeniero Informático

AUTOR

Diego Arturo Maguiño Valencia

ASESOR

Mag. Felix Arturo Oncevay Marcos

Lima, Agosto, 2021

AGRADECIMIENTOS

En primer lugar, deseo agradecer a mis padres por todo su apoyo a lo largo de mi carrera universitaria, además de ayudarme con distintos detalles cada vez que estuviera a su alcance.

También agradezco a mi asesor Arturo Oncevay, y a los profesores Marco Sobrevilla y Andrés Melgar por la oportunidad de participar en el proyecto Chana así como su guía para el desarrollo de la tesis.

Finalmente, agradezco a los demás miembros del proyecto Chana quienes me ayudaron con lo relacionado al idioma. Asimismo, al Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) que a través del proyecto "Una plataforma de software para la traducción automática de textos entre lenguas originarias de la Amazonía peruana y español" (225-2015 FONDECYT), dieron los recursos necesarios para lograr este proyecto.

PUBLICACIONES

El presente trabajo se realiza para optar por el título de Ingeniero Informático de la Pontificia Universidad Católica del Perú y como parte de este se ha realizado la siguiente publicación relacionada durante su elaboración, la cual ha sido incluida como anexo:

- WordNet-Shp: Towards the Building of a Lexical Database for a Peruvian Minority Language, Diego Maguiño Valencia, Arturo Oncevay, Marco Sobrevilla (LREC 2018) En este paper se presentó el proceso completo para construir la WordNet en Shipibo-Konibo. (Anexo 1)

RESUMEN

Este proyecto tiene como objetivo el desarrollo de una base de datos léxica basada en sinonimia (mejor conocida como WordNet) para la lengua Shipibo-Konibo. Se trabajó con el fin de generar recursos electrónicos para esta lengua que a pesar de ser la segunda más hablada en la amazonia posee escasos recursos lingüísticos. Se contó con el apoyo de lingüistas y un hablante nativo de Shipibo-Konibo durante el proceso de desarrollo y para la validación del mismo.

Como base se usó un diccionario escaneado en Shipibo-Konibo y la WordNet en español disponible a través de Internet. Para que la lectura del diccionario fuera posible, se desarrolló un algoritmo con este fin, el cual separaba las palabras del diccionario y las guardaba de forma ordenada en una base de datos. Entre los datos guardados por cada término se encuentran sentidos, glosa en español, categoría gramatical y ejemplos de uso.

Una vez que ya se disponía de la base de datos del diccionario, este se usó como entrada para el algoritmo de creación de la WordNet en Shipibo-Konibo. Este algoritmo consiste en tomar la glosa y ejemplos de uso de cada sentido de cada término del diccionario y compararla con todos los synsets de la WordNet en español para determinar con cual se encuentra más relacionado. Esto se calcula en base al modelo Word2Vec el cual es usado para agrupar palabras detectando similitudes en los vectores que las representan matemáticamente. Realizado el cálculo, está relación es guardada en una base de datos, una vez que se completa el algoritmo la base de datos resultante es la WordNet.

También se implementó una interfaz web de consulta así hacer posible el acceso a cualquier usuario. Este recurso es muy útil para facilitar tareas como la desambiguación, extracción de información y traducción automática gracias a la flexibilidad en las búsquedas. Al tener un carácter multilingüe, la WordNet ayudará no solo a preservar sino también expandir el alcance y la posibilidad de integrar a la lengua con otras personas interesadas.

Tabla de contenido

Índice de Figuras

Índice de Tablas

1. DEFINICIÓN DEL PROBLEMA

1.1. Problemática

El idioma hablado no es solo el principal medio de comunicación humana, sino también representa la identidad cultural y la autonomía. Si el mundo pierde una lengua, los recuerdos y las experiencias de esta cultura van con ella. Se estima que durante el próximo siglo aproximadamente la mitad de todos los idiomas que se conocen hoy se habrán extinto. [Crystal, 2000].

Con el objetivo de ayudar en la revitalización de las lenguas en peligro, que son generalmente lenguas de bajos recursos, muchos esfuerzos deben hacerse. Una forma es alentar a las generaciones más jóvenes a utilizar su lengua materna mediante la construcción de plataformas *e-learning* y creando juegos instructivos. La documentación oral puede ser utilizada para preservar la cultura de las lenguas en peligro de extinción; especialmente debido a que muchos de estos idiomas sólo son hablados. Estos tienen ricas tradiciones orales[1] con cuentos, refranes, canciones, cantos e historias, pero no hay forma escrita. Por lo tanto, la extinción de tales idiomas conducirá rápidamente a la aniquilación de su cultura [Allah y Boulaknadel, 2012].

Los idiomas que se encuentran entre los 10 primeros más hablados cuentan con la mayoría de recursos lingüísticos necesarios para que puedan ser procesados eficientemente, es decir, cuentan con mecanismos computacionalmente eficaces para la comunicación entre personas y máquinas por medio de lenguajes naturales. En especial, este es el caso del inglés, debido a que la mayoría del *software* usado en el mundo se encuentra en ese idioma. Sin embargo, hay idiomas donde no se cuentan con léxicos completos o herramientas de análisis [Krauss, 2007]. Esto ocurre con mayor frecuencia en lenguas de bajos recursos como es el caso de las lenguas amazónicas del Perú.

Muchas herramientas de Procesamiento de Lenguaje Natural (PLN) requieren o se benefician de recursos lingüísticos fiables, como léxicos y gramáticas. En particular, para tareas de PLN como desambiguación de palabras y traducción automática es útil un recurso digitalizado como un diccionario ya que se accede a la palabra y se ubican sus sentidos y definiciones. Asimismo, también se tienen los tesauros donde se ordenan las palabras por sinónimos sin definiciones.

Otro ejemplo de recurso léxico digital es la WordNet [WordNet Princeton University, 2016]. Esta es una base de datos que almacena palabras y las agrupa en sinónimos, más conocidos como *synsets (synonim set:* conjunto de sinónimos). La versión disponible en inglés contiene las palabras organizadas en aproximadamente 117,000 *synsets* y cada uno representa un sentido único. Asimismo, este recurso es

[1] Tradición Oral: Patrimonio inmaterial de una comunidad y que se manifiesta de formas hablada. Por ejemplo, cuentos, mitos, leyendas, poesía, cantos, etc.

muy útil al servir de soporte para tareas como la desambiguación, extracción de información y traducción automática ya que se puede buscar por palabra (se puede diferenciar sus respectivos sentidos) y a la vez cada grupo de palabras está alineado con otro grupo que contenga el mismo significado pero en otro idioma. Al tener un carácter multilingüe, la WordNet ayudará no solo a preservar sino también a expandir el alcance de una lengua.

El presente proyecto tiene como objetivo el desarrollo de una WordNet para el Shipibo-Konibo alineada al Español. La elección de esta lengua se basa en que además de su importancia en el Perú, se cuenta con expertos en la Pontificia Universidad Católica del Perú (PUCP) avocados al estudio de esa lengua para incluir este proyecto en el desarrollo de un traductor automático Español – Shipibo-Konibo. Además, a pesar de ser la segunda más hablada en la Amazonía no posee recursos lingüísticos suficientes. Al inicio del proyecto para el Shipibo-Konibo, sólo se tiene un diccionario bilingüe donde se tienen las definiciones en Español.

Se trabajará con el fin de generar un recurso electrónico capaz de facilitar las tareas de PLN. Estas tareas serán parte de un proyecto de desarrollo de una plataforma de traducción automática entre Español y Shipibo-Konibo, ya que la desambiguación es un proceso intermedio crítico en la traducción. Una vez se cuente con un diccionario digital de Shipibo-Konibo (palabras y definiciones) se busca resolver el interrogante de que técnicas se deben usar para alinear y agrupar los sentidos de las palabras para armar este recurso. Finalmente, la falta de un modo de acceso para consultar información a un recurso lingüístico en la lengua Shipibo-Konibo es necesario para satisfacer las necesidades del trabajo futuro de PLN en dicha lengua.

1.2. Objetivo general

Desarrollar una base de datos léxica que contenga los sentidos de palabras (*synsets*) y sus relaciones semánticas correspondientes para la lengua Shipibo-Konibo.

1.3. Objetivos específicos

1. Digitalizar y pre-procesar un diccionario bilingüe Shipibo-Konibo – Español.

2. Implementar un algoritmo de clasificación que permita alinear cada sentido de cada palabra con el *synset* de igual significado en español.

3. Implementar una interfaz de comunicación que permita el acceso a los datos de la WordNet en Shipibo-Konibo.

1.4. Resultados esperados

Para objetivo específico 1:

1. Un algoritmo para separar los elementos del diccionario como términos, sentidos, categorías gramaticales y ejemplos de uso.
2. Una base de datos para almacenar los elementos del diccionario.

Para objetivo específico 2:

3. Un algoritmo de clasificación que permita ubicar el synset correspondiente en la WordNet de español a un sentido de cada palabra en Shipibo-Konibo.
4. Una base de datos para almacenar los synsets en Shipibo-Konibo y las alineaciones entre synsets de las WordNet en Shipibo-Konibo y Español.

Para objetivo específico 3:

5. Una estructura de la interfaz web para consultar synsets y el contenido.
6. Arquitectura de servicio web para acceder a la WordNet de Shipibo-Konibo.

1.5. Herramientas, métodos y procedimientos

A manera de resumen, en la tabla 1 se listarán las metodologías, métodos o procedimientos que se utilizaron para cada resultado esperado para luego explicar con más detalle cada uno:

Resultado esperado	Herramienta o recurso
Un algoritmo de para separar los elementos del diccionario como términos, sentidos, categorías gramaticales y ejemplos de uso.	NetBeans IDE, Diccionario bilingüe Shipibo-Konibo – Español, Java.
Una base de datos para almacenar los elementos del diccionario.	MySQL, Microsoft Excel, Diccionario bilingüe Shipibo-Konibo - Español, WordNet en español.
Un algoritmo de clasificación que permita ubicar el synset correspondiente en español a un sentido de cada palabra en Shipibo-Konibo.	Java, NetBeans IDE, base de datos del diccionario, Word2Vec, WordNet en español, métricas de validación.
Una base de datos para almacenar los synsets en Shipibo-Konibo y las alineaciones entre synsets en Shipibo-Konibo y español.	Java, NetBeans IDE, MySQL.
Una estructura de la interfaz web para consultar synsets y sus relaciones.	Java, NetBeans IDE.
Arquitectura de servicio web para acceder a la WordNet de Shipibo-Konibo.	Java, NetBeans IDE.

Tabla 1. Herramientas y recursos a usar

Recursos y Herramientas

1. Diccionario bilingüe Shipibo-Konibo - Español

Este diccionario digital contiene las palabras en Shipibo-Konibo recabadas y validadas manualmente por el equipo, entre ellos lingüistas. Las definiciones de las palabras están escritas en español y se indica la categoría gramatical de cada palabra. Además, cuenta con la traducción en español y ejemplos de uso para algunas palabras [Loriot et al., 1993].

Se tomará la definición de cada sentido por cada palabra encontrada en el diccionario que servirá como datos de entrada para el algoritmo de clasificación.

2. WordNet en Español

Se utilizará la versión en español como base para hacer el alineamiento. Se tiene la opción entre la MultiWordNet y *Multilingual Central Repository* (MCR) en sus versiones en español. La MultiWordNet versión 1.6 en español cuenta con 67 mil *synsets* [Bentivogli et al., 2012] mientras que la MCR versión 3.0 cuenta con 48 mil [Gonzalez y Rigau, 2013]. Ambas poseen todas las relaciones semánticas entre *synsets* que serán heredadas para Shipibo-Konibo al momento de completar la alineación. La diferencia está en la versión, la MCR tuvo su última actualización en una fecha mas reciente en comparación a MultiWordNet y el proceso para descargar

la versión en español es simple y no requiere registros como la MultiWordNet por lo que se decidió trabajar con la MCR a pesar de tener menor número de *synsets*.

Al aplicar el algoritmo se realizará la clasificación palabra por palabra del diccionario comparando con esta WordNet para que la WordNet resultante de Shipibo-Konibo quede alineada a la del español.

3. Métricas de Validación

Se tomará en cuenta dos métricas para verificar la eficiencia del algoritmo de alineamiento de *synsets*. La comparación de las métricas con los resultados obtenidos se realizará con un estándar de prueba (llamado también *gold standard*) que estará conformado por una muestra de *synsets* en Shipibo-Konibo alineados manualmente con sus similares en español. Cabe resaltar que estas métricas serán obtenidas para cada *synset* y finalmente se obtendrá la media de los resultados obtenidos para todos los *synsets*. Las métricas de precisión y sensibilidad son definidas a continuación:

$$\text{Precisión} = \frac{\textit{Número de palabras de un synset del gold standard que fueron seleccionados correctamente para ese synset}}{\text{Número de palabras seleccionadas por el método de alineamiento para un synset}}$$

$$\text{Sensibilidad} = \frac{\textit{Número de palabras de un synset del gold standard que fueron seleccionados correctamente para ese synset}}{\text{Número de palabras pertenecientes a un synset del } \textit{gold standard}}$$

4. Java

Es un lenguaje de programación y plataforma de computación que fue lanzado por primera vez por Sun Microsystems en 1995 (esta compañía luego fue adquirida por Oracle). Se caracteriza por que el código compilado funciona bajo cualquier sistema operativo y tipo de dispositivo. Además, es rápido, seguro y fiable. Es por estos motivos que para el algoritmo y todas las interfaces se hizo uso de este lenguaje [Java, 2016].

5. NetBeans IDE

Es el entorno de desarrollo oficial de la versión 8 del lenguaje de programación Java. Este entorno es libre y sin restricciones de desarrollo. Su uso es recomendable ya que es un entorno diseñado por Oracle, la misma compañía que le da mantenimiento al lenguaje Java. NetBeans cuenta con un entorno de depuración que permite analizar el código fuente.

6. MySQL

Es un sistema de gestión de base de datos relacional de código abierto. La información contenida en una base de datos MySQL se almacena en forma de tablas relacionadas. Estas bases de datos se utilizan frecuentemente para el desarrollo de aplicaciones Web.

Una base de datos MySQL se puede acceder (consultar) usando directamente lenguajes como C++, Java, Python entre otros lenguajes de programación. Para las

consultas se usa un subconjunto de los comandos estándar de Lenguaje de Consulta Estructurado conocido como SQL [MySQL, 2016].

En la base de datos se guardarán los *synsets* formados por el algoritmo.

Métodos y procedimientos

1. Construcción del Algoritmo de Alineamiento

Se tomará la definición de cada sentido por cada palabra encontrada en el diccionario de Shipibo-Konibo para ubicar el *synset* en español que represente el mismo concepto. Una vez que todos los sentidos hayan sido alineados, los grupos resultantes de palabras en Shipibo-Konibo serán los *synsets*.

2. Validación de una muestra de la WordNet del Shipibo-Konibo

Para realizar la validación, primero, un grupo de lingüistas alinearán manualmente una muestra de *synsets* en español con sus respectivos en Shipibo-Konibo, generando así un estándar de prueba o *gold standard*. Finalmente, Los *synsets* generados del alineamiento automático realizado por el algoritmo descrito en el paso anterior serán comparados con los *synsets* del *gold standard*, considerando como métricas de evaluación a la precisión y sensibilidad. Esto nos dará una medida de calidad de la WordNet generada.

3. Construcción del Servicio Web

La interfaz a construir será similar a la WordNet de Princeton [Fellbaum, 1998]. El servicio será implementado de modo que la WordNet pueda ser integrada en el traductor. En la Figura 1 se muestra la interfaz de consulta de la WordNet de Princeton. En la misma se puede ver los diferentes sentidos para la palabra *star* (en español se podría traducir como "estrella"). Primero se observan los sentidos que son sustantivos. Se da una defición de la palabra como un cuerpo celeste que irradia energía. Luego se tiene sl *synset* donde las palabras tienen el mismo significado de estrella al tomarse el sentido de campeón o as. En la segunda parte se tiene la palabra con los sentidos que son verbos. Uno de los sentidos es el de aparición en un show y da un ejemplo de un actor que aparece en una película.

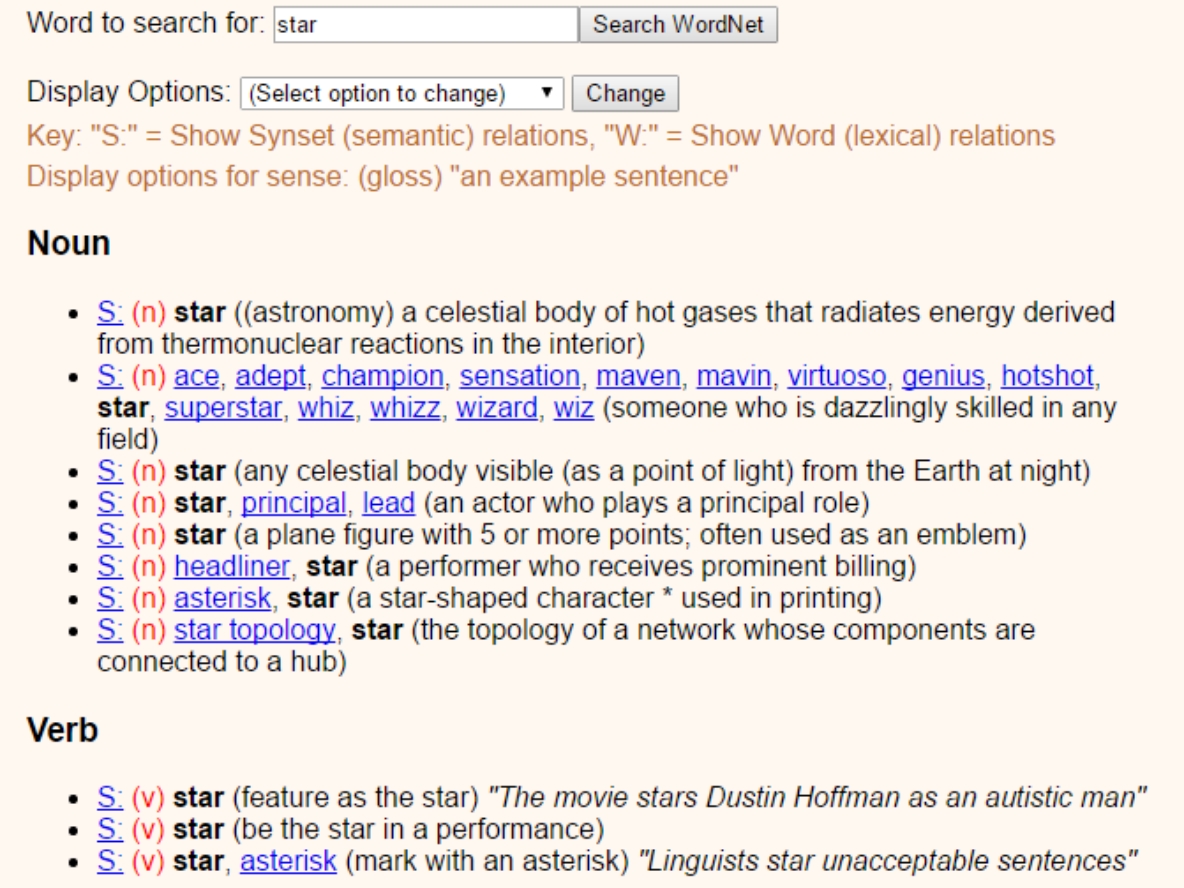


Figura 1: Interfaz de una consulta realizada en la WordNet en ingles de Princeton

1.6. Alcance

Para el recurso lingüístico a construir en este proyecto se desarrolló un algoritmo que permita la clasificación y alineamiento entre *synsets* en Español y Shipibo-Konibo. Las relaciones semánticas serán heredadas automáticamente de la WordNet en español al hacer el alineamiento.

Además, se tendrá como dato de entrada un diccionario digitalizado en Shipibo-Konibo. Para la validación de este diccionario, se contará con lingüistas y hablantes de esta lengua. Asimismo, se implementará una interfaz web para consulta de la WordNet en Shipibo-Konibo similar a otras que se pueden acceder a través de internet.

El proyecto se enfocará en la lengua del Shipibo-Konibo gracias a que se cuenta con el diccionario bilingüe y la disposición de expertos en el tema. El algoritmo de clasificación se podría aplicar a otras lenguas, además del Shipibo-Konibo, con la condición de que se cuente con un diccionario y que las definiciones de las palabras estén en español.

El resultado del proyecto será un recurso digital cuyo nombre estándar es WordNet el cual será para la lengua Shipibo-Konibo. Este recurso es muy útil al facilitar tareas como la desambiguación del sentido de las palabras, extracción de información y traducción automática gracias a la flexibilidad en las búsquedas de palabras. Al tener un carácter multilingüe gracias a la alineación con la WordNet en español, no solo se ayudará a preservar sino también expandir el alcance y la

posibilidad de integrar a la lengua Shipibo-Konibo con otras estudios o proyectos de investigación .

1.6.1. Limitaciones

El diccionario que se tendrá como entrada será completado y validado de forma manual por lingüistas y otras personas asociadas, por lo que se depende del correcto ingreso de palabras incluyendo sentidos, categoría gramatical y ejemplos de uso (si los hubiera).

Para este proyecto se desarrollará un algoritmo de clasificación para los sentidos de cada palabra encontrada en un diccionario limitándose al contenido del mismo.

El no estar muy familiarizado con la lengua Shipibo-Konibo demanda que las validaciones manuales sean realizadas con la ayuda de un hablante por lo que se depende de su presencia.

1.6.2. Riesgos

Riesgo identificado	**Impacto en el proyecto**	**Medidas correctivas para mitigar**
Retraso en la digitación y validación del diccionario electrónico en Shipibo-Konibo.	Retraso en las pruebas generales del algoritmo de clasificación.	Priorizar la corrección y validación manual del texto desde el inicio del proyecto.
No contar con la presencia del hablante en Shipibo-Konibo.	Retraso general validación de los resultados.	Adecuar el horario de consulta con el hablante con su disposición.

Tabla 2. Riegos del proyecto

1.7. Justificación

Las tradiciones orales como cuentos, leyendas e historias pueden ser utilizadas para enriquecer la cultura de una lengua pero no es suficiente para su preservación e integración con su entorno. Se necesitan recursos lingüísticos físicos y digitales con el fin de ayudar en la revitalización de las lenguas en peligro, que son generalmente lenguas de bajos recursos, como el caso del Shipibo-Konibo [Allah y Boulaknadel, 2012].

Actualmente, el uso de recursos lingüísticos digitales (como diccionarios) es ampliamente usado para fines didácticos e informativos en todo el mundo. Además, ayuda a preservar y expandir el alcance de su uso más allá de su comunidad, como potencial desarrollo de PLN que se le pueda dar a una lengua que dispone de estos recursos.

La solución propuesta es útil para tareas como la desambiguación del sentido de las palabras. Esta tarea será parte de un proyecto de desarrollo de una plataforma de traducción automática entre español y Shipibo-Konibo, ya que la desambiguación es un proceso intermedio importante y crítico en la tarea de traducción. Esta solución se desarrollará para la lengua Shipibo-Konibo porque además de su importancia para el Perú, se cuenta con expertos en la PUCP dedicados al estudio de esta familia lingüística.

2. Marco Conceptual

A continuación se presentarán conceptos directamente relacionados con la temática de este proyecto para asegurar su entendimiento. Se tratará la definición de lingüística computacional, desambiguación y sus características.

2.1. Conceptos Generales

- **Procesamiento del Lenguaje Natural (PLN)**

Área que maneja el procesamiento computacional de información, expresada en lenguaje natural, con el fin de implantar a las computadoras con la capacidad de comprender textos escritos por humanos y producirlos en una lengua familiar para que los humanos puedan entenderlos [Jurafsky y Martin, 2009].

Para eso se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio de lenguajes naturales.

Según Carbonell [Carbonell, 1994] los objetivos del PLN se basan en tres tipos de aplicaciones:

- 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 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 significativos, la elaboración de resúmenes, etc. Las actuales investigaciones en el campo del PLN intentan abordar estos problemas.
- Traducción automática: El objetivo original del PLN toma una vez más la delantera en cuanto a resultados científicos recientes, avances tecnológicos y aplicaciones prácticas. Diversos sistemas multilingües eficaces de traducción automática ya están siendo explotados industrialmente y continuarán evolucionando de manera rápida en un futuro inmediato.

Entre las dificultades encontradas en el área de PLN destacan dos: Ambigüedad de sentidos y variedad de léxicos. El lenguaje natural es localmente ambiguo, y la resolución de ambigüedades es necesaria para un procesamiento eficaz. Otra dificultad es la extensión y variedad de léxicos. Se debe tener registro de las miles de palabras y sus respectivos sentidos de la lengua objetivo. Además, hay palabras del mismo idioma que no se usan o tienen otro sentido según la comunidad.

- **Lengua de bajos recursos**

Es una lengua para la que no existe la capacidad de automatizada de tecnologías del lenguaje humano. Históricamente, el desarrollo de la tecnología para la explotación automatizada de materiales en idiomas extranjeros ha requerido un esfuerzo prolongado y una inversión de esfuerzo para obtener datos de gran tamaño. Los métodos actuales pueden requerir varios años y una inversión de dinero muy alta por cada lengua, sobre todo para construir material traducido o transcrito. Como resultado, existen sistemas de tecnología lingüística humanos

principalmente para los idiomas de uso generalizado o de alta demanda. Para el resto, este objetivo es muy difícil [Onyshkevych, 2014].

En el Perú, podemos considerar como lenguas de bajos recursos a las amazónicas, entre ellos el Shipibo-Konibo que será la lengua objetivo en este proyecto. De acuerdo al censo del 2007, se estima que la población shipiba es de 23,000 personas hasta un máximo de 30,000.

- **Base de datos léxica**

Una base de datos es una colección organizada de información digital [Ramakrishnan y Gehrke, 2000]. Al ser léxica, la información almacenada puede incluir los sentidos, la categoría léxica y sinónimos de las palabras, así como las relaciones semánticas y fonológicas entre diferentes palabras o conjuntos de palabras.

2.2. Conceptos relacionados a recursos y modelos lexicales

- **Diccionario electrónico**

Recopilación digital de un gran conjunto de palabras del mismo idioma e incluye por cada una todos sus sentidos y categoría gramatical. En otras palabras es un diccionario almacenado de forma digital. Ejemplo de una palabra encontrada en un diccionario de idioma español [RAE, 2016]:

ambiguo:

n°	categoría	sentido
1	Adjetivo	Dicho especialmente del lenguaje: Que puede entenderse de varias formas.
2	Adjetivo	Dicho de una persona: Que, con sus palabras o comportamiento, vela o no define claramente sus actitudes u opiniones.
3	Adjetivo	Incierto, dudoso.

- **Tesauro**

Un tesauro es una obra de consulta que lista palabras agrupadas de acuerdo a la similitud de significado (puede contener sinónimos y antónimos), en contraste con un diccionario, que proporciona definiciones para las palabras, y en general los enumera en orden alfabético [Tesauro, 2016]. El objetivo principal de este tipo de obras de referencia es "ayudar al usuario a encontrar la palabra o palabras, por el cual una idea puede ser mejor expresada" - Peter Mark Roget, arquitecto del tesauro más conocido en el Idioma en Inglés [Roget, 1991]. Por ejemplo, si se busca la palabra 'dato' en el tesauro en línea *Open Thesaurus* se tiene lo siguiente:

1. antecedente, apunte, circunstancia, documento, nota, noticia, referencia, reseña

2. cantidad, cifra, número
3. aspecto, elemento, factor, punto

Cada uno de estos tres conjuntos de palabras representa un sentido diferente para una misma palabra. Cada palabra puede tener más de un sentido pero hay uno que es sinónimo por lo cual pertenece a dicho grupo.

- **Word2Vec**

Es una red neuronal de dos capas que usa representaciones distribuidas de texto para capturar similitudes entre conceptos [DLJ, 2016]. Su entrada es un corpus de texto y su salida es un conjunto de vectores llamados vectores de características. Word2Vec crea vectores que representan numéricamente las palabras de un texto. Luego se toman estos vectores para compararlos semánticamente entre sí.

Entre las aplicaciones de Word2Vec se encuentran comparar entidades similares o clasificarlas. Un ejemplo de aplicación es representar nombres de ciudades y países. Por ejemplo, al usar el modelo, Francia estará más cerca de Paris que de otras ciudades como Berlín o Madrid.

2.3. WordNet

Es una base de datos léxica en un idioma determinado donde sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos (***synset*s**), cada uno expresando un concepto diferente. Cada término del *synset* tiene su respectiva definición (puede ser más de una) y ejemplos de su uso. Los *synset*s están unidos entre sí por medio de relaciones conceptuales semánticas y léxicas [Fellbaum, 1998].

La WordNet idealmente es libre y públicamente disponible para su descarga a través de internet. Su estructura hace que sea una herramienta útil para la lingüística computacional y herramienta de procesamiento de lenguaje natural.

WordNet superficialmente se parece a un tesauro, en el que grupos de palabras se juntan basándose en sus significados. Sin embargo, hay algunas diferencias importantes. En primer lugar, WordNet articula no sólo las cadenas de las formas de las letras sino también significados específicos de palabras. Como resultado, las palabras que están en estrecha proximidad entre sí en la red son semánticamente desambiguadas. En segundo lugar, las etiquetas de WordNet las establecen relaciones semánticas entre las palabras, mientras que los grupos de palabras en un diccionario de sinónimos no sigue ningún patrón distinto de explícita similitud de significado [WordNet Princeton University, 2016]. En el caso del español se cuenta, entre otras, con la MultiWordNet [Girardi, 2002].

En la Tabla 3 se presentan los 4 sentidos de la palabra ‘estrella’ extraidos de la MultiWordNet presentando 1 ejemplo por cada sentido:

Synset	Sentido	Ejemplo
{estrella}	Cuerpo celeste de gases calientes que irradia energía derivada de las reacciones termonucleares en el interior.	La estrella más próxima a la Tierra es el sol.
{crac, diestro, as, estrella, genio}	Alguien con mucha habilidad para algo.	Cristiano Ronaldo es la estrella de su equipo.
{estrella, protagonista}	Persona que interpreta un rol importante	El hombre araña es la estrella en esta película.
{estrella}	Figura plana de 5 puntos.	¿Me podrías dibujar una estrella?

Tabla 3: Sentidos de la WordNet multilingüe para la palabra "estrella". Extraído de MultiWordNet.

- **Relaciones entre palabras:**

Sinonimia: Es una relación semántica de identidad o semejanza de significados entre determinadas expresiones o palabras (llamadas sinónimos). Por lo tanto, sinónimos son expresiones o palabras que tienen un significado similar o idéntico entre sí y pertenecen a la misma categoría gramatical. Por ejemplo, sinónimos de desastre son calamidad, devastación y ruina [Miller et al., 1990].

- **Relaciones entre *synset*s:**

Hiperonimia: Término general que puede ser utilizado para referirse a la realidad nombrada por un término más específico. Por ejemplo, **ser vivo** es hiperónimo para los términos **planta** y **animal** [Miller et al., 1990] Está relación aparece entre *synset*s de sustantivos.

Hiponimia: Se le denomina a la palabra que posee todos los rasgos semánticos de otra más general (su hiperónimo) pero que en su definición añade otras características semánticas que la diferencian de ésta. Por ejemplo, los hipónimos de **día** son: **lunes**, **martes**, **miércoles**, etc. [Miller et al., 1990]. Está relación aparece entre *synset*s de sustantivos.

Meronimia: Relación semántica no simétrica entre los significados de dos palabras dentro del mismo campo semántico. Se denomina merónimo a la palabra cuyo significado constituye una parte del significado total de otra palabra. Por ejemplo, **dedo** es un merónimo de **mano** y **mano** es merónimo de **brazo** [Miller et al., 1990]. Esta relación se encuentra entre *synset*s de sustantivos.

Holonimia: Noción semántica que se opone a meronimia, del mismo modo en que se oponen el todo y la parte. Así, por ejemplo, **BICICLETA** es un holónimo mientras que **sillín**, **pedal** y **aro** son merónimos [Miller et al., 1990]. Esta relación se encuentra entre *synset*s de sustantivos.

Troponimia: Relación entre verbos que denota una manera particular de hacer una cosa. La troponimia es a los verbos como la hiponimia es a los sustantivos. Por ejemplo: **decir**, **contar** sería tropónimo de **verbalizar**, **gritar**, **susurrar** [Miller et al., 1990].

Antonimia: Son palabras que tienen un significado contrario entre sí. Deben pertenecer, al igual que los sinónimos, a la misma categoría gramatical. Por ejemplo, antónimos de **alegría** son: **tristeza, depresión, melancolía** [Miller et al., 1990]. Los adjetivos están organizados mediante antonimia. Por ejemplo, bonito y feo.

2.4. Conceptos relacionados a la ambigüedad y corpus

- **Ambigüedad**

Se le llama ambiguo a una idea cuando existe más de una estructura lingüística para ella; es decir, que puede tomarse de diferentes formas dependiendo de cómo está redactada [Gupta et al., 2007].

- **Tipos de Ambigüedad:**

Léxica: Cuando una palabra tiene varios sentidos. Ejemplo: banco. Tiene 2 sentidos: Entidad financiera y objeto para sentarse [Piruzelli y da Silva, 2010].

Sintáctica: Se da cuando una frase puede interpretarse de diferentes maneras. Ejemplo: “Se debe limpiar aquí.” No se sabe si es una orden para alguien que necesita limpiarse o ese lugar necesita una limpieza [Piruzelli y da Silva, 2010].

Fonética: Cuando hay más de una forma de componer un conjunto de sonidos en palabras. Ejemplo: ¿Me diste la caja? y ¿Mediste la caja? [Piruzelli y da Silva, 2010].

Funcional: Cuando se usa un término con doble función gramatical. Ejemplo: he vuelto a ver (antes no veía y ahora sí; o bien, me he dado una vuelta para ver cómo continúan las cosas por aquí) [Piruzelli y da Silva, 2010].

Morfológica: Se da cuando coinciden en una frase dos formas de un mismo verbo. Ejemplo: Pedro y yo escribimos un cuento (no se sabe si lo hemos escrito ya o lo estamos escribiendo) [Piruzelli y da Silva, 2010].

Pragmática: Depende del contexto del lenguaje y del hablante en un momento dado. Ejemplo: golpeó el armario con el bastón y lo rompió (no sabemos si se rompió el bastón o el armario) [Piruzelli y da Silva, 2010].

Semántica: Se da cuando afecta a un elemento de la frase que puede ser interpretado de diversos modos. Ejemplo: Pedro quiere pelearse con un francés (no sabemos si se trata de cualquier francés o de uno en particular) [Piruzelli y da Silva, 2010].

La ambigüedad léxica será la más frecuente ya que se trabaja con palabras individuales que pueden tener varios sentidos. Para resolver esta dificultad existe la tarea de desambiguación del sentido de las palabras que permite identificar el sentido de una palabra dentro de un contexto.

- **Corpus**

Es definido como un conjunto de datos lingüísticos (pertenecientes al uso oral o escrito de una lengua, o dos) sistematizados siguiendo determinados criterios, suficientemente extensos en amplitud y profundidad, de manera que sean representativos del uso lingüístico o de alguno de sus ámbitos, dispuestos de tal modo que pasan a ser procesados por computador, con la finalidad de propiciar resultados útiles para análisis [Sánchez, 1995].

3. Estado del arte

3.1. Introducción

La selección de los trabajos relacionados se realizó usando como metodología la revisión sistemática. Se eligió este método debido a la ventaja que ofrece al sintetizar investigaciones realizadas anteriormente, minimizando la parcialidad gracias al proceso estructurado que sigue [Kitchenham, 2004].

El objetivo de esta revisión es dar a conocer los diferentes aspectos relacionados con la construcción, uso o mejora de WordNets para diferentes lenguas tanto si son catalogados como lenguas de bajos recursos o no. Para tal fin, se planteó la siguiente pregunta de investigación: ¿Cómo se han desarrollado o mejorado WordNets en el mundo?

3.2. Método usado en la revisión del estado del arte

En base a la pregunta, se elaboró una cadena que fue usada para la búsqueda primaria. Está conformada por los siguientes elementos:

TITLE-ABS-KEY("WordNet" AND ("construction" OR "making" OR "building" OR "implementation" OR "enhancing" OR "adding")) AND PUBYEAR > 2010

En la presente revisión sistemática se usó esta cadena en la página de librería digital Scopus. El análisis de la inclusión estuvo basado en el abstract, en el mismo título y año de publicación. Además, se tomaron 2 papers de las conferencias LREC 2008-2014 (Conferencia Internacional sobre Recursos de Lenguaje y Evaluación) debido a que al ser una conferencia sobre recursos lingüísticos se encontró varios papers donde se detallaba el uso, mejora o desarrollo de una WordNet. Según las indicaciones, las publicaciones debían ser no mayores a 5 años de antigüedad por lo que se buscó a partir del año 2011.

De esta forma fue posible recolectar varios artículos con diferentes formas de construir una WordNet en diferentes países. De la búsqueda en Scopus se obtuvo 12 resultados de los cuales 6 fueron usados. Los otros 6 fueron descartados debido a que su objetivo no fue el desarrollo de una WordNet, sino al uso de la misma en diversas aplicaciones. En algunos casos la construcción se llevó a cabo en su totalidad, en otros se dio el primer paso pero aún queda mucho por hacer.

En las siguientes subsecciones se procederá a detallar cada estudio recolectado.

3.3. Estudio N° 1: El uso de WordNet para la construcción de WordNets

Este artículo [Farreres et al., 1998] resume un conjunto de metodologías y técnicas para la construcción rápida de WordNets multilingües. La versión en inglés se utiliza como modelo para las demás en varias etapas del proceso de construcción.

El enfoque central es aplicado para la construcción de fragmentos sustanciales de WordNets de forma ordenada siguiendo la WordNet modelo. Se hacen 2 distinciones en el método de desarrollo: una para verbos y otra para sustantivos. En el caso de los verbos la mayor parte se hizo manualmente y la validación fue automatizada. Para el caso de sustantivos, se comparó una lista de conceptos en inglés con otra de 360 conceptos en español (al construir un modelo en español) preparados previamente de forma manual para alinearlas. Los conceptos faltantes se agregaron manualmente.

A partir de esta lista de conceptos alineados manualmente el método inicia seleccionando los principales términos (más relevantes) de una categoría semántica diferenciando géneros. Esto se realiza escogiendo conceptos únicos evitando sinónimos y antónimos. Luego, usando estas palabras se construye la taxonomía completa de una semántica primitiva (semántica compuesta de unidades cuyo significado no se puede descomponer en otras unidades). Se obtuvo un 99% de aciertos al usar este método.

El método se aplicó satisfactoriamente para construir la WordNet en español y en catalán. Se aplicó un conjunto de técnicas complementarias para vincular las palabras en español y catalán recogidas de varios diccionarios bilingües (en el caso de sustantivos) y léxicos (caso de los verbos) con la WordNet en Inglés. Mediante la metodología descrita anteriormente ya es posible construir taxonomías precisas de diccionarios electrónicos monolingües (están correctamente clasificados).

3.4. Estudio N° 2: Mejora de la WordNet japonesa

Este proyecto toma como base la WordNet en japonés que consta de 51 mil *synsets.* El objetivo fue desarrollar 3 métodos para extenderla.

El primer objetivo fue incrementar la cobertura. Una forma de hacerlo es corregir manualmente los *synsets* que se obtuvieron de forma automática. La otra forma a la que le ponen mayor interés, es agregando conceptos que no son expresados por la WordNet en inglés de Princeton pero que sí pueden aparecer en algunos textos.

El segundo comprendió el proceso de anotación de cuatro textos. Los dos primeros son traducciones de textos en inglés anotados en la WordNet (como definiciones), el tercero es el periódico japonés que forma el Corpus de Kyoto y el cuarto es un corpus abierto de frases bilingües en inglés-japonés.

El tercer objetivo es relacionar la WordNet con otros recursos como a un léxico japonés y a una colección de figuras de la librería OCAL (*Open Clip Art Library*). De esta forma si se busca un concepto como **herbívoro** la búsqueda retorna adicionalmente una figura de un herbívoro alimentándose.

Para lograr estos objetivos se cuenta con una interfaz gráfica como se muestra en la Figura 2 donde se puede buscar una palabra o código ili del synset directamente.

También se selecciona el idioma a buscar. Da el resultado tanto en japonés como en inglés.

Figura 2: Interfaz de usuario para ingresar datos a la WordNet Japonesa

Se espera que con estas mejoras la WordNet en japonés se convierta en un recurso útil no sólo para el procesamiento del lenguaje natural, sino también para el aprendizaje del lenguaje y para la investigación lingüística [Bond et al., 2009].

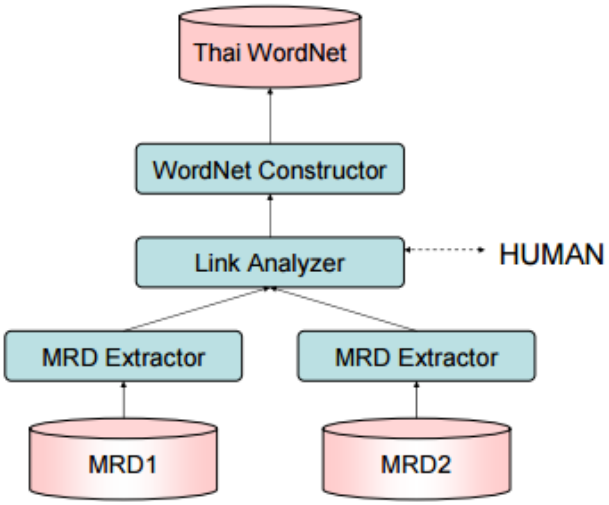
3.5. Estudio N° 3: Construcción de una WordNet para los verbos persas

Este artículo [Rouhizadeh et al., 2007] reporta un proyecto en desarrollo para construir una WordNet para los verbos persas. Considerando que la mayoría de los verbos en persa son compuestos (formado por más de un verbo. Ejemplo: "había estado") y altamente relacionados con otras partes del habla, las WordNets de verbos, sustantivos, adjetivos y adverbios persas tienen que ser construidos de forma paralela o en todo caso contar con mucha comunicación entre los autores. Para poder llevar a cabo este proyecto se contó con una base de datos en línea con más de 16 millones de palabras en persa actual que se usó como entrada. Realizaron una interfaz gráfica para ingresar y validar los datos.

Se plantea que a medida que vaya creciendo, esta WordNet sea evaluada en los próximos años de la siguiente manera: en primer lugar, se comparan los resultados con 3 diccionarios bilingües confiables. En segundo lugar, algunos expertos humanos deberían comprobar y evaluar los *synsets*. En tercer lugar, una vez terminado, utilizar la WordNet en algunas aplicaciones y evaluar los resultados. Por último, la WordNet tiene que ser comparada con otros léxicos construidos con otro enfoque.

3.6. Estudio N° 4: Construcción de WordNet tailandés basado en diccionarios electrónicos

En este trabajo [Sathapornrungkij y Pluempitiwiriyawej, 2005], los autores proponen un método semiautomático para construir una WordNet para el idioma Tailandés usando como recursos la WordNet del inglés (WordNet-Pr) y un diccionario tailandés. Para conseguir su objetivo, los autores diseñaron un sistema, llamado WordNet Builder, que soporte cada etapa del proceso de construcción. La arquitectura general del sistema es mostrada en la Figura 3.



MRD: Machine readable dictionary: diccionario electrónico

Figura 3: Esquema del constructor

La figura 3 presenta una visión general del sistema de generador WordNet. Tiene como entrada a los diccionarios de WordNet en inglés (MRD 1) y Lexitron (MRD 2) y como salida la WordNet tailandés en sí. Los componentes del sistema incluyen los extractores de MRD, el analizador de enlace (*link analyzer*) y el constructor WordNet. El analizador de enlace busca crear los enlaces entre palabras para crear los *synsets* de modo que queden alineados con sus análogos en inglés. Se toma en cuenta que una palabra al tener más de un sentido puede tener varias relaciones con otras palabras. Se verifica si los enlaces formados son correctos mediante un modelo de clasificación estadística. Esta verificación reduce el tiempo de intervención humana. Finalmente, luego de ser verificada ya está disponible para uso y mejora.

3.7. Estudio N° 5: Construcción de la WordNet de griego clásico

Este proyecto [Bizzoni et al., 2014] describe el proceso a seguir para construir una WordNet para griego clásico. En este caso para armar los *synsets* las palabras fueron extraidas de un diccionario griego-inglés partiendo del supuesto de que las palabras griegas traducidas por la misma palabra o frase en inglés tienen una alta probabilidad de ser sinónimos o al menos semánticamente estrechamente relacionados. También corresponde una alineación de sentidos con la WordNet en inglés de Princeton.

Si una palabra se consideraba inadecuada para incluirla en un *synset* pero está relacionada semánticamente (hipónimo, antónimo), entonces esta relación es insertada manualmente. Asimismo, una vez que se completan los *synsets*, estos son revisados manualmente para validarlos. Se contó con una interfaz web para los pasos anteriores.

Luego de la verificación, la WordNet griega fue lanzada *online* con su respectiva interfaz web. Aún se viene trabajando en su ampliación léxica y en la implementación de nuevas características como búsqueda bilingüe (buscar desde la WordNet en griego palabras en otro idioma y mostrar la relación) y el permitir que los usuarios contribuyan al seleccionar el mejor sentido para una palabra según un contexto especifico. Además, ya se cuenta con validación por un usuario maestro y restricciones de acceso y edición.

3.8. Estudio N° 6: Relación semántica basada en corpus para la Construcción del WordNet polaco

Este proyecto [Broda et al., 2008] concentra sus esfuerzos en construir una WordNet para el idioma polaco. Para iniciar, agruparon material léxico teniendo en cuenta las relaciones semánticas y categorías gramaticales.

Para armar los *synsets*, el método usado consiste en asignar una medida a la relación semántica que existe entre palabras como verbos o adjetivos. Las restricciones usadas sintácticamente se despliegan en un amplio corpus morfológicamente escrito de Polonia.

Para evaluar el método se realizó una prueba de similitud con la WordNet y se reforzó la validación con el apoyo de evaluadores humanos. Un lexicógrafo también evaluó manualmente una muestra adecuada de sugerencias. Los resultados se comparan favorablemente con otros métodos conocidos de adquisición de relaciones semánticas.

3.9. Estudio N° 7: Adición de relaciones semánticas a la WordNet Rumana

En este proyecto [Mititelu, 2012] el autor presenta retos planteados por el sistema rumano y su metodología para la identificación de palabras derivadas en la WordNet rumana. En base a estos retos, se plantea marcar las relaciones de derivación (o morfológicas) entre las palabras existentes en la WordNet rumana y agregarles una etiqueta semántica que tiene validez en varios idiomas. Asimismo, también se desea agregar relaciones semánticas.

La WordNet rumana cuenta con 57895 *synsets* donde se establecen 120198 relaciones. Estos resultados se obtuvieron del trabajo en diversos proyectos nacionales e internacionales entre 2000 y 2011, con los últimos desarrollos dentro de un proyecto en donde ya se documentó toda la información correspondiente.

Para lograr los objetivos, el método usado consiste en que primero se juntan pares (palabras de la WordNet rumana y sus afijos) mediante heurísticas para luego verificarlos tanto de forma automática como manual. En cuanto a prefijos, se obtuvieron 2862 pares. Tras la adición de sufijos, se obtuvieron 13556 pares. La explicación viene del hecho de que el rumano tiene un mayor número de sufijos que de prefijos por lo que se les pone mayor énfasis. Los miembros de los pares correctos están unidos entre sí y la relación se asocia a una etiqueta semántica siempre que sea necesario. Se demostró que esta etiqueta tiene validez entre lenguajes. De esta forma se contribuye al aumento del número de las relaciones entre *synsets*. Además, palabras que pertenecen a una misma familia léxica se identifican más fácilmente.

Se identificaron tres niveles en los que es importante marcar las relaciones semánticas. En primer lugar, a nivel monolingüe (netamente enfocado a la lengua rumana), la densidad de las relaciones en una WordNet puede aumentar entre las palabras con un sentido, pero sobre todo entre palabras de varios sentidos.

En segundo lugar, a nivel multilingüe, las etiquetas semánticas asociadas con las relaciones de derivación se establecen a nivel de *synsets*, de modo que los conceptos podrían ser transferidos de una WordNet a otra, siempre que estén alineadas entre sí. A mayor número de WordNets con tales relaciones, más estudios comparativos se pueden hacer. Se puede analizar cómo una cierta relación semántica se realiza morfológicamente en varios idiomas, qué expresan los afijos, etc.

Tercero, a nivel de aplicaciones, una WordNet enriquecida con estas relaciones descritas se convierte en una base de conocimientos útiles para diversas tareas tales como la búsqueda de respuestas, recuperación de información o traducción automática.

Se plantea que en el futuro se podría adaptar estas herramientas para marcar estas relaciones en el momento en que se implementan nuevos *synsets* en rumano de forma automática.

3.10. Estudio N° 8: Desarrollo automático de Wordnets de idiomas de bajos recursos utilizando desambiguación lingüística

Este proyecto [Taghizadeh y Faili, 2016] tiene como premisa que las WordNets son un recurso eficaz para el procesamiento del lenguaje natural y la recuperación de información, especialmente para tareas de procesamiento semántico. Sin embargo, el desarrollo automático de WordNets para lenguas de bajos recursos no ha sido bien estudiado. Es por ello que este proyecto tiene como objetivo el desarrollo de un algoritmo EM (Expectation–maximization) para crear WordNets de alta calidad y de gran escala para lenguas de bajos recursos. El método propuesto permite desarrollar una WordNet usando solo un diccionario bilingüe y un corpus monolingüe. En este artículo se toma como base la WordNet en inglés y se busca desarrollar este recurso para la lengua persa.

En general y sin tener en cuenta el enfoque adoptado, el paso principal hacia la construcción de una WordNet es generar los *synsets*. Primero la WordNet se inicializa con *synsets* previamente definidos. Para cada *synset*, todas las traducciones

de palabras en inglés se extraen del diccionario bilingüe y los vínculos entre las traducciones y los *synsets* se establecen. Dado que los diccionarios traducen palabra por palabra, las traducciones son ambiguas. Por lo tanto, la tarea es anotar enlaces y encontrar las incorrectas.

La idea propuesta es utilizar un algoritmo iterativo que encuentra el óptimo local del problema con pocas iteraciones en un tiempo razonable. En primer lugar, para cada palabra persa en el corpus, todas las traducciones son extraídas del diccionario bilingüe. A continuación, todos los *synsets* de las traducciones inglesas son considerados como los *synsets* candidatos para la palabra persa.

Una puntuación se calcula para cada par de palabras en persa y *synsets* utilizando el algoritmo EM. En esta etapa, utilizan un método de desambiguación, en el que la frecuencia de co-ocurrencia de pares de palabras en la lengua persa se ha utilizado para eliminar la ambigüedad de palabras de un corpus. Los resultados experimentales mostraron que la precisión de este método varía para diferentes categorías gramaticales. La máxima precisión se muestra para los adjetivos, que es del 89,7%; al lado de los adverbios, que es del 65,6%; y la precisión es más baja para los sustantivos al 61,6%. El problema de este enfoque es que se necesita un corpus de gran escala que normalmente no se tiene en lenguas de bajos recursos ya que de esto depende la calidad de la WordNet resultante. Debido a esta razón es que se hace un cambio en esta etapa.

El algoritmo EM debe encontrar la probabilidad de mapear cada palabra en el idioma de destino a cada uno de sus *synsets* candidatos. Si un *synset* candidato representa un sentido correcto de una palabra en el idioma de destino, se espera que este sentido se produzca en un corpus que contiene esa palabra. Así que los datos observados son las palabras de un corpus en el idioma de destino; la parte invisible de cada dato es la etiqueta del sentido de las palabras en la WordNet. El algoritmo EM cambia entre dos etapas: En la primera se declara una distribución aproximada de los datos que faltan dados los parámetros. La segunda es la búsqueda de mejores parámetros dados a la aproximación. La precisión fue de 90% según juicio manual pero de 18% según una fuente externa. El número es bajo debido a la importancia del número de palabras a considerar para la elaboración de la WordNet.

3.11. Conclusiones sobre el estado del arte

Luego de presentar distintos proyectos de la construcción de una WordNet en distintas realidades se puede observar que estas nuevas WordNets se basan en la original en ingles de Princeton a la cual toman como patrón. Se busca que las nuevas WordNets tengan una estructura similar, tengan un método para armar los *synsets* y que estos queden alineados con sus análogos en inglés. También se observa en todos los casos que se necesita algún recurso léxico de entrada propio del idioma sobre el que se quiere desarrollar.

De este modo, mediante este proyecto se desea construir una WordNet en Shipibo-Konibo para contribuir con esta comunidad de Perú en su integración con el estado, otras comunidades y con todo el mundo a través del internet.

4. Extracción de datos del diccionario de sentidos

4.1. Introducción

El presente capítulo trata sobre los proceso de obtención y extracción de datos de un diccionario Shipibo-Konibo – Español digitalizado para guardar los datos contenidos. De esta manera, quedarían listos para guardar estos datos temporalmente en una base de datos (será nombrada como base de datos del diccionario) para luego clasificarlos en el *synset* que les corresponda tomando como base la WordNet en español y así formar la WordNet en Shipibo-Konibo.

A continuación se explicará cómo se logró dar solución al primer objetivo específico: Digitalizar y pre-procesar un diccionario bilingüe Español – Shipibo-Konibo.

Se tienen dos resultados esperados: i) Algoritmo para separar los elementos del diccionario como términos, sentidos, categorías gramaticales y ejemplos de uso. ii) Una base de datos para almacenar los elementos del diccionario.

4.2. Resultado N° 1: Algoritmo para separar los elementos del diccionario como términos, sentidos, categorías gramaticales y ejemplos de uso.

El diccionario en físico ha sido proporcionado en formato digital (PDF) por el grupo de investigación a cargo del proyecto FONDECYT del traductor automático entre Shipibo-Konibo y Español. Este diccionario también se encuentra como imagen en la web del Instituto Lingüístico de Verano (*Summer Institute of Linguistics* - SIL[2]).

Dado que el diccionario se encontraba en formato PDF[3] escaneado como imágenes, se tuvo que realizar un procesamiento del archivo para poder extraer las definiciones que contenía. Junto con el grupo de investigación, se dividió el diccionario entre 10 miembros del grupo en archivos de texto plano para proceder a su corrección manual. Esto ocurrió debido a las fallas de la digitalización OCR[4], ya que hubieron muchos errores en caracteres como signos de puntuación o letras con tildes por lo que la corrección era necesaria. El error más común fue el confundir signos de puntuación como corchetes, llaves y paréntesis. Luego se juntaron todos los archivos para tener el diccionario completo como texto.

Luego de la corrección del diccionario, se codificó un algoritmo en lenguaje Java que identifique los patrones encontrados en el diccionario para extraer la información de todas las entradas. En la introducción del diccionario se explican las reglas que componen la estructura general para cualquier palabra pero no se mencionan todas las excepciones lo que dificultó esta tarea. A continuación se presenta la estructura general del diccionario:

[2]SIL: http://www.peru.sil.org/

[3]PDF: *Portable Document Format*. Formato de Documento Portátil que se ve como un documento impreso.

[4]OCR: *Optical Character Recognition*. Reconocimiento óptico de caracteres.

- Término introductorio: el término se encuentra como forma básica de la palabra
- Referencia en lugar de todo lo demás (casos de subentradas)
- Variante del término introductorio (opcional)
- Categoría gramatical
- Parte principal: Forma no básica del término introductorio. Para el caso de un sustantivo, pronombre, adjetivo o adverbio se presenta con prefijo, sufijo u otra variante mientras que para un verbo se tiene su participio pasado.
- Variante de la parte principal (opcional)
- Etimología (opcional): Explica el origen castellano o quechua del término introductorio o de la subentrada en caso exista.
- Sentidos (por lo menos un sentido)
    - Glosas (por lo menos una glosa por sentido)
    - ilustración verbal (se refiere a un ejemplo de uso, opcional)
    - nota de uso (Texto aclarativo sobre el uso de una palabra, opcional)

- Párrafo de sinónimos (opcional): Una lista de sinónimos escritos del término introductorio.
- Lista de subentradas: (opcional)
    - Cada una tiene por lo menos una definición.
    - Puede tener una ilustracion verbal.
    - Hay 2 tipos de subentrada:
        - Sustantivos compuestos. Especies de animales.
        - Frases que incluyen la palabra del artículo.

Por ejemplo, al buscar la palabra '*sároranti*' en el diccionario se tiene lo siguiente:

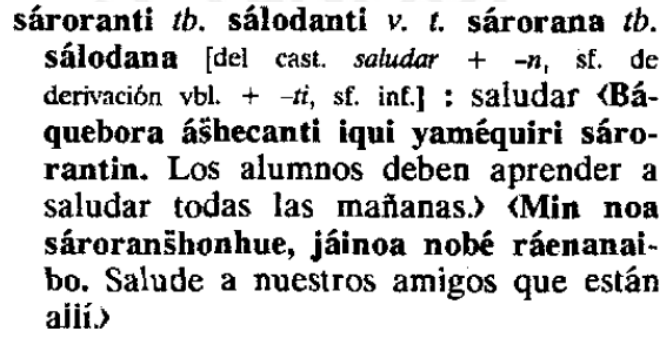
- Término introductorio: *sároranti*
- Referencia: no se encuentra
- Variante del término introductorio: *sálodanti*
- Categoría gramatical: verbo transitivo
- Parte principal: *sárorana*
- Variante de la parte principal: *sálodana*
- Etimología: [del cast. saludar + -n, sf. de derivación vbl. + -ti, sf. inf.]
- Sentidos:
    - Glosas: saludar.
    - ilustración verbal: <Báquebora áxecanti iqui yaméquiri sárorantin. Los alumnos deben aprender a saludar todas las mañanas.>
    - nota de uso: no se encuentra

- Párrafo de sinónimos: sinón: *johuéati*.
- Lista de subentradas: no se encuentran subentradas.

Por cada término que el algoritmo procesó se obtuvo una línea como salida que constaba de todos sus datos identificados y separados en campos por un símbolo especial (se hizo uso del símbolo /). De esta forma se facilitó la inserción a la base de datos ya que se conoce a priori lo que significa cada campo. Si está vacío no se considera y si no lo está se insertará el dato en la tabla respectiva. La estructura usada de la salida es:

- Se inicia con estos 9 campos que son únicos: *Término introductorio/ referencia/ tipo de variante del termino/ variante/ categoría gramatical/ parte principal/ tipo de variante de parte principal/ variante de parte principal/ etimología/*
- Ya que un término puede tener más de un sentido, se necesita guardar los datos de cada uno. Se comprobó que el máximo número de sentidos de una palabra encontrada en el diccionario es 12. Por lo tanto, este patrón de campos se repite 12 veces seguidas al ser el número máximo de sentidos: *6 glosas/ 2 Nota de uso/ 3 Ejemplos/.* Da un total de 11 campos por cada sentido.
- Al último quedan 3 campos que son únicos: *Párrafo de sinónimos/ término padre/ tipo de subentrada.*

Las subentradas de un término son representadas en una línea aparte. Si es una entrada principal, entonces los 2 últimos campos quedan vacíos. Si es una subentrada se tiene el término principal como termino padre y según este se obtiene el tipo de subentrada. Por ejemplo, en el diccionario se tiene lo siguiente:



sároranti *tb.* **sálodanti** *v. t.* **sárorana** *tb.* **sálodana** [del cast. *saludar* + *-n*, sf. de derivación vbl. + *-ti*, sf. inf.] : saludar ‹**Báquebora áshecanti iqui yaméquiri sárorantin.** Los alumnos deben aprender a saludar todas las mañanas.› ‹**Min noa sároranshonhue, jáinoa nobé ráenanaibo.** Salude a nuestros amigos que están allí.›

Figura 4: Imagen escaneada del diccionario en formato pdf

La salida del término '*sároranti*' mostrado sería:

sároranti//tb./sálodanti/v. t./sárorana///del cast. saludar + -n, sf. de derivación vbl. + -ti, sf. inf./: saludar////////Báquebora áxecanti iqui yaméquiri sárorantin. Los alumnos deben aprender a saludar todas las mañanas./Min noa sároranxonhue, jáinoa nobé ráenanaibo. Salude a nuestros amigos que están allí.//
//johuéati//

Hay algunos separadores juntos ya que entre ellos se encontrarían otros elementos si los hubiera. Para el caso presentado solo hay una definición por lo que se encuentran muchos separadores juntos ya que se espera hasta 12 sentidos y el término mostrado tiene solo uno.

4.3. Resultado N° 2: Base de datos del diccionario

Figura 5: Vista completa de la base de datos del diccionario.

Se necesitaba contar con una estructura de base de datos que permita almacenar los datos del diccionario digitalizado que ya fueron extraídos por el algoritmo presentado en el primer resultado.

Este resultado tuvo como finalidad construir una base de datos con una estructura que permita que sus sentidos sean utilizados para luego almacenarlos siguiendo la estructura de una WordNet. En la Figura 5, se pueden apreciar seis tablas, las cuales conforman la vista completa del diccionario en la base de datos del diccionario. A continuación, se realizará una descripción de las tablas que lo conforman:

La tabla “Termino” es un maestro que contiene información de cada una de las entradas del diccionario que pueden ser palabras, prefijos o sufijos, como la etimología, categoría gramatical y variantes del término.

- idTermino: llave primaria.
- termino: entrada o término introductorio
- etimología: descripción
- abrevVariante: categoría gramatical
- variante: variante del término
- abrevPartePrincipal: tipo de variante
- partePrincipal: parte principal de la entrada
- abrevVariantePP: tipo de variante de parte principal
- variantePP: variante de parte principal
- parrafoSinonimos: sinónimos asociados a la entrada
- tipoReferencia: tipo de referencia
- referencia: referencia a otro termino en caso de ser subentrada

La tabla “Relacion” muestra la relación que tiene cada una de las entradas con sus subentradas, mostrando la relación entrada-subentrada como una relación padre-hijo, respectivamente.

- padreIdTermino: id del termino padre

- hijoIdTermino: id del termino subentrada

La tabla “Definicion” indica el número de definiciones o sentidos equivalentes de la palabra en Shipibo-Konibo en español que tiene una entrada del diccionario.

- idDefinicion: llave primaria
- definicionNro: Número para diferenciar distintas definiciones de un mismo término
- termino_idTermino: Término asociado

La tabla “Glosa” muestra la información de cada uno de los sentidos equivalentes que tiene cada definición de la palabra en Shipibo-Konibo, que puede ser una sustitución de una palabra en castellano o una explicación del sentido con una frase en castellano, especificando también su categoría gramatical en cada caso.

- idGlosa: llave primaria
- glosa: cadena que contiene la glosa
- especif-tb: si es tb. Significa que es otra forma de expresar la misma idea. Si es especif. se refiere a algo más específico.
- catGramatical: categoría gramatical.
- definicion_idDefinicion: llave foránea. Id de definición que contiene a esta glosa.

La tabla “Ilustracion” registra los ejemplos o ilustraciones verbales que puede tener cada una de las definiciones de las entradas, que puede ser una palabra, frase u oración que se ofrece como ejemplo.

- idIlustracion: llave primaria
- ilustracion: cadena que contiene la ilustración o ejemplo.
- definicion_idDefinicion: llave foránea. Id de definición que contiene a esta ilustración.

La tabla “Nota” muestra las notas de comentario que puede tener una definición o un término directamente. Estas pueden referirse a una orientación a la gramática referente a la palabra, a su uso regional o cultural. En el caso de los afijos funciona como sustituto de la definición.

- idNota: llave primaria
- nota: cadena que contiene la nota
- definicion_idDefinicion: llave foránea. Id de definición que contiene a esta nota.
- termino_idTermino: llave foránea. Id de término que contiene a esta nota.

5. Algoritmo de clasificación de *synsets*

5.1. Introducción

El presente capítulo aborda la implementación de un algoritmo de clasificación que permita construir toda la estructura interna de la WordNet.

A continuación se explicará cómo se logró dar solución al segundo objetivo específico: Implementar un algoritmo de clasificación que permita alinear cada sentido de cada palabra del diccionario Shipibo-Konibo con el *synset* de igual significado en español.

Se tienen dos resultados esperados: i) Algoritmo de clasificación que permita ubicar el *synset* correspondiente en español a un sentido de cada palabra en Shipibo-Konibo. ii) Base de datos para almacenar los *synsets* en Shipibo-Konibo, las relaciones existentes entre ellos y las alineaciones entre *synsets* en Shipibo-Konibo y español; es decir, la WordNet.

5.2. Resultado N° 3: Algoritmo de clasificación que permita ubicar el *synset* correspondiente en español a un sentido de cada palabra en Shipibo-Konibo.

Este algoritmo permite crear la estructura de la WordNet para que luego sea almacenada de forma permanente en una base de datos.

En primer lugar, fue necesario contar con la WordNet en español. Como se detalló en la sección de Recursos y herramientas, se utilizó la *Multilingual Central Repository* (MCR), la cual servirá como base para realizar la clasificación y está disponible para ser descargada de forma libre[5].

Luego, se deben modelar las clases necesarias para manejar tanto las palabras del diccionario en Shipibo-Konibo como los *synsets* de la WordNet en español. Para ello, se crearon las clases Término y Sentido. Un término puede tener varios sentidos y cada sentido puede tener varias glosas asociadas, las cuales son oraciones o frases que expresan una definición. Las glosas también pueden dar más detalle o usar sinónimos de otras glosas para expresar el sentido Por ejemplo, el término 'tratar' tiene los siguientes sentidos obtenidos de la WordNet en español:

- Cuidar: [proveer tratamiento para; "El doctor trató la lesión de mi pierna"]
- Procesar – Alisar: [Sujetos a un proceso o tratamiento, a menudo con el fin de leer para algún propósito; "Queso tratado o procesado"; "Agua tratada"]
- Manejar: [Realizar operaciones matemáticas y lógicas en (datos) de acuerdo con las instrucciones programadas con el fin de obtener la información requerida; "Los resultados de las elecciones aún estaban siendo tratados cuando dio su discurso de aceptación"]
- Abordar: [Ocuparse de manera verbal o por alguna forma de expresión artística; "Este libro se ocupa de incesto"; "El curso trató sobre toda la civilización occidental"]

[5] Descarga disponible en: http://adimen.si.ehu.es/web/MCR/. Consultado el 28 de Setiembre de 2016.

Una vez extraída toda la información que se necesitaba de la base de datos del diccionario (términos, sentidos y glosas) se necesitó un modelo para comparar adecuadamente las glosas de las palabras en Shipibo-Konibo y obtener de alguna forma la cercanía de significado entre los sentidos de cada palabra en Shipibo-Konibo y los datos de la WordNet en español. Para determinar a que *synset* pertenece cada sentido se tomó en cuenta términos, glosas y ejemplos de cada *synset* en español. Para su representación se utilizó el modelo Word2Vec, el cual será detallado a continuación:

Para este proyecto, el modelo usado es el de Word2Vec que ha sido detallado en la sección 2. Este modelo fue entrenado en base a un corpus general [Cadellino, 2016] no anotado de la lengua española compuesto por aproximadamente 1.5 mil millones de palabras. El corpus fue creado compilando varios recursos de la lengua española que se pueden encontrar en Internet. A continuación se listan algunos de ellos:

- Colección de textos jurídicos en español de la Unión Europea
- Documentos de las Naciones Unidas
- Partes de corpus en español en otros idiomas
- Artículos de la Wikipedia en español

Word2Vec es usado para agrupar palabras detectando similitudes en los vectores que las representan matemáticamente. Esta similitud se puede obtener al calcular el coseno de dos palabras representadas como vectores. Word2Vec puede usarse en aplicaciones que involucren gustos, gráficos de medios sociales, traducción automática y otras series verbales o simbólicas en el que los patrones se pueden discernir [DLJ, 2016].

En cuanto al procesamiento del corpus se procedió a reemplazar todos los caracteres no alfanuméricos con espacios en blanco excepto comillas y guiones ya que se usan para ciertas palabras según el idioma para evitar errores y trabajar solo con las palabras. Luego todos los múltiples espacios en blanco con sólo un espacio en blanco. La capitalización de las palabras permaneció sin cambios. El corpus original tenía la siguiente cantidad de datos:

- Un total de 1420 millones de palabras
- 46 millones de oraciones
- 3.8 millones de palabras únicas

Después, se aplicó el modelo skip-gram de Word2Vec, el cual realiza la clasificación de una palabra en base a otra palabra en la misma frase. Más precisamente, se utiliza cada palabra como una entrada a un clasificador de lineal y predice las palabras dentro de un rango determinado antes y después de la palabra actual. Se encontró que el aumento del rango tomado aumenta la complejidad computacional pero tiene el beneficio de que mejora la calidad de los vectores de palabras resultantes. Esto se debe a que las palabras más distantes son por lo general menos relacionadas con la palabra actual, así se le da menos peso a las palabras distantes mediante el muestreo de menos de dichas palabras en los ejemplos de entrenamiento [Mikolov et al., 2013].

Para aplicar Word2Vec se filtraron las palabras con menos de 5 ocurrencias y se descartaron las 273 palabras más comunes ya que se consideran irrelevantes para entender el contexto. Algunos ejemplos de estas palabras son: ya, que, para, por, entre otras. Finalmente, se obtuvieron los siguientes valores:

- Un total de 771 millones de palabras
- 1 millón de palabras únicas.

Con el modelo de Word2Vec entrenado, se procedió con el algoritmo de clasificación. A continuación se presenta el pseudocodigo del algoritmo:

```
1. Para cada término en Shipibo-Konibo
2.    Para cada sentido de un término en Shipibo-Konibo
3.          máximo = 0;
4.          Para cada synset de la WordNet en Español
5.                similitud = Hallar_Similitud_Word2Vec (sentido, synset)
6.                Si (similitud > máximo) entonces
7.                      máximo = similitud
8.                      synset_encontrado = sentido.ObtenerSynset()
9.          fin_para
10.         insertar_Base_Datos (sentido, synset_encontrado)
11.   fin_para
12. fin_para
```

Según el pseudocódigo, este algoritmo funciona de la siguiente manera: Cada sentido de cada palabra encontrada en el diccionario de Shipibo-Konibo fue comparada con cada *synset* de la WordNet en español mediante un modelo de Word2Vec previamente entrenado para obtener la similitud, que fue expresada como un número decimal. Para este cálculo se tomó la misma palabra, las glosas y ejemplos correspondientes, si los hubiera, de cada palabra en español. Se filtraron las glosas para tomar solo las palabras de las categorías que surgen con mayor frecuencia y que permiten entender el contexto. Estas categorías son sustantivos y verbos; y en menor medida, adjetivos y adverbios. Por ejemplo, para el caso de un sustantivo se tomaron los sustantivos y verbos de su glosa (debido a la fuerte conexión que existe entre ambas categorías). Se tomó cada palabra una sola vez eliminando repeticiones ya que podría darse el caso de que se repitan los términos o palabras contenidas en las glosas. Además, se usó un Tree-tagger [Schmid, 1994] para extraer las categorías gramaticales y lemas de las palabras. Por ejemplo, el proceso para la palabra "*manéxti"* se tomó el sentido cuya glosa es "*amarrar los pelos de la cabeza o la coronilla con otros pelos de su misma cabeza*" sería de la siguiente manera al pertenecerle a un verbo:

{*Amarrar, los, pelos, de, la, cabeza, o, la, coronilla, con, otros, pelos, su, misma, cabeza, a*}

{*Amarrar, pelos, cabeza, coronilla, pelos, cabeza*} - se filtraron solo verbos y sustantivos

{Amarrar, pelo, cabeza, coronilla} - se lematizaron las palabras y borraron repeticiones

La similitud fue expresada como un número decimal entre -1 y 1 (al ser resultado de un coseno). Si el resultado es mayor significa que la relación entre palabras es más cercana. Por lo tanto, para decidir a cuál *synset* corresponde la palabra en Shipibo-Konibo se toma el de mayor similitud. En la Figura 6 se observa un esquema de este proceso. Se toma cada sentido de las palabras en Shipibo-Konibo (como "*tiscóti*" en la figura) y se procede a comparar su similitud con cada *synset* en español. Se obtiene un resultado por cada *synset* y se toma el mayor porque significa mayor similitud. En el caso de la Figura 6 sería S1 {derramar, salpicar} ya que el resultado es mayor que los demás.

Cada vez que se encuentre el *synset* correspondiente a una palabra en Shipibo-Konibo se insertó en la base de datos todo lo relacionado a ella siguiendo el estándar usado por MCR. También se guardaron las relaciones con otros *synsets* que fueron heredadas de la WordNet en español. Esta relación se mantiene en la mayoría de los casos pero de igual forma se hará una verificación manual.

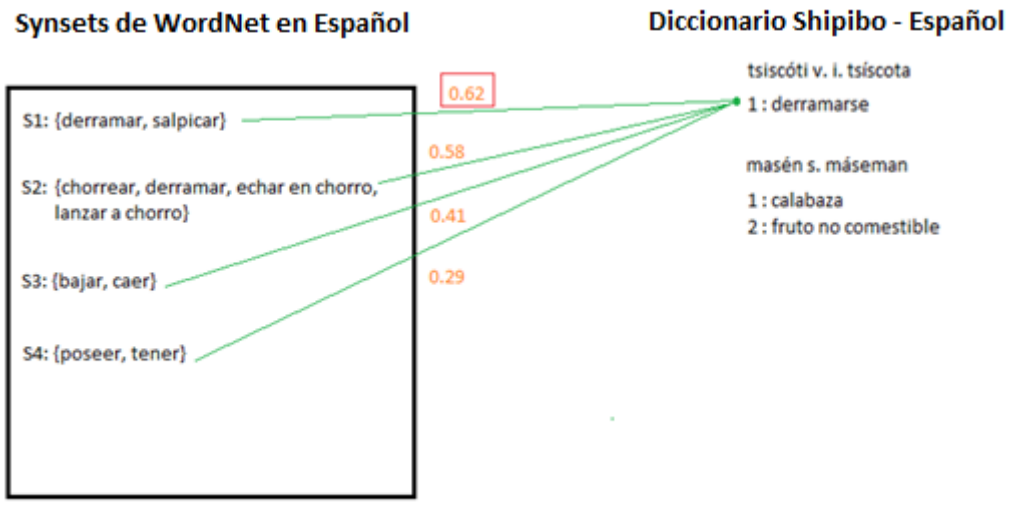


Figura 6: Esquema del algoritmo de clasificación.

Una vez que se ejecutó el algoritmo con dos palabras en Shipibo-Konibo como muestra, se obtuvo el *synset* en español más cercano el cual fue hallado por el algoritmo de clasificación descrito. En la Figura 7 se muestra la clasificación de dos palabras. Por ejemplo, *mánocoxoti* (se traduce a español como limpiar) fue clasificado correctamente a uno de los *synsets* donde se encuentra la palabra limpiar

en el mismo sentido que ordenar o arreglar. También se muestran todas las palabras asociadas al *synset*, así como su código ili (usado como estándar internacionalmente).

```
Palabra en shipibo: mánocoxoti
Primera palabra en español: limpiar
Codigo Synset: spa-30-00181664-v
Contenido Synset: {arreglar, despejar, limpiar, ordenar, recoger, retirar}

Palabra en shipibo: cahuánti
Primera palabra en español: pasar
Synset: spa-30-02230772-v
Contenido Synset: {ceder, dar, entregar, pasar, traspasar}
```

Figura 7: Resultados de ejecución de código en Netbeans

5.3. Resultado N° 4: Base de datos para almacenar los *synsets* y alineaciones.

Una vez implementado y ejecutado el algoritmo de clasificación, se necesitó contar con una estructura de base de datos que permita almacenar los datos de las palabras y *synsets* que componen la WordNet. Los nombres de tablas y campos siguen el estándar usado por la MCR. Entre los datos se encuentra el código estándar (ili: *interlingual index*) que se le da al *synset* para identificar a un mismo *synset* en diferentes idiomas. Por ejemplo, 'spa-30-00001740' y 'eng-30-00001740' hacen referencia a un mismo *synset* (se distingue porque el número dentro del código es el mismo) pero el primero está en español (empieza con spa) y el segundo en inglés (empieza con eng). Para el Shipibo-Konibo se usará la abreviación shk.

En la Figura 8, se pueden apreciar cuatro tablas más importantes que conforman la WordNet. En ellas se puede identificar las palabras, sus sentidos o ejemplos si los hubiera y a cual *synset* pertenecen. A continuación, se realizará una breve descripción de las tablas que lo conforman:

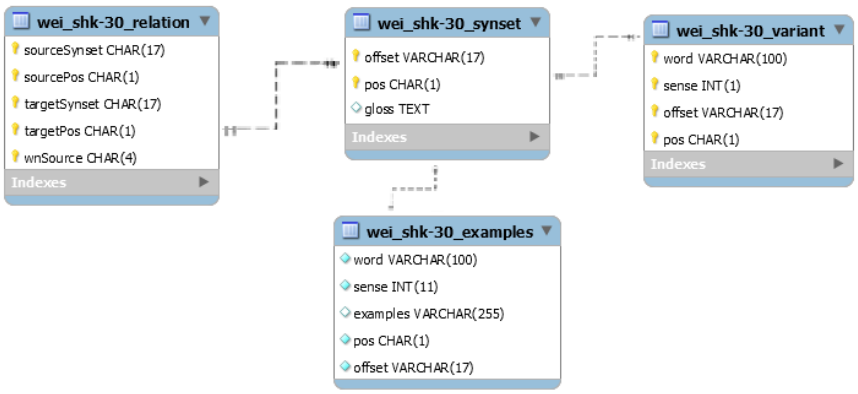


Figura 8: Vista parcial de la WordNet para Shipibo-Konibo

Tabla wei_shk-30_synset: Indica la relación de *synsets* identificados por un código único. Cada uno tiene una categoría gramatical y glosa asociada a dicho *synset*. Las categorias gramaticales son: sustantivos (n), verbos (v), adjetivos (a) y adverbios (adv). A continuación, se presenta un ejemplo en la Tabla 4 de una fila registrada en la tabla wei_shk-30_synset:

Synset	Categoría	Glosa
shk-30-01009240-v	v	expresar con palabras

Tabla 4: Ejemplo de fila registrada en la tabla de *synsets*

Tabla wei_shk-30_examples: Muestra ejemplos del uso de los términos. No es necesario que todos los términos tengan un ejemplo pero facilita para entender el uso de una palabra. Cada término o palabra puede tener varios ejemplos. En la tabla 5 se presenta un ejemplo de fila para la tabla de ejemplos:

Palabra	Sentido	Ejemplo	Categoría	Synset
yóiti	1	Janra ea yóique: -Cátima. El me dijo: -No te vayas.	v	shk-30-01009240-v

Tabla 5: Ejemplo de fila registrada en la tabla de ejemplos

Tabla wei_shk-30_variant: Indica cada sentido de cada término. Se identifica por número de sentido y *synset* correspondiente. A continuación, se presenta un ejemplo en la Tabla 6 de una fila registrada en la tabla wei_shk-30_variant:

Palabra	Sentido	Synset	Categoría
yóiti	1	shk-30-01009240-v	v
yóyo	1	shk-30-01009240-v	v

Tabla 6: Ejemplo de fila registrada en la tabla de variantes

Tabla wei_shk-30_relation: Integra las relaciones que existen entre los *synset*. A continuación, se presenta un ejemplo en la Tabla 7 de una fila registrada en la tabla wei_shk-30_relation:

Synset origen	Categoría Origen	Synset destino	Categoría Destino
shk-30-00001740-a	a	shk-30-05200169-n	n
shk-30-00001740-a	a	shk-30-05616246-n	n

Tabla 7: Ejemplo de fila registrada en la tabla de relaciones

Después de haber procesado la base de datos de la WordNet del Shipibo-Konibo, se procedió a obtener las estadísticas de la misma.

Del total de 5800 palabras almacenadas, 4815 corresponden a las principales clases gramaticales, que son sustantivos, verbos, adjetivos y adverbios. Las 985 restantes son sufijos, prefijos, conjunciones, preposiciones, entre otras clases gramaticales. El número de palabras por sentido por cada clase gramatical de la WordNet en Shipibo-Konibo se muestra en la Tabla 8:

Categoría \ N° Sentidos	Sustantivos	Verbos	Adjetivos	Adverbios
1	2231	1453	357	96
2	174	266	62	11
3	59	48	13	2
4	14	16	1	0
5	6	3	0	0
6	2	0	0	0
7	0	0	0	1
Total	2486	1786	433	110

Tabla 8: Número de palabras por sentido para cada categoría gramatical en la WordNet del Shipibo-Konibo

De los términos procesados, se obtuvieron 5134 *synsets* en total, siendo 2528 pertenecientes a los sustantivos, 1967 a los verbos, 518 a los adjetivos y 121 a los adverbios.

A fin de verificar la dimensión y distribución de la WordNet del Shipibo-Konibo, fue realizada una comparación con la WordNet del Español. Para realizar la comparación se mostrarán los mismos datos de la WordNet en español. La cantidad de palabras por sentido por cada clase gramatical en la WordNet del español es mostrada en la Tabla 9:

Categoría N° Sentidos	Sustantivos	Verbos	Adjetivos	Adverbios
1	18147	3582	3958	398
2	5935	1607	861	209
3	1538	613	238	53
4	583	243	77	12

5	180	121	24	3
6	84	43	15	1
>6	84	42	7	1
Total	26551	6251	5180	677

Tabla 9: Número de palabras por sentido para cada categoría gramatical en la WordNet del Español

Para realizar la comparación, las distribuciones de la cantidad de palabras por sentido por cada clase gramatical en las WordNet del Shipibo-Konibo y el Español son presentadas en las Figuras 9 y 10.

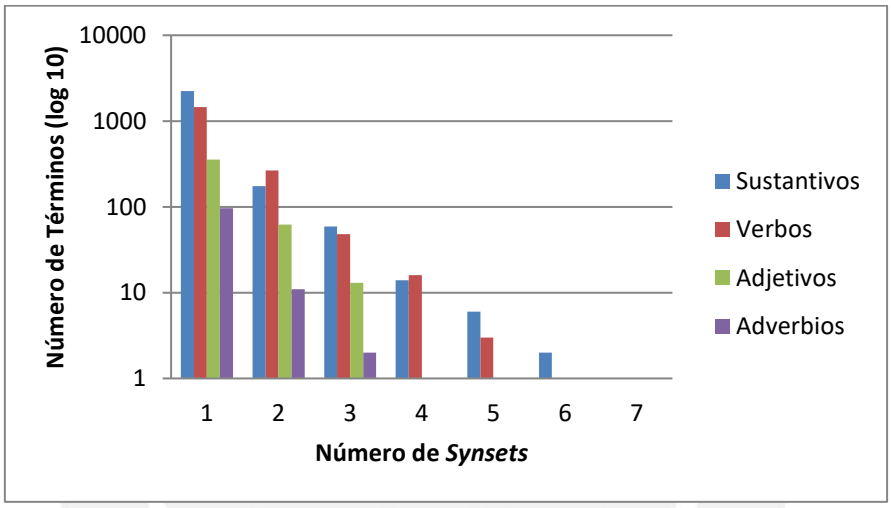


Figura 9: Distribución de las palabras por *synset* por cada clase gramatical en el Shipibo-Konibo (escala logarítmica).

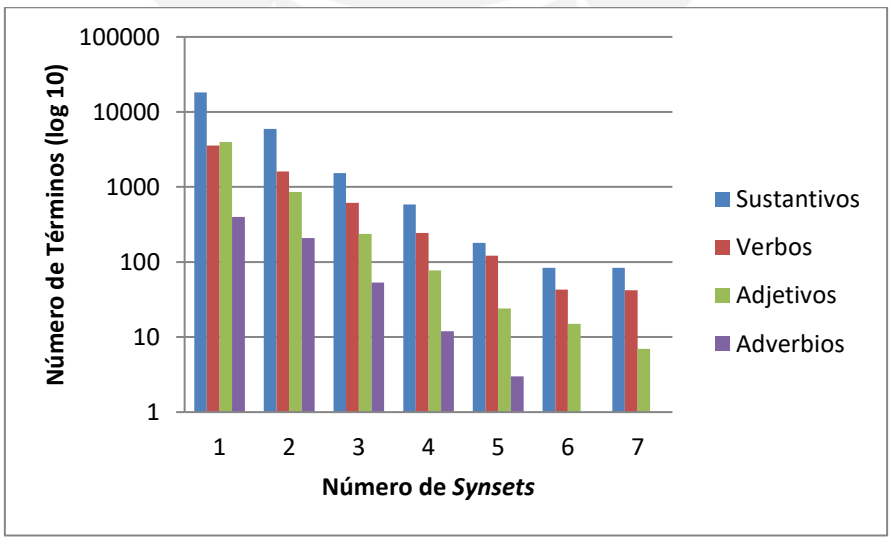


Figura 10: Distribución de las palabras por *synset* por cada clase gramatical en el Español (escala logarítmica).

Pese a las diferencias entre las cantidades de palabras (o términos) y en la cantidad de *synsets*, existen similitudes en la distribución de las palabras por sentidos en cada clase gramatical. Por ejemplo, la cantidad de palabras que poseen un sentido es mayoritaria en todas las clases tanto en el español como en el Shipibo-Konibo. Un detalle a resaltar es que en el caso del Shipibo-Konibo, los verbos poseen una mayor variación de significados comparándolo con el Español donde los sustantivos tienen la mayor variación de significados entre las clases gramaticales. Con estos resultados se podrá realizar la comparación con el gold standard mencionado anteriormente.

- Gold Standard

El *gold standard* o estándar de oro fue hecho por lingüistas y un hablante nativo de Shipibo-Konibo quienes seleccionaron un grupo de palabras extraídas de WordNet en español y pusieron todos los sinónimos posibles para cada una de ellas. De esta manera, se formaron manualmente un centenar de *synsets*, separados por categorías gramaticales divido de la siguiente manera: 75 formados por sustantivos, 15 por verbos, 7 por adjetivos y 2 por adverbios. El número de *synsets* por categoría gramatical es proporcional al de la WordNet en español.

- Precisión y Sensibilidad

Para el cálculo de las métricas respecto al *gold standard*, primero se obtuvieron la precisión y sensibilidad por cada *synset*, luego el promedio de estos datos es el resultado final. Los resultados globales obtenidos son de una precisión de **0.37** y sensibilidad de **0.30**. Por categoría gramatical, los resultados son los siguientes:

Categoria	Precisión	Sensibilidad	F-Score
Sustantivos	0.37	0.36	0.36
Verbos	0.44	0.17	0.25
Adjetivos	0.33	0.22	0.25
Adverbios	0.00	0.00	-

Tabla 10: Resultados desagregados de precisión y sensibilidad por categoría gramatical

# Palabras	Precisión	Sensibilidad	F-Score
1	0.36	0.36	0.36
2	0.38	0.19	0.25
3+	0.43	0.11	0.17

Tabla 11: Resultados desagregados de precisión y sensibilidad por número de palabras por *synset*

Los resultados se explican por el tamaño de la muestra del *gold standard* y por la MCR ya que se encuentran términos cuyo sentido es el mismo pero se han asignado a diferentes *synsets*. Destaca el caso de los adverbios donde el número de aciertos es cero. Esto se debe a la baja cantidad de adverbios, lo cual a su vez se debe a la baja frecuencia en su uso.

En cuanto a las métricas, la precisión nos dice que el método ha clasificado correctamente una tercera parte de los términos mientras que la sensibilidad indica cuántos términos totales del *synset* fueron identificados. Según la tabla 10, las métricas tienen a ser más precisas a mayor frecuencia de la categoría gramatical correspondiente. En cuanto a la tabla 11, mientras hayan más palabras de donde acertar, la precisión puede ser mayor pero si hay más palabras a acertar entonces la sensibilidad baja porque no se llega a cubrir todas. Con estos resultados, es posible concluir que el método tuvo una mayor precisión con los verbos pero los sustantivos y adjetivos tienen un mejor balance entre precisión y sensibilidad.

6. Comunicación con la WordNet

6.1. Introducción

Este capítulo abarca los detalles de cómo se realizará la comunicación entre un usuario y la WordNet dependiendo de la forma de acceso. La diferencia consiste en si el recurso se encuentra disponible de forma independiente o si depende del proyecto de traducción. Si un usuario accede a la WordNet por internet de forma independiente entonces usará una interfaz web. Será posible acceder desde el proyecto del traductor automático que obtendrá los datos solicitados a través de un servicio web y no necesariamente se tendrá que acceder a través de la interfaz web ya mencionada.

A continuación se explicará cómo se logrará dar solución al tercer objetivo específico: Implementar una interfaz de comunicación que permita el acceso a los datos de la WordNet en Shipibo-Konibo. Aún se encuentra en desarrollo por lo que solo se mostrarán esquemas.

Se tienen dos resultados esperados: i) Estructura de la interfaz web para la consulta de *synsets* y sus relaciones. ii) Arquitectura de servicio web para acceder a la WordNet.

6.2. Una estructura de la interfaz web para la consulta de *synsets* y sus relaciones.

La estructura la conforman 2 secciones. En la sección de entrada se espera que el usuario ingrese una palabra en Shipibo-Konibo para ser ubicada en la WordNet. Al hacer clic en el botón Buscar se mostrará en la segunda sección toda la información disponible sobre dicha palabra como se observa en la Figura 11.

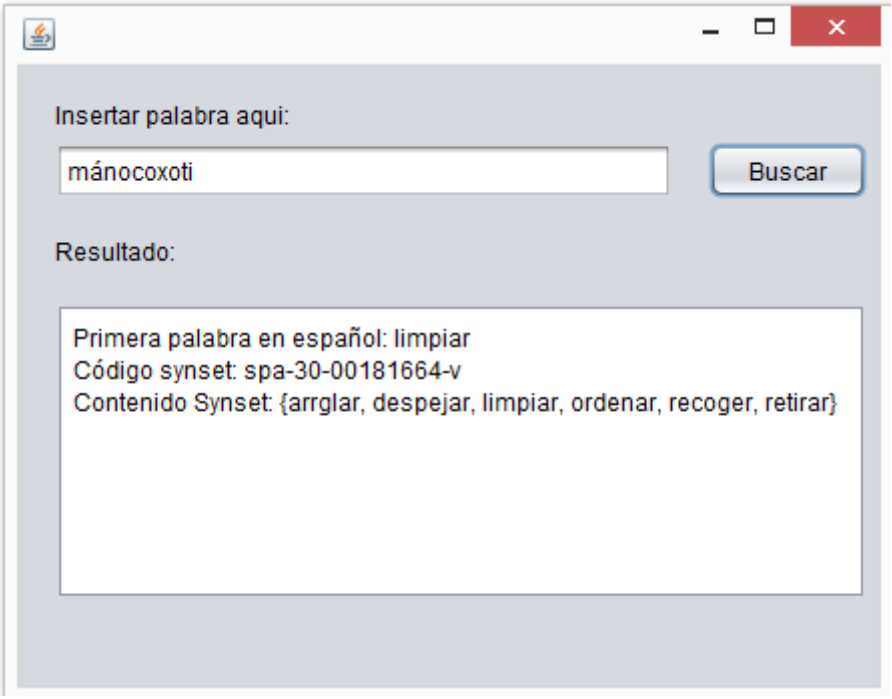


Figura 11: Resultado de la búsqueda.

6.3. Arquitectura de servicio web para acceder a la WordNet.

En cuanto a la integración de la WordNet con el proyecto del traductor automático, se desarrollará un servicio web para facilitar la comunicación entre el usuario y el servidor del traductor como mostrado en la Figura 12.

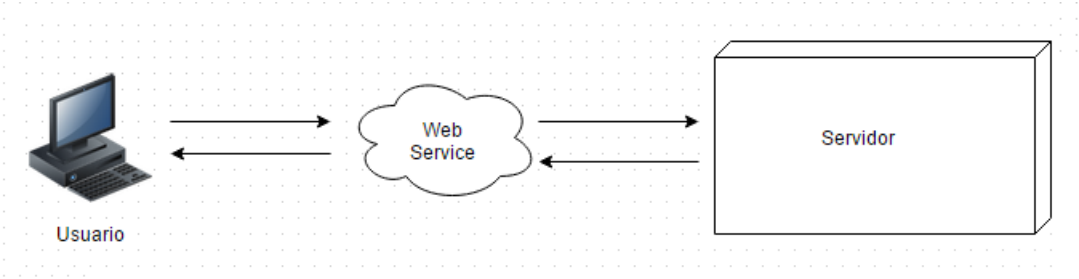


Figura 12: Diagrama de arquitectura del servicio web.

7. Conclusiones

El presente proyecto tuvo como objetivo la construcción de un recurso léxico basado en sinonimia para el Shipibo-Konibo que hace uso del estándar internacional para ubicar traducciones en otros idiomas en base al código de los synsets.

El proceso como se realizó este fue el siguiente: Se realizó un pre-procesamiento del diccionario para extraer datos de los términos y sus sentidos. Se construyó el algoritmo usando la WordNet en español para ubicar cada sentido de cada palabra en Shipibo-Konibo con su respectivo *synset* en español. Además, se implementó una interfaz de consulta para la WordNet disponible via web.

Se realizó una evaluación manual de la calidad de los *synsets* con la ayuda de linguistas y armar el estándar de prueba (gold standard). Las relaciones entre *synsets* fueron heredados de la WordNet en español pero falta una validación manual de la misma. Los resultados mostraron que existe una cierta similitud en la distribución de los sentidos del Shipibo-Konibo y Español lo que dado que ocurren muchas palabras con un solo sentido y se demuestra que a mayor número de sentidos, la cantidad de palabras disminuye considerablemente.

El producto final almacena todas las palabras incluyendo sus sentidos y está alineado a la WordNet en español (MCR). Está disponible vía web para facilitar y motivar su uso.

8. Trabajos Futuros

Se espera que se posible la implementación de funcionalidad para que los usuarios puedan agregar información a la WordNet en Shipibo-Konibo.

En la WordNet de inglés y español los sentidos están ordenados por frecuencia de uso. En el caso de Shipibo-Konibo es necesario que se cuente con un corpus anotado con los sentidos extraídos para así poder ordenar por frecuencias. Al no contar con un corpus los sentidos están en orden en el que se encuentran en el diccionario Español - Shipibo-Konibo.

Además, se requiere mayor validación ya que no se verificó las relaciones entre *synsets* porque se heredaron de la WordNet en español.

Referencias bibliográficas

[Allah y Boulaknadel, 2012] Allah, F. A., y Boulaknadel, S. (2012). Toward computational processing of less resourced languages: Primarily experiments for Moroccan Amazigh language. INTECH Open Access Publisher.

[Bentivogli et al., 2012] Bentivogli, L., Magnini, B., y Loinaz, I. A. Linguistic Processors and Infrastructure, 2012.

[Besacier et al., 2013] Besacier, L., Barnard, E., Karpov, A., y Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85-100.

[Bizzoni et al., 2014] Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., y Crane, G. R. (2014). The Making of Ancient Greek WordNet. In LREC (Vol. 2014, pp. 1140-1147).

[Bond et al., 2009] Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., y Kanzaki, K. (2009, August). Enhancing the japanese wordnet. In Proceedings of the 7th workshop on Asian language resources (pp. 1-8). Association for Computational Linguistics.

[Broda et al., 2008] Broda, B., Derwojedowa, M., Piasecki, M., y Szpakowicz, S. (2008, May). Corpus-based Semantic Relatedness for the Construction of Polish WordNet. In LREC.

[Cadellino, 2016] Cristian Cardellino: Spanish Billion Words Corpus and Embeddings (March 2016). Disponible en: http://crscardellino.me/SBWCE/

[Carbonell, 1994] Carbonell, Jaime. "El procesamiento del lenguaje natural, tecnología en transición." Actas del Congreso de la Lengua Española: Sevilla, 7 al 10 octubre, 1992. Instituto Cervantes, 1994.

[Crystal, 2000] Crystal, D. (2000). Language death. Ernst Klett Sprachen.

[DLJ, 2016] Deep Learning for Java (2016) Disponible en: http://deeplearning4j.org/

[Farreres et al., 1998] Farreres, X., Rigau, G., y Rodriguez, H. (1998). Using wordnet for building wordnets. arXiv preprint cmp-lg/9806016.

[Fellbaum, 1998] Fellbaum, C. WordNet: an electronic lexical database, Cambridge, MA, USA, 1998.

[Girardi, 2002] Christian Girardi. MultiWordNet. Disponible en http://multiwordnet.fbk.eu/online/multiwordnet.php.

[Gonzalez y Rigau, 2013] Gonzalez-Agirre, A., y Rigau, G. (2013). Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository. Linguamática, 5(1), 13-28.

[Java, 2016] Java (2016). java.com: Java. [online] Disponible en: http://www.java.com

[MySQL, 2016] Mysql.com. (2016) MySQL [online] Disponible en: https://www.mysql.com

[Gupta et al., 2007] Gupta, S., Purver, M., y Jurafsky, D. (2007, June). Disambiguating between generic and referential you in dialog. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 105-108). Association for Computational Linguistics.

[Jurafsky y Martin, 2009] Jurafsky, D., y Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2009.

[Kitchenham, 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1-26.

[Krauss, 2007] Krauss, Michael E. "Keynote-mass language extinction and documentation: The race against time." The vanishing languages of the Pacific rim (2007): 3-24.

[Loriot et al., 1993] Loriot, J., Lauriault, E., y Day, D. (1993). Diccionario Shipibo-Konibo - Castellano. Instituto Lingüístico de Verano.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., y Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database*. International journal of lexicography, 3(4), 235-244.

[Mititelu, 2012] Mititelu, V. B. (2012). Adding Morpho-semantic Relations to the Romanian Wordnet. In LREC (pp. 2596-2601).

[Onyshkevych, 2014] Darpa.mil. (2014). Low Resource Languages for Emergent Incidents (LORELEI). [online] Disponible en: http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents.

[Piruzelli y da Silva, 2010] Piruzelli, M. P. F., y da Silva, B. C. D. (2010). Estudo exploratório de informações lexicais relevantes para a resolução de ambiguidades lexical e estrutural. Anais do Encontro do Círculo de Estudos Linguísticos do Sul, Universidade do Sul de Santa Catarina.

[RAE, 2016] Rae.es (2016) Real Academia Española. Disponible en http://www.rae.es/

[Ramakrishnan y Gehrke, 2000] Ramakrishnan, R., y Gehrke, J. (2000). Database management systems. Osborne/McGraw-Hill.

[Roget, 1991] Roget, P. M. (1991). Roget's Thesaurus of English Words and Phrases. TY Crowell Company.

[Rouhizadeh et al., 2007] Rouhizadeh, M., Shamsfard, M., y Yarmohammadi, M. A. (2007) Building a WordNet for Persian Verbs.

[Sagot, 2010] Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In 7th international conference on Language Resources and Evaluation (LREC 2010).

[Sánchez, 1995] Sánchez, A. (1995). Definición e historia de los corpus. CUMBRE–Corpus Linguistico de Espanol Contemporaneo. Madrid: SGEL.

[Sathapornrungkij y Pluempitiwiriyawej, 2005] Sathapornrungkij, P., y Pluempitiwiriyawej, C. (2005). Construction of Thai WordNet lexical database from machine readable dictionaries. Proc. 10th Machine Translation Summit, Phuket, Thailand.

[Schmid, 1994] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

[Taghizadeh y Faili, 2016] Taghizadeh, N., & Faili, H. (2016). Automatic Wordnet Development for Low-Resource Languages using Cross-Lingual WSD. Journal of Artificial Intelligence Research, 56, 61-87.

[Tesauro, 2016] Plinio Conti. Tesauro en español. Disponible en http://openoffice-es.sourceforge.net/thesaurus/ Consulta: 16/04/2016

[Uszkoreit, 1996] Uszkoreit, H. (1986, August). Categorial unification grammars. In Proceedings of the 11th coference on Computational linguistics (pp. 187-194). Association for Computational Linguistics.

[WordNet Princeton University, 2016] Princeton University, Office of Communications, Mahlon Lovett. Disponible en https://wordnet.princeton.edu/

Anexo 1: WordNet-Shp: Towards the Building of a Lexical Database for a Peruvian Minority Language

WordNet-Shp: Towards the Building of a Lexical Database for a Peruvian Minority Language

Diego Maguiño-Valencia, Arturo Oncevay-Marcos and Marco A. Sobrevilla Cabezudo
Research Group on Artificial Intelligence (IA-PUCP)
Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú
{dmaguino,arturo.oncevay,msobrevilla}@pucp.edu.pe

Abstract

WordNet-like resources are lexical databases with highly relevance information and data which could be exploited in more complex computational linguistics research and applications. The building process requires manual and automatic tasks, that could be more arduous if the language is a minority one with fewer digital resources. This study focuses in the construction of an initial WordNet database for a low-resourced and indigenous language in Peru: Shipibo-Konibo (shp). First, the stages of development from a scarce scenario (a bilingual dictionary shp-es) are described. Then, it is proposed a synset alignment method by comparing the definition glosses in the dictionary (written in Spanish) with the content of a Spanish WordNet. In this sense, word2vec similarity was the chosen metric for the proximity measure. Finally, an evaluation process is performed for the synsets, using a manually annotated Gold Standard in Shipibo-Konibo. The obtained results are promising, and this resource is expected to serve well in further applications, such as word sense disambiguation and even machine translation in the shp-es language pair.
Keywords: WordNet, lexical database, minority language

1. Introduction

The building of digital linguistic resources is a great support for endangered languages, as they help to preserve relevant information and knowledge related, not only for the language itself, but also for the community whose speak it. Nevertheless, if those resources are not developed to be able for further analysis and research, they may be insufficient to assist in the preservation efforts (Berment, 2002).

In that context, computational linguistics is a research area that aims to understand linguistic phenomena, in an automatic way, through the processing and exploiting either linguistic corpora or language patterns from large amounts of data. In order to achieve that goal, the corpus must be in a machine-readable format, and might include structured information and linguistic meta-data that helps to automatically understand patterns from the language.

Among the most important lexical and structured resources, the WordNet is included (Fellbaum, 1998). This resource could be used as a thesaurus for different languages, and its exploitation might ease more complex tasks such as word sense disambiguation or even machine translation.

For that reason, this article describes the building of an initial version of a WordNet for an endangered language, which faces additional problems caused by the low-density of digital resources. The language case study is Shipibo-Konibo (SHP), one of the 47 indigenous languages spoken in Peru, specifically in the Amazonian region and has over 23,000 speakers. (Ministerio de Educación, Perú, 2013). Like most of it peers, SHP is classified as a minority language from both a social and a computational perspective (Forcada, 2006).

The paper is organized as follow. Section 2 defines shortly what a WordNet is. Next, Section 3 presents works regarding the main aspects in the building of WordNet-like resources for other languages. After that, the construction of the WordNet for Shipibo-Konibo is detailed in Section 4. Additionally, there is an evaluation process in Section 5. Finally, conclusions and future work are discussed.

2. WordNet

A WordNet is a lexical database in an specific language. WordNet contains words grouped into synonyms sets (synsets) where each synset represent a different sense and is identified by a interlingual index (ILI) that allows to work in multilingual contexts. In addition to having a synonyms set, each synset may show a definition (gloss) and also some use examples. The synsets are connected between them through semantic and lexical relationships like hypernym, hyponym, and others (Fellbaum, 1998).

There are shallow similarities between a WordNet and a thesaurus, which different sets of terms are grouped based on a meaning-similarity criteria. On the other hand, the labels of the Wordnet are defined by the semantic relationship between the words or entries, while the clusters of words in a synonym dictionary may not follow any distinctive pattern of explicit meaning similarity (Miller, 1995).

For the Spanish language, there are two main multilingual options that are vastly used: MultiWordNet (Emanuele et al., 2002) and Multilingual Central Repository (MCR) (Gonzalez-Agirre et al., 2012). Despite the fewer amount of synsets contained in the latter repository, MCR will be used in the study because its more recent updates.

Finally, a WordNet is established ideally as a free and open-source resource, so the goal is similar for the Shipibo-Konibo WordNet.

3. Related Works

The section describes the different aspects related to the development or improvement of WordNet-like resources for different languages, whether they are minority ones or not.

Farreres et al. (1998) presented a semi-automatic approach to address development of new WordNets, using the English version as a model. There is a manual alignment for

verbs and nouns, while the validation step was automated. The latter process chose the most relevant terms and develop a complete taxonomy of semantic primitives. The proposed method was applied to build both Spanish and Catalan WordNets, using bilingual dictionaries and lexicons as the main sources.

Semi-automatic procedures were also proposed for building WordNets in other languages. For Persian, there were special considerations regarding the verb composition, as they are formed by more than 2 verbs usually (Rouhizadeh et al., 2008). Besides, an Ancient Greek WordNet was developed from a Greek-English dictionary, by supposing a semantic closeness regarding the terms translated (Bizzoni et al., 2014).

Thai WordNet was built using an own system called WordNet Builder (Sathapornrungkij and Pluempitiwiriyawej, 2005). The WordNet in English and machine readable dictionaries (MRD) were the main input sources, and both were connected trough a Link Analyzer for synset match. The validation process was performed by a statistical classifier, reducing human intervention.

In the Polish version, the validation process was performed with a similarity measure with the English WordNet, enhanced by human evaluation later (Broda et al., 2008). For the study, the synsets were built using a metric for semantic relation between different terms and POS-tags.

There were other studies focused in the improvements of previous WordNets. Mititelu (2012) proposed an addition of morpho-semantic relations for the Romanian WordNet. Heuristics were applied to group the terms and affixes in pairs. Then, these pairs were semantically tagged in three levels: monolingual, multilingual and for different NLP applications. Likewise, Bond et al. (2009) addressed improvements in the Japanese WordNet by increasing the vocabulary coverage, annotating more bilingual English-Japanese texts, and connecting the WordNet with different resources such as lexicons or image repositories.

Finally, Taghizadeh and Faili (2016) focused their efforts in building WordNet for low-resource scenarios. Using an Expectation-Maximization (EM) algorithm plus a cross-lingual word sense disambiguation method, a high quality WordNet could be built for Persian. Other positive aspect was the use of minimal resources, such as a bilingual dictionary and a monolingual textual corpus.

4. WordNet-Shp

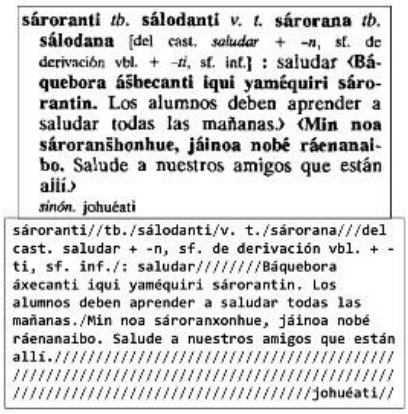
This section includes the steps followed for the development of the initial WordNet-Shp, a WordNet-like resource for Shipibo-Konibo. There are two main phases. The first phase consisted in the digitalization and pre-processing of a bilingual dictionary shp-es (Spanish). The second one consisted on the synset alignment task by using a similarity metric (provided by word2vec) with the definition glosses in the dictionary and the Spanish WordNet of Multilingual Central Repository (MCR)(Gonzalez-Agirre et al., 2012). The words of different glosses are compared against each other to obtain an average for each word obtained from the gloss of the dictionary and then averaged to obtain a final result the synset. The highest result is chosen.

The processed dictionary, the aligned synsets, and the Gold Standard for evaluation which consisted of a hundred synsets (see Subsection 5.1) are available in a project site[1].

4.1. Pre-Processing a Dictionary

An algorithm was built to automatically extract the words from an old-fashioned Spanish-Shipibo bilingual dictionary (Lauriout et al., 1993). For each word entry, an structured output is obtained, which includes several fields.First, there are 9 unique fields: term, reference, type of variant of the term, variant, grammatical category (POS-tag), main part, type of variant of the main part, variant of the main part, etymology. Then, given that a term might have more than one sense, each sense must be stored separately. So it was confirmed that the maximum number of different senses per word in the dictionary was 12. Each sense includes 6 glosses, 2 usage notes and 3 examples. Thus, it makes a total of eleven fields for every sense. Finally, there are 3 unique fields at the end of the entry: Synonymous paragraph, parent term, sub-entry type. The latter two only applies when the term is a sub-entry and is related to a parent main entry.

For instance, the parsed output for the term *sároranti* would be structured as in Figure 1.



sároranti *tb.* **sálodanti** *v. t.* **sárorana** *tb.* **sálodana** [del cast. *saludar* + *-n*, sf. de derivación vbl. + *-ti*, sf. inf.] : saludar ‹**Báquebora ášbecanti iqui yaméquiri sárorantin.** Los alumnos deben aprender a saludar todas las mañanas.› ‹**Min noa sároranšhonhue, jáinoa nobé ráenanaibo.** Salude a nuestros amigos que están allí.›

sinón. johuéati

```
sároranti//tb./sálodanti/v. t./sárorana///del
cast. saludar + -n, sf. de derivación vbl. + -
ti, sf. inf./: saludar////////Báquebora
áxecanti iqui yaméquiri sárorantin. Los
alumnos deben aprender a saludar todas las
mañanas./Min noa sároranxonhue, jáinoa nobé
ráenanaibo. Salude a nuestros amigos que están
allí.////////////////////////////////////////
////////////////////////////////////////////
//////////////////////////////////johuéati//
```

Figure 1: (Top) Original dictionary entry for *sároranti*. (Bottom) Parsed output for the entry. There are many separators ('/') together due to the possibility of extracting other elements between them. In this case, there is only one word sense, so there are a lot of separators together at the end, since up to 12 senses are expected as maximum.

The total amount of entries stored is about 5800, including 4815 terms from the main grammatical word classes (nouns, verbs, adjectives and adverbs). The remaining 985 are suffixes, prefixes, conjunctions, prepositions, among other categories. A distribution of the main word entries and their respective amount of senses in the Shipibo-Konibo Wordnet is presented in Table 1.

[1]WordNet-Shp data available in: chana.inf.pucp.edu.pe/resources/wordnet-shp

#s. \ POS	Nouns	Verbs	Adj.	Adv.
1	2 231	1 453	357	96
2	174	266	62	11
3	59	48	13	2
4	14	16	1	0
5	6	3	0	0
6	2	0	0	0
7	0	0	0	1
Total	2 486	1 786	433	110

Table 1: Distribution of words with and without ambiguous senses found in the Shipibo-Konibo dictionary: Number of senses (#s.) per Part-of-Speech (POS) tag

4.2. Synsets Alignment

The alignment algorithm focuses in comparing the glosses of the word entries in Shipibo with the data of a Spanish WordNet, in order to obtain the closeness between the meanings among them. To define which synset each word sense belongs to, all the synsets in Spanish (terms, glosses and examples) were taken into account.
For this research, a word2vec model (Mikolov et al., 2013) was trained based on a general corpus (without annotation) of the Spanish language composed by approximately 1.4 billion words (Cardellino, 2016). The corpus was created by compiling various resources of the Spanish language that can be found on the Internet. Some of them are listed below:

- Collection of legal texts in Spanish from the European Union
- United Nations documents
- Parts of corpus in Spanish in other languages
- Protected articles from the Wikipedia in Spanish

With the trained word2Vec model, the classification algorithm for the synset alignment was the next step. The pseudo-code is presented below:

Algorithm 1 Synset alignment

for $i = 1$ to $|V|$ **do**
  for $j = 1$ to $|s_{v_i}|$ **do**
    $max_{ij} = 0$;
    for $k = 1$ to $|W_{es}|$ **do**
      $sim_{ijk} = word2vec\_similarity(s_{jv_i}, W_{es_k})$
      if $sim_{ijk} > max_{ij}$ **then**
        $max_{ij} = sim_{ijk}$
      end if
    end for
    $insert\_to\_BD(s_{jv_i}, W_{es_{argmax(k)}})$
  end for
end for

Where $|V|$ is the size of vocabulary V, $|s_{v_i}|$ is the number of different senses for the word s_{v_i}, and $|W_{es}|$ is the size of entries in the Spanish WordNet (W_{es}). word2vec_similarity is obtained by calculating the cosine between the two vectors.
Regarding the pseudocode, this algorithm works as follows: Each sense of each word found in the Shipibo dictionary was compared to each synset of the WordNet in Spanish using the word2vec model previously trained to obtain a similarity metric. This measurement was expressed as a decimal number.
For the calculation, it was taken the same word, glosses and corresponding examples, if any, of each Spanish word found in the Shipibo gloss. The glosses were filtered to consider only the words of the categories that arise more frequently and that allow to understand the context. These categories are nouns and verbs; and to a lesser extent, adjectives and adverbs.
For example, in the case of a noun, the nouns and verbs of its gloss were considered (due to the strong connection between the two categories). Each word was taken only once, eliminating repetitions because it could be the case that the terms or words contained in the glosses would be repeated.
Additionally, the Tree-tagger (Schmid, 1995) was used to extract grammatical categories and lemmas from words in Spanish. For instance, for the verb *manéxti* ("clean" in English), which was taken in the sense whose gloss is "to tie the hairs of the head or crown with other hairs of the same head", the process would be as follows:

- Tie the hairs of the head or the crown with other hairs his own head to
- Tying, hair, head, crown, hair, head - Only verbs and nouns
- Tie, hair, head, crown - Lemmas were extracted and word repetitions were erased

The similarity was expressed as a decimal number between -1 and 1 (being the result of a cosine). The greater the result, the relationship between words would be closer. Therefore, to decide which synset corresponds to the word in Shipibo, the synset with the greater similarity is considered. Each time the synset corresponding to a word in Shipibo was found, everything related to the term was inserted in the database following the standard used by MCR (Gonzalez-Agirre et al., 2012).
Once the algorithm with two words was executed in Shipibo as a sample, the nearest Spanish synset was obtained, and that was found by the classification algorithm described. For example, *mocoxoti* (which means "clean") was correctly classified into one of the synsets where the word "clean" (in Spanish) is found in the same sense as to sort or arrange. It also shows all the words associated with the synset, as well as its ili code (used as the international standard).

5. Evaluation

To carry out the evaluation a gold standard was needed. It was prepared manually by a group of linguists and native speakers of Shipibo. The evaluation metric in this study is the accuracy, which is calculated after testing the classification algorithm with the gold standard.

Category	# Synsets
Nouns	105
Verbs	43
Adjetives	8
Adverbs	2

Table 2: Current state of the WordNet-Shp

5.1. Gold Standard

The gold standard was made by linguists and a native Shipibo-Konibo speaker. They selected a group of words extracted from the WordNet in Spanish and put all the possible synonyms for each one. In this way, a hundred synsets were formed manually, separated by grammatical categories as thus: 76 formed by nouns, 15 by verbs, 7 by adjectives and 2 by adverbs.

5.2. Results

All the words of the gold standard were evaluated to analyze if the retrieved synset is really the corresponding one. The total accuracy obtained with the gold standard was 32.8%. Some samples of classified synsets are presented below:

- Synset: spa-30-03571439-n Word: *chachí* (injector). Gloss: injector. Synset Result: spa-30-03571439-n
- Synset: spa-30-05599617-n Word: *cói* (chin). Gloss: protruding part of the jaw. Synset Result: spa-30-05598147-n
- Synset: spa-30-03343853-n Word: *tóoati,tsakati* (gun) Gloss: portable weapon. Synset Result: spa-30-00001740-n

In the first sample, there is a match because the original and the retrieved synset are the same. In the second example, the synsets do not match so the classification was labeled as incorrect. However, in the synset result (spa-30-05598147-n) the word "nose" shows up, which is a part of one's face so it's related to what we were originally looking for (chin). In the third and final example, the classification was not correct either but there is a lexical gap because the word *tsakati* wasn't found in the dictionary as a entry.
The precision obtained might be due to the following points:

- The obtained synset makes sense but it's not the same ILI. This could happen because the wordnet is very refined
- Some or several words used in the gloss of a particular word do not bear much relation with it
- Minor errors in the dictionary processing

Finally, after processing words from the dictionary that were included in the gold standard and an extra hundred (200 in total), the current state of the Wordnet in Shipibo by number of synsets is presented in Table 2.

6. Conclusions and Future Work

This study aimed the development of a new lexical and synonym-based resource for the Shipibo-Konibo language. Likewise, this repository uses the international standard for locating translations in other languages based on the synsets codification. For this purpose, a bilingual dictionary was pre-processed for extracting information of word entries and their senses in Shipibo-Konibo. After that, using the Spanish WordNet, an algorithm aligned each word sense in Shipibo-Konibo with its Spanish peer. The existing relationships in the Spanish WordNet were considered in order to be inherited in the Shipibo-Konibo repository.
Regarding the evaluation, there was a manual analysis of the synset quality. This was supported by professional linguists, and the output was the development of the Gold Standard. The results of the alignment showed a close similarity in the word sense distribution between Shipibo-Konibo and Spanish. This is caused mainly by the high presence of unique-sense words. Besides, as the number of senses is higher, the amount of words decreases considerably.
Finally, the developed resource stores all the words including their different senses in Shipibo-Konibo. Also, there is a web interface under development for querying the entries. All of these resources will be available in the following link: https://github.com/iapucp/wordnet-shp-lrec2018
As future work, we want to improve our algorithm of synset alignment and to include other relations between synsets, like hypernyms or hyponyms. We believe that it will not be difficult because it is possible to bring these relations from WordNets in other languages (like Spanish).

Acknowledgments

We highly appreciate the linguistic team effort that made possible the creation of this resource: Dr. Roberto Zariquiey, Alonso Vásquez, Gabriela Tello, Renzo Ego-Aguirre, Lea Reinhardt and Marcela Castro. We are also thankful to our native speakers (Shipibo-Konibo) collaborators: Juan Agustín, Carlos Guimaraes, Ronald Suárez and Miguel Gomez. Finally, we gratefully acknowledge the support of the "Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica" (CONCYTEC, Peru) under the contract 225-2015-FONDECYT.

7. Bibliographical References

Berment, V. (2002). Several directions for minority languages computerization. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pages 1–5. Association for Computational Linguistics.

Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., and Crane, G. R. (2014). The making of Ancient Greek WordNet. In *LREC*, volume 2014, pages 1140–1147.

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., and Kanzaki, K. (2009). Enhancing the japanese wordnet. In *Proceedings of the 7th workshop on Asian language resources*, pages 1–8. Association for Computational Linguistics.

Broda, B., Derwojedowa, M., Piasecki, M., and Szpakowicz, S. (2008). Corpus-based semantic relatedness for the construction of Polish WordNet. In *LREC*.

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings, March.

Emanuele, P., Luisa, B., and Christian, G. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet, Mysore, India*.

Farreres, X., Rigau, G., and Rodriguez, H. (1998). Using wordnet for building wordnets. *arXiv preprint cmp-lg/9806016*.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Forcada, M. (2006). Open source machine translation: an opportunity for minor languages. In *Proceedings of the Workshop "Strategies for developing machine translation for minority languages", LREC*, volume 6, pages 1–6.

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

Lauriout, E., Day, D., and Loriot, J. (1993). Diccionario Shipibo-Castellano.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Ministerio de Educación, Perú. (2013). *Documento nacional de lenguas originarias del Perú*. MINEDU. URI: http://repositorio.minedu.gob.pe/handle/123456789/3549.

Mititelu, V. B. (2012). Adding morpho-semantic relations to the Romanian Wordnet. In *LREC*, pages 2596–2601.

Rouhizadeh, M., Shamsfard, M., and Yarmohammadi, M. A. (2008). Building a WordNet for Persian verbs. In *in the Proceedings of the Fourth Global WordNet Conference (GWC'08). The Fourth Global WordNet Conference, 2008*. Citeseer.

Sathapornrungkij, P. and Pluempitiwiriyawej, C. (2005). Construction of Thai WordNet lexical database from machine readable dictionaries. *Proc. 10th Machine Translation Summit, Phuket, Thailand*.

Schmid, H. (1995). Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Taghizadeh, N. and Faili, H. (2016). Automatic Wordnet development for low-resource languages using cross-lingual WSD. *J. Artif. Intell. Res.(JAIR)*, 56:61–87.