Linha D'Água: São Paulo, v. 34, n. 02, p. 47-64, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p47-64

Artículo / Artigo / Article

Documentación de lenguas amenazadas en la época digital

Documentação de línguas ameaçadas na era digital

Endangered Language Documentation in the Digital Age

Mika Hämäläinen*
mika.hamalainen@helsinki.fi
https://orcid.org/0000-0001-9315-1278

Jack Rueter**
jack.rueter@helsinki.fi
https://orcid.org/0000-0002-3076-7929

Khalid Alnajjar***
khalid.alnajjar@helsinki.fi
https://orcid.org/0000-0002-7986-2994

Resumen

Presentamos nuestra infraestructura para la documentación de lenguas urálicas, que consiste en herramientas para redactar diccionarios de tal forma que las entradas sean estructuradas en el formato XML (Extensible Markup Language). Desde los diccionarios en XML podemos generar código para analizadores morfológicos que son útiles para todo tipo de actividades de PLN. En este artículo mostramos las ventajas que una documentación digital y legible por máquina tiene. Describimos, también, el sistema en el contexto de lenguas urálicas amenazadas.

Palavras-chave: Diccionarios digitales; Procesamiento de lenguajes naturales; Lenguas urálicas; Documentación lingüística; Infraestructura abierta.

* Universidad de Helsinki, Facultad de Humanidades, Departamento de Humanidades Digitales, Helsinki, Finlandia.

** Universidad de Helsinki, Facultad de Humanidades, Departamento de Humanidades Digitales, Helsinki, Finlandia.

*** Universidad de Helsinki, Facultad de Humanidades, Departamento de Humanidades Digitales, Helsinki, Finlandia.

Resumo

Apresentamos a nossa infraestrutura para documentação de línguas urálicas, a qual consiste num conjunto de ferramentas para redigir dicionários de modo que as entradas sejam estruturadas num formato XML (Extensible Markup Language). A partir de dicionários XML, podemos criar códigos para analisadores morfológicos, que são úteis a todos os tipos de atividades de processamento de língua natural. Neste artigo, demonstramos as vantagens da documentação digital legível por máquina e descrevemos o sistema no contexto das línguas urálicas ameaçadas.

Palavras-chave: *Dicionários digitais; Processamento de linguagens naturais; Línguas urálicas; Documentação linguística; Infraestrutura aberta.*

Abstract

We present our infrastructure to document Uralic languages, which consists of tools to write dictionaries so that entries are structured in XML (Extensible Markup Language) format. From dictionaries in XML, we can generate code for morphological analysers useful for all kinds of NLP tasks. In this article, we show the advantages of digital and machine-readable documentation. We also describe the system in the context of endangered Uralic languages.

Keywords: *Digital Dictionaries; Natural Language Processing; Uralic Languages; Linguistic Documentation; Open Infrastructure.*

Introducción

La mayoría de las lenguas habladas en el mundo están amenazadas y están en peligro de extinción. Su documentación y revitalización son de altísimo valor cultural, por lo cual han recibido mucha atención académica en varias disciplinas como en la antropología, tipología, lexicografía y lingüística computacional. No obstante, los recursos producidos en cada proyecto no serán necesariamente publicados de forma abierta ni para la comunidad de los hablantes nativos ni para el uso de otros proyectos científicos.

El objetivo de nuestro artículo es describir nuestra abierta infraestructura para documentar lenguas minoritarias. Presentamos nuestras experiencias con las siguientes lenguas urálicas: sami de skolt (sms), erzya (myv), moksha (mdf), komi ziriano (kpv) y komi permio (koi). Como pertenecen a la rama urálica, son lenguas que exhiben una amplia riqueza morfológica lo que hace su tratamiento automático un desafío para los métodos modernos apoyados en el aprendizaje automático. Sin embargo, realizando la documentación lingüística de forma estructurada que permite la lectura automática, es posible crear los recursos computacionales necesarios para el PLN (procesamiento de lenguajes naturales) al mismo tiempo con la documentación lingüística.

Estamos a punto de comenzar a trabajar con la lengua apurinã (apu), lo que nos permite reflejar nuestro contexto urálico desde una perspectiva más amplia, y aumenta la relevancia de

nuestro trabajo en el contexto latinoamericano. Describimos, así, cómo nuestra infraestructura puede funcionar en contextos fuera de lo urálico.

La documentación lingüística es un campo de estudio académico que se ha desarrollado considerablemente en las últimas décadas. Su objetivo es proporcionar un registro completo de las prácticas lingüísticas características de una comunidad de habla determinada (HIMMELMANN, 1998). El objetivo de la documentación del lenguaje es crear el registro más completo posible de la comunidad de habla tanto para las futuras generaciones como para la revitalización del idioma. El resultado de dicho trabajo se manifiesta como un *corpus* lingüístico u otro tipo de colección de material. Estos datos son una documentación del idioma que puede analizarse y estudiarse de diversas maneras. La cuestión si los materiales actualmente documentados realmente describen el uso del lenguaje de una comunidad de habla con éxito puede ser discutible, y al menos este objetivo nunca podrá alcanzarse por completo. Sin embargo, especialmente en la época actual con lenguas en peligro de extinción, estos materiales, a menudo, constituyen los únicos recursos disponibles en estos idiomas.

Si los materiales de documentación lingüística deberían ser accesibles y cómo deberían ser distribuidos, ha sido un tema de debate. Creemos que es importante entender que esto también es una cuestión de granularidad, y la pregunta no es necesariamente si los materiales son accesibles, sino más bien qué partes deberían permitir qué tipo de acceso. Existen razones buenas para mantener materiales culturalmente sensibles disponibles solo para los grupos específicos. Al mismo tiempo, siempre hay materiales en cualquier idioma que son más neutrales y que los propios autores quieren hacer accesibles. Especialmente para trabajos escritos y publicados, siempre puede ser posible negociar una publicación con licencias abiertas modernas, lo que también permitiría la reutilización de los mismos materiales en diferentes propósitos de investigación abierta.

Estos materiales son particularmente importantes cuando desarrollamos herramientas de PLN, porque este trabajo puede beneficiarse mucho de los recursos más transparentes y accesibles que posible. En las secciones siguientes analizaremos ejemplos de dicho trabajo, incluido el contexto de los treebanks de dependencias universales. Hay que enfatizar que la tecnología abierta desarrollada en una infraestructura abierta también puede usarse para procesar materiales que están disponibles solo para un investigador en particular o miembros individuales de la comunidad. Por lo tanto, la infraestructura abierta beneficia tanto a los entornos de uso abiertos como cerrados, mientras que una infraestructura cerrada, posiblemente, solo beneficia a los grandes actores comerciales.

1 Estudios relacionados

Hay varios proyectos individuales en diferentes partes del mundo que trabajan con diccionarios en línea para lenguas amenazadas. Sin embargo, muchos proyectos tienen una lengua en su enfoque y trabajan sin conocer otros proyectos con otras lenguas amenazadas. Esto

ha llegado en una situación en que los investigadores resuelven el mismo tipo de problemas individualmente para su lengua de interés. En este apartado, presentamos algunos proyectos digitales.

El trabajo con idiomas en peligro de extinción en América del Norte ha demostrado que se debe proporcionar herramientas de aprendizaje al principiante en un idioma. Las comunidades son pequeñas y la falta de familiaridad con la tradición lexicográfica puede fácilmente ser perjudicial para la experiencia de aprendizaje del principiante. No se puede esperar que el estudiante de un nuevo idioma sepa dónde se encuentra una entrada del diccionario ni que adopte automáticamente la ortografía normativa. Cuando el usuario del idioma carece del teclado o del conocimiento para escribir correctamente, las estrategias de relajación ortográfica se pueden implementar en soluciones que dominan la morfología para móviles y en línea. El conocimiento morfológico y la relajación de ortografía se utilizan para atender a los principiantes en lenguas tsimshiánicas y salishanas en el uso de diccionarios y el PLN (LITTELL et al., 2017).

En un frente completamente separado, también se ha trabajado para proveer a la comunidad de Yupik de la isla de St. Lawrence el acceso sin obstáculos a materiales lingüísticos en línea. Esto ha sido posible utilizando un diccionario morfológicamente consciente. En su sistema, se ha introducido una estrategia de múltiples métodos de entrada que atienden a diferentes sistemas de escritura (HUNT et al., 2019). El trabajo aquí está hecho a medida, y se mantiene un fuerte vínculo entre un idioma y su comunidad. Estos idiomas en peligro de extinción se incluyen en la categoría de lenguas de bajos recursos.

Lo problemático es que "lengua de bajos recursos" es un término que se utiliza para casi cualquier idioma con menor presencia en Internet que el inglés. Lenguas como el hindi (IRVINE y CALLISON-BURCH, 2014), el árabe (CHEN et al., 2018) o bien el persa (AHMADNIA et al., 2017) son consideradas a menudo lenguas de bajos recursos en el mundo de PLN, aunque tienen millones de hablantes. En el trabajo de Nasution et al., (2018), por lo contrario, las lenguas malasias son relativamente pequeñas en comparación con las lenguas mayoritarias que las rodean. El enfoque consiste en trabajar simultáneamente con un grupo de idiomas muy relacionados en una infraestructura multilingüe e independiente del idioma. Los autores analizan el uso de las entradas de un diccionario bilingüe y explican la dificultad de seleccionar los diccionarios bilingües adecuados para comenzar la documentación.

Una de las infraestructuras más ambiciosas para la documentación de lenguas minoritarias desde el punto de vista de la lingüística computacional es, sin duda, la de Giella (MOSHAGEN et al., 2014). Su infraestructura está basada en dos componentes principales: los transductores TEF (transductores de estados finitos) y diccionarios en XML. Los transductores son una forma de documentar la morfología de una lengua de manera computacional. Es decir, son colecciones de reglas sobre cómo el sistema morfológico de una lengua funciona. Estas reglas pueden usarse directamente para un análisis automático de texto y para conjugar lemas en sus variantes morfológicas.

Se utilizan los transductores y los diccionarios para herramientas como corrección ortográfica en *Word*[1], predicción de texto en teclados de *Android* e iOS[2], sistemas interactivos para aprender lenguas (BONTOGON et al., 2018) y diccionarios en línea (RUETER y HÄMÄLÄINEN, 2017). Nuestra infraestructura está basada en Giella, lo que nos permite sincronizar los datos entre las dos infraestructuras. Esto significa que los avances en la documentación lingüística en nuestra infraestructura pueden usarse directamente en las herramientas producidas en Giella.

2 Nuestra infraestructura para la documentación lingüística

Giella requiere una competencia relativamente alta en programación para poder redactar diccionarios y programar reglas morfológicas en los transductores, y al mismo tiempo, requiere buenos conocimientos en la lengua que está en curso de la documentación. Su infraestructura es demasiado complicada incluso para los que han estudiado informática, y por lo tanto no es accesible para una comunidad fuera de los que colaboran directamente con Giella. Por este motivo, nuestra infraestructura tiene varias interfaces para distintos tipos de usuarios; tanto para usuarios que no tienen conocimientos suficientes para escribir XML o programar transductores como para desarrolladores que quieren utilizar las herramientas sin saber cómo compilarlas desde cero.

2.1 Diccionarios XML en Línea

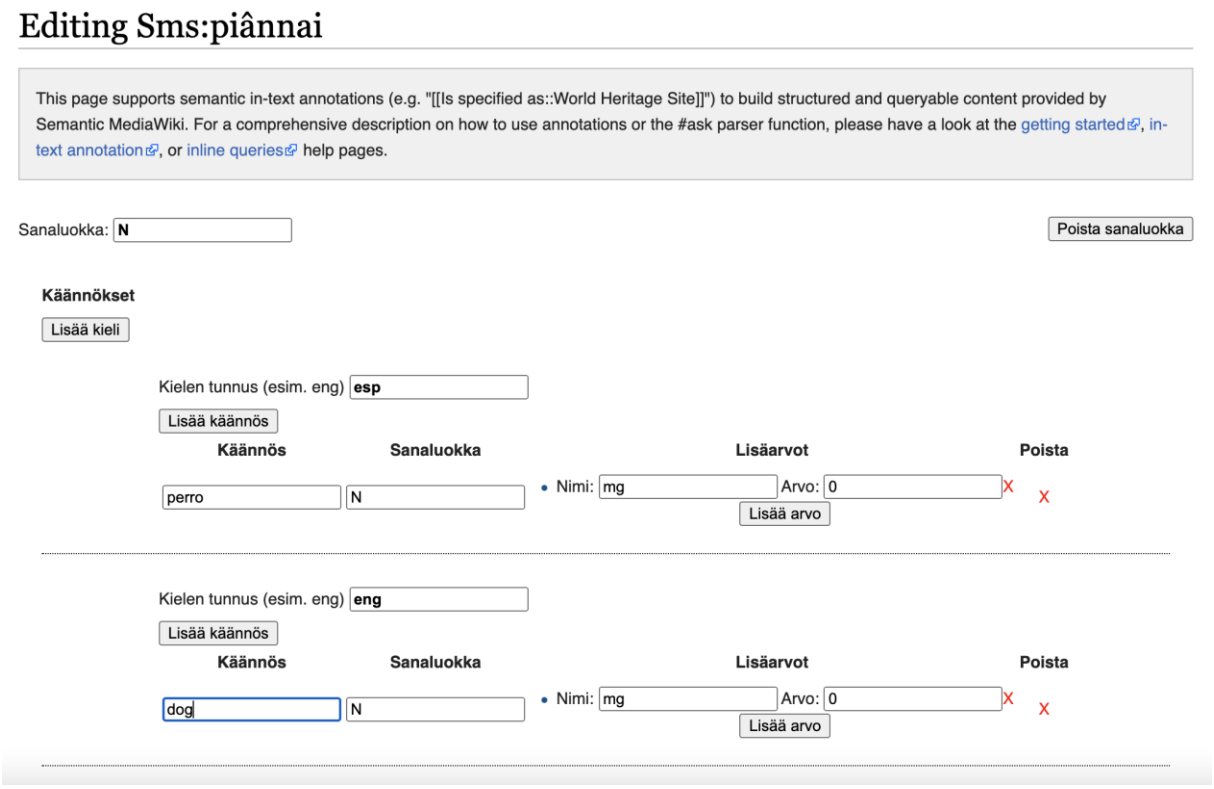
Un paso muy importante en la documentación de una lengua minoritaria es el trabajo lexicográfico. Esto resulta en un diccionario que puede ser útil tanto para los hablantes nativos como para los que quieran aprender el idioma. Nosotros guardamos los diccionarios en el formato XML muy estructurado. Eso quiere decir que todo tipo de metadatos están en sus propios campos en vez de estar guardados en varios partes de una entrada lexicográfica de forma no estructurada. Esto es importante ya que no solo queremos guardar los diccionarios para el uso de un ser humano, sino también queremos que sean legibles de forma automática.

Nuestro sistema *Akusanat*[3] (HÄMÄLÄINEN y RUETER, 2019a) está basado en *MediaWiki* y permite visualizar el contenido de los diccionarios XML para todo tipo de usuarios. Los datos del *MediaWiki* están sincronizados con los archivos XML utilizando el control de versiones *Git*. Esto significa que, si alguien modifica una entrada lexicográfica en Akusanat, estos cambios resultarán en un cambio en el diccionario XML almacenado en *GitHub*. Si alguien modifica los diccionarios XML directamente, Akusanat descargará los

[1] Disponible en: http://divvun.no/korrektur/korrektur.html. Accedido en: 11 jul 2021
[2] Disponible en: http://divvun.no/keyboards/mobileindex.html. Accedido en: 11 jul 2021
[3] Disponible en: https://akusanat.com. Accedido en: 11 jul 2021

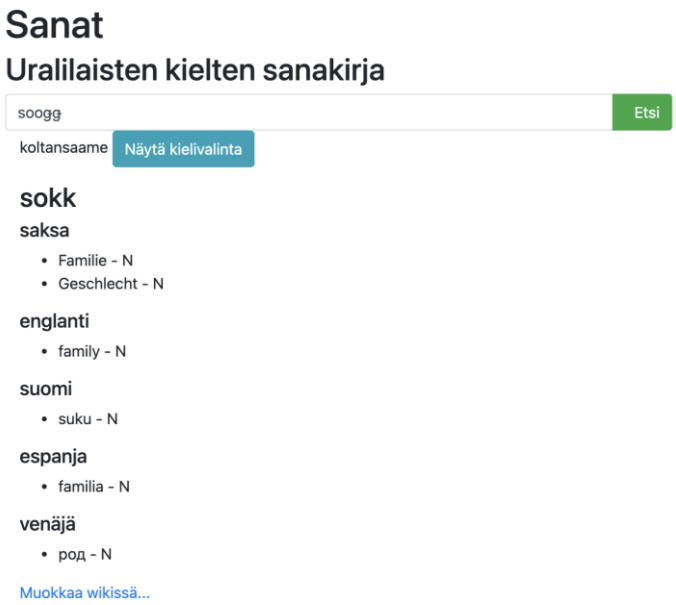
nuevos cambios desde *GitHub* y actualizará su base de datos de forma automática. Esto hace posible que usuarios avanzados puedan editar los archivos XML directamente con su herramienta favorita y que los usuarios menos avanzados puedan hacer cambios en línea con una interfaz. Akusanat no deja a los usuarios modificar la sintaxis *Wiki* directamente, sino que muestra un formulario que asegura que los cambios siguen siendo estructurados y compatibles con XML (Fig. 1).

Figura 1: El formulario en Akusanat para editar la entrada *piânnai* (perro) en sami de skolt



Para la búsqueda utilizamos los transductores para procesar el input del usuario. Esto quiere decir que el usuario puede buscar una palabra en cualquiera de sus conjugaciones morfológicas, ya que el TEF puede lematizar palabras de forma automática. También es posible buscar palabras escritas de manera errónea. Los transductores contienen información sobre los errores de ortografía más comunes en cada lengua, lo que nos permite resolver el lema, aunque la palabra no haya sido escrita según la norma ortográfica. Esto es importante en el caso de las lenguas con las que trabajamos, ya que las normas ortográficas no son tan establecidas como en el caso de lenguas mayoritarias.

Figura 2: La interfaz para realizar búsquedas en Akusanat



La Figura 2 muestra la interfaz para buscar palabras en el diccionario. En el ejemplo, el término de búsqueda es la palabra en sami de skolt *soogg* que es el genitivo de la palabra *sokk* que significa *familia*. Nuestro sistema lematiza el término de búsqueda automáticamente con el TEF de sami de skolt, y muestra la entrada para el lema *sokk* al usuario.

La idea de utilizar *MediaWiki*, y sobre todo *Semantic MediaWiki*, para crear diccionarios no es nueva, ya que ya existen varios proyectos que utilizan la tecnología como su base (MULJADI et al., 2006; BON y NOWAK, 2013; DUEÑAS y GÓMEZ, 2015). Aunque, sin duda, *MediaWiki* tiene sus ventajas, en práctica nosotros hemos tenido que programar nuestras propias extensiones *MediaWiki* para añadir la funcionalidad necesaria; el formulario para editar, la sincronización de *MediaWiki*-XML, la búsqueda con los transductores etc. El problema que hemos experimentado muchas veces es que el funcionamiento interno de *MediaWiki* cambia demasiado a menudo. Esto significa que, si queremos mantener el *MediaWiki* actualizado con las últimas actualizaciones de seguridad, tenemos que hacer muchos cambios en nuestro código fuente para mantener el funcionamiento de nuestras extensiones con la nueva versión de *MediaWiki*. Aun así, seguimos utilizando y desarrollando Akusanat[4] por el momento, ya que ofrece un entorno sencillo para los usuarios. En el siguiente apartado, describimos el otro sistema que estamos desarrollando, y que, algún día, podrá sustituir Akusanat.

2.2 Redacción de Diccionarios

En este apartado, describimos el sistema Ve′rdd[5] (ALNAJJAR et al., 2020). El sistema funciona con los mismos diccionarios XML que Akusanat y puede usarse en línea de la misma

[4] El código fuente está disponible en https://github.com/mikahama/akusanat. Accedido en: 11 jul 2021

[5] Disponible en: https://akusanat.com/verdd/. Accedido en: 11 jul 2021

manera. La diferencia está en el enfoque del sistema. Ve′rdd no es un sistema para visualizar las entradas lexicográficas para un usuario final, sino un sistema creado específicamente para redactar diccionarios tanto digitales como impresos. Para realizar el sistema, hemos colaborado con un grupo de profesionales que trabajan con diccionarios impresos.

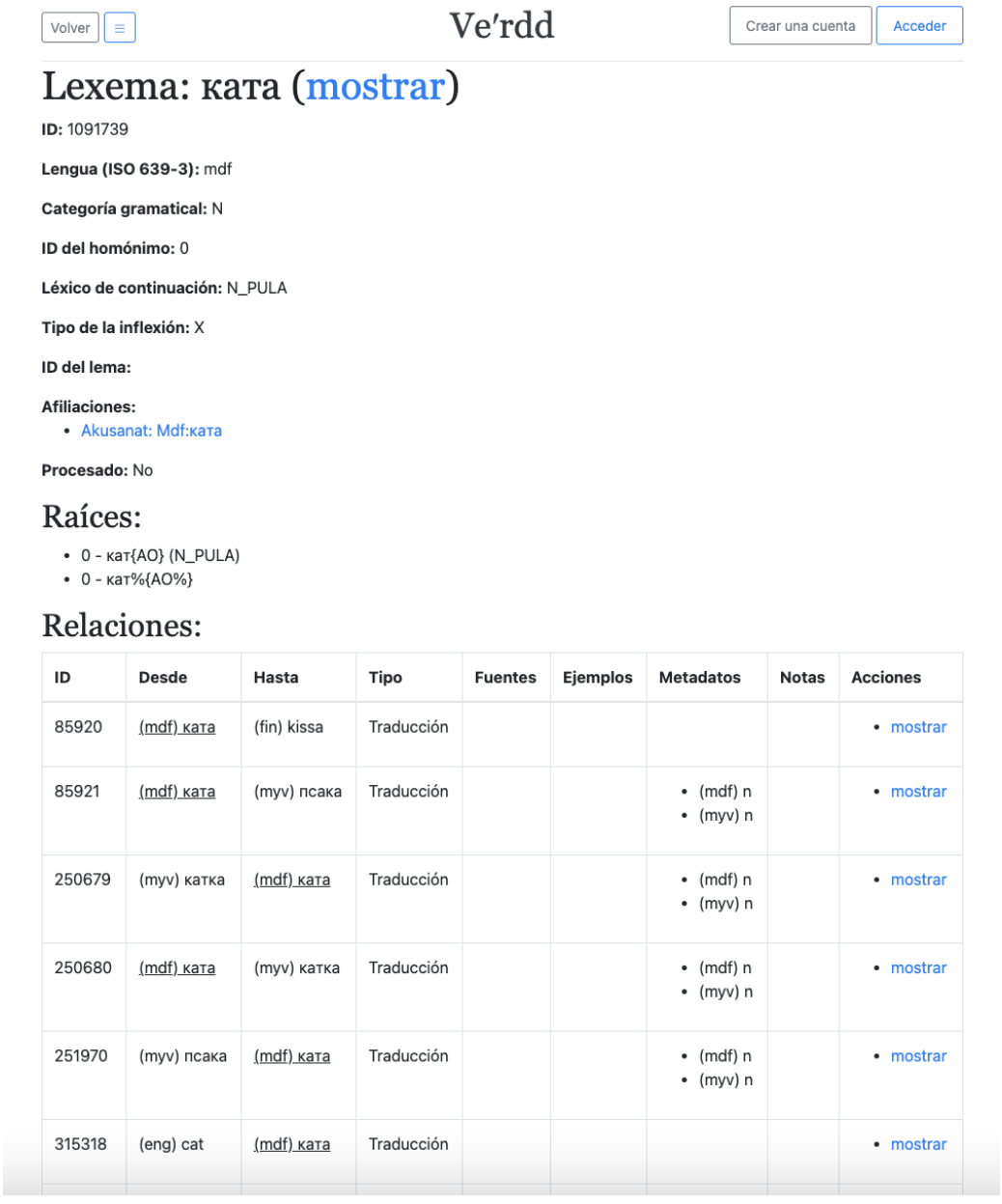
Con las lenguas con las que trabajamos, la documentación lexicográfica no empieza desde cero, ya que tanto las lenguas sami habladas en los países nórdicos como las lenguas pérmicas y mordvínicas habladas en Rusia han recibido mucha atención por su documentación durante el siglo pasado. Por ejemplo, para el sami de skolt existe el diccionario de Sammallahti & Mosnikoff (1991), y existen varios estudios sobre las lenguas mordvínicas (AASMÄE et al., 2016; GRÜNTHAL, 2016) y pérmicas (HAMARI, 2011; KLUMPP, 2016). Si hay diccionarios en forma digital, existen en un formato sin estructurar como un archivo *Word*, CSV o bien PDF producido con un sistema de ROC (reconocimiento óptico de caracteres). Por este motivo, Ve′rdd incluye funcionalidad para importar datos lexicográficos de formatos sin estructurar. Hemos prestado mucha atención en la calidad de la conversión, ya que, en el caso de nuestras lenguas, sobre todo, en el caso de sami de skolt, es muy frecuente que exista el mismo carácter con muchas codificaciones diferentes. Por ejemplo, ′ (U+02B9 modificador de letra prime) es un carácter muy común en sami de skolt, pero por la razón del teclado finlandés, es a menudo escrito como ' (U+0027 apóstrofo) o bien ´ (U+00B4 acento agudo). Ve′rdd está programado para tomar en cuenta los caracteres posibles de la lengua e intentar a corregir los caracteres erróneos automáticamente.

Figura 3: La interfaz para realizar búsquedas y filtrar entradas léxicas en Ve′rdd

La Figura 3 muestra la interfaz para realizar búsquedas y filtrar palabras en Ve′rdd. La interfaz está diseñada para apoyar el flujo de trabajo del editor del diccionario. Por ejemplo, es

posible mostrar solamente las entradas sin procesar. Esto significa entradas que nadie ha verificado después de importar los datos desde un formato sin estructurar. Para facilitar el desarrollo de los transductores es también posible ordenar y filtrar las palabras según el léxico de continuación. El léxico de continuación es una forma de expresar que una palabra se conjuga del mismo modo que otras palabras con el mismo léxico de continuación.

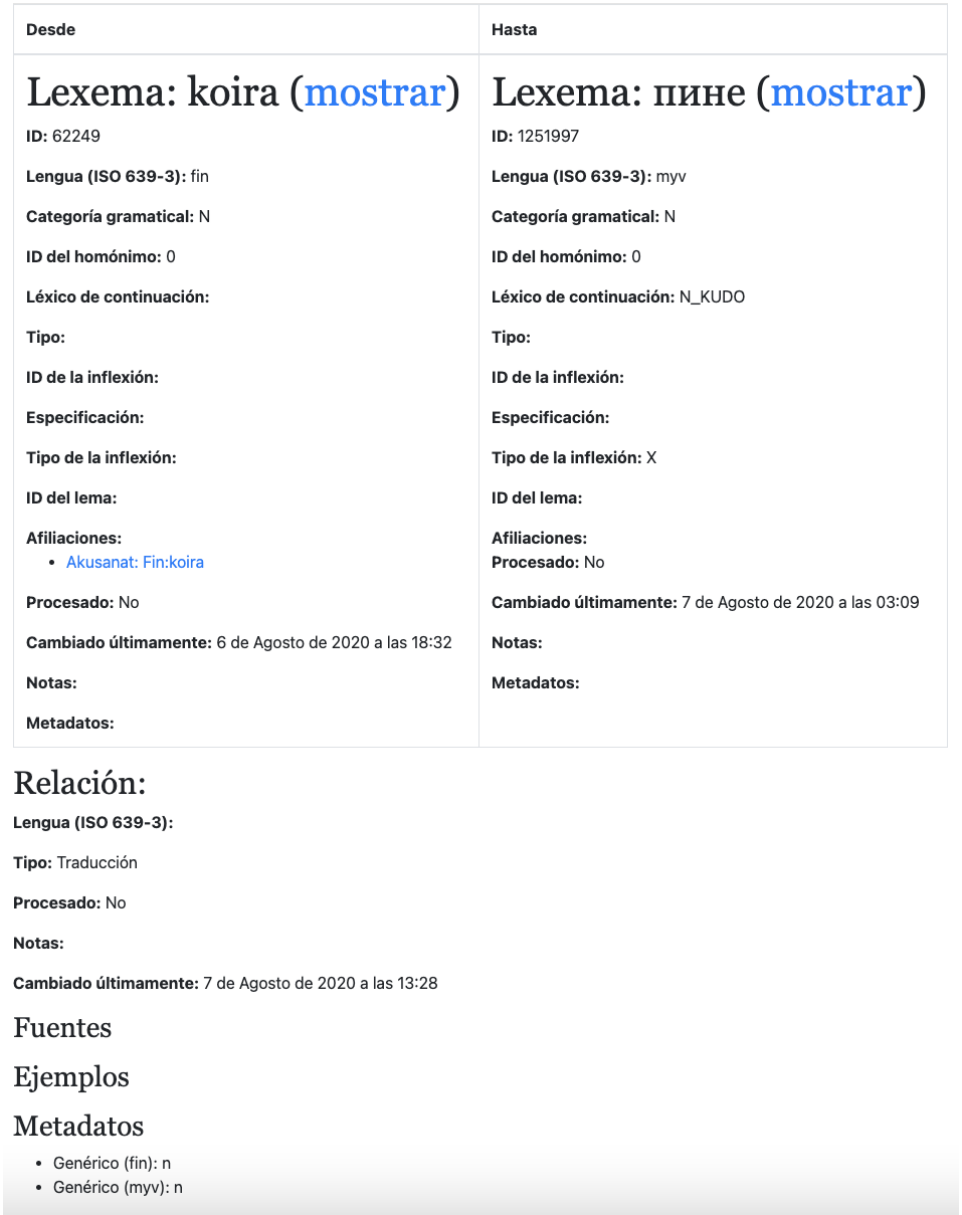
Figura 4: La interfaz para editar entradas léxicas en Ve′rdd



Aparte de solamente buscar y filtrar entradas léxicas, es importante tener la posibilidad de editarlas. La Figura 4 muestra la interfaz para inspeccionar una entrada en el diccionario. Si el usuario está conectado con su cuenta, además de ver, puede editar la información de la entrada. Ve′rdd está diseñado para ser una herramienta para diccionarios multilingües, por eso

una entrada está conectada a otras entradas en el sistema. En la figura, se pueden ver relaciones de tipo traducción que conectan una palabra a sus traducciones en otras lenguas. También es posible definir otro tipo de relaciones entre lenguas como relaciones etimológicas. Las relaciones pueden existir entre las palabras de la misma lengua, por ejemplo, es posible indicar palabras compuestas o bien derivaciones con las relaciones. Como los transductores contienen información derivativa, Ve′rdd añade automáticamente este tipo de relaciones al importar un diccionario sin estructurar.

Figura 5: La interfaz para comparar dos entradas relacionadas en Ve′rdd



Ve′rdd puede visualizar la relación entre dos palabras enlazadas con algún tipo de relación para verificar que una palabra en una lengua está enlazada al homónimo correcto en otra lengua (Figura 5). También es posible editar el tipo de la relación o bien borrar las relaciones innecesarias.

En todo momento, Ve′rdd tiene la posibilidad de exportar el diccionario en formatos distintos. Los más importantes para nosotros son el XML de Giella que puede utilizarse para generar los transductores y el código Latex. Desde el código Latex, es posible generar un PDF para imprimir el diccionario. El formato Latex hace posible cambiar el estilo del diccionario sin cambiar el contenido, si hay cambios en Ve′rdd, es posible actualizar el contenido del diccionario sin cambiar el estilo definido en Latex. Esta funcionalidad ha sido un objetivo importante para nosotros ya que el trabajo hecho en Ve′rdd no debería únicamente servirse para lo digital sin que también para editar diccionarios imprimidos.

2.3 Recursos para el PLN

Nuestros sistemas para editar diccionarios son directamente útiles para el desarrollo de los transductores ya que podemos exportar el léxico en el formato necesario para HFST (LINDÉN et al., 2013). HFST es la herramienta que utilizamos para crear los transductores. Nosotros disponemos de transductores para el sami de skolt (RUETER y HÄMÄLÄINEN, 2020), erzya y moksha (RUETER et al., 2020) y las lenguas komi. Los transductores pueden utilizarse para lematizar palabras, analizar su morfología o bien generar formas conjugadas. Estos transductores son difíciles de compilar para personas que no trabajan con los transductores a menudo. Por este motivo, nosotros compilamos todos los transductores cada noche y los distribuimos mediante nuestra página web[6]. No sólo compilamos nuestros transductores sino todos los transductores para todas las lenguas en la infraestructura Giella.

Sin embargo, queda difícil usar los transductores como tal. Por este motivo, hemos desarrollado una librería *Python* llamada UralicNLP (HÄMÄLÄINEN, 2019). Con la librería, es posible descargar transductores y diccionarios compilados, y usarlos directamente en *Python*. Fig 6 muestra cómo utilizar nuestros transductores desde *Python*. En la segunda línea de código, se analiza la palabra *шляпа* (sombrero) en erzya (myv). El resultado indica que la palabra es un nombre (+N) indefinido (+Indef) en el singular (+Sg) del nominativo (+Nom). En la cuarta línea generamos la forma conjugada de la misma palabra en plural (+Pl). El resultado es la palabra plural *шляпат*.

Figura 6: Un ejemplo del uso de UralicNLP

```
>>> from uralicNLP import uralicApi
>>> uralicApi.analyze("шляпа", "myv")
[('шляпа+N+Sg+Nom+Indef', 0.0)]
>>> uralicApi.generate("шляпа+N+Pl+Nom+Indef", "myv")
[('шляпат', 0.0)]
>>>
```

[6] Disponible en: https://models.uralicnlp.com/nightly/. Accedido en: 11 jul 2021

Los transductores producen todas las interpretaciones posibles de una palabra. En el caso de las lenguas urálicas, existe mucha homonimia en la conjugación. Esto significa que, si utilizamos los transductores con un texto corriente, no podemos lematizar las palabras en su contexto ya que los transductores producen todos los lemas posibles. Por este motivo, utilizamos desambiguadores con gramática de restricciones (KARLSSON, 1990) basados en la herramienta llamada VISL CG-3 (BICK y DIDRIKSEN, 2015). Las reglas de gramática de restricciones eliminan morfologías que no son posibles en la oración, y resultan en una oración morfológicamente desambiguada.

Figura 7: Un ejemplo del uso del desambiguador de komi-ziriano

```
>>> from uralicNLP.cg3 import Cg3
>>> oracion = "Ныв ёртыслы гижис письмö"
>>> cg = Cg3("kpv")
>>> print(cg.disambiguate(oracion.split(" ")))
Warning: Line 6 had empty tag.
[('Ныв', [<ныв – N, Sg, Nom, <W:0.000000>>]), ('ёртыслы', [<ёрт – N, Sg, Dat, Px
Sg3, So/PC, <W:0.000000>>]), ('гижис', [<гижны – V, TV, Ind, Prt1, Sg3, <W:0.000
000>>]), ('письмö', [<письмö – N, Sg, Nom, <W:0.000000>>])]
>>>
```

En la Figura 7, podemos ver cómo se utilizan los desambiguadores de gramática de restricciones mediante UralicNLP. En la tercera línea se inicializa el objeto de desambiguación para el komi-ziriano (kpv) y en la cuarta línea se llama el método de desambiguación del objeto con una oración. El resultado contiene la forma de la palabra en la oración, su lematización y su morfología para cada palabra de la oración.

Aparte de los diccionarios estructurados y las herramientas basadas en reglas, disponemos de treebanks de las dependencias universales para el sami de skolt, moksha, erzya (RUETER y TYERS, 2018), komi-ziriano (PARTANEN et al., 2018) y komi-permio (RUETER et al., 2020). Estos treebanks contienen anotaciones sintácticas con las etiquetas morfológicas de las dependencias universales. Con los últimos treebanks, también hemos añadido las etiquetas morfológicas que producen los transductores para facilitar el uso de los dos recursos juntos.

3 Los Beneficios de la Documentación Digital

La documentación digital nos ha permitido utilizar los últimos métodos en el mundo de PLN para aumentar automáticamente los datos que tenemos en los diccionarios. Cómo todos los diccionarios XML con los que trabajamos son multilingües, el primer paso que hemos dado con la tecnología de PLN ha sido la predicción de traducciones (HÄMÄLÄINEN et al., 2018). La idea ha sido la siguiente: si el diccionario de sami de skolt contiene traducciones al finés, alemán e inglés, y el diccionario de erzya contiene traducciones al finés, inglés, ruso y francés, entonces, con esta información, debería ser posible deducir automáticamente traducciones de sami de skolt al ruso y francés y de erzya al alemán dada la existencia de dos lenguas en común:

finés e inglés. Con un modelo probabilístico hemos aumentado la cantidad de traducciones en los diccionarios de sami de skolt, erzya, moksha y komi-ziriano. Después de obtener los resultados automáticos, hemos comprobado las traducciones de forma manual antes de incluirlas en los diccionarios XML.

Como las redes neuronales exigen una gran cantidad de datos para ser entrenados, es habitual creer que su uso no es posible en el caso de las lenguas en peligro. Nosotros hemos tomado la perspectiva que podemos generar la cantidad de datos necesaria para una red neuronal con nuestras herramientas morfológicas. Utilizando los treebanks y los transductores, hemos generado datos para entrenar una red neuronal para realizar desambiguación en vez de utilizar la gramática de restricciones para erzya y komi-ziriano (ENS et al., 2019). La idea ha sido generar todos los análisis posibles para las palabras en los treebanks y entrenar la red neuronal para desambiguar los análisis con el análisis del treebank. También hemos podido utilizar las redes neuronales para aumentar las relaciones etimológicas en el diccionario de sami de skolt (HÄMÄLÄINEN y RUETER, 2019b).

Nuestras herramientas son compatibles con la infraestructura de Giella. Esto ha hecho posible utilizar nuestros diccionarios y transductores directamente en su plataforma en línea para aprender idiomas (ANTONSEN y ARGESE, 2018), en los teclados de *Android* y *iPhone* y en la corrección ortográfica para *Word* y *OpenOffice* desarrollados por Divvun[7] en colaboración con Giella.

La documentación digital tiene claramente sus beneficios, ya que podemos realizar aprendizaje automático con diccionarios estructurados y transductores de morfología. Por este motivo el proyecto conducido en la universidad de Oulu para redactar el nuevo diccionario finés-sami de skolt ha optado por utilizar Ve′rdd para crear el diccionario. Hemos trabajado juntos con los empleados del proyecto para aumentar la funcionalidad de nuestro sistema. Ve′rdd ha hecho posible el trabajo simultáneo de los editores que, sin Ve′rdd, hubiesen utilizado Excel y Word para su trabajo. Esto habría significado una posibilidad pérdida de producir un diccionario estructurado para el interés de PLN y un diccionario impreso al mismo tiempo.

Como hemos desarrollado todos los recursos y herramientas de forma abierta, incluso investigadores ajenos han empezado a utilizar los recursos mediante UralicNLP. La librería de *Python* ha sido utilizada, entre otros, por Creutz y Sjöblom (2019) para corregir textos escritos por hablantes no-nativos. Avikainen (2019) ha utilizado la librería para investigar noticias en periódicos antiguos digitalizados y Rämö (2020) para generar títulos de noticias de forma automática.

[7] Disponible en: http://divvun.no/. Accedido en: 11 jul 2021

Conclusiones y el trabajo futuro

En este artículo hemos presentado nuestras soluciones para la documentación digital de lenguas urálicas amenazadas. Como trabajamos con muchas lenguas a la vez en colaboración con otro proyecto de infraestructura para lenguas minoritarias, Giella, podemos diseñar nuestras herramientas e infraestructura de tal modo que comenzar el trabajo con una nueva lengua no requiere que inventemos la rueda de nuevo, sino que podemos incorporar la lengua fácilmente en todas las herramientas compatibles. Esto quiere decir, que con el trabajo lexicográfico en Ve'rdd, podemos crear transductores de forma fácil. Y, al tener un transductor, los recursos de la lengua ya pueden utilizarse para tareas más complejas como teclados, corrección ortográfica y todo tipo de tareas de PLN mediante UralicNLP.

A la hora de escribir este artículo, hemos comenzado a trabajar con la lengua apurinã. Gracias a los esfuerzos de su documentación lingüística (FACUNDES, 2000), podemos expresar sus reglas en el formalismo de TEF. Hemos comenzado a incorporar sus materiales lexicográficos en nuestra infraestructura. Aunque el trabajo aún está en sus fases iniciales, las experiencias por ahora han sido positivas. La morfología de apurinã es muy distinta a la morfología urálica, pero la robustez de la tecnología de TEF nos permite modelar su morfología.

Estamos muy interesados en colaborar con la documentación digital de todo tipo de lenguas amenazadas. Desde el punto de vista del PLN, los recursos multilingües son más útiles para todo tipo de tareas que los recursos mono- o bilingües. Como ya hemos visto en este artículo, si tenemos diccionarios multilingües en el mismo sistema, podemos aumentar la cantidad de traducciones que tienen de una forma automática.

Todas las herramientas y recursos que hemos descrito en este artículo están disponibles de forma abierta en *GitHub*[8]. En nuestro trabajo, siempre utilizamos licencias abiertas y almacenamos datos de forma permanente en Zenodo.

Referencias

AASMÄE, N.; PAJUSALU, K.; KABAJEVA, N. Gemination in the Mordvin Languages. *Linguistica Uralica*, 52(2). 2006. Disponible en: https://www.ceeol.com/search/article-detail?id=396961. Accedido en: 11 jul 2021

AHMADNIA, B.; SERRANO, J.; HAFFARI, G. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. En *Proceedings of RANLP*. 2017 p. 24-30. DOI: 10.26615/978-954-452-049-6_004. Accedido en: 11 jul 2021

ALNAJJAR, K.; HÄMÄLÄINEN, M.; RUETER, J.; PARTANEN, N. Ve'rdd. Narrowing the Gap between Paper Dictionaries, Low-Resource NLP and Community Involvement. En *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*. 2020 p. 1-6. DOI: 10.18653/v1/2020.coling-demos.1. Accedido en: 11 jul 2021

[8] Los enlaces están disponibles en https://uralicnlp.com/. Accedido en: 11 jul 2021

ANTONSEN, L.; ARGESE, C. Using authentic texts for grammar exercises for a minority language. En *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018)*. Linköping Electronic Conference Proceedings. 2018 p. 1–9. Disponible en: https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=152&Article_No=1 Accedido en: 11 jul 2021

AVIKAINEN, J. *A Method for Wavelet-Based Time Series Analysis of Historical Newspapers*. Universidad de Helsinki. Tesina de Master. 2019. Disponible en: https://helda.helsinki.fi/handle/10138/310021. Accedido en: 11 jul 2021

BICK, E.; DIDRIKSEN, T. Cg-3—beyond classical constraint grammar. En *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. 2015. p. 31-39. Disponible en: https://aclanthology.org/W15-1807. Accedido en: 11 jul 2021

BON, B.; NOWAK, K. Wiki lexicographica. Linking medieval latin dictionaries with semantic MediaWiki. En *Electronic lexicography in the 21st century: thinking outside the paper:* proceedings of the eLex 2013 conference, Estonia, 2013, p. 407-420. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=4565204. Accedido en: 11 jul 2021

BONTOGON, M.; ARPPE, A.; ANTONSEN, L.; THUNDER, D.; LACHLER, J. Intelligent Computer Assisted Language Learning (ICALL) for nêhiyawêwin: An In-Depth User-Experience Evaluation. En *Canadian Modern Language Review,* 74(3). 2018. p. 337-362. DOI: https://doi.org/10.3138/cmlr.4054. Accedido en: 11 jul 2021

CHEN, X.; SUN, Y.; ATHIWARATKUN, B.; CARDIE, C.; WEINBERGER, K. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6. 2018. p. 557-570. Disponible en: https://arxiv.org/abs/1606.01614. Accedido en : 11 jul 2021

CREUTZ, M.; SJÖBLOM, E. E. Toward automatic improvement of language produced by non-native language learners. En *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*. 2019. p. 20-30. Disponible en: https://aclanthology.org/W19-6303. Accedido en: 11 jul 2021

DUEÑAS, G.; GÓMEZ, D. A bilingual dictionary with Semantic Mediawiki: The language Saliba's case. En The *4th International Conference on Language Documentation and Conservation (ICLDC)*. 2015. Disponible en: http://hdl.handle.net/10125/25338. Accedido en : 11 jul 2021

ENS, J.; HÄMÄLÄINEN, M.; RUETER, J.; PASQUIER, P. Morphosyntactic Disambiguation in an Endangered Language Setting. En *22nd Nordic Conference on Computational Linguistics (NoDaLiDa): Proceedings of the Conference.* 2019. p. 345-349. Disponible en: https://aclanthology.org/W19-6139. Accedido en: 11 jul 2021

FACUNDES, S. D. S. The language of the Apurinã people of Brazil. Buffalo: State University of New York at Buffalo (Dissertation). 2000. Disponible en: http://www.etnolinguistica.org/tese:facundes-2000. Accedido en: 11 jul 2021

GRÜNTHAL, R. Transitivity in Erzya: Second language speakers in a grammatical focus. En *Mordvin languages in the field*. Finno-Ugrian Society. 2016. p. 291-318. Disponible en:

https://researchportal.helsinki.fi/en/publications/transitivity-in-erzya-second-language-speakers-in-a-grammatical-f. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M. UralicNLP: An NLP Library for Uralic Languages. *Journal of open source software*, 4(37). 2019. Disponible en: https://joss.theoj.org/papers/10.21105/joss.01345. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M.; RUETER, J. An open online dictionary for endangered Uralic languages. En *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography*, 111. 2019a. Disponible en: http://hdl.handle.net/10138/305873. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M.; RUETER, J. Finding Sami Cognates with a Character-Based NMT Approach. En *Proceedings of the 3rd Workshop on Computational Methods in the Study of Endangered Languages: (Volume 1) Papers.* 2019b. p. 39-45. Disponible en: https://aclanthology.org/W19-6006. Accedido en: 11 jul 2021

HÄMÄLÄINEN, M.; TARVAINEN, L. L.; RUETER, J. Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018 p. 862-867. Disponible en: https://aclanthology.org/L18-1138. Accedido en: 11 jul 2021

HAMARI, A. The abessive in the Permic languages. En *Suomalais-Ugrilaisen Seuran Aikakauskirja*, 2011(93). 2011. p.37-84. DOI: https://doi.org/10.33340/susa.82172. Accedido en: 11 jul 2021

HIMMELMANN, N. P. Documentary and descriptive linguistics. *Linguistics*, 36. 1998. p. 161-196. DOI: https://doi.org/10.1515/ling.1998.36.1.161. Accedido en : 11 jul 2021

HUNT, B.; CHEN, E.; SCHREINER, S. L.; SCHWARTZ, L. Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 2019 pp. 122–126. DOI: 10.18653/v1/N19-4021. Accedido en: 11 jul 2021

IRVINE, A.; CALLISON-BURCH, C. Hallucinating phrase translations for low resource mt. En *Proceedings of the Eighteenth Conference on Computational Natural Language Learning.* 2014. p. 160-170. DOI: 10.3115/v1/W14-1617. Accedido en: 11 jul 2021

KARLSSON, F. Constraint Grammar as a Framework for Parsing Unrestricted Text. En *Proceedings of the 13th International Conference of Computational Linguistics, Vol. 3*. 1990. p. 168-173. DOI: https://doi.org/10.3115/991146.991176. Accedido en : 11 jul 2021

KLUMPP, G. Semantic functions of complementizers in Permic languages. En *Complementizer Semantics in European Languages*, 2016. p. 529-586. DOI: https://doi.org/10.1515/9783110416619-016. Accedido en: 11 jul 2021

LINDÉN, K.; AXELSON, E.; DROBAC, S.; HARDWICK, S.; KUOKKALA, J.; NIEMI, J.; PIRINEN, T.; SILFVERBERG, M. HFST - A System for Creating NLP Tools. En *Systems and Frameworks for Computational Morphology. Communications in Computer and Information Science*. 380. Humboldt-Universität in Berlin: Springer. 2013. p. 53-71. DOI: 10.1007/978-3-642-40486-3_4. Accedido en: 11 jul 2021

LITTELL, P.; PINE, A.; DAVIS, H. Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. En *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics. 2017. p. 141–150. DOI: 10.18653/v1/W17-0119. Accedido en: 11 jul 2021

MOSHAGEN, S.; RUETER, J.; PIRINEN, T.; TROSTERUD, T.; TYERS, F. M. Open-source infrastructures for collaborative work on under-resourced languages. En *Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*. 2014. p. 71-77. Disponible en: http://www.syros.aegean.gr/users/spyrosv/papers/ccurl14.pdf#page=78. Accedido en: 11 jul 2021

MULJADI, H.; TAKEDA, H.; KAWAMOTO, S.; KOBAYASHI, S.; FUJIYAMA, A. Towards a Semantic Wiki-Based Japanese Biodictionary. En *Proceedings of the First Workshop on Semantic Wikis - From Wiki to Semantics*. 2006. Disponible en: http://www-kasm.nii.ac.jp/papers/takeda/06/muljadi06eswc.pdf. Accedido en: 11 jul 2021

NASUTION, A.H.; MURAKAMI, Y.; ISHIDA, T. Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages. En *Proceedings of the 11th Language Resources and Evaluation Conference. European Language Resource Association.* 2018. Disponible en: https://aclanthology.org/L18-1536. Accedido en: 11 jul 2021

PARTANEN, N.; BLOKLAND, R.; LIM, K.; POIBEAU, T.; RIESSLER, M. The first Komi-Zyrian universal dependencies treebanks. En *Second Workshop on Universal Dependencies (UDW 2018)*. 2018. p. 126-132. DOI: 10.18653/v1/W18-6015. Accedido en: 11 jul 2021

RÄMÖ, M. (Re)lexicalization of auto-written news with contextual and cross-lingual word embeddings. Universidad de Helsinki. Tesina de Master. 2020. Disponible en: https://helda.helsinki.fi/handle/10138/321924. Accedido en: 11 jul 2021

RUETER, J.; HÄMÄLÄINEN, M. FST Morphology for the Endangered Skolt Sami Language. En *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020).* 2020. p. 250-257. Disponible en: https://aclanthology.org/2020.sltu-1.35. Accedido en: 11 jul 2021

RUETER, J.; HÄMÄLÄINEN, M.; PARTANEN, N. Open-Source Morphology for Endangered Mordvinic Languages. En *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS).* The Association for Computational Linguistics. 2020. p. 94–100. DOI: 10.18653/v1/2020.nlposs-1.13. Accedido en: 11 jul 2021

RUETER, J. M.; HÄMÄLÄINEN, M. Synchronized Mediawiki based analyzer dictionary development. En *3rd International Workshop for Computational Linguistics of Uralic Languages Proceedings of the Workshop*. 2017. DOI: 10.18653/v1/W17-0601. Accedido en: 11 jul 2021

RUETER, J.; PARTANEN, N.; PONOMAREVA, L. On the questions in developing computational infrastructure for Komi-Permyak. En *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages.* 2020 p. 15-25. Disponible en: https://aclanthology.org/2020.iwclul-1.3. Accedido en: 11 jul 2021

RUETER, J. M.; TYERS, F. M. Towards an open-source universal-dependency treebank for Erzya. En *Proceedings of International Workshop for Computational Linguistics of Uralic Languages.* 2018. DOI: 10.18653/v1/W18-0210. Accedido en: 11 jul 2021

SAMMALLAHTI, P.; MOSNIKOFF, J. Suomi-Koltansaame sanakirja. LÄÄ′DD-SÄÄ′m SÄÄ′NNÊ′RJJ Ohcejohka: Girjegiisá Oy. 1991.

Recibido: 29/01/2021.

Aceptado: 05/04/2021.