The Digital Language Diversity Project

Kit de Supervivencia Lingüística Digital del Euskera

Recomendaciones del DLDP para mejorar la Vitalidad Digital del euskera

Erasmus+

Edición

Kit de Supervivencia Lingüística Digital del Eusker — Recomendaciones del DLDP para mejorar la Vitalidad Digital del euskera

Autores:
Klara Ceberio Berger, Antton Gurrutxaga Hernaiz, Paola Baroni, Davyth Hicks, Eleonore Kruse, Valeria Quochi, Irene Russo, Tuomo Salonen, Anneli Sarhimaa, Claudia Soria

Este trabajo se ha llevado a cabo en el marco del Proyecto de Diversidad Lingüística Digital (www.dldp.eu/es), financiado por la Unión Europea dentro del Programa Erasmus+ (Acuerdo de Financiación nº 2015-1-IT02-KA204- 015090).

© 2018

Este trabajo está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

Diseño de portada: Eleonore Kruse

Aviso legal
Esta publicación refleja los puntos de vista de los autores y la Agencia y la Comisión Nacional Erasmus+ no son responsables del uso de la información que contiene.

www.dldp.eu
www.facebook.com/digitallanguagediversity
dldp@dldp.eu
www.twitter.com/dldproject

Vista rápida de las recomendaciones

Capacidad digital		
Indicador		**Recomendación**
Disponibilidad de recursos lingüísticos		Desarrollar recursos lingüísticos intermedios y avanzados
		Elaboración de diccionarios: diversidad, tamaño, especialización y divulgación
		Aumentar el tamaño y la diversidad del corpus
		Recabar datos lingüísticos disponibles públicamente desde las redes sociales
		Utilizar herramientas para el análisis del corpus y alimentar tu diccionario con los datos sobre la lengua en uso
		Primeros pasos hacia la síntesis y reconocimiento del hablan

Presencia y uso digital		
Indicador		**Recomendación**
Uso para la comunicación electrónica		Valorar la utilización de las lenguas regionales o minoritarias para la comunicación interpersonal
Disponibilidad de los medios de Internet		Incrementar la cantidad de contenido y diversificar los tipos de medios de Internet
		Incrementar la cantidad de contenido de texto (páginas web, blogs, foros)
		Crear o alimentar un archivo web de documentos y registros
		Transmitir online utilizando herramientas de software gratuitas
		Registrar historias digitales en tu propio idioma
		Promover iniciativas de subtitulación
Wikipedia		Eleva tu Wikipedia a un nivel mayor
		Iniciativas para aumentar el tamaño y la calidad de Wikipedia

Rendimiento digital		
Indicador		**Recomendación**
Disponibilidad de los servicios de Internet		Expandir el abanico de posibilidades para utilizar servicios de Internet en tu idioma
		Recabar información y experiencias de la comunidad de usuarios de tu lengua regional o minoritaria para determinar cuáles son los servicios más importantes y los más utilizados
		Valorar la utilización de la lengua del usuario en los negocios
		Desarrollar aplicaciones para smartphones
Redes sociales localizadas		Iniciativas para localizar las interfaces de usuario de las redes sociales
Software localizado: sistemas operativos y software básico		Reforzar las iniciativas para localizar los softwares gratuitos o privativos de carácter general más utilizados en la comunidad lingüística
		Considerar los videojuegos como una valiosa oportunidad de revitalización
Servicios de Traducción Automática		Aumentar el número de combinaciones lingüísticas; intentar incluir el inglés, si no está ya incluido

1. El euskera: un breve perfil del idioma

El euskera es la única lengua no indoeuropea en Europa occidental, y como lengua aislada no tiene ningún parentesco conocido con ninguna otra lengua. La clasificación más aceptada distingue los siguientes dialectos de este a oeste del País Vasco: vizcaíno, guipuzcoano, labortano, altonavarro, bajonavarro y suletino (Hualde y Ortiz de Urbina, 2003)[1].

El trabajo para establecer un estándar comenzó en la década de 1960 y las primeras decisiones de la Real Academia de la Lengua Vasca (Euskaltzaindia) se publicaron en 1968 (Conferencia de Arantzazu). A partir de entonces, el proceso de estandarización fue muy rápidamente.

A pesar de que surgieron algunas dificultades al inicio del proceso, esta estandarización ha tenido éxito. Está considerado como bien establecido y aceptado en su mayoría por la sociedad vasca (Hualde y Zuazo, 2007)[2].

La lengua vasca es hablada por una población de alrededor de 800.000 hablantes a ambos lados del Pirineo occidental: en cuatro provincias del País Vasco sur (Araba, Vizcaya, Gipuzkoa y Navarra) en el norte de España, y en los tres territorios históricos en el norte del País Vasco (Lapurdi, Baja Navarra, y Zuberoa) en el suroeste de Francia. Estas siete provincias se llaman Euskal Herria (País Vasco).

En la actualidad, el vasco es cooficial con el español en la Comunidad Autónoma del País Vasco (CAV), que comprende las provincias de Araba, Vizcaya y Gipuzkoa. También tiene un estatus oficial más restringido en Navarra. Vasco no tiene estatus oficial en el País Vasco Norte.

Según la ONU, el vasco se encuentra en una posición débil en el CAV y especialmente en Navarra; está en grave peligro de extinción en la parte francesa. En la CAV parece estar haciendo algún progreso, donde su supervivencia parece estar asegurada.

En cuanto a la educación, la situación también es muy diferente según el área (Gardner y Zalbide, 2005)[3]. Mientras que la mayoría de los escolares en la CAV y en el norte de Navarra se educan en euskera o en programas bilingües español-euskera, en la parte sur de Navarra y en la parte francesa se enseña menos en euskera.

A nivel universitario la situación es bastante similar. La situación es muy diferente según el área y dependiendo de la universidad. En la Universidad del País Vasco (UPV-EHU), por ejemplo, los estudiantes pueden estudiar parcial o totalmente en euskera, pero en otras universidades las posibilidades de estudiar en euskera son mucho menores que español o en francés, aunque se esté haciendo un esfuerzo importante para cambiar esta situación.

La presencia del vasco en los medios de comunicación en el País Vasco es todavía bastante restringida. Hay un solo periódico en vasco, Berria, que se distribuye en todo el País Vasco, así como muchos otros periódicos locales en euskera que se publican a diario. También hay algunas revistas semanales especializadas y diversas que publican sus contenidos solo en euskera. Cabe destacar el creciente número de periódicos digitales que han surgido en los últimos años.

Actualmente hay dos radios públicas que emiten exclusivamente en euskera para el País Vasco en su conjunto (Euskadi Irratia y Euskadi Gaztea). Sin embargo, al mismo tiempo, hay un número significativo de estaciones de radio territoriales, regionales y municipales que emiten en vasco.

El vasco no se usa mucho en la televisión. Hay tres canales principales que emiten exclusivamente en euskera: ETB1, ETB3 y Hamaika Telebista. También hay algunos canales de televisión locales en vasco.

1 Hualde, J .I . & Ortiz de Urbina, J . (eds .) (2003). *A grammar of Basque*. (Mouton Grammar library, 9 .) Berlin: Mouton de Gruyter.
2 Hualde, J .I . & Zuazo, K . (2007). The standardization of the Basque language. In *Language Problems and Language Planning*. 31(2):142-168.
3 Gardner, N . & Zalbide, M . (2005). Basque Acquisition Planning. In *International Journal of the Sociology of Language*, 174, 55-72.

En cuanto a las tecnologías del lenguaje, hay una serie de productos, tecnologías y recursos creados para el euskera. Hay varios diccionarios digitales, corpus y herramientas para síntesis de voz y reconocimiento de voz, así como correctores ortográficos y gramaticales. También se han desarrollado herramientas de traducción automática para español y euskera (Hernáez et al., 2012)[4].

4 Hernáez, I., Navas, E., Odriozola, I., Sarasola, K., Díaz de Ilarraza, A., Leturia, I., Díaz de Lezana, A., Oihartzabal, B., Salaberria, J. (2012). Euskara Aro Digitalean – Basque in the Digital Age. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg. http://www.meta-net.eu/whitepapers/volumes/basque

2. Vitalidad digital del euskera

2.1 Evaluación de la vitalidad digital del euskera

Para evaluar la vitalidad digital del euskera, hemos aplicado la Escala de Vitalidad Lingüística Digital, que se describe en el documento "How to use the Digital Language Vitality Scale" (Ceberio et al., 2018). Para recabar la información necesaria para aplicar la escala, hemos combinado el conocimiento de expertos, ofrecido por el socio vasco Elhuyar Fundazioa y varios asesores del DLDP, con el conocimiento de los hablantes, recabado a través de un cuestionario online.

Basándonos en nuestra investigación, el euskera puede situarse en el nivel "En desarrollo" (4) (Developing), alcanzando una puntuación de 4,8. El nivel "En desarrollo" se describe a continuación:

> "Una lengua que se encuentra en el nivel "En desarrollo" (Developing) de la vitalidad digital es utilizada para la comunicación y las redes sociales, pero es posible que su frecuencia sea aún ocasional. Puede que estén disponibles algunos medios y servicios digitales, así como una Wikipedia de mediano tamaño. Existen recursos lingüísticos básicos, y puede haber evidencia de otros recursos más avanzados. Al menos una de las redes sociales y sistemas operativos utilizados por la comunidad de los hablantes está localizada. Puede que haya un servicio o herramienta de traducción automática, al menos para una combinación lingüística."

La siguiente tabla muestra los indicadores para la vitalidad digital del idioma y la puntuación que se le ha asignado al euskera[5]:

Evidencia de conectividad	2
Alfabetización digital	4
Codificación	6
Población alfabetizada digitalmente	6
Disponibilidad de recursos lingüísticos	5
Uso para la comunicación electrónica	5
Uso en las redes sociales	5
Disponibilidad de medios de comunicación en Internet	5
Wikipedia	5
Disponibilidad de los servicios de Internet	5
Redes sociales localizadas	5
Software localizado	5
Herramientas/servicios de traducción automática	4
Dominio específico de Internet	6

2.2 Uso digital: cómo, con qué frecuencia y por qué utiliza la gente el euskera en la web

Los resultados del estudio muestran que el euskera está en forma digitalmente hablando, y que se utiliza activamente. Los encuestados tienen una alta competencia lingüística y un buen conocimiento de las herramientas y recursos digitales que existen. No obstante, tal y como se apunta al inicio del estudio, el perfil de los encuestados no refleja del todo los diferentes perfiles sociolingüísticos de los hablantes. En cuanto a la edad de los encuestados, la mayoría tiene más de 30 años. Se considera que la población más activa en el mundo digital es aquella que se encuentra en la franja de 15-34 años[6]; por eso, no podemos descartar la posibilidad de que este grupo esté infrarrepresentado en este estudio.

Un gran número de encuestados afirman que utilizan el euskera regularmente en Internet (páginas web, comunicación electrónica, blogs, Wikipedia, etc.). Esto también cuenta para las redes sociales, especialmente para Facebook y Twitter, que muestran una actividad significativa en euskera. Ya que estos tipos de redes se asocian mayormente a registros informales, esto refuerza la idea de que el euskera se utiliza y es necesario en la comunicación online oral y escrita del día a día, lo cual supone un signo innegable de vitalidad, y esto es un factor importante para la supervivencia

5 Para aclaración de los indicadores y la puntuación: Ceberio et al. 2018. "How to use the Digital Language Vitality Scale". DLDP. http://wp.dldp.eu/wp-content/uploads/2018/08/HowToUseTheDLVS.pdf

6 https: //www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/

digital de una lengua. Por el contrario, el euskera se utiliza menos en LinkedIn, el servicio de red social orientado a los negocios y al empleo. Esto puede ser el reflejo de la situación de la lengua vasca, donde el uso ha crecido en los entornos familiares e informales, pero hace falta más trabajo en contextos profesionales y formales.

Con respecto a servicios e interfaces digitales localizados en euskera, cabe destacar que la mayoría de los encuestados sabe de la existencia de versiones en euskera de los servicios mencionados. No obstante, pese a saber de su existencia, algunos de los encuestados no utilizan el euskera en sus dispositivos, aplicaciones o software. Esto puede deberse a muchas razones, pero este estudio nos muestra los hábitos digitales del hablante de euskera, teniendo en cuenta que: cerca de un tercio de los encuestados cree que es más fácil utilizar las herramientas en castellano, la forma de buscar e instalar software en euskera no es tan fácil como en otros idiomas, y, como consecuencia, el usuario tiene que hacer un esfuerzo adicional.

Algunos de los encuestados piden una página donde estén recogidos todos los recursos disponibles en euskera y listos para que los usuarios los descarguen. Además, se ha mencionado que es necesario divulgar la información sobre los servicios e interfaces existentes en euskera, especialmente entre la gente joven.

Por último, quisiéramos mencionar que hay una demanda para más productos de entretenimiento en euskera, especialmente para la gente joven. La mayoría de la gente está consumiendo juegos de ordenador o de móvil en otros idiomas, porque encontrarlos en euskera les resulta un poco más difícil.

3. Recomendaciones para mejorar la vitalidad digital

El Kit de Supervivencia Lingüística Digital es un instrumento cuyo objetivo es permitir a los hablantes y a las comunidades de las lenguas regionales y minoritarias autoevaluar el grado de vitalidad de su lengua y conocer qué tipo de acciones e iniciativas concretas pueden mejorar ese grado de vitalidad. Este documento está dedicado a ese segundo objetivo, y adaptado al caso particular del euskera.

En este conjunto de recomendaciones, sugerimos algunas acciones que se pueden adoptar (sobre todo a nivel popular) para hacer avanzar a una lengua hacia los siguientes pasos para la vitalidad digital.

Las recomendaciones están distribuidas en tres secciones, cada una relativa al tipo de indicador de la vitalidad digital.

3.1 Tres tipos de indicadores de la vitalidad digital

3.1.1 Capacidad digital

La capacidad digital, para nosotros, significa cuánto apoyo infraestructural y tecnológico tiene una lengua para que pueda funcionar en el mundo digital. Un prerrequisito es que la lengua debe tener, al menos, un sistema de escritura, porque, en su defecto, es imposible que funcione en el mundo digital. Para que una comunidad utilice una lengua en el mundo digital, tienen que cumplirse una serie de condiciones básicas, como el acceso a una conexión de Internet y la alfabetización digital. De igual manera, la existencia y la disponibilidad de recursos lingüísticos determina en gran medida la funcionalidad de una lengua en contextos digitales. Por ejemplo, funcionalidades como los correctores ortográficos de los smartphones, en principio, pueden potenciar el uso de la lengua permitiendo teclearla más fácil y rápidamente. La capacidad digital de una lengua solo se refiere a su potencial de ser utilizada en el mundo digital, pero en ningún caso garantiza que la comunidad vaya a utilizarla. Así sucede, por ejemplo, con muchas lenguas regionales o minoritarias de Europa: aunque muchas de esas lenguas cumplen los requisitos de la capacidad digital, en muchos casos se utilizan poco en comparación con la lengua oficial de los países en los que se hablan. Un escaso uso digital puede deberse a otros factores, como sumisión psicológica, falta de competencia en el lenguaje escrito, falta de espacios digitales (foros, blogs...) en los que se utiliza la lengua, etc.

3.1.2 Presencia y uso digital

Una vez garantizado el nivel infraestructural de la capacidad digital, es posible utilizar la lengua en una gran variedad de medios y para muchos fines. El segundo grupo de indicadores (del 6 al 9) se refiere a cómo y cuánto se utiliza una lengua en el mundo digital: si se utiliza (y hasta qué punto) para comunicarse, para producir contenido creativo o para fines educativos o de entretenimiento, entre otros. El denominador común de este grupo de indicadores es que están relacionados con la creación de contenido digital en esa lengua, se utilice para comunicar o para otros fines. Como hemos dicho, los indicadores están ordenados de manera que sugieren una cierta progresión ascendente: los mensajes de texto, la mensajería y los correos electrónicos se consideran funciones más básicas que, por ejemplo, escribir artículos de Wikipedia o desarrollar e-books o videojuegos en esa lengua. Sin embargo, eso no significa que ese orden tenga que considerase como una escalera en la que subir peldaño a peldaño. Por eso, no es obligatorio que tengamos Wikipedia para empezar a producir o localizar videojuegos. La función comunicativa se considera más básica que otras, según Gibson [8]. Estos usos digitales de la lengua abarcan también una progresión de usos más privados de la lengua a usos más públicos y, a menudo, oficiales. Hemos destacado cuatro indicadores para esta clase: *uso para la comunicación electrónica, uso en las redes sociales, disponibilidad de los medios de Internet y Wikipedia*.

3.1.3 Rendimiento digital

El rendimiento digital agrupa los indicadores que muestran qué se puede hacer con una lengua en el mundo digital. Esta es otra manera de ver hasta qué punto se utiliza una lengua en los diferentes medios. Esta perspectiva se centra más en los fines para los que se utiliza la lengua que en el abanico de medios disponibles en los que se utiliza. Para este grupo, hemos identificado cinco indicadores: *disponibilidad de los servicios de Internet, redes sociales localizadas, software localizado, herramientas y servicios de traducción automática.*

3.2 Estructura de las recomendaciones

Cada una de las recomendaciones está estructurada de la siguiente manera:

- Un texto explicativo para motivar y describir la recomendación.
- Destinatarios a los que va dirigida la recomendación: desde particulares, grupos de usuarios y asociaciones hasta organizaciones e instituciones, así como grupos de investigación, desarrolladores de software y empresas.
- Ejemplos: casos de éxito o de interés en los que se han llevado a cabo las iniciativas propuestas en la recomendación o que ilustran cómo se puede implementar.
- Para leer más: artículos, entradas de blog o trabajos académicos que proporcionan información adicional sobre la recomendación.
- Módulo relacionado en el Programa de Formación (PF) que contiene información relevante sobre la recomendación.

4. Capacidad digital

Línea principal de acción: Preparar tu lengua para el entorno digital

- Como competencia básica, promueve la alfabetización en la lengua regional o minoritaria.
- Garantiza una conectividad buena y actualizada y un acceso a Internet generalizado.
- Promueve una competencia digital (media-alta) de los hablantes de la lengua regional o minoritaria (usuarios digitales potenciales).
- Desarrolla recursos y herramientas lingüísticas, implicando a diferentes agentes (comunidades de usuarios, grupos de investigación, empresas, agentes políticos).

Disponibilidad de recursos lingüísticos

Los recursos lingüísticos son una condición fundamental para el desarrollo de aplicaciones informáticas lingüísticas más avanzadas.

Aunque sea habitual el uso del término recursos lingüísticos, hay que tener en cuenta que las herramientas también están incluidas en esa denominación. Entre los recursos se encuentran diccionarios, colecciones de textos o corpus, gramáticas, bases de datos léxicas y terminológicas, así como bases de conocimiento y ontologías. Entre las herramientas se encuentran correctores ortográficos y de estilo, analizadores morfosintácticos, etiquetadores gramaticales (POS taggers) y analizadores sintácticos, extractores de terminología y de expresiones multipalabra (MWE), traductores automáticos y herramientas de síntesis y reconocimiento del habla.

Para evaluar el grado de vitalidad de una lengua en relación con los recursos lingüísticos, en el documento "How to use the Digital Language Vitality Scale" (Ceberio et al., 2018) hemos clasificado los recursos y las herramientas en tres niveles: recursos básicos, intermedios y avanzados.

- Básicos: diccionarios electrónicos monolingües y bilingües, corpus digital (<100 millones de palabras), corrector ortográfico.
- Intermedios: diccionario monolingüe basado en el corpus (>100 millones de palabras), corpus paralelos, web corpus, software de extracción de términos, etiquetado gramatical, traducción automática básica (basada en reglas), síntesis de habla.
- Avanzados: corpus extensos (más de mil millones de palabras), corpus multilingües, análisis sintáctico, WordNet, procesamiento semántico, traducción automática avanzada (estadística, híbrida, neural), reconocimiento de voz.

Evidentemente, esta clasificación solo es orientativa, y no debe ser considerada de manera demasiado estricta, pero puede ser útil para estructurar los valores de este indicador y para organizar las recomendaciones de abajo.

R3 Desarrollar recursos lingüísticos básicos

R3.4 Utilizar herramientas como las concordancias para realizar consultas en los corpus

Como se ha mencionado, es muy importante que una lengua disponga de recursos como corpus de textos.

Existen algunas herramientas simples pero eficaces, los programas de concordancias, que podrían ser muy útiles para obtener información sobre las palabras y su uso, y que pueden utilizarse con texto sin procesar, es decir, con texto que no ha sido analizado para añadir información lingüística sobre las palabras. Se puede obtener información sobre la frecuencia de las palabras, las combinaciones más frecuentes o colocaciones, y utilizar los resultados para elaborar diccionarios y realizar otras tareas de procesamiento del lenguaje.

No obstante, para explotar este tipo de recurso de la manera más eficiente, lo ideal sería que dichos textos estuvieran procesados lingüísticamente. Una palabra o lema puede presentarse en diferentes formas o tokens (componentes léxicos), como diferentes formas de género, formas de plural, verbos irregulares o, en algunos idiomas, formas inflexionales. Así, para obtener datos de calidad, es recomendable tener el texto analizado, al menos a un nivel básico; por ejemplo, saber qué lema corresponde a cada palabra del texto, y cuál es su categoría gramatical. Avanzando un poco más, a veces el análisis morfológico es indispensable, porque, a menudo, las lenguas regionales y minoritarias son morfológicamente ricas, lo que muchas veces conlleva una escasez de datos.

Un etiquetador gramatical (POS tagger) es una herramienta que realiza esta tarea automáticamente, y supone un hito importante en el desarrollo tecnológico de una lengua. Si en tu lengua existen este tipo de herramientas, se recomienda utilizarlas, y consultar el corpus una vez que se haya etiquetado.

Destinatarios: particulares, colectivos, grupos de investigación

Ejemplos:

» AntConc 3.4.0 Tutorial 1: Getting Started – YouTube
» TextSTAT – Tutorial - YouTube

Para leer más:

» AntConc: A freeware corpus analysis toolkit for concordancing and text analysis
» TextSTAT – Simple Text Analysis Tool
» What are the most useful programs for forming text corpus or dictionary?

Módulo relacionado en PF: 6

R4 Desarrollar recursos lingüísticos intermedios y avanzados

En este nivel, es probable que los recursos básicos ya estén a disposición de los usuarios, y que ya se hayan desarrollado algunos recursos intermedios. En este punto, es hora de completar el conjunto de los recursos intermedios y de plantearse el desarrollo de los recursos y herramientas avanzados.

R4.1 Elaboración de diccionarios: diversidad, tamaño, especialización y divulgación

En el nivel de desarrollo, es probable que haya más de un diccionario electrónico disponible, y que la mayoría de ellos sean bilingües; estos, a menudo, están disponibles online.

Dependiendo de la situación concreta de la lengua, se pueden llevar a cabo diferentes acciones, ya sea para crear nuevos diccionarios o enriquecer los existentes.

Las recomendaciones pueden formularse en torno a estas cuatro palabras clave: diversidad, tamaño, especialización y divulgación. En otras palabras, una lengua que se encuentra en el nivel de desarrollo necesita diccionarios bilingües y monolingües, de diferentes tamaños y cobertura, y dirigidos a diferentes tipos de usuarios y fines (estudiantes, aprendices, traductores o especialistas del sector). Además, es fundamental que dichos diccionarios estén disponibles en Internet, y, especialmente, que tengan apps para consultas.

La elaboración de diccionarios es una tarea para expertos, pero un hablante competente de la lengua puede participar, al menos, en algunas tareas de un proyecto de elaboración de un diccionario. Hoy en día, la tecnología ofrece la posibilidad de crear, editar y publicar diccionarios de manera más fácil que hace, por ejemplo, veinte años.

No obstante, cierto tipo de proyectos requieren de una infraestructura humana, tecnológica y financiera, que puede ser difícil de conseguir para las lenguas minoritarias. Por eso, tradicionalmente, los proyectos se han llevado a cabo con la participación o promoción de las instituciones o de editores privados, que tienen un modelo de negocio basado en la venta de versiones impresas.

Todo ello está cambiando rápidamente, ya que las ventas de diccionarios en papel están disminuyendo año tras año, tanto que no se considera realista pensar que el mercado pueda ahora financiar la publicación de un diccionario, mucho menos la de uno monolingüe, en una lengua minoritaria.

Así, la elaboración de diccionarios se basa en el apoyo público, el micromecenazgo y el trabajo colaborativo, sobre todo en el caso de las lenguas minoritarias.

Por eso, la primera recomendación es lanzar iniciativas para implicar a las instituciones y a los grupos de hablantes en tales proyectos.

Abajo figuran algunas referencias de herramientas que pueden facilitar el trabajo de preparar y publicar diccionarios.

Destinatarios: colectivos, empresas, instituciones.

Ejemplos:

Diccionarios lexicográficos (monolingües o bilingües):

» FieldWorks (FLEx Quick Tour on Vimeo)
» Lexonomy (Gentle introduction to Lexonomy)
» Tshwanalex (TLex Suite: Dictionary Compilation Software)[1]

Specialized, technical or terminological projects:

» TermWiki In My Language | TermWiki.com
» TermKate — An integral web platform for the creation and publishing of terminological dictionaries

Dictionary apps:

» SIL Dictionary app builder

Para leer más:

» Kroskrity, P.V. (2015). Designing a dictionary for an endangered language community: Lexicographical deliberations, language ideological clarifications. University of Hawaii Press.
» Garret, A. (2018). Online dictionaries for language revitalization, to appear in The Routledge handbook of language revitalization, edited by Leanne Hinton, Leena Huss, and Gerald Roche (Routledge, 2018).
» Kotorova, E. (2016). Dictionary for a Minority Language: the Case of Ket. In Proceedings of the XVII EURALEX International Congress.

Módulo relacionado en PF: 6

R4.2 Aumentar el tamaño y la diversidad del corpus

En una lengua que se encuentra en el nivel de desarrollo de la vitalidad digital, seguramente haya uno o más recursos de corpus. El siguiente paso es aspirar a un corpus de un tamaño considerable, y a empezar a preparar la construcción de corpus bilingües o paralelos y corpus especializados.

Aunque, en un sentido amplio, cualquier colección de textos podría, en principio, considerarse como corpus, muchos consideran que se deben reunir ciertas condiciones para ello. En primer lugar, el objetivo de ser útil para obtener datos sobre el uso de la lengua en general (corpus de referencia), o sobre el uso en un cierto campo, género, registro o época (corpus especializado). Este objetivo es uno de los temas más discutidos en el diseño de corpus, y está relacionado con el tamaño y la representatividad de la muestra, o, al menos, con el equilibrio entre los diferentes tipos de texto incluidos. En segundo lugar, hoy en día es muy improbable que un corpus no esté en formato digital. Finalmente, en el nivel básico, un corpus contiene texto sin procesar, pero, especialmente

1 No es gratuita, pero existe una Licencia de Lenguas Académicas o en Peligro de Extinción.

en el caso de las lenguas que tienen una morfología rica, se suele considerar prácticamente esencial que el corpus esté lingüísticamente etiquetado, incluyendo, por ejemplo, información sobre el lema, la categoría gramatical y el caso de cada componente léxico. Se pueden encontrar recomendaciones para el diseño y construcción de corpus en Corpus design criteria (Atkins et al., 1992) y en Corpus Linguistics: corpus building principles (Markus Dickinson, 2015); para la codificación, véase 15 Language Corpora - The TEI Guidelines - Text Encoding Initiative.

Un corpus que reúna dichas características es una fuente rica de información, que puede utilizarse o consultarse para diferentes fines: análisis lingüístico del uso de la lengua, desarrollo de otros recursos y herramientas, testeo de hipótesis y entrenamiento de herramientas para el procesamiento automático del lenguaje. Por ejemplo, el análisis de corpus ha sido muy útil para elaborar diccionarios: podemos informarnos sobre nuevas palabras, cuáles son las palabras más utilizadas o las variantes de una palabra, o qué adjetivos son los más utilizados con un sustantivo dado.

Sin embargo, el cumplimiento de dichas condiciones, especialmente de la primera, puede ser un reto incluso para las lenguas mayoritarias, que no suelen tener pocos recursos. Construir semejante corpus es un trabajo considerable que requiere de un esfuerzo colectivo y que exige una inversión de tiempo y dinero que puede sobrepasar los medios de algunas comunidades de lenguas regionales o minoritarias.

Uno de los principales problemas es cómo obtener un número considerable de textos en formato digital. Negociar con editores y autores puede convertirse en una ardua tarea, y las obras no siempre están disponibles en un formato que se pueda incorporar en el corpus a bajo coste. Una posibilidad que puede ayudarnos a superar este obstáculo es adoptar un enfoque basado en la web, o tomar la web como corpus. En Internet, los textos ya están digitalizados, pero la viabilidad de la estrategia depende no solo de la cantidad de contenido web que hay en la lengua regional o minoritaria, sino también de nuestra capacidad de encontrarlo y recopilarlo. Por ejemplo, debemos identificar las páginas que están escritas en nuestra lengua meta. Si no dispones de los medios para desarrollar una herramienta para eso, existen varias herramientas que facilitan dicha tarea, tales como TextSTAT y BootCaT. La primera utiliza un sistema de rastreo, y hay que especificar la URL que se quiere captar. Para identificar qué páginas web tienen contenido en la lengua con la que se está trabajando, puede ser útil pedir a la comunidad lingüística que te proporcionen las URLs de las páginas específicas conocidas que hay en esa lengua. BootCat ofrece una funcionalidad adicional: puedes especificar una serie de palabras origen, y, utilizando Bing como motor de búsqueda, la herramienta combina dichas palabras para realizar diversas búsquedas y recopilar los documentos recuperados. Puedes utilizar palabras que sean exclusivas de tu lengua o que pertenezcan a un campo concreto (para construir corpus específicos de dicho campo).

Otra manera complementaria de recopilar contenidos son los repositorios de textos. Este enfoque está estrechamente ligado a la recomendación Crear un archivo web de documentos y registros, en la sección Presencia y uso digital.

Al hablar de un tamaño considerable para un corpus monolingüe y general de una lengua regional o minoritaria, podríamos marcarnos 100 millones de palabras como reto deseable. Aunque se hayan utilizado corpus más pequeños en los estudios y proyectos de lenguas regionales o minoritarias, y difícilmente podamos cuestionar su utilidad, hay una razón de peso para insistir en la necesidad de que el corpus sea grande. La calidad de las herramientas de procesamiento de lenguajes naturales basadas en datos que se entrenan en los datos lingüísticos, a menudo en forma de texto sin procesar o de texto comentado, depende, en gran medida, del tamaño del corpus. Los corpus también son útiles para evaluar las herramientas lingüísticas en las versiones de referencia, que están revisadas por humanos y que son extremadamente valiosas.

Los corpus paralelos son colecciones de textos que son traducciones mutuas (en realidad, un texto origen y su traducción). Los corpus paralelos son un recurso muy valioso para los traductores, pero también para elaborar diccionarios bilingües, ya que podemos saber cómo se ha traducido una palabra, expresión o término dado. El tipo más básico de corpus paralelo es una colección de documentos bilingües. Si sabemos qué documentos son traducciones mutuas, podemos decir que el corpus está alineado a nivel de documento. Pero es mucho más útil obtener corpus paralelos alineados a nivel de oración. Un procedimiento ideal para construirlos es coger las memorias de traducción (MT) como punto de partida. Las MTs son el resultado del proceso de traducción hu-

mana cuando se ha utilizado una herramienta de Traducción Asistida por Ordenador (TAO). Si no, podemos alinear documentos bilingües automáticamente utilizando herramientas como Wordfast Aligner o hunalign – sentence aligner. Otra posibilidad es detectar textos bilingües en Internet. Esta no es una tarea fácil, pero existen algunas experiencias en las lenguas regionales o minoritarias que pueden servir de inspiración; por ejemplo, Bitextor, ILSP Focused Crawler y PaCo2. Otra posible fuente de corpus paralelos son las secciones de artículos de las wikipedias de diferentes idiomas, que son traducciones mutuas. Las wikipedias no son idénticas. En términos estrictos, se consideran como corpus comparables. Pero sabemos que tienen cierto contenido en común, y se han desarrollado diversas técnicas para encontrarlos y alinearlos.

Finalmente, los corpus especializados son una fuente fiable de datos para elaborar diccionarios especializados o terminológicos. Como en el caso de los corpus generales, la manera tradicional de construir corpus especializados es recopilar textos y documentos que pertenecen a un campo concreto, pero también podemos optar por un enfoque en el que el corpus es la web, que se ha demostrado que es una fuente adecuada de datos (véase, por ejemplo, Gurrutxaga et al., 2010). Para construir desde la web un corpus especializado de un campo, el método de las *palabras* origen puede ser muy útil. BootCaT ofrece esa posibilidad.

Destinatarios: colectivos, grupos de investigación, organizaciones.

Ejemplos:

- BootCaT – Simple Utilities to Bootstrap Corpora and Terms from the Web || BootCaT front-end tutorial
- TextSTAT – Simple Text Analysis Tool
- Building your own corpus – TextSTAT and AntConc
- ILSP Focused Crawler - ILSP NLP
- Bitextor: the automatic bitext generator
- hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene
- W2c- large multilingual corpus | Martin Majlis - Academia.edu
- TenTen Corpus Family – Wikipedia

Para leer más:

- Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, 7(1), 1-16.
- Gurrutxaga, A., Leturia, I., Pociello, E., San Vicente, I. & Saralegi, X., (2010). Exploiting the Internet to build language resources for less resourced languages. In *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages LREC 2010*, Valetta, Malta.
- Lynn, T., Foster, J., Dras, M. & Dhonnchadha, E.U. (2012). Active Learning and the Irish Treebank, In *Proceedings of the Australasian Language Technology Association Workshop 2012* (pp. 23-32), Dunedin, NZ.
- San Vicente, I. & Manterola, I. (2012). PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. In LREC (pp. 1-6).
- Smith, J.R., Quirk, C. & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403-411). Association for Computational Linguistics.

Módulo relacionado en PF: 6

R4.3 Recabar datos lingüísticos disponibles públicamente desde las redes sociales

El DLDP recomienda utilizar al máximo los mensajes y publicaciones de las redes sociales escritas en las lenguas regionales o minoritarias y que sean públicamente accesibles. Especialmente en Facebook, existen páginas públicas específicas que promueven debates sobre temas culturales y lingüísticos relacionados con la lengua minoritaria, donde la gente nombra los objetos de una imagen para comparar si todos utilizan la misma palabra, o comparten poemas y canciones.

Al considerar los datos de las redes sociales, se pueden seguir diferentes caminos:

- Un corpus formado con textos rastreados de Internet podría integrarse con datos de las redes sociales;
- Un conjunto de datos de datos textuales de las redes sociales puede utilizarse para analizar hasta qué punto son uniformes y coherentes las diferentes variantes de una misma lengua;
- Los desarrolladores pueden utilizar el mismo conjunto de datos para construir software que lematice y comente a nivel de categoría gramatical.

Un paso obligatorio para poder asumir este reto sería comprobar detenidamente la licencia de los datos.

Este tipo de textos a menudo se considera ruidoso, ya que es un contenido generado por los usuarios y no editado, que contiene faltas de ortografía o de gramática. Tales textos constituirían un tipo específico de corpus basado en este carácter lingüístico, pero es útil para otros fines (análisis de la evolución del uso de la lengua a lo largo del tiempo, fotografía en tiempo real del uso de la lengua, análisis de la alternancia de código y otros fenómenos lingüísticos interesantes).

Destinatarios: grupos de investigación, desarrolladores de software

Ejemplos:

- The Corpus of Facebook Welsh Language Texts
- indigenoustweets.com

Para leer más:

- Honeycutt, C. & Cunliffe, D. (2010). The use of the Welsh language on Facebook: An initial investigation. In *Information, Communication & Society*, 13(2), pp.226-248.
- Lynn, T., Scannell, K. & Maguire, E. (2015). Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets. In ACL-IJCNLP 2015, p.1.
- Lackaff, D. & Moner, W. J. (2016). Local languages, global networks: Mobile design for minority language users. *In Proceedings of the 34th Annual International Conference on the Design of Communication* (SIGDOC '16).

Módulo relacionado en PF: 6

R4.5 Utilizar herramientas para el análisis del corpus y alimentar tu diccionario con los datos sobre la lengua en uso

Existen algunas herramientas simples pero eficaces, los programas de concordancias, que podrían ser muy útiles para obtener información sobre las palabras y su uso, y que pueden utilizarse con texto sin procesar, es decir, con texto que no ha sido analizado para añadir información lingüística sobre las palabras. Puedes obtener información sobre la frecuencia de las palabras y las combinaciones más frecuentes o colocaciones, y utilizar los resultados para elaborar diccionarios y realizar otras tareas de procesamiento del lenguaje. Ya se ha proporcionado información más detallada y referencias sobre estas herramientas en la recomendación R3.1 Utiliza herramientas como las concordancias para realizar consultas en los corpus.

Pero si tienes un etiquetador gramatical en tu lengua, el potencial para obtener datos sobre el uso de la lengua es mucho mayor. Puedes obtener las frecuencias de uso no solo de las formas de palabras en el texto (tokens), sino también de entradas de diccionario (lemas); puedes encontrar nuevas palabras no recogidas en el diccionario; o puedes investigar cómo se combinan las palabras para formar expresiones, colocaciones o términos complejos. Para ello, necesitas un sistema para realizar consultas en los corpus, o una herramienta para analizar el corpus (en corpus-analysis.com, encontrarás diversas herramientas de ese tipo).

Destinatarios: grupos de investigación, desarrolladores de software.

Ejemplos:

- Corpkit: a tool for investigating text

» The IMS Open Corpus Workbench (CWB)
» NoSketch Engine
» BlackLab - An open source corpus search engine
» Unitex - GramLab

Para leer más:

» The IMS Open Corpus Workbench (CWB) Corpus Encoding Tutorial
» 5 Reasons: Anke Luedeling on "Corpus Linguistics"

Módulo relacionado en PF: 6

R4.6 Primeros pasos hacia la síntesis y reconocimiento del hablan

La síntesis del habla es una simulación del habla humana generada por ordenador. Es una de las dos principales tareas de las tecnologías del habla (la otra es el reconocimiento de voz). Cuando una lengua empieza a adentrarse en el mundo de estas tecnologías, lo normal es empezar con la síntesis, porque esta tarea es más fácil que el reconocimiento de voz.

El desarrollo de las tecnologías del habla abre un amplio abanico de posibilidades para aplicaciones, entre las que se encuentran, principalmente, sistemas para personas con discapacidades (lectores de pantalla para personas con visión limitada, sintetizadores del habla para personas con deficiencia del habla, o sistemas de reconocimiento para personas con discapacidades que no pueden utilizar dispositivos informáticos convencionales). Por otro lado, se ha demostrado que los sintetizadores del habla son extremadamente valiosos para el aprendizaje online cuando los estudiantes no tienen acceso a los hablantes nativos. Por ejemplo, esto ha sido especialmente eficaz para los estudiantes extranjeros de irlandés a quienes la ortografía irlandesa les resulta difícil de pronunciar. Además, la síntesis es necesaria, junto con el reconocimiento, como componente de un sistema en el que el usuario puede comunicarse utilizando el lenguaje natural con el ordenador u otros dispositivos móviles.

El desarrollo de las tecnologías del habla es una tarea que requiere un nivel técnico avanzado. Por eso, esta recomendación está dirigida a grupos de investigación y desarrolladores de software. Sin embargo, ciertas tecnologías de síntesis (tales como la síntesis concatenativa) utilizan muestras de habla real, y, en este punto, es importante contar con la participación de hablantes que puedan proporcionar locuciones para crear una base de datos del habla, o que incluso puedan contribuir con su propia voz.

Destinatarios: grupos de investigación, desarrolladores web

Ejemplos:

» The MARY Text-to-Speech System (MaryTTS)
» Festcat - Catalan Speech Synthesis
» eSpeakNG
» abair.ie – The Irish Language Synthesiser
» Aholab TTS
» The Welsh National Language Technologies Portal - Speech
» Mozilla Common Voice

Para leer más:

» Speech Synthesis for Minority Languages: A Case Study on Scottish Gaelic
» Issues in Porting TTS to Minority Languages
» Frisian TTS, an example of bootstrapping TTS for minority languages
» Chasaide, A. N., Chiaráin, N. N., Wendler, C., Berthelsen, H., Murphy, A., & Gobl, C. (2017). The abair initiative: Bringing spoken irish into the digital space. In *Proceedings of Interspeech 2017*.

5. Presencia y uso digital

Línea principal de acción: Promover el uso y la creación y compartición de contenido

- Encuentra y prueba maneras de animar a la gente a que utilicen su lengua regional o minoritaria en comunicaciones electrónicas privadas y en las redes sociales.
- Promueve la creación de este tipo de contenidos: páginas web, blogs, foros, así como radio y televisión de Internet;
- Iniciativas para subir y compartir medios en las lenguas regionales o minoritarias;
- Realizar una colaboración abierta para la subtitulación;
- Wikipedia: crear, editar, corregir, actualizar.

5.1 Uso para la comunicación electrónica

R5 Valorar la utilización de las lenguas regionales o minoritarias para la comunicación interpersonal

Con respecto a la comunicación interpersonal digital, existen diferentes tipos de comunicación electrónica. La mensajería instantánea, por ejemplo, tiene muchas funciones en común con la oralidad. Así, en un entorno digital, una lengua que no se utiliza en la comunicación electrónica es una lengua que no se habla. Otro aspecto importante de la comunicación electrónica es que se utiliza en registros más formales, como el correo electrónico en un contexto profesional.

Destinatarios: particulares, colectivos, organizaciones, instituciones.

Ejemplos:

- Tribo, a loita normalizadora e cotiá entre os máis pequenos
- Campaign to use Welsh in the European Football Cup
- Mintzanet: an Option for Practicing the Basque Language Anywhere in the World

5.2 Disponibilidad de los medios de Internet

R7 Incrementar la cantidad de contenido y diversificar los tipos de medios de Internet

La presencia en Internet de contenido en una lengua concreta es un indicador muy significativo de su vitalidad. El contenido de la web está a disposición de todos, y una lengua que tiene poco contenido en forma de páginas web, blogs, foros, audio o vídeo carece de visibilidad y presencia en el entorno digital.

Se podrían adoptar algunas medidas efectivas para incrementar la cantidad de tipos de medios que ya están presentes en la lengua, y para incorporar nuevos tipos.

R7.1 Incrementar la cantidad de contenido de texto (páginas web, blogs, foros)

La presencia en la web de este tipo de contenido es básica y esencial para cualquier lengua. En el caso de la mayoría de lenguas regionales y minoritarias que han logrado una presencia considerable en Internet, gran parte del contenido web pertenece a esas categorías.

Por eso, cualquier cosa que promueva y facilite la creación y publicación de este tipo de contenido beneficiará a la lengua y a sus usuarios.

Hoy en día, existen muchas aplicaciones y servicios que facilitan esta tarea. Entre los nombres conocidos, podemos mencionar WordPress, Blogger y Wix, pero puedes encontrar muchas más opciones en las referencias de la sección "Para leer más".

Destinatarios: particulares, colectivos, organizaciones.

Ejemplos:

» The Maya Tz'utujil initiative: the Maya Tz'utujil initiative is a space on Facebook, Twitter and WordPress to teach, learn, and broadcast the Tz'utujil language, which primarily covers the geographic national area of Guatemala, and includes multiple collaborative partners who have become fully linked to the digital initiative.
» Basque Language on the Web: Making an Impact

Para leer más:

» 6 Differences Between Blogging in a Minority Language versus English
» How to Choose the Best Blogging Platform in 2018 (Compared)
» The 16 best free blogging platforms
» Best 10 Free Website Builders | 2018's Best Website Builders

Módulo relacionado en PF: 5

R7.2 Crear o alimentar un archivo web de documentos y registros

Para una lengua que tiene una actividad digital muy escasa, se puede crear un archivo web de grabaciones o documentos. Quienes no hablan la lengua también pueden contribuir. Ese recurso puede ser útil para restablecer la lengua, o para animar a los hablantes interesados pero inseguros a empezar a utilizar la lengua.

» Create and feed repositories of texts in your language
» Diversify the types of media in your language, including audio and video content
» Contribute to existing repositories, such as Wikitongues

Destinatarios: collectives, organisations, institutions, individuals.

Ejemplos:

» Indigitization – Toolkit for the Digitization of First Nations Knowledge
» Endangered Languages Documentation Programme (ELDP)
» Greenstone Digital Library Software
» Wikitongues
» BasaBALIWiki
» Yadiko Ukruri initiative 'jitómagaro uai'
» Cultural Codex
» AIKUMA Project - preserving endangered languages

Para leer más:

» Nichols, D.M., Witten, I.H., Keegan, T.T., Bainbridge, D. & Dewsnip, M. (2005). Digital libraries and minority languages. In *New Review of Hypermedia and Multimedia*, 11(2), 139-155
» Nenek – A cloud-based collaboration platform for the management of Amerindian resource languages

Módulo relacionado en PF: 5

R7.3 Transmitir online utilizando herramientas de software gratuitas

Una emisora de radio online puede ser un medio eficaz para difundir una lengua y hacerle recuperar la visibilidad, en especial para comunidades dispersas. La disponibilidad de software libre lo hace relativamente fácil. Puedes inspirarte con las historias recogidas más abajo.

Destinatarios: particulares, colectivos, organizaciones.

Ejemplos:

» Brazil's First Indigenous Online Radio Station Uses Digital Media to Promote Native Languages and Communities
» Meet the Young Ecuadorians Behind the First Kichwa-Language Radio Show in the US

Para leer más:

» Icecast, a streaming media (audio/video) server that can be used to create an Internet radio station.
» Top 5 Free Tools to Live Stream Your Event Online – Capterra Blog

Módulo relacionado en PF: 5

R7.4 Registrar historias digitales en tu propio idioma

Así se define la narración digital (Barret, 2005):

> "La narración digital es la expresión moderna del arte antiguo de la narración. Las historias digitales adquieren su poder mezclando imágenes, música, narrativa y voz; así, otorgan una profunda dimensión y un color intenso a los personajes, las situaciones, las experiencias y las percepciones."

La narración digital se está utilizando, entre otros, en la educación (Hoven, 2009):

> "La narración digital se está utilizando cada vez más como medio para fomentar el aprendizaje reflexivo y/o evaluación en muchos programas de enseñanza de lenguas y de formación de profesores de lenguas en todo el mundo. La narración digital puede proporcionar a los profesores y estudiantes de lenguas muchos recursos para ayudar a los estudiantes a identificarse con las lenguas y culturas extranjeras o con las primeras lenguas (L1) y culturas (C1) que se están perdiendo, y a sentirse más cómodos para utilizar estas lenguas para fines reales."

Desde esa perspectiva, puede ser útil a efectos de revitalización lingüística.

Destinatarios: particulares.

Ejemplos:

» Digital Storytelling: What it is… And… What it is NOT
» How Storytelling Can Do Wonders in Blogging

Para leer más:

» Frequently-Asked Questions about Digital Storytelling
» Hoven, D. (2009). Digital Storytelling in Indigenous Education: Internet Technologies to (Re-) Establish L1 and C1 Literacy and Fluency. In *Internet-Based Language Learning: Pedagogies and Technologies*, p.47.

Módulo relacionado en PF: 5

R7.5 Promover iniciativas de subtitulación

En el entorno digital, la subtitulación de películas y vídeos se ha beneficiado de las posibilidades de trabajar de manera colaborativa. Existen muchos proyectos en los que voluntarios traducen los subtítulos de una película a cada vez más lenguas. Esta es una oportunidad para las lenguas regionales y minoritarias. Algo que antes era difícil y caro ahora puede hacerse realidad gracias al trabajo colectivo de los hablantes de las lenguas regionales y minoritarias, ya sea para poder ver películas en versión original con subtítulos en su lengua regional o minoritaria, o para dar mayor proyección a las obras originales escritas en la lengua regional o minoritaria, subtitulándolas en idiomas de mayor difusión.

Destinatarios: particulares, colectivos.

Ejemplos:

» Amara: Caption, Subtitle and Translate Video
» PerMondo – Introduction to subtitling
» TED translations
» Contribute translated content - YouTube Help

Para leer más:

» Crowdsourcing Subtitles for Endangered Languages
» Review: Amara is a Web-based service that lets anyone transcribe and translate online video
» How Crowdsourced Video Translation Works: Webinar Q&A with Amara
» Is crowdsourcing translation a threat or an opportunity for the audiovisual market?
» Azpitituluak, a Project for Basque Subtitles
» Dowling, M., Lynn, T. & Way, A. (2017). A Crowd-sourcing Approach for Translations of Minority Language User-Generated Content. In Proceedings of 1st Workshop on Social MT, Prague, Czech Republic.

Módulo relacionado en PF: 5

5.3 Wikipedia

Es muy probable que haya una wikipedia en tu lengua, pero es posible que sea pequeña y que los hablantes no la utilicen o no sepan de su existencia. Teniendo en cuenta que Wikipedia es un recurso muy representativo de la vitalidad de una lengua, es importante hacerla crecer y mejorarla. En la Lista detallada de wikipedias, puedes comprobar si hay una wikipedia en tu lengua, y, en ese caso, cuáles son sus cifras más relevantes, como el número de artículos o de usuarios activos.

En caso de que no haya ninguna wikipedia en la lengua, se recomienda iniciar un proyecto y/o traducir la interfaz de usuario de Wikipedia.

R9 Eleva tu Wikipedia a un nivel mayor

R9.2 Iniciativas para aumentar el tamaño y la calidad de Wikipedia

Ya existe una wikipedia en la lengua, seguramente de mediano tamaño (entre 10.000 y 100.000 artículos), pero, teniendo en cuenta que Wikipedia es un recurso muy representativo de la vitalidad de una lengua, es importante hacerla crecer en número y extensión de los artículos, y mejorar su calidad.

Wikipedia es un recurso que incluye muchos tipos de contenido, y, por tanto, ofrece la oportunidad de trabajar en diferentes áreas y aspectos para mejorarlo. A continuación, analizaremos las siguientes características relativas a los artículos:

» **Número**
Cualquier aumento debería considerarse como positivo, pero, si se trata de un reto para una

lengua En Desarrollo, alcanzar o incluso aproximarse a las 100.000 entradas sería también un gran logro. Sobra decir que superar ese hito debería considerarse como un éxito de primer orden.

» **Tipo**
Animar a los usuarios a incluir contenido local o regional en la wikipedia de la lengua regional o minoritaria. Para atraer a más usuarios, una opción es diferenciar la wikipedia de tu lengua regional o minoritaria como recurso donde encontremos información que no hay en otras wikipedias, o que es más detallada y elaborada. Pero, especialmente en este punto, esto no debería conllevar no seguir incluyendo artículos de carácter más general, en áreas de la ciencia, las humanidades, la historia, el arte o cualquier otro tipo de contenido que puede ser relevante para los usuarios y su vida cotidiana (páginas conocidas sobre famosos, marcas, deportes), sobre todo si queremos animar y alentar a la siguiente generación de hablantes a utilizar estos recursos digitales. Sin ese tipo de información, la lengua corre el riesgo de reproducir en Wikipedia la situación sociolingüística de diglosia que sufre en la sociedad, y de no poder contribuir a darle la vuelta. Debe ser una estrategia mixta. Una búsqueda sobre los artículos más leídos y a los que más se ha contribuido en Wikipedia ayudará a marcar las pautas para este método.

» **Extensión**
Para una lengua que se encuentra en el nivel de desarrollo de la vitalidad digital, merece la pena empezar a considerar la extensión de los artículos como parámetro de gran importancia. Si queremos que los usuarios que entran en Wikipedia para buscar información enciclopédica, más allá de la mera definición de la entrada, hagan eso en la Wikipedia de la lengua, es fundamental dotar a los artículos de una extensión que pueda satisfacer mínimamente dicha necesidad, aun sabiendo que siempre encontrarán más datos en wikipedias de las lenguas mayoritarias.

» **Calidad lingüística**
En algunas wikipedias de lenguas minoritarias, existe una gran preocupación en torno a la calidad lingüística de los artículos. Es muy importante encontrar un equilibrio: si la calidad es insuficiente, ello puede perjudicar al prestigio de Wikipedia en esa lengua, pero, si se hace demasiado hincapié en exigir un alto nivel de calidad, puede que ciertos usuarios se desanimen y no participen. Por ejemplo, los autores de los artículos de la wikipedia en sardo pueden elegir entre tres variantes (Limba Sarda Comuna, Logudoresu y Campidanesu), y se proporciona un enlace a un corrector ortográfico externo para dichas variantes, con la sugerencia de evaluar la coherencia del texto antes de publicarlo en Wikipedia.

Maneras de involucrar a los usuarios en la edición de Wikipedia:

» Organizar un editatón en los festivales locales, y promover proyectos de traducción en la educación, también para mejorar la alfabetización digital.
» Hacer que los institutos de secundaria se comprometan para educar a sus alumnos en la edición de Wikipedia (una clase que cree una página sobre tu municipio o club deportivo, por ejemplo).
» Organizar un seminario para formar a los formadores (a través de la lengua regional o minoritaria), y establecer los términos correctos. Ofrecer un seminario a los nuevos editores en su lengua regional o minoritaria es más atractivo que hacerlo en la lengua mayoritaria alternativa.
» Identificar las páginas wiki más conocidas (en general), y destacarlas como artículos prioritarios para que los editores los traduzcan.
» Intentar involucrar a las autoridades locales en la promoción de Wikipedia a través de campañas que fomenten su uso, ofrecer apoyo a las comunidades de contribuidores de Wikipedia, y ayudar con la edición y la corrección.

Destinatarios: particulares, colectivos.

Ejemplos:

» Celtic Knot - Wikipedia Language Conference
» Basque Wikimedians User Group
» Collaboration with Wikipedia in 2016 and 2017
» Catalan Wikipedia - Creation

Para leer más:

» Galipedia, the Wikipedia in Galician, is now 15 years old (English version by Google Translate from the original in Spanish)
» The Basque Wikipedia, Local Knowledge Gone Global (and back)
» Wikipedia in Catalan, leader in the 1,000 most important articles (English version by Google Translate from the original in Catalan)

Módulo relacionado en PF: 4

6. Rendimiento digital

Línea principal de acción: Crear oportunidades para hacer cosas digitalmente en tu lengua

- » Promover la demanda de servicios de Internet en las lenguas regionales o minoritarias
- » Localización de software y de interfaces de usuario
- » Servicios de Traducción Automática
- » Obtener un dominio específico

6.1 Disponibilidad de los servicios de Internet

R10 Expandir el abanico de posibilidades para utilizar servicios de Internet en tu idioma

Para que una lengua sea digitalmente operativa, es fundamental que puedas "hacer cosas" en la web. Estamos hablando de servicios como la banca online, salud, compras, turismo, cultura o noticias.

R10.1 Recabar información y experiencias de la comunidad de usuarios de tu lengua regional o minoritaria para determinar cuáles son los servicios más importantes y los más utilizados

A una lengua que se encuentra en el nivel 4 se le presupone cierta funcionalidad en Internet. Si la lengua posee cierto nivel de reconocimiento oficial, es probable que la administración pública ofrezca servicios online en dicha lengua. Una posible línea de trabajo es expandir y reforzar esa área, pero, al mismo tiempo, trabajar para que el sector privado se conciencie de la conveniencia de dirigirse a sus clientes y usuarios en la lengua regional o minoritaria. En cualquier caso, nuestra acción debería orientarse hacia las necesidades de los hablantes. Por eso, es muy importante detectar qué servicios quisieran utilizar los hablantes en su lengua regional o minoritaria.

Por tanto, antes de lanzar una iniciativa para animar a la administración o a las empresas para que ofrezcan sus servicios en tu lengua, es necesario disponer de información precisa sobre la situación real y las necesidades de los usuarios para establecer ciertas prioridades.

Una manera de obtener esa información es realizar una encuesta entre los usuarios. Puedes tomar como punto de partida el cuestionario desarrollado en el proyecto del DLDP, y traducirlo a tu lengua.

Una vez que hayas recibido los resultados de la encuesta, puedes organizar una campaña para tener en tu lengua los servicios más deseados por los hablantes. La mayoría de los servicios están bajo la responsabilidad de autoridades locales o centrales. Contacta con los responsables de dichos servicios. Pregunta si está disponible algún responsable lingüístico.

Para concienciar al responsable de los servicios, deberías recalcar la importancia de la lengua local para el bienestar y la inclusión de la gente.

Destinatarios: colectivos, organizaciones.

Ejemplos:

- » The 'Survey on Digital Fitness' | the Digital Language Diversity Project
- » Paricio Martín, S.J. & Martínez Cortés, J.P. (2014). El uso del aragonés en Internet y las nuevas tecnologías: herramientas y repercusión. In *Actas II Jornadas Aragonesas de Sociología*. pp. 105-120. Zaragoza, 2014.

Para leer más:

» Why language matters for the Millennium Development Goals - Unesco

R10.2 Valorar la utilización de la lengua del usuario en los negocios

La globalización ha recalcado aún más la importancia de la lengua en el comercio. Cada vez es más evidente que los consumidores online de la mayoría de los países se muestran más cómodos con las páginas web que están en su propia lengua. Eso también es así en el caso de las lenguas regionales o minoritarias, más aún si somos conscientes de los factores emocionales que desempeñan un papel importantísimo en la imagen que una marca busca crear para los consumidores. Es por eso que el factor lingüístico merece ser tenido en cuenta a la hora de diseñar una estrategia de negocio.

Destinatarios: empresas, organizaciones.

Ejemplos:

» Let languages shout out your business benefits
» The Benefits of Translating into Minority Languages – Translation

Para leer más:

» Language Means Business
» The Importance of Language in Global E-Commerce | TransPerfect
» The Value of Language in e-Commerce white paper
» Cunliffe, D., Pearson, N. & Richards, S. (2010). E-commerce and Minority languages: a Welsh perspective. *Language and the Market, Palgrave Macmillan, Basingstoke*, pp.135-147.

R10.3 Desarrollar aplicaciones para smartphones

Las aplicaciones para smartphones son una atractiva manera de transmitir contenido para una multitud de fines. Dado que los smartphones están tan extendidos, una app puede ser una manera relativamente fácil para que una lengua sea apreciada por un público más amplio.

Deberías considerar promocionar el desarrollo de apps gratuitas para el aprendizaje de idiomas, por ejemplo. En esta área, hay muchos ejemplos bonitos y replicables, tales como Wahzhazhe (disponible para iOS y Android) y Speak Mohawk (versiones de iOS y Android), provistos por First Nations, o la app del diccionario de rawang.

Hacer palabras, frases sencillas y saludos es una divertida manera de sumergir a alguien en una lengua y una cultura, y servirá también como una excelente introducción a la singularidad de la cultura específica que emana de la lengua. Puede haber funcionalidades adicionales, como la grabación de palabras y frases para permitir al usuario practicar el idioma. Suelen tener buena acogida los juegos y tests que permiten a los estudiantes evaluar sus habilidades. Cuanta más sumersión y compromiso implique la experiencia, mejor.

Por desgracia, desarrollar apps es caro en comparación con las páginas web. Pero existen opciones para crear *apps de tarjetas mnemotécnicas* utilizando plataformas existentes como Memrise o Anki, y requieren mucha menor inversión por parte de la comunidad lingüística. Este método puede generar contenido interesante para aprender con el móvil, pero quizás sin el prestigio de una app independiente. Ejemplo de clases de la lengua lakota en Memrise: https://www.memrise.com/courses/english/lakota

Destinatarios: particulares, colectivos, organizaciones, desarrolladores de software.

Ejemplos:

» How Technology is Saving Native Tribe Languages

» Wahzhazhe, an Osage language app for phones and tablets
» Six Nations school launches app that teaches people to speak Mohawk

6.2 Redes sociales localizadas

R11 Iniciativas para localizar las interfaces de usuario de las redes sociales

Ya hemos mencionado el importante papel que desempeñan las redes sociales en la vitalidad de una lengua. Aunque la interfaz de usuario esté en tu lengua, no es necesariamente una condición para que interactúes utilizando tu lengua en las redes sociales, pero es un factor que tiene cierta influencia en el uso de la lengua, y que puede ayudar a más usuarios a empezar a utilizarla.

Por desgracia, en los últimos años, han disminuido las opciones que ofrecían antes empresas como Facebook, Twitter o Google para localizar sus interfaces (véase más abajo Scannell, 2012).

Es necesario también recalcar la importancia que tiene la calidad de la traducción en el prestigio de la lengua y en el valor efectivo que tiene esto para los usuarios; esto, posiblemente, pueda aumentar el uso de la lengua. Además, debería fijarse como objetivo terminar la traducción.

Destinatarios: colectivos, organizaciones, instituciones.

Ejemplos:

» Sa tradutzione de Facebook in sardu

Para leer más:

» Losse, K. (2008). Facebook: Achieving quality in a crowd-sourced translation environment
» Scannell, K. (2012). Translating Facebook into endangered languages. In Proceedings of the 16th Foundation for Endangered Languages Conference (pp. 106-110).

Módulo relacionado en PF: 3

6.3 Software localizado: sistemas operativos y software básico

Tener la posibilidad de utilizar software en tu lengua es un signo de prestigio, indica que puedes vivir en tu lengua en el mundo digital, y todo ello tiene mucho valor, ya que anima a los usuarios a elegir su lengua regional o minoritaria como lengua de comunicación y trabajo.

R13 Reforzar las iniciativas para localizar los softwares gratuitos o privativos de carácter general más utilizados en la comunidad lingüística

Probablemente, una lengua en situación de desarrollo tendrá localizados, como mínimo, un sistema operativo de escritorio y otro móvil (ya sea abierto o comercial), y algún software de carácter general (un procesador de texto y un navegador).

Es hora de ir más allá de estas aplicaciones básicas y expandir el abanico de posibilidades para utilizar software localizado en la lengua, sea un software libre y de código abierto o un software privativo.

Para lograr la máxima eficiencia y centrar el esfuerzo en localizar el software más utilizado, es importante conocer las necesidades de los usuarios de la lengua regional o minoritaria, así como aprender del camino recorrido por lenguas regionales o minoritarias más avanzadas.

En caso del software de código abierto, se puede iniciar un esfuerzo comunitario para localizar software libre de carácter general más utilizado en la comunidad lingüística. Algunos ejemplos de ese tipo de iniciativas son Softcatalà (catalán), Meddal (galés), Librezale (euskera), An DROUIZIG (bretón) y Softaragonés (aragonés). Para la localización de software comercial (Windows, MacOS,

MS Office), las autoridades locales desempeñan un papel fundamental como interlocutores con las empresas.

Destinatarios: colectivos, organizaciones, instituciones.

Ejemplos:

» How to Localize Software: 10 Dos and Don'ts for a Watertight
» Software Localization Process

Para leer más:

» Localization 101: A Beginner's Guide to Software Localization
» How to Localize your Software, App or Game : 7 Best Practices
» How To Localize an Android Application

R14 Considerar los videojuegos como una valiosa oportunidad de revitalización

Los videojuegos ofrecen a los usuarios un alto nivel de interactividad y, por tanto, de implicación con una lengua. Desarrollar o localizar videojuegos en lenguas locales es una excelente manera de darles un nuevo impulso, así como de mostrar a las generaciones más jóvenes que la lengua está viva y en forma para el mundo moderno.

Como es de imaginar, el mercado para los videojuegos en lenguas menos utilizadas es bastante limitado. Las grandes empresas no invierten en mercados pequeños; por eso, la responsabilidad recae de nuevo en los activistas y entusiastas, como Gwenn Meynier, quien tradujo al bretón el juego Steredenn, o Frédéric Antonpietri y Fabien Mariani, quienes crearon el juego Winterfall en corso. En el País Vasco, Game Erauntsia es un grupo de jugadores vascos de videojuegos que hacen campaña por utilizar el euskera en el mundo de los videojuegos.

Ejemplos:

» Steredenn
» Winterfall

Para leer más:

» Get inspired by the story of Kisima Ingitchuna, the first video game in Inupiaq language that, in addition, drawn from indigenous culture for its story and characters: Showcasing Alaska's Inupiat culture through gaming
» PES 2016 Introduces the First Welsh Language Video Game Box Art
» The video game: A challenging universe for minority languages
» Conquering digital worlds in Scottish Gaelic

6.4 Servicios de Traducción Automática

La disponibilidad de servicios de traducción automática está considerada como indicador de un uso digital extendido y activo, ya que presupone un gran abanico de herramientas y recursos. Desde el punto de vista de la usabilidad digital de la lengua, la disponibilidad de una traducción automática fiable para una lengua es un signo de que la lengua ha adquirido un nivel bastante alto de presencia e importancia digital.

En entornos donde una lengua minoritaria se tiene que mover e intentar sobrevivir como una herramienta para la comunicación, ciertos usos de la traducción automática pueden llevar a situaciones que deben tratarse con cuidado. Un sistema que tenga un mínimo de calidad puede ser útil y efectivo para la asimilación, es decir, para producir textos comprensibles, incluso si no son adecuados para la publicación (divulgación). Pero es difícil en cualquier lengua, pero, sobre todo, en las lenguas regionales o minoritarias, alcanzar la calidad requerida para la divulgación sin que el

texto tenga que ser editado por un traductor. Si no se tiene en cuenta esta diferencia, puede darse un abuso de la traducción automática, en detrimento, por desgracia, de la propia lengua regional o minoritaria. Dicho abuso de la traducción automática ha provocado algunas malas experiencias (por ejemplo, el caso de Wikipedia en cheroqui) y críticas (véase este artículo de M. Bauer: When things are way, way, WAY worse than you thought they might get, y este otro de M.B. Měchura: Do minority languages need machine translation?). Por eso, es imprescindible que los usuarios potenciales de la traducción automática sean conscientes de estas limitaciones, para extender el buen uso de esta tecnología. Véase, por ejemplo, esta Nota de Aviso para la Traducción Automática, que recalca este punto para el caso del galés.

En el nivel 4 de vitalidad digital, es muy probable que la lengua disponga de un servicio o herramienta online, con al menos una lengua emparejada con la lengua regional o minoritaria, y disponible en al menos una dirección.

Merece la pena recalcar aquí la necesidad de que la administración pública coordine los datos bilingües. Muchos gobiernos no son conscientes del valor de los datos bilingües para ayudarles a desarrollar herramientas de traducción que les ayuden a satisfacer las necesidades de traducción (si la lengua goza de estatus oficial). Normalmente, se exige a los tecnólogos de la lengua que asesoren a este respecto, y que recalquen la necesidad de una centralización coordinada de las traducciones, para asegurar la reutilización de los datos (memorias de traducción) y el fácil acceso a los datos bilingües a fin de construir sistemas de traducción automática. Si se ajusta específicamente para la administración pública, se puede obtener un resultado de calidad razonable; esto reduciría después el esfuerzo general de los traductores (editar contenido repetitivo en lugar de empezar de cero)..

R17 Aumentar el número de combinaciones lingüísticas; intentar incluir el inglés, si no está ya incluido

En el nivel En Desarrollo de vitalidad, es muy probable que la lengua tenga un sistema de traducción automática basado en reglas. En ese caso, una primera opción para diversificar las combinaciones lingüísticas sería reutilizar los datos y herramientas lingüísticos (diccionarios morfológicos, etiquetadores gramaticales) desarrollados para una combinación en el desarrollo de un sistema para nuevas combinaciones. Esto, evidentemente, limita el número de lenguas a las que podemos extender nuestro sistema, ya que tienen que ser un tanto próximas (no muy diferentes en cuanto a morfología y sintaxis). No obstante, también se han desarrollado sistemas de traducción automática basados en reglas para lenguas muy diferentes, como español-euskera (Matxin).

Por otra parte, puede ocurrir que la lengua cuente con una cantidad de corpus paralelo nada desdeñable, lo cual le permitiría entrar en el mundo de la traducción automática estadística. Algunos experimentos recientes con lenguas regionales o minoritarias en el campo de la traducción automática neural han dado resultados prometedores. Quizás así pueda la traducción automática para las lenguas regionales o minoritarias dar un salto cualitativo, y hacer que dichas lenguas estén más cerca de ser utilizadas como herramienta diaria en los servicios de traducción.

Destinatarios: grupos de investigación, desarrolladores de software.

Ejemplos:

- Matxin: an open-source transfer machine translation engine
- Automatic translation - Llengua catalana - Gencat.cat
- Oersetter: Frisian-Dutch Statistical Machine Translation
- You can now translate Wikipedia articles from Spanish into Basque, thanks to an open source machine learning tool

Para leer más:

- Popović, M., Arcan, M. & Klubička, F. (2016). Language related issues for machine translation between closely related South Slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)* (pp. 43-52).
- Dowling, M., Cassidy, L., Maguire, E., Lynn, T., Srivastava, A. & Judge, J. (2015). Tapadóir: Developing a statistical machine translation engine and associated resources for Irish. In *Proceedings*

of the The Fourth LRL Workshop: Language Technologies in support of Less-Resourced Languages, Poznan, Poland.

» Screen, B. (2017). Machine Translation and Welsh: Analysing free Statistical Machine Translation for the professional translation of an under-researched language pair. *Journal of Specialized Translation*, 28.

» Mayor, A., Alegria, I., De Ilarraza, A.D., Labaka, G., Lersundi, M. & Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine translation*, 25(1), p.53.

» Gompel, M.V., van den Bosch, A.P.J. and Dijkstra, A. (2014). Oersetter: Frisian-Dutch statistical machine translation. Ljouwert: Fryske Akademy

» Etchegoyhen, T., Martínez, E., Azpeitia, A., Labaka, G., Alegria, I., Cortes, I., Jauregi, A., Ellakuria, I, Martin, M. & Calonge, E. (2018). Neural Machine Translation of Basque. In *21st Annual Conference of the European Association for Machine Translation* (p. 139).