http://dx.doi.org/10.30920/letras.91.134.9

Lingüística computacional para la revitalización y el poliglotismo

Computational linguistics for revitalization and polyglotism

Luis Camacho Caballero
Pontificia Universidad Católica del Perú, Lima, Perú
Contacto: camacho.l@pucp.pe
https://orcid.org/0000-0001-6569-550X

Rodolfo Zevallos Salazar
Pontificia Universidad Católica del Perú, Lima, Perú
Contacto: qichwa@pucp.pe
https://orcid.org/0000-0003-0192-7740

Resumen

A pesar de las leyes existentes, en la práctica el Estado peruano ignora la multiculturalidad y se comporta como una entidad monolingüe y monocultural. Dado que este paradigma equivocado todavía vigente, el Estado no ha invertido lo suficiente para desarrollar las habilidades lingüísticas con el fin de servir a todos los ciudadanos por igual. Las consecuencias de ello son la falta de fomento, la discriminación y finalmente el aislamiento que lleva a la extinción de las lenguas autóctonas. Nuestra iniciativa es cambiar el paradigma equivocado, despertar el orgullo nacional por nuestras raíces nativas y hacerlo en tres frentes: demostrar que nuestras lenguas se pueden usar en el mundo tecnológico moderno al igual que las lenguas bien establecidas, demostrar que nuestras lenguas pueden portar cultura y entretenimiento bajo los cánones contemporáneos y demostrar que nuestras lenguas aportan valor económico a la nación, lo que justifica su preservación más allá del derecho. En este documento se describe una hoja de ruta para el desarrollo de la lingüística computacional de idiomas infrasoportados que todavía son hablados por millones de hablantes. Tal es el caso del quechua, aimara, guaraní, náhuatl, mixteco, otomí, quiché, maya o zapoteco. Debido a la masiva presencia de los hablantes de estas lenguas en el entorno urbano y a su uso habitual de Internet y telefonía móvil, se apuesta por la construcción de corpus de estas lenguas vía *crowdsourcing online*.

Palabras clave: Planificación lingüística; Tecnología del lenguaje; Lenguas en peligro de extinción; Lenguas como recurso

Abstract

Despite existing laws, in practice, the Peruvian State ignores multiculturalism and behaves as a monolingual and mono-cultural

organization. Since this misguided paradigm is still in place, the state has not invested enough to develop language skills to serve all citizens equally. The consequences of this are the lack of promotion, discrimination and finally the isolation that leads to the extinction of our indigenous languages. Our initiative is to change the wrong paradigm, to awaken national pride for our native roots, and to do it on three different ways: to demonstrate that our languages can be used in the modern technological world as well as well-established languages, to demonstrate that our languages can carry culture and entertainment under contemporary canons and to demonstrate that our languages provide economic value to the nation, which justifies their preservation beyond rights. This document describes a roadmap for the development of computational linguistics of under-supported languages that are still spoken by millions of speakers. Such is the case of languages such as: Quechua, Aymara, Guaraní, Nahuatl, Mixtec, Otomi, Quiche, Mayan or Zapotec. Due to the massive presence of the speakers of these languages in the urban environment and their habitual use of the Internet and mobile telephony, we are committed to build corpora of these languages via online crowdsourcing.

Keywords: Language planning; Language Technology; Endangered Languages; Language Economics; Language as a Resource

Recibido: 10.07.20 Aceptado: 17.08.20

1. Introducción

En el Perú existen 48 lenguas nativas que aún están vivas pero amenazadas. Todas estas lenguas están en riesgo de extinción. Los expertos señalan que el proceso de sustitución es irreversible a menos que surjan políticas y herramientas disruptivas (Adelaar, 2014).

Dentro de las Tecnologías de la Información y Comunicación (TIC) y con la etiqueta Tecnologías del Lenguaje Humano (TLH) hay una cantidad significativa de herramientas computacionales para el procesamiento del lenguaje. Así, debe destacarse la lingüística computacional como la herramienta potencial para la revitalización de las lenguas nacionales peruanas, pues la falta de dicho soporte impide el crecimiento de estas lenguas y su uso productivo en Internet (y en cualquier sistema electrónico). Hoy en día, las computadoras son indispensables para comunicarse verbalmente o por escrito. Herramientas de procesamiento de texto, diccionarios electrónicos y sistemas avanzados de procesamiento de voz como sintetizadores (generación de voz) o transcriptores

(reconocimiento de voz) están disponibles para los idiomas oficiales de países desarrollados (inglés, mandarín, japonés, etc.).

Sin embargo, hay más de 6900 lenguas en el mundo y solo una pequeña fracción posee los recursos necesarios para la implementación de las tecnologías del lenguaje humano. De tal forma, las TLH actualmente están relacionadas solo con lenguas para las cuales hay grandes recursos disponibles o que repentinamente son de interés debido a la escena económica o política. Desgraciadamente, hasta el momento la mayoría de las lenguas de los países en desarrollo o de las minorías étnicas reciben muy poca atención.

Procesar una nueva lengua a menudo conduce a desafíos inéditos (sistemas fonológicos especiales, problemas de segmentación de palabras, estructuras gramaticales borrosas, lenguaje no escrito, etc.). La falta de recursos requiere, por su parte, innovación en las metodologías de recolección de datos o modelos para los cuales la información es compartida entre varias lenguas. Además, algunos aspectos sociales y culturales relacionados con el contexto de la lengua traen problemas adicionales: lenguas con muchos dialectos en diferentes regiones, cambio de una lengua a otra dentro del discurso (*codemixing*) o presencia masiva de hablantes no nativos. Para desarrollar sistemas para dichas lenguas, uno debe tomar prestados recursos y conocimientos de otras lenguas similares, que requieren la ayuda de dialectólogos (encontrar índices de proximidad entre idiomas), y fonetistas (mapa de los inventarios fonéticos entre la lengua de bajos recursos y la lengua de mayores recursos).

La investigación lingüística de las lenguas nacionales peruanas solo ha logrado la creación de los respectivos alfabetos oficiales (Ministerio de Cultura, 2020). Lamentablemente, todavía no han sido asignados suficientes fondos gubernamentales para el mantenimiento y la difusión de ese conocimiento y mucho menos para la digitalización moderna de estas lenguas. Por tal motivo, el nivel de informatización de los idiomas peruanos es extremadamente bajo, por lo que todos estos idiomas son considerados infrasoportados; ello significa que cumplen una o más de las siguientes características: carencia o poca difusión de un sistema de escritura única, falta de presencia en Internet, falta de una masa crítica de expertos lingüistas y ausencia de recursos electrónicos (corpus

monolingües, diccionarios electrónicos bilingües, discursos transcritos en base de datos, diccionarios de pronunciación o vocabularios específicos).

El procesamiento computacional de la lengua quechua recién comenzó alrededor de 2010; hasta la fecha, solo hay tres grupos de investigación visibles en Internet: Hinantin, Qichwa 2.0 y Siminchikkunarayku. Hasta donde se conoce, Anette Ríos (2016) hizo el mayor procesamiento computacional para el quechua sureño. También es destacable la aplicación QichwaDic que permite buscar en varios diccionarios bilingües español-quechua digitalizados.

Con respecto a otros idiomas nativos americanos, puede citarse el gran trabajo de construcción de recursos lingüísticos del idioma shipibo-konibo realizado por el Grupo de Inteligencia Artificial de la Pontificia Universidad Católica del Perú (IAPUCP). Esta investigación es parte de la labor de Siminchikkunarayku, que recopila y mantiene toda la investigación realizada previamente, produce nuevos productos computacionales y hace incidencia política para conseguir que las autoridades nacionales reconozcan la riqueza potencial que traería el florecimiento de nuestras lenguas autóctonas.

2. Tecnología del lenguaje humano

Las tecnologías de la información y comunicación están ahora preparándose para la próxima revolución. Más allá de las computadoras personales, redes, miniaturización, multimedia, dispositivos móviles y computación en la nube, la próxima generación brindará un *software* que servirá a los usuarios mucho mejor porque conocerá, hablará y entenderá su idioma.

Concebimos la tecnología del lenguaje como la aplicación funcional de la lingüística computacional, dirigida a analizar y generar el lenguaje de diversas maneras y para una variedad de propósitos. Traducción, análisis automático del lenguaje o los rotuladores morfosintácticos son solo ejemplos de tales aplicaciones prácticas. Los pioneros de tales desarrollos son el servicio en línea gratuito Google Translate que actualmente traduce 104 idiomas, el supercomputador de IBM llamado Watson y el asistente móvil de Apple Siri para teléfonos iPhone que puede reaccionar a los comandos de voz y responder preguntas en inglés, alemán, francés, japonés y algunos idiomas más.

La próxima generación de TIC dominará el lenguaje humano a tal punto que los usuarios podrán comunicarse utilizando la tecnología en su propio idioma. Los dispositivos podrán encontrar automáticamente las noticias y la información más importantes de todo el mundo simplemente utilizando comandos de voz fáciles de usar. La tecnología del lenguaje podrá traducir automáticamente o ayudar a los intérpretes, resumir conversaciones y documentos, además de apoyar a los usuarios en tareas de aprendizaje.

En cuanto a la evolución humana, el habla y la mímica son las formas más antiguas y naturales de comunicación, pero la información compleja y la mayoría del conocimiento humano se almacena y transmite a través de la escritura. La tecnología del lenguaje vincula el lenguaje a diversas formas de conocimiento, independientemente de los medios (voz o texto) en los que es expresado usando diccionarios, reglas de gramática y semántica. Las tecnologías de voz y texto se superponen e interactúan con otras tecnologías multimedia y de comunicación multimodal tales como: aprendizaje de idiomas asistido por computadora, generación de información estructurada (IE), respuesta interactiva, traductor automático, corrección ortográfica, búsqueda web, resumen de textos, sintetización de la voz y transcripción.

3. Corpus lingüístico

La documentación de idiomas es un subcampo emergente de la lingüística aplicada. Respecto a herramientas colaborativas e interfaces para transcribir, archivar y buscar grabaciones multimedia, la documentación ha hecho enormes progresos. Sin embargo, paradójicamente, este campo raramente ha considerado aplicar métodos automatizados para anotar datos más eficientemente con calidad y en cantidad; ello es necesario para fundar una nueva y mejor investigación lingüística de idiomas infrasoportados basada en corpus.

Todos los esfuerzos combinados entre la tecnología del lenguaje y la documentación del lenguaje pueden ser claramente rentables tanto para investigaciones teóricas basadas en corpus como para la planificación lingüística y la revitalización de idiomas en peligro de extinción. Mientras que los documentalistas lingüísticos proporcionan corpus y el análisis lingüístico necesarios para el modelamiento computacional de los idiomas en cuestión,

los ingenieros lingüísticos aplican los métodos formales descriptivos y/o estadísticos para la programación de reglas gramaticales y léxicas legibles por computadora con el fin de crear herramientas informáticas para usuarios finales (Blokland, 2015). La documentación del lenguaje hablado puede aumentar el tamaño del conjunto de datos utilizado en la investigación llevada a cabo por lingüistas computacionales. Por otro lado, la tecnología del lenguaje puede crear herramientas que analicen el corpus hablado de una manera mucho más efectiva, y así permitir crear mejores descripciones y anotaciones lingüísticas. Ello permite lidiar con la transcripción de conjuntos de datos más grandes porque el procesamiento puede automatizarse y realizarse mucho más rápidamente que de la lenta y tradicional forma manual.

El campo de la lingüística computacional se ha consolidado gracias a la explotación de corpus más grandes y mejor anotados. No obstante, falta mucho más corpus, carencia que genera uno de los principales cuellos de botella para el procesamiento computacional del lenguaje. En tal sentido, producir con calidad y mantener dichos recursos es una tarea compleja que requiere tiempo, una cantidad considerable de fondos económicos y la cooperación de varios expertos (Woodbury, 2014). Mientras que "*big data*" es una tendencia cada vez mayor, producir corpus continúa siendo una tarea detrás de los reflectores en el procesamiento del lenguaje natural, en particular para idiomas diferentes del inglés. Así, muchas lenguas y sus dialectos poseen un menor número de datos casi siempre insuficientes para ser procesados por metodologías de aprendizaje automático.

Un corpus paralelo voz-texto típico "independiente del hablante" requiere un conjunto suficientemente diverso de hablantes, al menos 50 personas diferentes (Barnard, Davel & Van Hereden, 2009). El corpus debe ser fonéticamente equilibrado, lo que implica la distribución correcta de las palabras y que los discursos sean pronunciados por varones y mujeres de una amplia gama de edades. Huang y Zhang (2015) describen cómo debe construirse el corpus. Sin recursos, el único enfoque para hacerlo a gran escala es el *crowdsourcing*. De esta manera, se han propuesto varios incentivos para convencer a los hablantes nativos de participar en el proceso de producción de corpus (Benjamin, 2016). En el caso

de la transcripción, estos incentivos se han utilizado con cierto éxito (Parent & Eskenazi, 2010); no obstante, el número de idiomas infrasoportados para los cuales hay suficientes transcriptores disponibles es bastante limitado y puede ser muy diferente de un idioma a otro (Gelas, Abate, Besacier & Pellegrino 2011).

Cuando un corpus se desarrolla desde el principio, la tarea de transcripción se puede simplificar ya que es posible emplear instrucciones predeterminadas. Este beneficio debe sopesarse contra la carga adicional de solicitar y registrar locutores. En este caso, la recopilación de datos generalmente comienza con el acopio de un corpus de texto (que, insistimos, solo es posible si existe un sistema de escritura adecuadamente estandarizado). De este corpus, se extrae una colección de indicaciones y se presenta a los hablantes seleccionados un menú de grabación. Aunque la verificación sigue siendo necesaria para garantizar que los hablantes digan las palabras deseadas, los métodos automáticos han demostrado ser bastante exitosos y eficientes para este propósito; por ejemplo, en un sistema ASR se arranca con un corpus sin procesar (De Vries, Davel, Badenhorst, Basson, Barnard & De Waal, 2014), asumiendo que todas las indicaciones se registraron correctamente, y luego este sistema se usa para identificar de forma iterativa las expresiones equivocadas y mejorar la precisión del sistema ASR.

Para el proceso de grabación en sí mismo, a pesar de que se necesita calidad, la cantidad también es relevante, pero los recursos son limitados. Incluso si hubiera muchos hablantes disponibles, difícilmente podrían registrarse sus voces de manera profesional debido a los costos de movilidad de las personas y el alquiler del estudio de grabación. Como opción, a menudo se han empleado servicios telefónicos controlados por menús (también conocidos como servicios interactivos de respuesta de voz). Las hojas de instrucciones se distribuyen a los hablantes seleccionados, luego ellos deben llamar a un número gratuito y son guiados para registrar esas indicaciones en orden. La amplia disponibilidad de teléfonos inteligentes en la presente década ha llevado a varios grupos a desarrollar aplicaciones que brindan lo mejor de ambos mundos: el contacto personal de trabajadores de campo y la automatización del proceso de grabación (Bird, 2018; Bird, Hanke, Adams & Lee, 2014); los trabajadores de campo pueden gestionar varios teléfonos simultáneamente, lo que permite la recopilación de voz de muchos hablantes en un tiempo relativamente corto.

Volviendo al Perú, Roberto Zariquiey y su equipo establecieron que solamente 25 dialectos de idiomas peruanos autóctonos cuentan con una documentación moderna; es decir, que tienen bases de datos que incluyen audio, video, conversaciones naturales y transcripciones; pero incluso en estos casos la documentación es incompleta (Zariquiey et ál., 2019, p. 52). De esta forma, siguen siendo lenguas con escasos recursos lingüísticos pero claramente en una situación bastante mejor que la del resto de lenguas autóctonas peruanas. Para idiomas que tienen los mismos o mejores recursos lingüísticos que estos 25 dialectos referidos, y donde además el grueso de su población de hablantes nativos cuenta con acceso a telefonía móvil e Internet, altamente expuestos a la influencia de las cadenas de radio y televisión nacionales, planteamos que pueden ser construidos eficientemente corpus de gran escala basados en tres pilares: la automatización de la recolección, *crowdsourcing* y la masificación. Dados todos los requisitos, esta propuesta solamente tendría sentido para el caso de 5 idiomas: ashaninka, awajún, aimara, quechua sureño (Ayacucho, Cusco) y shipibo-konibo.

Con respecto al primer pilar, debido a una sinergia existente, la documentación del lenguaje hablado puede aumentar el tamaño del conjunto de datos utilizado en la investigación llevada a cabo por lingüistas computacionales. Por otro lado, la tecnología del lenguaje puede crear herramientas que analicen el corpus hablado de una manera mucho más efectiva, y así permitir crear mejores descripciones y anotaciones lingüísticas. Ello facilita lidiar con la transcripción de conjuntos de datos más grandes porque el procesamiento puede automatizarse y realizarse mucho más rápidamente que con la lenta y tradicional forma manual. De esta forma, colectar el número de voces y textos necesarios es una tarea extenuante para un pequeño número de personas. Aquí surge el segundo pilar: la partición de la tarea en una gran cantidad de pequeñas porciones y que el procesamiento de cada porción no represente mayor esfuerzo facilitaría que la tarea sea realizada por una gran cantidad de voluntarios sin descuidar la calidad. A esto se llama *crowdsourcing*.

Creemos que el tercer pilar es la mayor innovación presentada en este campo en el presente siglo. Proponemos SIMINCHIKKUNARAYKU MARATHON, una campaña mediática que alcance prácticamente a toda la nación peruana y que 1) anime a los ciudadanos hablantes nativos de los 5 mencionados

idiomas a grabar sus voces y 2) avive el interés en estos idiomas de los ciudadanos que no son hablantes nativos pero que con altísima probabilidad sí hablaron sus padres o abuelos. La tarea sería automatizada por una aplicación móvil que las personas usarían para repetir y grabar sus voces siguiendo unas frases piloto. La aplicación móvil ya existe y se llama HUQARIQ; sin embargo, muchas mejoras deben ser implementadas para que se convierta en una herramienta de escala industrial completamente útil. Entre estas mejoras se necesita: garantizar la identidad del usuario mediante una conexión con el Registro Nacional de Identidad y Estado Civil (Reniec), verificar que el usuario pronuncie correctamente el idioma seleccionado, crear un "balanceador de carga" para que todas las frases se graben aproximadamente la misma cantidad de veces y garantizar la seguridad de la información mediante una arquitectura de transmisión de cuatro capas: Flask, Gunicorn, Gunicorn supervisor y Nginx.

En tal sentido, debe incrementarse la cantidad y calidad de las frases piloto (prompts); las nuevas se tomarán de 1) el diccionario de Diego González de Holguín para quechua sureño, pero con la escritura estándar actual, 2) el diccionario del Ministerio de Educación del Perú para el quechua central y 3) el diccionario del Ministerio de Educación del Perú para el aimara. Se construirán un total de 5000 frases por cada uno de los tres supradialectos las cuales, en su conjunto, incluirán todos los fonemas de cada uno de ellos. Cada frase durará alrededor de 5 segundos; luego, cada serie durará alrededor de 7 horas. El usuario no estará expuesto a las 5000 frases, sino solo a 240 (1200 segundos, 20 minutos); según nuestra experiencia, esperamos que cada usuario tarde 1,5 horas en completar la tarea.

Una cuarta actividad se realizará en paralelo antes de la fecha central: el control sobre la emisión de la publicidad. El objetivo de esta actividad es asegurar que la publicidad esté llegando y especialmente comprobar que esté consiguiendo primero la curiosidad y luego el involucramiento de la población. Para ello se necesita lograr el compromiso voluntario de las empresas de radiodifusión y de telecomunicaciones para usar gratuitamente su infraestructura, así como involucrar a personalidades para que manifiesten públicamente su apoyo a la campaña. Tan importante como lo primero es definir indicadores que permitan determinar el cumplimiento del objetivo.

4. Planificar para que la lingüística computacional sea una verdadera herramienta de revitalización

Cada idioma representa un desafío particular para la lingüística computacional, pero finalmente el abordaje de cada uno es el mismo. De esta manera, la ruta que debe recorrerse para el desarrollo computacional de las lenguas americanas autóctonas es prácticamente igual al que se seguiría para idiomas de otros continentes. Ello es una ventaja, pues existe mucha documentación de la experiencia que ya se está viviendo en Europa con respecto a sus lenguas autóctonas

Como factor clave, debe enfatizarse el desarrollo de las herramientas básicas y los datos lexicográficos y lingüísticos, los cuales se conocen como recursos lingüísticos. Si estos están poco desarrollados, o no existen, es imposible desarrollar la tecnología del lenguaje. Tales datos pueden ser colecciones grandes y estructuradas de texto, grabaciones de audio o glosarios que han sido adaptados para su uso en tecnología del lenguaje. Así, debe utilizarse el pequeño trabajo previo y es imperativo invertir más en el desarrollo de los recursos lingüísticos. Por ejemplo, las herramientas básicas en tecnología del lenguaje son transcriptores y sintetizadores de código abierto que tratan con el lenguaje cotidiano y pueden ser adoptados para un uso específico. También lo son herramientas para analizar el habla y la pronunciación, o las herramientas de soporte necesarias para usuarios finales. Igual lo pueden ser los sistemas generales de traducción automática. Es vital que las herramientas sean abiertas y accesibles para todos, por lo que cualquiera que quiera desarrollar soluciones que incorporen tecnología del lenguaje para el quechua podría utilizar estos recursos sin tener que llevar a cabo investigación y desarrollo básicos que requieren mucho tiempo.

En Europa, la iniciativa *Cracking the Language Barrier* actualmente reúne a casi todos los proyectos europeos de investigación e innovación, así como a organizaciones comunitarias afines que trabajan en o con tecnologías multilingües, áreas cercanas o temas estrechamente relacionados. En esta iniciativa global, los miembros colaboran en su objetivo conjunto de superar cualquier tipo de barreras de lenguaje y comunicación con la ayuda de sofisticadas tecnologías de lenguaje. Entre las áreas de colaboración están los documentos de estrategia (como la Agenda Estratégica para el Mercado Único Digital Multilingüe), tareas

científicas compartidas y campañas de evaluación, gestión de datos, repositorios de recursos y tecnología, al igual que eventos y actividades de difusión. META-NET, uno de los miembros de la iniciativa, persigue el financiamiento para un gran proyecto insignia llamado *Human Language Project* cuyos objetivos son: 1) producir bases de datos lingüísticos, 2) procesar datos correspondientes a una gran diversidad de idiomas, y 3) convertir esos datos en conocimiento avanzado y aplicaciones de lingüística computacional. En ese camino, META-NET mantiene actualizada la agenda estratégica de investigación en lingüística computacional de idiomas europeo (Rehm, 2018).

Para que esta ruta sea recorrida se necesita crear la demanda y generalizar la expectativa de que cada lengua debería tener una existencia digital. Hasta ahora, no hay ningún esfuerzo para crear conciencia pública sobre la equidad lingüística, por lo que las personas piensan que los buenos recursos lingüísticos son tan probables como ganar la lotería, y por lo tanto no vale la pena intentar conseguirlos. Las personas que no saben que es posible construir recursos para sus lenguas sin duda no lo exigen. Para la mayoría de la gente, la tecnología es algo que se toma tal y como se presenta, no se les ocurre preguntar a los desarrolladores por nuevas características. Sin ejercer poder económico y sin el apoyo político para levantar la demanda, las comunidades lingüísticas ni siquiera sueñan en una presencia significativa en la esfera digital.

El desarrollo de la tecnología es relativamente oneroso, pero el costo de la pérdida de oportunidades y el mantenimiento de prácticas que se están volviendo rápidamente obsoletas es mucho más caro. La elección real radica en aceptar que para una mejor calidad de vida, el costo inherente de no usar la mejor tecnología disponible es muy alto. Invertir en el desarrollo tecnológico aumentará la competitividad de nuestra economía, sociedad e idiomas. Para lograr que la Comunidad Andina sea una opción en el mundo tecnológico, debemos asegurarnos que el público, las empresas y las instituciones puedan usar la tecnología del lenguaje e implementar soluciones sin ser obstaculizados por el desarrollo de una infraestructura complicada y costosa.

5. Conclusiones

Más allá de los esfuerzos persistentes en el desarrollo en tecnología del lenguaje, para demostrar que los recursos digitales pueden ser construidos apenas los

fondos estén disponibles, los activistas deben seguir tres estrategias: incidencia, asociación internacional y búsqueda de incentivos a la innovación.

La primera estrategia hacia la inclusión digital es una intensa incidencia. Los ciudadanos comunes no pueden exigir servicios lingüísticos, pero sus gobiernos sí pueden. Sin embargo, para hacer tales demandas, los burócratas necesitan convencerse de que ellas son razonables y alcanzables.

En segundo lugar, lo que se puede intentar es una alianza de organizaciones con ideas afines, por ejemplo, dentro de un gran portafolio de proyectos como el *Human Language Project.* La agrupación de recursos puede generar costos mucho más bajos por idioma, creando economías de escala que podrían inclinar la balanza hacia el apoyo financiero. Tal alianza probablemente habilite un ambiente en el que prospere el procesamiento computacional de los idiomas en riesgo, y por lo tanto finalmente aparezca una propuesta de valor atractiva para que las agencias financieras respalden completamente la creación de una infraestructura de datos lingüísticos para idiomas en riesgo.

En tercer lugar, cuando los políticos y el pueblo alcancen el acuerdo de que las lenguas tienen valor e interés, los fondos se pueden movilizar para generar innovación en tecnología del lenguaje. Es importante apoyar y asegurar la participación de empresas nacionales que practiquen la innovación en tecnología del lenguaje y/o que puedan utilizar herramientas de la tecnología del lenguaje para mejorar sus servicios o producción. Estas compañías crearán soluciones de acuerdo con la necesidad de la sociedad. La base fundamental establecida en este programa les permitirá realizar esas soluciones. Ello debe fomentarse a través de un programa de incentivos, así como una buena interacción y cooperación entre los participantes en las etapas finales. Siminchikkunarayku movilizará activamente a todos los actores identificados y fomentará la agrupación de estos para que ataquen proyectos específicos, creando así una oportunidad para el surgimiento de asociaciones locales que participen en proyectos internacionales. Por otro lado, con gran determinación debe perseguirse que las grandes corporaciones internacionales inviertan en la construcción de los recursos lingüísticos del quechua y demás idiomas americanos.

Finalmente, la producción de contenido abundante en cada lengua también será un gran desafío. Esto no es sencillo, porque hay muchas más lenguas

que activistas, investigadores o modelos de negocio. Tradicionalmente se han propuesto varios incentivos para que los miembros de las comunidades lingüísticas participen en el proceso de producción de recursos lingüísticos digitales. Lo primero es la creación de herramientas que hagan sus vidas más fáciles, por ejemplo, produciendo publicidad radial de los productos que compran. En segundo lugar, está la producción de audiolibros y juguetes para sus hijos, que se puedan usar en la educación inicial. Lo tercero es la generación de servicios culturales para las poblaciones migrantes que desean volver y/o mantener el vínculo con sus lugares de origen. En cuarto lugar, están las recompensas intrínsecas, como el orgullo de ver el idioma propio crecer en Internet o el reconocimiento dentro de las redes sociales para quienes toman un papel activo en el avance del idioma.

Agradecimientos

Este artículo ha sido producido en el marco del proyecto Lurin Qichwa Corpus financiado por la Pontificia Universidad Católica del Perú, beca de investigación CAP DGI 2020 ID 809

Referencias bibliográficas

Adelaar, W. F. H. (2014). Endangered languages with millions of speakers: Focus on Quechua in Peru. *JournaLIPP*, 3, 1-12. https://lipp.ub.uni-muenchen.de/lipp/article/view/393

Barnard, E., Davel, M., Van Heerden, C. (Septiembre de 2009). ASR corpus design for resource-scarce languages. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Congreso llevado a cabo en Brighton, Reino Unido. http://doi.org/10.13140/RG.2.1.1824.2000.

Benjamin, M. (2016). Digital language diversity: Seeking the value proposition. En C. Soria et ál. (Eds.), *CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity* (pp. 52-58). Eslovenia: LREC. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf

Bird, S. (2018). Designing Mobile Applications for Endangered Languages. En K. L. Rehg y L. Campbell (Eds.), *The Oxford Handbook of Endangered*

Languages. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190610029.013.40

Bird, S., Hanke, F. R., Adams, O. y Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. En *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages* (pp. 1-5). Baltimore: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-2201

Blokland, R., Fedina, M., Gerstenberger, C., Partanen, N., Riebler, M. y Wilbur, J. (2015). Language documentation meets language technology. *First International Workshop on Computational Linguistics for Uralic Languages. Septentrio conference series*. https://doi.org/10.7557/5.3457

De Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., Barnard, E., De Waal, A. (2014). A smartphone-based asr data collection tool for under-resourced languages. *Speech communication, 56*, 119-131. https://doi.org/10.1016/j.specom.2013.07.001

Gelas, H., Abate, S. T., Besacier, L., Pellegrino, F. (2011). Quality Assessment of Crowdsourcing Transcriptions for African Languages. *INTERSPEECH, 12th Annual Conference of the International Speech Communication Association.* Florencia, 3065-3068. https://www.researchgate.net/publication/221478079_Quality_Assessment_of_Crowdsourcing_Transcriptions_for_African_Languages

Ministerio de Cultura (2020). Base de Datos de Pueblos Indígenas u Originarios. https://bdpi.cultura.gob.pe/

Parent, G., Eskenazi, M. (2010). Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. *2010 IEEE Spoken Language Technology Workshop.* Berkeley, 312-317. https://doi.org/10.1109/SLT.2010.5700870

Rehm, G. (2018). The META-NET strategic research agenda for language technology in europe: An extended summary. En G. Rehm, F. Sasaki, D. Stein y A. Witt (Eds.), *Language technologies for a multilingual Europe: TC3 III* (pp. 19-41). Berlín: Language Science Press. http://doi.org/10.5281/zenodo.1291926

Ríos, A. (2016). A basic language technology toolkit for quechua. *Sociedad Española para el Procesamiento del Lenguaje Natural, 56*, 91-94.

http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5291

Wang, D. y Zhang, X. (2015). Thchs-30: A free chinese speech corpus. arXiv preprintarXiv:1512.01882

Woodbury, A. C. (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language documentation and description*, *12*, 19-36.

Zariquiey, R., Hammarström, H., Arakaki, M., Oncevay, A., Miller, J., García, A. y Ingunza, A. (2019). Obsolescencia lingüística, descripción gramatical y documentación de lenguas en el Perú: hacia un estado de la cuestión. *Lexis*, *43* (2), 271-337. https://doi.org/10.18800/lexis.201902.001