LINGUATEC:Desarrollo de recursos lingüi'sticos para avanzar en la digitalizacio'n de las lenguas de los Pirineos LINGUATEC:Development of linguistic resources to advance the digitisation of the languages of the Pyrenee

Itziar I. Aldabe, Josu Aztiria, Francho Beltrán, Myriam Bras, Klara Ceberio, Itziar Cortes, Jean-Baptiste Coyos, Benaset Dazeas, Louise Esher, Gorka G. Labaka, et al.

▶ **To cite this version:**

Itziar I. Aldabe, Josu Aztiria, Francho Beltrán, Myriam Bras, Klara Ceberio, et al. (Dir.). LINGUATEC:Desarrollo de recursos lingüi'sticos para avanzar en la digitalizacio'n de las lenguas de los Pirineos LINGUATEC:Development of linguistic resources to advance the digitisation of the languages of the Pyrenee. 2019. artxibo-02494778

HAL Id: artxibo-02494778

https://artxiker.ccsd.cnrs.fr/artxibo-02494778v1

Submitted on 29 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los Pirineos

LINGUATEC: Development of linguistic resources to advance the digitisation of the languages of the Pyrenees.

Itziar Aldabe[6], Josu Aztiria[3], Francho Beltrán[1], Myriam Bras[2], Klara Ceberio[3], Itziar Cortes[3], Jean-Baptiste Coyos[4], Benaset Dazeas[5], Louise Esher[2], Gorka Labaka[6], Igor Leturia[3], Kepa Sarasola[6], Aure Séguier[5], Jean Sibille[2]
[1]ARAGON TURISMO [2]CNRS-UNIV TOULOUSE 2 [3]ELHUYAR
[4]EUSKALTZAINDIA [5]LO-CONGRÉS [6]UPV/EHU
Información de contacto: j.aztiria@elhuyar.eus
https://linguatec-poctefa.eu/

Abstract: The goal of the project is to develop, test and disseminate new innovative linguistic resources, tools and solutions for a better digitalization level of the Aragonian, Basque and Occitan languages. As a result, we will obtain, among others, (1) a road map of Aragonian Digitalization, (2) new monolingual and bilingual lexicons and morphosyntactic and syntactic analysers for Occitan, (3) a Northern Basque speech recognition system, and several linguistic tools as well as (4) new innovative solutions for Aragonian, Basque and Occitan.
Keywords: Machine Translation, Languages with limited resources, Bilingual Corpora

Resumen: El objetivo del proyecto es desarrollar, probar y difundir nuevos recursos, nuevas herramientas y aplicaciones lingüísticas innovadoras para mejorar el nivel de digitalización del aragonés, vasco y occitano. Resultados esperados: (1) Hoja de ruta para la digitalización del aragonés, (2) Nuevos recursos lingüísticos, (3) Herramientas lingüísticas desarrolladas (síntesis de voz occitana, aragonesa y vasca del País Vasco francés, detector de texto occitano y variantes del occitano, mejora de la traducción automática del francés al occitano, del castellano vasco, del castellano al aragonés, (4) Aplicaciones innovadoras desarrolladas en los idiomas de los Pirineos.
Palabras clave: Traducción automática, Idiomas con recursos limitados, Corpus bilingüe

1 Introducción

El proyecto LINGUATEC[1] se desarrolla dentro del marco de cooperación interregional POCTEFA (Programa INTERREG V-A España-Francia-Andorra) está financiado con fondos FEDER de la Comunidad Europea y cofinancia proyectos de cooperación transfronteriza diseñados y gestionados por actores de ambos lados de los Pirineos y de las zonas litorales que participan en el Programa desarrollo inteligente, sostenible e integrador del territorio.

Este proyecto también se enmarca perfectamente en los objetivos del informe del Parlamento Europeo aprobado en septiembre de 2018 sobre "La igualdad lingüística en la era digital"[2]. El informe fue aprobado por amplia mayoría y entre otras recomendaciones pide crear o ampliar proyectos de diversidad lingüística digital, que investiguen las necesidades digitales de todas las lenguas europeas, incluyendo desde aquellas con muy pocos hablantes hasta las que cuentan con gran número de hablantes, con el fin de combatir la brecha digital y contribuir a preparar a dichas lenguas para el futuro digital sostenible"También pide a las administraciones "que mejoren el acceso a los servicios e informaciones en línea en diferentes lenguas, especialmente en el caso de servicios en regiones transfronterizas". Con ello el informe pretende reducir las desigualdades entre lenguas y comunidades lingüísticas, fomentar el acceso equitativo a los servicios y estimular la movilidad en Europa de las empresas, los ciudadanos y los trabajadores, así como garantizar la creación de un mercado único digital multilingüe inclusivo".

Siguiendo los objetivos del marco POCTEFA y el informe del Parlamento Europeo el proyecto LINGUATEC está desarrollando nuevos productos de tecnología de la lengua para mejorar el

[1]https://linguatec-poctefa.eu/
[2]http://www.europarl.europa.eu/doceo/document/A-8-2018-0228_ES.html

nivel de digitalización del aragonés, vasco y occitano. Además de la creación de nuevos recursos y herramientas lingüísticas tales como léxico monolingüe y bilingüe, sistemas de análisis morfosintácticos y sintácticos, sistemas de traducción automática o sistemas de reconocimiento de voz. Dentro del proyecto se crearán 6 aplicaciones en las que esos recursos y herramientas se puedan utilizar, y a ser posible de forma coordinada para todas las lenguas.

Este artículo comienza presentando brevemente el estado de las lenguas con las que trabajaremos en el proyecto y con la presentación de los socios del consorcio. El tercer apartado describe los recursos y herramientas a desarrollar. El cuarto presenta las 6 aplicaciones informáticas que se desarrollarán aprovechando los recursos creados. Finalmente se describe el plan de diseminación para el año 2020 que será el último del proyecto.

2 Idiomas y Grupos Participantes

El consorcio agrupa tres socios que trabajan con el euskera, dos con el occitano y uno con el aragonés. La situación actual de estos tres idiomas no es la misma. La red META-NET elaboró en 2012 una serie de 32 "libros blancos" que mostraban grandes diferencias de desarrollo entre los diferentes idiomas con respecto a las tecnologías de la lengua[3]. Según esos estudios en un claro primer nivel se encuentra el inglés; el español y el francés aparecen en un segundo nivel. El libro blanco The Basque Language in the Digital Age concluye que el euskera necesita de mucha más investigación fundamental y recursos para su supervivencia digital (Hernáez et al., 2012). El occitano y el aragonés no llegaron a considerarse dentro del grupo de las 32 lenguas de estudio, lo cual confirma que su situación actual es aún más debil que la del euskera. Una hipótesis de partida en este proyecto es el pensar que la experiencia ontenida en la digitalización para el euskera puede ser útil a la hora de realizar pasos en la digitalización del occitano y del aragonés.

2.1 Elhuyar Fundazioa (líder del proyecto)

La Fundación Elhuyar es una entidad sin ánimo de lucro fundada en 1972 para la promoción y difusión del euskera y de la ciencia y la tecnología (libros revistas, diccionarios...). Desde 2002 su unidad de I+D ha participado en múltiples proyectos para desarrollar aplicaciones relacionadas con diccionarios y corpus, con correctores ortográficos y de estilo, tecnologías para la traducción, herramientas de extracción terminológica, gestión de la información y del conocimiento, tecnologías del habla y enseñanza de idiomas. En 2018 ha creado una aplicación para la gestión multilingüe en sistemas de contenidos (CMS) (Cortes et al., 2018), y un exitoso sistema de traducción neuronal (Etchegoyhen et al., 2018). También gestiona una app en código abierto de traducción automatica [4].

2.2 Lo Congrès Permanent de la Lenga Occitana

LO CONGRÈS es, junto con Elhuyar Fundazioa, la entidad de la que parte la idea y coordina el proyecto. LO CONGRÈS es el organismo interregional para la regulación del idioma occitano. Su objetivo es contribuir a la vitalidad y el desarrollo del occitano mediante la producción de herramientas computacionales (lexicografía, lexicología, terminología, neología, fonología , ortografía, gramática y toponimia).

2.3 UMR 5263 - CNRS (Toulouse)

El laboratorio CLLE-ERSS, UMR 5263 CNRS y Universidad Tolosa 2 Joan Jaurès. En el proyecto LINGUATEC aportará sus competencias en lingüística occitana y francesa y en el tratamiento automático de las lenguas. Cuenta con experiencia en la construcción de recursos para el occitano (Bernhard et al., 2018).

2.4 Aragón Turismo

La Sociedad De Promoción y Gestión del Turismo Aragonés, SLU, o TURISMO DE ARAGÓN, es la entidad encargada por el Gobierno de Aragón de la promoción turística y del patrimonio regional, incluyendo la lengua aragonesa. Tiene mucha experiencia en actividades de promoción y difusión. Ha desarrollado proyectos de nuevas tecnologías en colaboración con el Instituto Tecnológico Aragonés (ITA) [5].

2.5 Euskaltzaindia

La Academia de la Lengua Vasca Euskaltzaindia, entidad oficial encargada de cuidar y normalizar el uso del idioma vasco, tiene experiencia en bases de datos lexicográficas, corpus lingüísticos, en indexación y clasificación de textos para búsqueda, y además, con la ayuda de este tipo de tecnología, para difundir su servicio de consulta

[3] http://www.meta-net.eu/whitepapers/overview

[4] http://www.mitzuli.com/en/

[5] https://www.itainnova.es/es

sobre el uso del euskara en la parte norte del País Vasco.

2.6 Ixa Taldea (UPV/EHU)

El Grupo Ixa de la Universidad Del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU) viene trabajando en tecnologias de la lengua para el euskera desde 1988. Entre sus 23 productos registrados se destacan los siguientes: (i) el corrector ortográfico Xuxen[6], comercializado desde 1984; (ii) el sistema de traducción automática Opentrad-Matxin, en explotación por la empresa Elhuyar[7]; (iii) la base de conocimiento Multilingual Central Repository[8], que aglutina, entre otros, los wordnets del español, catalán, gallego y euskera; (iv) el corpus etiquetado de Ciencia y Tecnología, realizado en colaboración con Elhuyar[9]; (v) el corpusetiquetado de referencia para el EPEC[10]; (vi) la herramienta de desambiguación de sentidos y entidades basada en grafos UKB[11]; y (vii) las cadenas de procesamiento de texto Ixa-pipes[12] e Ixa-Kat[13].

Su papel en el proyecto se centra en el asesoramiento basado en la experiencia previa en el tratamiento de lenguas de escasos recursos (Alegria and Sarasola, 2017), en la mejora de los sistemas de traducción automática existentes en el proyecto y en la creación inicial de un sistema de traducción para el par de lenguas euskera-francés.

3 Desarrollo de recursos y herramientas lingüísticas para facilitar el tratamiento automatizado de las lenguas

Se pretende en primer lugar establecer un marco metodológico común para el desarrollo de la informatización en estas lenguas de los Pirineos. Para ello hemos realizado una puesta en común sobre las estrategias de digitalización de cada lengua y hemos acordado las prioridades para el desarrollo de recursos lingüísticos. Adicionalmente también estamos desarrollando herramientas lingüísticas para avanzar en la interoperabilidad de las lenguas de los Pirineos.

Como resultado de esta puesta en común se establecieron las siguientes prioridades, estableciéndose la tarea de cada socio:

Occitano:

- LO-CONGRÉS Detector textual del occitano
- LO-CONGRÉS Detector textual de variantes del occitano
- LO-CONGRÉS Mejora de la traducción automática en francés occitano
- LO-CONGRÉS Síntesis vocal en occitano,
- CNRS Léxico monolingüe occitano: Colección de formas flexionadas.
- CNRS Léxico bilingüe occitano/otras lenguas
- CNRS Análisis morfosintáctico
- CNRS Análisis sintáctico

Aragonés:

- TURISMO DE ARAGÓN Mejora de la traducción automática en español-aragonés
- TURISMO DE ARAGÓN Síntesis vocal en aragonés

Euskera:

- ELHUYAR Síntesis de voz en euskera de Iparralde (Iparrahotsa 2.0)
- ELHUYAR Reconocimiento de voz en euskera: reconocer y clasificar palabras
- ELHUYAR Mejora de la traducción automática para el par español-euskera
- UPV/EHU Mejora de la traducción automática español-euskera
- UPV/EHU + ELHUYAR Investigación sobre traducción euskera-francés

4 Desarrollo de Aplicaciones Innovadoras en el Ámbito Lingüístico

Este proyecto pretende desarrollar una serie de aplicaciones innovadoras en el ámbito lingüístico. Con ello queremos facilitar la interoperabilidad entre estas lenguas de los Pirineos. Las aplicaciones previstas y el socio responsable de cada una de ellas son las siguientes:

- LO CONGRES Barra descargable de traducción automática para sitios web
- ELHUYAR Aplicación de Traducción Automática para CMS

[6] http://xuxen.eus/es/xuxen5
[7] http://matxin.elhuyar.eus/
[8] http://adimen.si.ehu.eus/web/MCR
[9] http://www.ztcorpusa.eus/aurkezpena.htm
[10] http://www.ixa.eus/epec-dep-deskarga
[11] http://ixa2.si.ehu.eus/ukb
[12] http://ixa2.si.ehu.eus/ixa-pipes/
[13] http://ixa2.si.ehu.eus/ixakat/

- ELHUYAR App de traducción automática entre las lenguas de los Pirineos: Euskara-Francés, Euskara-Español , Francés-Occitano y Español-Aragones
- ELHUYAR App de traducción automática entre las lenguas de los Pirineos: Euskara-Francés, Euskara-Español , Francés-Occitano y Español-Aragones
- EUSKALTZAINDIA Manual del Vasco Unificado: .Euskara Eskuz Esku Digitala".
- TURISMO DE ARAGÓN Diccionario Online del Aragonés.
- LO CONGRÉS Buscador Semántico Multilingüe.
- Medición de la Vitalidad del occitano, euskara y aragonés

5 Difusión de resultados

Durante el año 2020 se llevarán a cabo actividades con el objeto de dar a conocer los resultasos a entidades, organizaciones, empresas, profesionales e investigadores del área POCTEFA y de otras regiones europeas, que trabajan en el ámbito de las lenguas y las tecnologías lingüísticas o que realizan aplicaciones multilingües para servicios de e-administración, e-salud, e-justicia, e-educación o e-cultura.

Aparte de varios talleres que cada socio organizará a nivel local, también se organizará un seminario Europeo sobre Tecnologías de la Lengua y contribución al Desarrollo Económico. Su objetivo será la ayuda práctica a otros agentes ajenos al proyecto para que puedan incorporar la tecnología obtenida. Esperamos firmar acuerdos Universidad–Empresa para la comercialización de soluciones.

La colaboración transfronteriza permitirá transferir conocimientos y desarrollar soluciones lingüísticas con potencial de mercado, que beneficien a profesionales de las lenguas y faciliten el acceso público multilingüe a contenidos. Estamos dando pasos para avanzar en el desarrollo de un clúster transfronterizo de tecnologías lingüísticas.

Reconocimientos

La investigación llevada a cabo en este proyecto se lleva a cabo como parte del proyecto "LINGUATEC: Desarrollo de la cooperación transfronteriza y transferencia de conocimiento en tecnologías de la lengua"(POCTEFA EFA227/16, FEDER), financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional (FEDER).

Referencias

[Alegria and Sarasola2017] Alegria, I. and K. Sarasola. 2017. Language technology for language communities: An overview based on our experience. In *Çommunities in Control: Learning tools and strategies for multilingual endangered language communities", CinC 2017. October 19-21 2017, Alcanena / Portugal. Full corrected version: FEL XXI Alcanena 2017 Communinities in Control. Foundation for Endangered Languages, DIDLeS, SOAS World Languages Institute and Mercator Research Centre. pp. 91-97, ISBN: 978-0-9560210-9-0.*

[Bernhard et al.2018] Bernhard, D., A.-L. Ligozat, F. MARTIN, M. Bras, P. Magistry, M. Vergez-Couret, L. Steible, P. Erhart, N. Hathout, D. Huck, C. Rey, P. Reynés, S. Rosset, J. Sibille, and T. Lavergne. 2018. Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard. In *11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, May.

[Cortes et al.2018] Cortes, I., I. Leturia, I. Alegria, A. Astigarraga, and K. S. eta Manex Garaio. 2018. Massively multilingual accessible audioguides via cell phones. In *EAMT 2018 (Project/Product Track).*

[Etchegoyhen et al.2018] Etchegoyhen, T., E. Martinez, A. Azpeitia, I. Alegria, G. Labaka, A. Otegi, K. Sarasola, I. Cortes, A. Jauregi, I. Ellakuria, E. Calonge, and M. Martin. 2018. Quales: Estimación automática de calidad de traducción mediante aprendizaje automático supervisado y no-supervisado. *Procesamiento del Lenguaje Natural, vol. 61, pp. 143-146. ISSN: 1135-5948.*

[Hernáez et al.2012] Hernáez, I., E. Navas, I. Odriozola, K. Sarasola, A. Diaz de Ilarraza, I. Leturia, A. Diaz de Lezana, B. Oihartzabal, and J. Salaberria. 2012. *Euskara Aro Digitalean – The Basque Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.