Click here to be redirected to the virtual room of the Project Expo.

Project abbreviation: LINGUATEC

Project name: Development of cross-border cooperation and the transfer of knowledge in language technologies.

Project coordinator: Josu Aztiria Urtaran (j.aztiria@elhuyar.eus)

Project consortium:

- ELHUYAR FUNDAZIOA
- LO CONGRÈS PERMANENT DE LA LENGA OCCITANA
- UNIVERSIDAD DEL PAÍS VASCO / EUSKAL HERRIKO UNIBERTSITATEA (UPV/EHU
- CNRS (CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
- EUSKALTZAINDIA
- SOCIEDAD DE PROMOCIÓN Y GESTIÓN DEL TURISMO ARAGONÉS, SLU

Funding: Cooperation Programme Interreg V-A España-Francia-Andorra (Interreg POCTEFA); EFA 227/16 LINGUATEC; 797.875 €

Project duration: Project start: 01/01/2018 -Project end: 31/12/2020

Main key words: NLP, Cross-border cooperation, Neural Machine Translation, Languages with limited resources, Bilingual Corpora, Under-Resourced Language, Text-To-Sepeech, Specch-to-Text, Artificial Intelligence, Dictionaries, Digital Roadmap, tools.

Background of the research topic: The development of language technologies is uneven in the different languages of the Pyrenees: at the first level of digitisation are French and Spanish; at the second level are Catalan and Basque, which have a significant set of digital resources and tools; and at the third level are Occitan and Aragonese, which still have clear gaps in their digital development.

Given that the starting point in terms of the level of digitization of Basque, Occitan and Aragonese is very different, each language requires a series of specific resources and tools, which respond to the specific needs of the level of digitization of each one.

Goal of the project:

- Improve the technological capacity of Aragonese, Occitan and Basque-language with the development of language resources, tools and applications, and through technological cooperation and knowledge transfer between languages.
- The construction of a linguistic infrastructure based on Artificial Intelligence that guarantees the digital development of the languages of the Pyrenees.
- The Linguatec project can serve as a guide for other languages under-resourced and a low level of technological development.

Results:

- Digital Roadmaps:
    - Comparative report on the situation of Basque, Aragonese and Occitan.
    - Roadmap for the digitisation of Aragonese
- Resources:
    - Online dictionary of Aragonese.
    - Monolingual and Bilingual Occitan lexicon
    - Occitan morphosyntactic analysis
- Occitan syntactic analysis

Click here to be redirected to the virtual room of the Project Expo.

- Tools:
    - Speech recognition system for Basque-language.
    - TTS neural system to the Basque language
    - Improved automatic translation in the Spanish-Basque pair
    - TTS system in Aragonese.
    - Improved automatic translation in the Spanish-Aragonese pair.
    - TTS neural system in Occitan.
    - Occitan textual detector
    - Textual detector of Occitan variants
    - Improved machine translation in the French-Occitan pair

Project abstract: Cross-border cooperation will allow the transfer of knowledge and development of linguistic solutions with a potential market uptake, benefitting language professionals, easing access to multilingual contents, and fostering the development of a cross-border language technology cluster.

Publications:

LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los Pirineos . Aldabe, I, Aztiria, J., Beltrán, F., Bras, M., Ceberio K., Cortes, I., Coyos J.D., Dazeas B., Esher, L., Labaka, G., Leturia, I., Sarasola, K., Séguier A. and Sibille J.. In *Procesamiento del Lenguaje Natural (SEPLN)*, ISSN 1989-7553. 2019.

Neural Text-to-Speech Synthesis for an Under-Resourced Language in a Diglossic Environment:the Case of Gascon Occitan. Ander Corral, Igor Leturia, Aure Seguier, Michäel Barret2, Benaset Dazéas, Philippe Boula de Mareüil, Nicolas Quint. . Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 53–60. Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020.