IKER-GAITU: Research on Language Technology for Basque and Other Low-Resource Languages

Eneko Agirre[1], Itziar Aldabe[1], Xabier Arregi[1], Mikel Artetxe[1], Unai Atutxa[1], Ekhi Azurmendi[1], Iker De la Iglesia[1], Julen Etxaniz[1], Victor García-Romillo[2], Inma Hernaez-Rioja[2], Asier Herranz[2], Mikel Iruskieta[1], Oier López de Lacalle[1], Eva Navas[2], Paula Ontalvilla[1], Aitor Ormazabal[1], Naiara Perez[1], German Rigau[1], Oscar Sainz[1], Jon Sanchez[2], Ibon Saratxaga[2], Aitor Soroa[1], Christoforos Souganidis[2], Jon Vadillo[2] and Aimar Zabala[1]

[1]*HiTZ Basque Center for Language Technology - Ixa NLP Group, University of the Basque Country UPV/EHU*
[2]*HiTZ Basque Center for Language Technology - Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU*

Abstract
The general objective of the IKER-GAITU project is to research on language technology to increase the presence of Basque in the digital environment. It will be carried out between 2023 and 2025 thanks to a grant from the Department of Culture and Language Policy of the Basque Government. Current techniques require enormous amounts of textual and oral data per language. On the other hand, the data available for Basque and other low-resource languages might not be enough to attain the same quality as larger languages with the current technology. For this reason, it is essential to research on language technology, so that low-resource languages are present with the same quality as the rest of the languages in these technologies. IKER-GAITU pursues the following research objectives: 1. A system that automatically captures the level of Basque proficiency, written and oral; 2. Bring personalized voice technology to people with disabilities; 3. Spontaneous voice transcription, both when Basque and Spanish are mixed and when there are several speakers; 4. Textual conversational systems in Basque that match the quality of the most powerful large language models. In this project summary we present the results for the first year. More information at https://hitz.eus/iker-gaitu.

Keywords
Educational Applications, Speech Synthesis, Speech Recognition, Large Language Models, Low Resource Languages, Basque

1. Introduction

The general objective of the IKER-GAITU project is to research on language technology to increase the presence of Basque in the digital environment. It will be carried out between 2023 and 2025 thanks to a grant from the Department of Culture and Language Policy of the Basque Government.

The digital revolution has reached global languages, and to some extent also other languages such as Basque. Language technology is one of the most fruitful fields of Artificial Intelligence, which is having a profound impact on society. Examples of this impact are the current dialogue systems, both text-based, such as GPT, or voice-based, similar to Siri or Alexa. Unfortunately, such systems perform better in global languages such as Spanish or French, which can create a dangerous gap to the detriment of Basque.

Current techniques require enormous amounts of textual and oral data per language. On the other hand, the data available for Basque and other low-resource languages might not be enough to attain the same quality as larger languages with the current technology. For this reason, it is essential to research on language technology, so that low-resource languages are present with the same quality as the rest of the languages in these technologies.

IKER-GAITU pursues the following research objectives for the three-year period, organised in four working packages:

- A system that automatically captures the level of Basque competence, written and oral;
- Bring personalized voice synthesis technology to people with disabilities;
- Spontaneous voice transcription, both when Basque and Spanish are mixed and when there are several speakers;
- Textual conversational systems in Basque that match the quality of the most powerful large language models.

In all these objective systems, the aim is to achieve a quality sufficient to be integrated in applications that reach society as soon as possible. For this purpose the results, data and algorithms that are created are being distributed openly.[1]

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain
e.agirre@ehu.eus (E. Agirre)

[1]https://hitz.eus/iker-gaitu

For the first year, the specific objectives where the following: 1) A first prototype for checking the proficiency level of written Basque; 2) A freely available foundation model called Latxa, the largest language model built for Basque; 3) A first prototype of an open dialogue system.

In the following sections we present the work performed in each of the working packages.

2. Assessing the competence in Basque

This work package addresses the task of assessing the written and oral competencies of Basque language learners, according to the levels established in the CEFR standard (Common European Framework of Reference for Languages) for evaluation.

The idea is to automatically evaluate written documents or audios generated by the students. As a first approach, we will develop a binary classifier, which, given a text or an audio, will determine whether or not it would reach the C1 level. We propose the development of a neural classifier, for which it is essential to have labeled data, either text or audio, with which to train the system.

HABE [2], an association dedicated to the teaching of Basque and in charge of the evaluation of official tests, has numerous data in different formats: electronic documents, manuscripts and recordings or audios. These files are labeled with their respective evaluation, corresponding to the CEFR level.

In the design of the system, two autonomous processes have been distinguished: the one corresponding to the evaluation of the written texts and the one corresponding to the oral exercises.

As far as written texts is concerned, the initial task involved the extraction of editable text from a corpus of 157,268 manuscripts for communication levels B1, B2, C1 and C2 in Basque. This transcription of handwritten text is a complicated task, since all the manuscripts are by different authors and the aim is to faithfully collect the original text without any kind of correction.

The neural evaluator will be trained on these transcribed documents. However, without being waiting for the transcription work to be completed, we carried out a set experiments leveraging the available exercises in electronic format, about 800 exercises in all. On the one hand, we fine-tuned the RoBERTa encoder for Basque [1][3] as regression model and a binary classifier. On the other hand, we experimented with a preliminary version of Latxa, using it in as fine-tuned classifier and as prompting-based model, with only 250 examples in the latter case.

Regarding the experimental setting, we propose two scenarios for the evaluation of these models. In the first one, we have randomly distributed the training, development and test partitions, whereas in the second one, we distributed the exercises by years. The random distribution implies that texts from the same period of analysis can appear in both the training and test partitions.

The best results obtained in these preliminary experiments are close to 80% F1score in the case of random distribution and 66% F1score when the distribution is done annually. There is no doubt that these results are strongly conditioned by the scarcity of data. Our hypothesis is that the performance and generalization ability of the classifier will improve by training with more text. In the meantime we continue to investigate techniques to improve learning with little data.

With regard to the oral part of the C1 level assessment, we have begun to compile the resources and define the environment needed to perform an oral evaluation. First of all, we defined a specific setup for the oral test: it would consist on an oral test to be guided by the computer. The student will perform the test individually and independently, answering to the questions by voice. Thus, the format of the evaluation procedure does not match those been used so far over more than 15 years. Current standard tests are conducted by one or two examiners who evaluate two candidate students simultaneously, generally by giving the students to make a monologue on a particular theme.

Through HABE, recordings of the C1-level tests of the last 15 years have been obtained for a total of about 150 hours (448 audio files containing all recordings of test sessions). In addition to the recordings and the label 'pass/fail', the scores obtained by each student according to five criteria are available: fluidity, richness, correctness, suitability and coherence. Thus, the recordings can be used to train and test a classifier of C1-level (pass/fail classification), if the audio segments corresponding to each student are extracted, transcribed and annotated.

However, the quality and characteristics of the compiled audio files are diverse and pilot experiments of automatic transcription gave very poor results. For this reason a sample set of 16 files have been selected, representing the different recording quality found in the files. The aim is to establish labelling conditions so that students' audios can be extracted to be used in the training to train the classification system. Presently, we are preparing the process of annotating the students interventions and their corresponding transcriptions. In a near future, we hope to use those audios to train a C1 Level classifier and to perform experiments where the automatic transcription will also be used for a final decision.

[2] Helduen Alfabetatze eta Berreuskalduntzerako Erakundea.

[3] https://huggingface.co/ixa-ehu/roberta-eus-euscrawl-large-cased

3. Personalised voice synthesis

The aim of this work package is to advance the development of synthesis technologies that facilitate communication between people with oral disabilities with special focus on Basque. To this end, the following objectives shall be pursued:

- To develop the technology necessary to create personalized voices
- To implement voice models in environments with limited capacity (embedded models)

Voice personalization requires the use of a TTS neural architecture that having been trained with a high amount or audio data, it is able to adapt to a new voice using only a few recordings from the new speaker. A number of architectures based on Tacotron2 [2], FastSpeech2 [3] and VITS [4] have been evaluated and its performance tested in a variaty of experiments. Additionally, several vocoders have been trained, mostly based on the HiFi-GAN architecture [5], to minimize computational load without diminishing the quality of voice obtained.

Several models have been trained with recordings obtained from voice talents, using different architectures. Several experiments have also been conducted to evaluate different training and fine-tuning strategies. In addition, techniques for adapting systems to language dialects have been investigated using embedding. In this context, a research stay has been held at the OFAI (Vienna, Austria) and a publication at an international congress [6]. In the future these techniques will be used to adapt the system to most relevant Basque dialects.

Another area of application of these technologies has been the development of personalized voices for children. Special efforts have been done to obtain recordings of children's quality voices, to create different synthetic voices for children.

Finally, two neural architectures using Tacotron2 and VITS have been chosen to automate the personalization process from a small set of recordings and to be implemented to run on different platforms: Linux, Windows and Android. To do this, different options were explored and neural models were transferred to the ONNX platform. Using ONNX, new neural networks have been integrated into Aholab's AhoTTS synthesizer.

Thus, the following specific objectives have been achieved by 2023:

- For voice synthesis, three SoA architectures have been studied that allow personalization. The best among them has been chosen. The selection has been made taking into account the ability to personalize as well as the final quality of the synthetic voice.
- The voice models have migrated to ONNX (to be used in a mobile device) .
- For standard voices, three bilingual (ES/EU) voices (women, men and children) have been developed. These voices cannot be used for commercial purposes without explicit permission from the speaker.
- High-quality children's voices have been acquired and personalised for them. Those voices will be available on the voice bench.

The demonstrations of the work done can be found here.

4. Spontaneous speech transcription

The main objective of this work package is to obtain open ASR models able to perform transcription on bilingual (Basque-Spanish) environments. For the first year, the goal was to define the data sets required for bilingual automatic transcription and the sources from which they should be extracted. Simultaneously, different architectures are being tested with the available data.

4.1. Prototypes

The available ASR architectures have been evaluated. Priority has been given to those who allow the open use of the models. The main architectures that have been investigated are the Conformer-CTC and the Conformer Transducer within the Nvidia-NeMo framework [4].

4.2. Data

A number of audio and text databases have been used to train the prototypes. As for the audio data used to train the models, the databases in Basque used have been the last two versions from Mozilla Common Voice (16 and 17) [7], OpenSLR, and the recordings from the Basque Parliamente recently published [5]. See Table 1 for the amount of hours for each dataset.

The language model used was generated mainly using Wikipedia and has about 27 M sentences.

4.3. Preliminary results

Many experiments have been performed using the described models and data. The best results so far have been obtained using the Nvidia NeMo conformer-ctc

[4] https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/asr/models.html

[5] https://huggingface.co/datasets/gttsehu/basque_parliament_1

Table 1
Number of hours in each audio dataset

Dataset	Train	Test	Dev
Common Voice 15	58 h	10.83 h	–
Common Voice 16	173.78	21.50	–
Open SLR	6.45	2	–
Basque Parliament	368	2.85	2.62

model fine-tuned with all the available audio data (CV 16) from a Spanish pretrained model [6]. This model obtained a WER of 2.22% when tested with CV16 test set, a WER of 9.31% when tested with OpenSLR test dataset, and 4.22% when tested with the Basque Parliament test dataset. The low WER obtained for the CV test dataset can be explained by the probable leak of sentences of CV in the training text set used to generate the Language Model. An study of contamination between the databases showed that 43.88% of the sentences in the CV16 test-set was included in the Language Model training set. On the other hand, only 0.29% leakage was found in OpenSLR and 0.00% in the Basque Parliament test-set. The described models are available at Hugging Face [7],[8].

5. Basque language model

This first year we built the Latxa model family, the largest and best-performing LLMs available for Basque. Latxa is a breed of domestic sheep native to the Basque Country, famous for its cheese.

Our Latxa is a family of Large Language Models (LLM) ranging from 7 to 70 billion model parameters based on Meta's LLaMA models. Current LLMs exhibit incredible performance for high-resource languages such as English, ChatGPT being the most popular example. But, in the case of Basque and other low-resource languages, their performance is close to a random guesser, widening the technological gap between high- and low-resource languages when it comes to digital tools. We present Latxa to overcome these limitations and promote the development of LLM-based research, innovation and products for the Basque language.

The Latxa family of models are pre-trained base LLM models, without further fine-tuning on user-oriented instructions or preferences. These models are thus not for direct use by the general public. These models are key to building successful NLP tools for Basque. We release these open models to be used by technicians that know how to include such base LLMs in final-user applications, or know how to adapt them via fine-tuning. We are already working on instruction-following models, but it is still an open research issue whether models usable by the general public with similar quality to GPT can be constructed for Basque. The models were developed using in-house GPUs, with the final models being trained on the Leonardo supercomputer at CINECA under the EuroHPC Joint Undertaking (project EHPC-EXT-2023E01-013).

For the corpora, we leveraged a new corpus comprising 4.3M documents and 4.2B tokens after deduplication and filtering.

Addressing the scarcity of high-quality benchmarks for Basque, we introduce four multiple-choice evaluation datasets: questions from official language proficiency exams; reading comprehension questions; trivia questions from five knowledge areas; and questions from public examinations.

In our extensive evaluation (cf. Figure 1), Latxa outperforms all previous open models we compare to by a large margin. In addition, it is better than GPT-3.5 Turbo, and better than GPT-4 Turbo in language proficiency and understanding, despite lagging behind in reading comprehension and knowledge-intensive tasks.

To assess the quality of the models, we thoroughly evaluated them on a suite of diverse and challenging tasks. The tasks evaluate the performance of the models for a variety of linguistic competences such as reading comprehension, common sense reasoning, sentiment analysis, stance detection, topic classification, correference, inference and word senses (see model cards in HuggingFace for more details on evaluation datasets and procedure). The results in the figure below show the performance of different models, with the average in the rightmost part. We tested the English LLaMA models as well as some of the best language models for Basque to date, allowing for head-to-head comparison with our models (three purple bars). The figure clearly indicates the superiority of our three models, as well as the improvement of results as we increase model size.

Latxa models inherit the LLaMA-2 License, which allows for commercial and research use. Although based on an English LLM, these models are intended to be used with Basque text; for any other language the performance is not guaranteed.

The corpora, models and evaluation benchmarks, together with the code, form an open language model and evaluation suite for Basque. The suite is described in [8], and is publicly available in HuggingFace[9], please refer to the model card for more technical information and to get started with the models.

In addition to Latxa, we have explored whether multilingual language models perform better when working

[6] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_conformer_ctc_large

[7] https://huggingface.co/HiTZ/stt_eu_conformer_transducer_large

[8] https://huggingface.co/HiTZ/stt_eu_conformer_ctc_large

[9] https://huggingface.co/collections/HiTZ/latxa-65a697e6838b3acc53677304

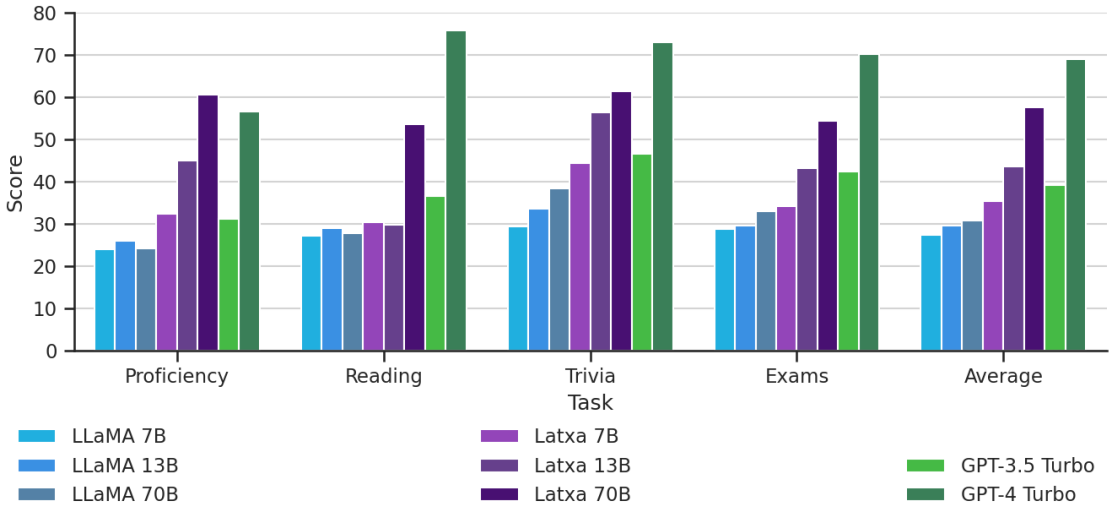

Figure 1: Results for Latxa on our new evaluation datasets alongside other available models.

in English than in Basque for the same tasks, even if machine translation is used in the process [9], whether small domain-adapted language models can be combined with large generalistic language models [10] and the effect of language typology when transferring knowledge from one language to the other on a wide range of experiments involving language models [11].

6. Conclusions

The objectives of the first year have been met. On the second year we are going to focus on data collection, as it is the key for the objectives set in the project. In parallel we plan to improve the models in all four working packages. In the case of the dialogue system, we will also focus on developing a instruction-tuned and aligned models.

7. Research Groups participating in the project

The project is carried out by HiTZ Basque Center for Language Technology[10] of the University of the Basque Country UPV/EHU, which comprises the following research groups.

Aholab Signal Processing Laboratory. Aholab[11] is the short name of the Signal Processing Laboratory of the University of the Basque Country (UPV/EHU). Aholab is a university research team and since 1995 it focuses its research in the areas of Text to Speech Conversion, Speech and Speaker Recognition and Speech Processing in general. The laboratory is part of the Basque Center for Language Technology (HiTZ) and the Department of Communications Engineering of the Faculty of Engineering of Bilbao (EIB).

Ixa NLP Group. Ixa[12] is a research group from the University of the Basque Country (UPV/EHU) that works in all areas of Natural Language Processing. Ixa is a multidisciplinary group with more than 25 years of experience, comprising computer scientists, linguists and other disciplines. The group is based on the Computer Science Faculty in San Sebastian and the Languages and Computer Systems department, but many members belong to other faculties and departments of the UPV/EHU. The group is part of the Basque Center for Language Technology (HiTZ).

Acknowledgments

This research project is funded by a grant from the Department of Culture and Language Policy of the Basque Government.

[10] https://hitz.ehu.eus
[11] https://aholab.ehu.eus/
[12] htpps://ixa.ehu.eus

References

[1] M. Artetxe, I. Aldabe, R. Agerri, O. Perez-de Viñaspre, A. Soroa, Does corpus quality really matter for low-resource languages?, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 7383–7390. URL: https://aclanthology.org/2022.emnlp-main.499. doi:10.18653/v1/2022.emnlp-main.499.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, Tacotron: Towards end-to-end speech synthesis, in: Proceedings of INTERPSEECH, ISCA, 2017, pp. 4006–4010.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech 2: Fast and high-quality end-to-end text to speech, arXiv preprint arXiv:2006.04558 (2020).

[4] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 5530–5540.

[5] J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, Advances in Neural Information Processing Systems 33 (2020) 17022–17033.

[6] L. Gutscher, M. Pucher, V. Garcia, Neural Speech Synthesis for Austrian Dialects with Standard German Grapheme-to-Phoneme Conversion and Dialect Embeddings , in: Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023), 2023, pp. 68–72. doi:10.21437/SIGUL.2023-15.

[7] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[8] J. Etxaniz, O. Sainz, N. Perez, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An open language model and evaluation suite for Basque, 2024. arXiv:2403.20266.

[9] J. Etxaniz, G. Azkune, A. Soroa, O. L. de Lacalle, M. Artetxe, Do multilingual language models think better in english?, 2023. arXiv:2308.01223.

[10] A. Ormazabal, M. Artetxe, E. Agirre, Comblm: Adapting black-box language models through small fine-tuned models, 2023. arXiv:2305.16876.

[11] M. Zubillaga, O. Sainz, A. Estarrona, O. L. de Lacalle, E. Agirre, Event extraction in basque: Typologically motivated cross-lingual transfer-learning analysis, in: Proceedings of the 15th Conference on Language Resources and Evaluation (LREC-Coling 2024), 2024.