Artificial Intelligence Research and Development
T. Alsinet et al. (Eds.)

doi:10.3233/FAIA240428

Pushing the Boundaries of Natural Language Processing (NLP): Enhancing Catalan Text Transcription Through Large-Scale Models in the Educational Field

Sergi RAMIREZ-MITJANS[a,b,1] , Huilin NI[a,b] , Jofre MOSEGUÍ[a,b] , Javier PUERTA[a,b] , Marcel VERA[a,b] and Karina GIBERT[a,b,2]
[a] *Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence (IDEAI-UPC)*
[b] *Bachelor's degree in Artificial Intelligence at Universitat Politècnica de Catalunya (UPC)*

ORCiD ID: Sergi RAMIREZ-MITJANS https://orcid.org/0000-0002-7782-3270
Karina GIBERT https://orcid.org/0000-0002-8542-3509

Abstract. This work delves into the intricate landscape of employing large language models for the transcription of Catalan texts within the realm of artificial intelligence. As it has been recently stated, Catalan is a medium language that has a very low presence in the digital frameworks and there is still space for improvement in the generation and maturity of the computational linguistics resources for the Catalan language [4]. Nowadays, in the middle of a disruptive digital transformation, the presence of all cultures and languages in Internet and digital frameworks becomes crucial. Generating digital contents in Catalan becomes a priority. Prioritizing the generation of digital contents in Catalan and being able to use Natural Language Processing or Speech recognition will enlarge the impact to other audiences. A central activity is transcription of Catalan videos into Catalan subtitles that eventually can be translated in other languages. Up to now, channels as popular as YouTube still cannot provide a transcription of a Catalan video to generate subtitles in Catalan, and this precludes the possibility to get the subtitles translated to other languages. In Catalonia, some specific policies promote the innovation and advances in the Catalan computational resources, like the [12], or the AINA project [8]. In this work, we analyse the state of the art on transcribing Catalan videos to generate subtitles in three languages (Catalan, Spanish and English). Eventually, the corpus analysed had some Spanish video that has also been considered. The paper rigorously tests a few resources to discern their efficacy in accurately transcribing Catalan text. Despite the promising capabilities of these models, our findings revealed a worse performance than expected, as these systems are still very sensitive to the characteristics of the speaker voice and speech. Common challenges included difficulties in handling Catalan-specific diacritics, idiosyncratic vocabulary, and nuances of regional dialects are identified. The paper describes the experimental setting where several tools have been tested and the results, providing some conclusions and diagnosis of the limitations and strengthnesses of the tested tools.

Keywords. Transcriptions, AI, Large Models, Natural Language Processing

[1] Corresponding Author: Sergi Ramirez, sergi.ramirez@upc.edu
[2] Corresponding Author: Karina Gibert, karina.gibert@upc.edu

1. Introduction

This work explores the use of large language models for transcribing Catalan texts in artificial intelligence. Catalan has minimal digital presence, requiring improved computational linguistics resources [4]. With the rise of online multimedia consumption, generating digital content in Catalan becomes essential. There is a growing demand for tools to transcribe and translate Catalan videos, enhancing their impact and reaching diverse audiences. Despite advancements, transcription tools for smaller languages like Catalan are underdeveloped. Platforms like YouTube do not support Catalan transcription, hindering subtitle generation and translation. Although Alexa started training in Catalan in 2021 [11], Amazon declared in 2023 that it will still remain unavailable [23]. Catalonia has enacted policies, such as the AINA project [8], to improve these resources, but challenges persist.

This study focuses on transcribing Catalan videos into subtitles and translating them into other languages. In the analysed corpus some videos were presented in Spanish, not in Catalan language, so the alternative path of generating Catalan from a Spanish transcription is also analysed. Catalan's linguistic complexity necessitates specialized tools. The paper tests models from Meta, OpenAI, Whisper, YouTube, Google translator among others to evaluate performances and robustness to regional dialects. The experimental setting and results highlight the limitations and strengths of the tested tools.

Additionally, the paper explores practical applications beyond entertainment, emphasizing educational contexts and promoting linguistic diversity. It includes an introduction to transcription and translation models, a historical overview, and results from transcribing and translating 102 Catalan videos for an AI training course. Conclusions and a critical analysis of the findings are provided.

2. State of Art

In recent years, the development and application of language models in NLP has significantly advanced, opening new possibilities in transcription and automatic translation of audiovisual content. For the specific case of Catalan, relevant research and developments contributed to the current state of these technologies. Speech recognition models appear in the 1950s – 1960s, where Bell Laboratories designed a system able to recognize a person's voice by dictating numbers. The IBM Shoebox (1961), considered the first transcriber with certain limitations of vocabulary, was presented at the New York World's Fair in 1964 [15]. In the 1970s, significant advances were made in the field of speech recognition thanks to the U.S. Department of Defence. Also, the "Harpy" speech system from Carnegie Mellon was developed, able to understand more than 1000 words, equivalent to the vocabulary of a three-year-old child. In the 1980s, the Hidden Markov Model (HMM) allowed to find patterns in sounds by estimating the probability that unknown sounds were words instead of using words as part of their identification. From the 1990s, more efficient machines and faster processors came, and speech processing improved. Currently, many tools from major technological companies such as Apple or Google offer tools for transcription and, audio translation [25].

2.1. Automatic Transcription of Videos in Catalan

In the field of automatic transcription of videos in Catalan, there has been a growing trend towards the use of language models as a key tool. These models, especially pre-trained ones like Bidirectional Encoder Representations from Transformers (BERT) [9], are proving to be effective in generating accurate and contextually relevant transcriptions of audiovisual content in Catalan. This approach has become particularly promising due to its ability to capture the linguistic and semantic complexities of the language under study, as well as its adaptability to a wide range of audiovisual contexts. There are studies [2] detailing the challenging task of speech-to-text transcription (STT) for languages such as Catalan and Spanish. Such research dates to 2014. Since then, there has been such evolution that we now have studies discussing new tools capable of transcribing the speech of a Catalan speaker into text. Recent research has explored how adapting pre-trained language models to Catalan can significantly improve the quality of automatic transcriptions. These studies [5] [10] [3] [18] highlighted the potential of language models to capture the meaning and intention more accurately behind Catalan, increasing precision and understandability.

Furthermore, efforts have been made to enhance these models' ability to handle the specific characteristics of Catalan, such as its unique grammar and vocabulary [16]. Adapting pre-trained language models through fine-tuning and customization techniques has proven to be an effective strategy for improving the accuracy and fluency of automatic transcriptions in Catalan.

In summary, the use of pre-trained language models like BERT in the automatic transcription of videos in Catalan represents a significant advancement in natural language processing technology. These models offer the ability to generate accurate and contextually relevant transcriptions in Catalan, which has important implications for a variety of applications, from accessibility to online audiovisual content indexing.

2.2. Automatic Translation of Videos in Catalan

Automatic Translation of Videos in Catalan is a research area in constant evolution that has seen significant advancements in recent years. Earlier methods, mainly statistical and rule-based systems, delivered acceptable but limited results in accuracy and fluency. The advent of Neural Language Models (NLMs), such as BERT [9], has significantly transformed this field by effectively capturing language complexities and contextual semantics. These neural models excel in generating precise and naturally fluent translations, addressing the unique challenges of translating audiovisual content in Catalan [28].

Beyond mere translation, these models also adapt cultural references, idioms, and specific grammatical structures of Catalan to ensure the original message is conveyed accurately across different cultural contexts. The Barcelona Supercomputing Centre (BSC) [7] has even developed models that consider the various dialects and variants of Catalan.

In summary, the evolution of neural language models has revolutionized automatic translation of videos in Catalan, providing precise and fluent translations. This advancement enhances accessibility to Catalan audiovisual content globally and promotes linguistic and cultural diversity, all while reducing the economic cost associated with manual translation.

3. Methodology

This section outlines our methodology for evaluating accurate transcription and translation of Catalan audiovisual content in e-learning. We assess both open-source and commercial solutions to determine the most effective tools for accurate and efficient transcription and translation. The process includes describing the experimental videos used, selecting and testing various tools, and providing a comparative analysis of their performance.

3.1. The experimental setting

The reference dataset comprises 122 pre-recorded AI course videos, each under 20 minutes, with technica content delivered by five female speakers. Additionally, 19 videos, around 8 minutes each, feature a Spanish speaker with a Latin American accent and low clarity.

Figure 1- Information for the 5 speakers.

Speaker	nr of videos	average length of	dialect	clarity of pronounce
1	42	10'	Central Catalan	High
2	20	14'	Central Catalan	High
3	11	12'	Central Catalan	Medium
4	18	9'	Central Catalan	Low
5	12	8'	Central Catalan	High

For testing, a sample of videos from all speakers was used to generate subtitles via tools like Whisper, NeMo, YouTube, SoftCatalà, Sonix, Transkriptor, Happy Scribe, and SeamlessM4T. After speaker review and corrections, subtitles for the remaining videos were created. Two methodologies were tested: generating voice-to-text translations to Spanish and English using Whisper and Google Translator, and text-to-text translations from Catalan subtitles to Spanish or English using NLLB and AINA. Human validation and corrections were required for all voice-to-text processes. Figure 2 shows the pipeline proposed in the project.

During the research step, all the tools were tested and modified under "Voice-to-text transcription Catalan" and "Machine text-to-text translation". The following sections describe the various tools used.

3.2. Automatic Transcription of Videos in Catalan

Next, the tools used in this research are briefly described.

3.2.1. Whisper OpenAI

Whisper [30], from OpenAI, uses Deep Learning (DL) and automatic speech recognition (ASR) to convert spoken language from audio and video into text. It analyses audio input, extracting linguistic features and patterns to generate textual transcriptions. Trained on vast amounts of data, Whisper delivers accurate and reliable transcriptions in many languages. It offers an intuitive API interface for easy integration into applications.

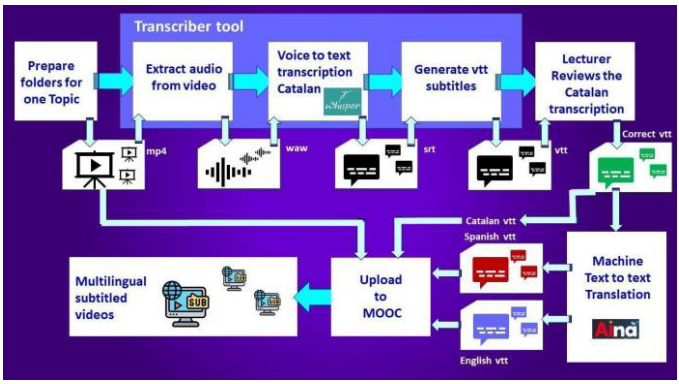

Figure 2 – Proposed pipeline for transcription of Catalan videos and subtitling in 3 languages

3.2.2. *NeMo NVIDIA*

Neural Modules (NeMo) [21], developed by NVIDIA, is an open-source toolkit designed for creating and training advanced conversational AI models, particularly for Automatic Speech Recognition (ASR) and Speech-to-Text (STT) tasks.

We focused on evaluating the three NeMo models for Catalan audio transcription:

1. Starting with the 'stt_ca_quartznet15x5' model, it uses the EncDecCTCModel architecture with a QuartzNet-like structure [17] of 15 blocks of 5 convolutional layers each, trained using Connectionist Temporal Classification (CTC).
2. The 'stt_ca_conformer_ctc_large' model, built on the EncDecCTCModelBPE architecture, integrates the Conformer NN architecture [13] and Byte Pair Encoding.
3. The 'stt_ca_conformer_transducer_large' model, which combines the Conformer architecture with a Transducer-based encoder, represents a significant step forward in ASR and STT technology compared with the previous two models.

3.2.3. *YouTube*

The YouTube [31] automatic subtitle generator offers as a convenient solution for transcribing audio and video content on the platform. Using advanced speech recognition algorithms, it automatically generates subtitles in multiple languages, enhancing accessibility and user experience. YouTube's systems allow users to specify regional variations, enhancing transcription accuracy. With five different variants for Spanish transcriptions such as Latin American, from Mexico, among others, users can choose the one that best matches the speaker's accent and dialect, and accuracy improves a lot.

Additionally, obtaining Spanish subtitles serves as an intermediary step in our workflow for transcribing them into Catalan, improving the further translation process.

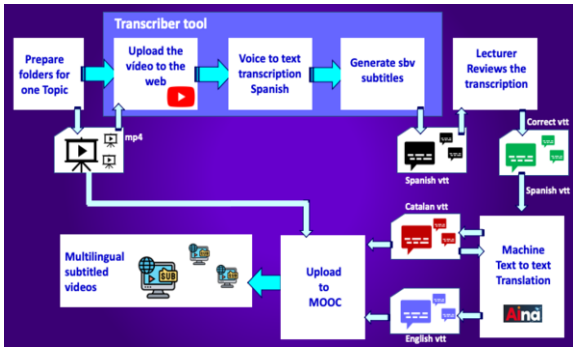

Figure 3 - Pipeline proposed in the project to use automatic transcription and translation subtitles for Spanish

3.2.4. SoftCatalà

SoftCatalà [26] is a Catalan non-profit organization focused on promoting the use of Catalan language in technology and digital communication. They contribute to various projects related to Catalan language tools, including language models like SoftCatalà/cat-ca_bert-base. This model, available on the Hugging Face model hub, is based on the BERT architecture and is specifically trained for various Catalan language processing tasks. It facilitates tasks such as text classification, sentiment analysis, and more, making it a valuable resource for developers working with Catalan language data in natural language processing applications.

3.2.5. Sonix

Sonix AI [25] is an AI-powered transcription and translation platform developed by Sonix. This platform utilizes advanced machine learning algorithms to accurately transcribe audio and video files into text in multiple languages. It offers a user-friendly interface, real-time collaboration features, and customizable settings to meet the needs of various users, including journalists, researchers, podcasters, and businesses. Sonix AI is designed to streamline the transcription process, saving users time and effort while ensuring high accuracy and reliability in converting spoken content into written text.

3.2.6. Transkriptor.com

Transkriptor.com [27] is an AI-powered transcription model that utilizes advanced machine learning algorithms to automatically transcribe speech into text files, supporting over 100 languages and dialects, including English, Spanish, and Chinese. Users can access Transkriptor.com through its web platform or integrate it into their applications for machine transcription. With its focus on accessibility, reliability, and efficiency, it competes with similar AI-driven transcription models like Happy Scribe, offering a compelling alternative for users seeking high-quality transcription services.

3.2.7. Happy Scribe

Happy Scribe [14] functions as a versatile platform for converting audio and video files into text, offering efficient and accurate transcription services. It works by utilizing advanced algorithms to accurately transcribe spoken context, catering to various languages and accents. Additionally, it provides subtitling capabilities, enabling users to create subtitles for videos based on the transcribed text. Users can upload their audio or video files to the platform, where the AI model generates transcriptions.

3.2.8. SeamlessM4T

SeamlessM4T [24], developed by Meta, is an all-in-one multimodal and multilingual translation model, revolutionizing communication by seamlessly integrating speech and text across languages. It offers comprehensive support for automatic speech recognition (ASR), speech-to-text (STT), text-to-speech (TTS), and text translation, eliminating the need for separate models.

3.2.9. FreeLing

FreeLing [22], developed by the TALP Research Center at the Universitat Politècnica de Catalunya (UPC) and led by Lluís Padró, offers a comprehensive suite of linguistic analysis tools. These include text tokenization, morphological analysis, POS tagging, and NER (named entity detection), supporting multiple languages such as English, Spanish, Catalan, and others. However, while FreeLing excels in linguistic analysis, it is not tailored for Automatic Speech Recognition (ASR) or Speech-to-Text (STT) tasks, making it unsuitable for our transcription needs.

3.3. Automatic Translation of Videos in Catalan

3.3.1. NLLB Meta

The Meta Neural Language Learning for Bilingual Model (NLLB) [20] by Meta (formerly Facebook), stands at the forefront of machine translation, aiming to seamlessly translate between various languages, including Catalan, Spanish, and English. Operating on neural machine translation (NMT) principles, it employs DL and a sophisticated architecture of interconnected ANN, including encoder and decoders. In the translation process, the encoder analyses the input text, encoding semantic and syntactic information into a hidden state, while the decoder generates translated output text using this hidden state, ensuring coherence and fidelity to original meaning.

3.3.2. AINA

The AINA project [8] stands as a pioneering endeavour dedicated to fostering the advancement of computational resources specifically for Catalan. The AINA project has developed AINA, a sophisticated translation tool designed to facilitate seamless and accurate translation between Catalan and other languages such as Spanish and English.

4. Application of machine transcription and translation to the videos of AI training course

4.1. Transcription

The tools presented in the previous section were tested on the random sample. Figure 4 shows the pros and cons of the several tools.

4.1.1. Whisper OpenAI

During our evaluation of transcription tools for Catalan videos, we encountered significant challenges with Whisper. Despite its advanced capabilities, Whisper's performance with Catalan audio proved inadequate, displaying numerous inaccuracies and mistakes in the transcriptions. This suggests a lack of integration or training for the intricacies of the Catalan language, resulting in subpar transcription quality.

Additionally, Whisper's performance with Spanish videos, particularly those featuring accents like a South American one, was suboptimal, struggling to capture nuances and variations in dialects and accents accurately, indicating insufficient training to handle diverse regional variations and accents in Spanish.

Software	Distributor	Interface	Transcription
Whisper	OpenAI	API	Inadequate transcription quality for Catalan videos, with numerous inaccuracies and mistakes. For Spanish videos we have a suboptimal performance, especially with different accents. Offers accurate punctuation signs and upper/lower cases.
NeMo stt_ca_quartznet15x5	NVIDIA	API	Rapid transcription but may compromise accuracy for both languages
NeMo stt_ca_conformer_ctc_large	NVIDIA	API	Shows superior performance in transcribing Catalan speech with complex linguistic features. For Spanish, the transcription is better from catalan
NeMo stt_ca_conformer_transducer_large	NVIDIA	API	Outperforms others in accuracy but is computationally heavier, for both languages
YouTube	Google (YouTube)	Drag and Drop in GUI	Unavailable for Catalan. Improved transcription accuracy observed for Spanish videos. Up to 5 Spanish accents available. Better transcription than others. Does not include punctuation signs or upper/lower cases in the generated subtitles.
happyscribe	Happy Scribe Ltd	Drag and Drop in GUI	Open version limited 30min/month
Sonix	Sonix	Drag and Drop in GUI	Only 30 minuts free credits
Softcatalà	Softcatalà	Drag and Drop in GUI	It's only available on-demand and limited videos bellow 2GB for Catalan and not available for Spanish
Buzz [29]	Github open source project	Local App	Use a local app to transcript the videos but it is required to install this app in your pc. Isn't portable when use multiples persons
oTranscribe+ [6]	BSC	Drag and Drop in GUI	This tool allows for subtitle creation but requires expert review. Is it the same as downloading the Hugging Face model
dictation.io [1]	envato elements	Drag and Drop in GUI	This tools don't transcript videos
Transkriptor.com	Transkriptor	Drag and Drop in GUI	Only private version is available
SeamlessM4T		API	Computationally expensive
Google Translation	Google	Drag and Drop in GUI	It only transcribes audio using the microphone and does not support uploading video or voice files. Therefore, it cannot be used for video transcription.

Figure 4 - Pros and cons of the several transcriptions tools.

4.1.2. NeMo NVIDIA

In the realm of ASR, the adage “speed versus accuracy” resonates profoundly. While the first model, mentioned in the section 3.2.2, offers swiftness, it may compromise precision. Conversely, the third model, mentioned in the section 3.2.2, though computationally intensive, stands as a testament to the pursuit of excellence, delivering unparalleled accuracy, even at the expense of computational resources. In this intricate dance between efficiency and efficacy, the choice ultimately rests on the task’s exigencies and the pursuit of transcription perfection.

It is worth noting that while these models produce outputs in plain text or JSON format, they do not inherently support the generation of subtitle file formats like VTT or SRT, rendering them unsuitable for our intended task.

4.1.3. YouTube

For videos recorded in Spanish. The presence of a heavy South American accent, along with low vocalization and tone, significantly affected Whisper’s transcription accuracy. To address this, we explored alternative methods, including YouTube’s automatic subtitle generator for Spanish-language videos. By selecting the appropriate Spanish variant matching the speaker’s accent, we achieved improved transcription accuracy compared to Whisper. Obtaining accurate Spanish subtitles was crucial as an intermediary step in transcribing videos into Catalan later, enhancing efficiency and precision in producing high-quality translations.

4.2. Translation

Testing all presented tools on the random sample elicits pros and cons (Figure 5)

Figure 5 - Pros and cons of the several translation tools.

Software	Distributor	Interface	Translation
NNLB	Meta	API	Translation quality impacted by limited integration and representation of Catalan language, resulting in inaccuracies.
Aina	BSC	API	Near-perfect translation accuracy for Catalan to Spanish, Catalan to English and Spanish to Catalan tasks, with a high automatic precision.
Google translate	Google	Drag and Drop in GUI	This tool allows you to upload a file and translate it. However, the uploading and downloading of files is not automatic, so a person is needed to perform this function. Furthermore, translations are not usually accurate and require human oversight for understanding.
happyscribe	Happy Scribe Ltd	Drag and Drop in GUI	The tool also allows for translation, but with a free account, only 30 minutes of video are permitted.
Youtube	Google	Drag and Drop in GUI	It allows for both transcription and translation processes to be carried out simultaneously. However, translations are often inaccurate because it fails to link phrases properly, lacking punctuation marks, verb conjugation, etc.

4.2.1. NLLB Meta

Despite its formidable capabilities in translating between many language pairs, our evaluation revealed notable challenges when applying to translate Catalan text. Catalan's limited digital integration and unique expressions present specific translation challenges. The model's performance is hindered by the scarcity of high-quality training data and linguistic resources, landing to inaccuracies. Consequently, the model cannot properly capture the nuances of Catalan language and culture, giving suboptimal translation.

4.2.2. AINA

We used AINA for translation tasks from Catalan to Spanish and Catalan to English. Notably, for videos recorded in Spanish, AINA also offers a model for translating from Spanish to Catalan, enhancing its versatility and utility. Our experience with AINA for translation tasks has been exceptionally positive. The translations produced by AINA are nearly perfect, achieving very high automatic accuracy rates. Remarkably, the quality of the translations is such that revision or manual intervention is often unnecessary.

5. Challenge and Future of Automatic Transcription of Videos in Catalan

This paper evaluates the strengths and limitations of many tools for voice transcription and translation in Catalan, specifically in the educational context of an AI training course. While a mostly automated pipeline was developed, significant manual intervention and human review of text is still required. Major corporations still do not support Catalan for subtitle generation and high-quality tools for Catalan, such as AINA, have only recently become available. Few days before this submission, AINA opened their transcription specialized in Catalan, which will be tested by future lines. In general, challenges still persist due to Catalan's dialectal variability and lexical richness. Improving the integration of NLP with speech recognition and audio signal processing is crucial for better transcription accuracy. Although Whisper is the best transcription tool, it still requires human review and does not perform well in Catalan. Similarly, text translation improves transcription but is not yet robust enough to be fully automated. Our analysis shows that AINA's text-to-text translation from Catalan to Spanish and English outperforms voice-to-text translation. In summary, while there has been significant progress in the automatic transcription and translation of videos in Catalan, many challenges remain. Human validation is essential to ensure quality.

Acknowledgements This research has been partially financed by the SGR2021-01532 funds from AGAUR, the predoctoral fellowship 2023-FISDU-00366 and Top Rosies Talent Project.

References

[1] [Agarwal 2024] Agarwal, A. (2024). Voice Dictation - Online Speech Recognition. https://dictation.io/
[2] [Anguera 2014] Anguera, X., et al (s/f). Audio-to-text alignment for speech recognition with very limited resources. https://xavieranguera.com/papers/IS2014_phonealignment.pdf
[3] [Armengol-Estapé 2021] Armengol-Estapé, J., et al. (2021). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. ACL-IJCNLP 2021 (pp. 4933–4946). Stroudsburg, PA, USA: Association for Computational Linguistics.
[4] [Bassa 2023] Bassa, M. (2023, julio 4). Almost 45% of Catalan speakers in Catalonia do not use Catalan to search on sites such as Google and YouTube. Fundació .cat
[5] [Bonafonte 1997] Bonafonte Cávez, A., et al (1997). A billingual texto-to-speech system in spanish and catalan. Procs. of EUROSPEECH '97, 2455–2458. WCL, University of Patras, Grece.
[6] [BSC] BSC. (s.d). oTranscribe. https://otranscribe.bsc.es/
[7] [BSC 2024] BSC. (2024, mayo 23). https://www.bsc.es/es/noticias/noticias-del-bsc/aina-impulsa-la-primera-solución-de-voz-que-incorpora-las-diferentes-variantes-del-catalán
[8] [BSC 2024b] BSC. (s/f). Projecte-aina (Projecte Aina). Recuperado el 13 de mayo de 2024, de Aina website: http://ttps://huggingface.co/projecte-aina
[9] [Devlin 2018] Devlin, J., et al (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. Recuperado de http://arxiv.org/abs/1810.04805
[10] [Elvira-García 2016] Elvira-García, W., et al(2016). A tool for automatic transcription of intonation: Eti_ToBI a ToBI transcriber for Spanish and Catalan. Language Res. and Evaluation, 50(4), 767–792.
[11] [Fàbregas 2021] Fàbregas, L (2021, noviembre 2019) El Govern recurre a expertos turcos para que Alexa y Siri hablen en catalán. The Objective Media, 19.11.2021
[12] [Generalitat 2022] Generalitat. (2022). Món digital i tecnologies de la llengua. llengua.gencat.cat/web/.content/temes/pacte-nacional-per-la-llengua/resums-tematics/6-digital.pdf
[13] [Gulati 2020] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition (No. arXiv:2005.08100). arXiv. https://doi.org/10.48550/arXiv.2005.08100
[14] [Happy Scribe] Scribe, H. (s/f). Happy scribe: Audio transcription & video subtitles. Recuperado el 14 de mayo de 2024, de Happy Scribe website: https://www.happyscribe.com/
[15] [IBM 1961] IBM. (s/f). (1961) Shoebox - IBM Archives (78-013). Recuperado de https://mediacenter.ibm.com/media/(1961)+Shoebox+-+IBM+Archives+(78-013)/0_4m2ynnkk
[16] [Kjartansson 2020] Kjartansson, O., et al (s/f). Open-Source High Quality Speech Datasets for Basque, Catalan, and Galician. Aclanthology.org/2020.sltu-1.3.pdf
[17] [Kriman 2019] Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., & Zhang, Y. (2019). QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions (No. arXiv:1910.10261). arXiv. https://doi.org/10.48550/arXiv.1910.10261
[18] [Külebi 2024] Külebi, B., et al ParlamentParla: A Speech Corpus of Catalan Parliamentary Sessions., de Aclanthology.org website: https://aclanthology.org/2022.parlaclarin-1.18.pdf
[19] [Mallafré 2000] Mallafré, J. (2000). Language Models and Catalan Translation. Selected Papers from the 4th Int'l Congress on Translation, Barcelona, 1998 págs. 141-151, 141–151.
[20] [Meta 2024] Meta. (s/f). Meta AI Research Topic - No Language Left Behind. Recuperado el 13 de mayo de 2024, de Meta.com website: https://ai.meta.com/research/no-language-left-behind/
[21] [NeMo 2024] NeMo - Automatic Speech Recognition. (s/f). Recuperado el 13 de mayo de 2024, de NVIDIA NGC Catalog website: https://catalog.ngc.nvidia.com/orgs/nvidia/collections/nemo_asr
[22] [Padro 2012] Padró, L. et al (2012) FreeLing 3.0: Towards wider multilinguality
[23] [Periodico 2023] Periódico, E. (2023, febrero 15). Amazon anuncia que Alexa seguirá sin hablar catalán: "No hay novedades". El Periódico
[24] [SeamlessM4T 2024] SeamlessM4T. (s/f). Recuperado el 14 de mayo de 2024, de Huggingface.co website: https://huggingface.co/docs/transformers/en/model_doc/seamless_m4t
[25] [Sonix 2024] Sonix. (s/f). A brief history of speech recognition. sonix.ai/history-of-speech-recognition
[26] [Softcatalà 2015] Softcatalà. (2015, agosto 20).: http://softcatala.org/
[27] [Transkriptor 2021] Transkriptor: Convert audio or video to text [Transcription]. (2021, abril 30). Recuperado el 14 de mayo de 2024, de Transkriptor! website: https://transkriptor.com/
[28] [Vázquez 2020] Vázquez, R., et al (2020, mayo). A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation. Aclanthology.org /2020.cl-2.5
[29] [Williams 2024] Williams, C. (2024). Introduction | Buzz. Buzz https://chidiwilliams.github.io/buzz/docs
[30] [Whisper] Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. (s/f).
[31] [Youtube 2024] YouTube. (s/f). Recuperado el 13 de mayo de 2024, de https://www.youtube.com