# Automatic Speech Recognition for Gascon and Languedocian Variants of Occitan

**Iñigo Morcillo, Igor Leturia, Ander Corral, Xabier Sarasola, Michäel Barret, Aure Séguier, Benaset Dazéas**

Orai NLP Technologies,
Usurbil, Basque Country, Spain
{i.morcillo, i.leturia, a.corral, x.sarasola}@orai.eus

Lo Congrès Permanent de la Lenga Occitana
Pau, Occitania, France
{m.barret, a.seguier, b.dazeas}@locongres.org

**Abstract**

This paper describes different approaches for developing, for the first time, an automatic speech recognition system for two of the main dialects of Occitan, namely Gascon and Languedocian, and the results obtained in them. The difficulty of the task lies in the fact that Occitan is a less-resourced language. Although a great effort has been made to collect or create corpora of each variant (transcribed speech recordings for the acoustic models and two text corpora for the language models), the sizes of the corpora obtained are far from those of successful systems reported in the literature, and thus we have tested different techniques to compensate for the lack of resources. We have developed classical systems using Kaldi, creating an acoustic model for each variant and also creating language models from the collected corpora and from machine translated texts. We have also tried fine-tuning a Whisper model with our speech corpora. We report word error rates of 20.86 for Gascon and 13.52 for Languedocian with the Kaldi systems and 16.37 for Gascon and 11.74 for Languedocian with Whisper.

**Keywords:** Automatic Speech Recognition, Occitan, Gascon, Languedocian, less-resourced languages

## 1. Introduction

### 1.1. The Occitan Language and its Dialectal Variety

Occitan is a Romance language spoken over a 180,000 km² area in France, Spain and Italy. It is a minority language used by several hundreds of thousands people (OPLO, 2020) and it has a low social status (despite having a prestigious literature from the time of Troubadours to the present day). It has little or no public recognition: regardless of being co-official along with Catalan and Spanish in the Val d'Aran (Catalonia, Spain), in Italy it only appears in the law for linguistic minorities protection and in France it receives financial support from some territorial communities and benefits from a new legal framework[1] for its transmission and spreading, but it remains unofficial.

Occitan language has no officially acknowledged and socially accepted standard. Conventionally, linguists divide it in six major dialects (Bec, 1986; Quint, 2014): Auvergnat, Languedocian, Limousin, Gascon (which has morphological and linguistic specific features because of an Aquitanic substrate it shares with Basque), Provençal and Vivaro-Alpine. Although several different spellings have been used in the course of its history, the spelling known as "classical" has become, in the past decades, the one used in the fields of teaching and communication. We can also note the existence of Mistralian spelling still being used, in addition to classical spelling, in the Provençal space. Even within a dialectal area, we can observe strong intradialectal variety (morphosyntactic as well as lexical and phonetic). For example, taking into account the intradialectal and interdialectal variations, the word 'braveja' (*he brags*) can be realised in any of the following ways in Occitan: /bra/'ved/zɔ; /bra/'βe/ʒɔ; /bra/'ve/dʒɔ; /bra/'βe/dʒɔ; /bra/'βe/tsɔ; /bʀa/'ve/dʒɔ; /bʀa/'ved/za; /bra/'βe/jɔ; /bra/'βe/jœ; /bra/'βœ/ʒœ; /bra/'βœ/jœ; /bra/'wœ/ʒœ; /bra/'wœ/jœ.

[1] Law of 21st May 2021 concerning patrimonial protection and promotion of regional languages.

### 1.2. Need for an Automatic Speech Recognition System for Occitan

It is essential for any language –and especially for minority and less-resourced languages– to develop language and speech technologies. In the context of a study for assessing the situation the Occitan language was in terms of language technologies, a 2015-2019 roadmap for the digital development (Gurrutxaga and Leturia, 2014) included various



*LREC-COLING 2024*, pages 1969–1978

NLP tools and resources to be worked on, such as dictionaries, corpora, spellchecking, machine translation and speech synthesis. Most of them have already been developed, and Occitan now has text corpora such as Bateloc (Bras and Vergez-Couret, 2008) or Lo Congrès' corpus; online dictionaries or lexica such as dicod'Òc, tèrm'Òc and Loflòc (Bras et al., 2017); a verb conjugator (vèrb'Òc); a PoS tagger (Urieli, 2013); a spell checker; a predictive keyboard (Congrès, 2018a,b); a machine translation system on the Apertium platform (Apertium, 2016) and text-to-speech (TTS) technology (Corral et al., 2020).

Regardless, the roadmap did not include an Automatic Speech Recognition (ASR) system due to its unfeasibility in the mentioned period of time and because of the higher priority of other more basic tools, but it did mention the compilation of the resources needed to build an ASR system. In this paper we describe how we set up such a system.

## 2. Related work

### 2.1. State of the art of ASR

Although until recently speech recognition was achieved through technologies such as Hidden Markov Models or HMM, today, just as most speech and language technologies, it is done through neural networks. In semi-classical ASR systems Recurrent Neural Networks (Rumelhart et al., 1986) are mainly used and specifically, those of the Long Short-Term Memory (LSTM) type (Hochreiter and Schmidhuber, 1997).

ASR systems have traditionally been divided into three modules: acoustic modelling, language modelling and post-processing of the final result. Nowadays, through the use of neural models it is possible to harness the full capacity of neural networks and develop end-to-end systems without the need to develop each of the modules (Graves and Jaitly, 2014). However, despite the promise of such systems, they do not currently offer the versatility and flexibility that modular systems such as the DNN-HMMs offer (Le et al., 2021; Prabhavalkar et al., 2023). Besides, end-to-end systems usually need much more transcribed audio to work well (Amodei et al., 2015; Fazel et al., 2021; Prabhavalkar et al., 2023).

The result of the ASR system is obtained through the interaction of the three modules. The acoustic model (AM) converts the input audio signal into its corresponding phonemes. The module is based on statistical systems to achieve a first approximation of phonemes and then applies neural architectures to obtain the final result. The language model (LM) is responsible for converting the sequence of phonemes into coherent text by using advanced LMs. Finally, the post-processing module is responsible for processing the text to obtain the final transcription applying punctuation, capitalization and extraction of numbers, dates, hours, acronyms, etc.

There are different software programs that allow both training and inference of speech recognition through neural networks, such as Sphinx, Kaldi (Povey et al., 2011) or DeepSpeech (Hannun et al., 2014), and also some pre-trained models that can be fine-tuned for new languages, such as Whisper (Radford et al., 2023). Kaldi is a modular system (acoustic modelling, language modelling, etc.), whereas DeepSpeech and Whisper are end-to-end.

### 2.2. ASR for less-resourced languages

Under-resourced languages are deemed challenging when creating ASR systems due to the scarce annotated data those languages have available. Consequently, various approaches are adopted for alleviating the scarcity problem. One such approach is cross-lingual transfer. For example, Khare et al. (2021) and Hou et al. (2021) have taken advantage of cross-lingual transfer learning for adapting data of a high resource language into a low resource target language, reporting considerable word error rate (WER) reductions. One can also jointly train data and create multilingual systems accompanied by both monolingual and multilingual LMs (Tan et al., 2014; Abate et al., 2020).

Another solution to the limited data problem is data augmentation. For instance, TTS systems have been successfully employed so as to augment the transcribed audio corpus (Rygaard, 2015; Ramazan and Yalçın, 2019; Huang et al., 2023; Bartelds et al., 2023).

With the aforementioned strategies it is possible to create more robust AMs for under-resourced languages. LMs can also benefit from augmentation techniques. Since LMs are derived from written text, some approaches focus on expanding the text corpus via machine translation (MT) before creating the LM (Nakajima et al., 2002; Jensson et al., 2008; Cucu et al., 2011; Punjabi et al., 2019). Others rely on LM adaptation (Nakajima et al., 2002; Jensson et al., 2009; Yılmaz et al., 2018) for better fitting the domain the ASR will be operating upon.

## 3. Resources and tools for an Occitan ASR

In order to build the desired transcription systems an annotated speech corpus is essential. We considered each of the Occitan variants as a separate language so two corpora are actually required. We also need text corpora for both variants for the pur-

pose of creating the LMs the Kaldi systems will be using. Moreover, as we contemplated the possibility of gathering an insufficient amount of text, we decided we would employ an MT system for augmenting the written corpus. Lastly, Kaldi leverages a pronunciation lexicon where each word of the corpus is mapped to its corresponding phonetic spelling. We will consequently require a tool that will normalise and phonetically spell the words. As for the Whisper systems, the transcribed speech corpora are enough to fine-tune the models. Corpora whose rights are not restricted are available for download and those whose rights holders have not allowed diffusion but who present a linguistic interest will be integrated in publicly available tools.

### 3.1. Transcribed Speech Corpora of Occitan

As we had no resources of our own, we called producers of audiovisual and written corpora in Occitan, with whom partnership agreements were drawn up to establish the conditions of use and respect for copyright. The quantity of the received material was unsatisfactory. We therefore decided to build an in-house online contribution platform, called ReVOc[2], as Mozilla's Common Voice option did not allow us to manage the internal dialectal variety of the language. On the platform, users could write themselves the sentences they wanted to record or generate random ones and, if it was necessary, correct them to adapt them to their dialectal varieties. The random sentences came from a small corpus gathered at the beginning of the project. We chose sentences from written press and books with an accessible language, from which we excluded sentences too short or too long according to the average length of all sentences. The overview of the participation data is displayed in Table 1. Apart from the online individual contributions, 10 speakers were recruited to contribute 5 hours each and thus obtain the quantity and diversity we were lacking (CNRS, 2022)[3].

The nature of the audio files is predominantly read speech, such as the sentences read from our online speech contribution platform and audio books, but the corpus also includes copyright-free resources, particularly Lingua Libre by Wikimédia (2016), which is a French participatory oral library, transcriptions of various videos (amateur and professional), transcriptions of radio broadcasts, audio podcasts, recordings made to train speech synthesis and a few collections. There was greater diversity in the nature and intra-dialectal variety of the audio for Gascon than for Languedocian.

As a result, we were able to collect a total of 126 hours of transcribed data for Gascon Occitan and 112 hours for Languedocian Occitan. The Gascon speech corpus is made up of nearly 91k utterances (about 960k words), whereas the Languedocian corpus has over 77k utterances (about 825k words).

### 3.2. Text Corpora of Occitan

Lo Congrès already had a digital textual corpus (Séguier Aure (2023a), Séguier Aure (2023b)), but it was far too small for the needs of developing speech recognition. Thanks to the partnership agreements established to create a large-scale shared corpus, we were able to obtain a total of about 2.6 million words for Gascon Occitan, and about 6.3 million for Languedocian Occitan. The vast majority of these are written press (traditional, but especially online) and literature, although there are also websites, a few blogs, various chronicles by individual authors and Wikipedia articles in Gascon. The very large difference in size between the two variants can be explained by the massive quantity of articles from the newspaper Jornalet[4] alone, representing 3.8 million words in Languedocian, with no Gascon equivalent. With the addition of the transcripts of the speech corpora, a total of 3.5 million words for Gascon and 7 million words for Languedocian were collected.

Even so, these corpora were far below our expectations in terms of size, which is why we decided to augment both corpora by using Revirada[5], an automatic translation system for Occitan and French.

### 3.3. Revirada, Machine Translation for Occitan

The Revirada Occitan-French and French-Occitan machine translator (for Occitan Gascon and Occitan Languedocian) was developed as part of the Poctefa Linguatec project (Linguatec[6]). It was built by *Lo Congrès* and *Elhuyar Foundation* based on the free and open source translator Apertium (Forcada et al., 2011), which is based on rules and lexicons. Many poorly endowed languages do not have the resources needed to train a machine translator based on machine learning which require large quantities of aligned parallel corpora (Forcada, 2006). This is the case for Occitan, even more so when considering each dialect as a separate language. The rule-based model also made it

[2] https://contribuir.locongres.com/revoc/

[3] CNRS – Service audiovisuel d'ARDIS (UAR2259). (2022, 7 octobre). Constitution d'un corpus TAL occitan : états des lieux et perspectives. [Vidéo]. Canal-U. https://www.canal-u.tv/135876.

[4] Jornalet. https://jornalet.com

[5] Revirada. https://revirada.eu.

[6] Linguatec. https://linguatec-poctefa.eu.

| Variant | #recordings | #hours | #users | Gender split | | Recordings by age | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Men | Women | age $\leq$ 15 | 15 $<$ age $\leq$ 35 | 35 $<$ age $\leq$ 60 | age $>$ 60 |
| Gascon | 23.5k | 51h | 364 | 48% | 52% | 1.4% | 24.0% | 64.5% | 10.1% |
| Lengadocian | 15.7k | 37h | | 62% | 38% | 0.4% | 44.4% | 43.2% | 12.0% |

Table 1: Participation data at the online recording platform ReVOc for obtaining part of the annotated speech corpus.

| Translation direction | Comprehensibility (%) | Correctness (%) | Faithfulness (%) |
|---|---|---|---|
| fr–oc$_{\text{Lang.}}$ | 94 | 91 | 94 |
| fr–oc$_{\text{Gasc.}}$ | 96 | 93 | 96 |
| oc$_{\text{Lang.}}$–fr | 78 | 68 | 79 |
| oc$_{\text{Gasc.}}$–fr | 71 | 63 | 73 |

Table 2: Evaluation results for the Revirada machine translator. The comprehensibility, correctness (how grammatically and stylistically correct the sentence is) and faithfulness are subjective indicators given by human evaluators.

possible to handle the intra-dialectal variety of Occitan. A great deal of work was done in the Occitan-French direction so that the translator would take into account all local variants of a word. Likewise, a lot of effort was put into the French-Occitan direction for producing a geographically coherent language.

Lo Congrès first enriched the Apertium translator's French-Occitan pair with royalty-free lexicons, grammatical disambiguation rules and lexical disambiguation rules. A private version of the translator was then created, to which copyrighted lexicons were added that only Lo Congrès has the right to use, but there is a public version as well (hectoralos et al. (2022)). An evaluation was carried out to assess the quality of the translations produced by Revirada. A web interface presented pairs of sentences translated by the machine translator and by a human and the evaluators were asked to rate the MT translation's comprehensibility and the grammatical and stylistic correctness. Then, they evaluated the faithfulness to the original French text. Table 2 shows the results obtained in the evaluation.

As previously mentioned, the Occitan corpora we gathered were considered insufficient. On account of this, we chose to augment the linguistic corpus with translated French journalistic texts as a means of expanding the lexicon and allowing the LM to hold more linguistic information. The sources are listed in Table 3, along with their sizes. The toponymic information contained in these texts should allow the ASR to recognise more region-specific words and names, although the focus remains on obtaining a generalist and versatile system. We discarded sentences that contain words not included in Revirada's dictionary and verbs that were not properly conjugated (which the system highlights) so as to produce as many lexically and grammatically correct sentences as possible. In this way, we obtained a corpus of over 500M words for each variant: around 521M words for Gascon and almost 503M for Languedocian.

### 3.4. Normaliser and Phonetiser

In order to build the neural TTS for Occitan (Corral et al., 2020), a tool that transforms a written Occitan text into its phonetised form (in the international phonetic alphabet), syllabified and accented, had been developed (Lo Congrès (2021)).This phonetiser is an algorithm based on a system of rules. The rules are based on the *Tableaux de relations graphie-phonie* by Romieu et al. (2014). The algorithm also includes a database with over 300,000 exceptions for words that do not follow the usual phonetisation rules, including foreign words and proper nouns.

This phonetiser is complemented by a normalisation algorithm. It spells out cardinal and ordinal numbers, acronyms, abbreviations, symbols, letters, dates, times, units of measurement, currencies, urls, e-mails, etc., so that they can be correctly phonetised.

We further improved this tool and used it to create the pronunciation dictionary, also known as the lexicon, that the ASR systems will be using. The phonetiser takes into consideration just one possible phonetic representation of a word for each dialect, so no intra-dialectal variety is introduced into the lexicon unless the text fed is written in an intra-dialectal variant. This will not be our case because it was settled that Occitan needed its standard spelling system and hence the transcripts would be written in standardised form. This simplifies the tool and its management. Nevertheless, we are aware that it could be detrimental to the system's performance, since we will lose intra-dialectal forms of pronouncing a word.

## 4. Experimental Setup

We carried out several experiments so as to assess the performance of the systems by measuring the WER. We trained AMs for each Occitan dialect and

| Newspaper | URL | Sentences | Words |
|---|---|---|---|
| *La République des Pyrénées* | https://www.larepubliquedespyrenees.fr | 1.96M | 35.75M |
| *Sud Ouest* | https://www.sudouest.fr | 8.87M | 171.81M |
| *La Dépêche* | https://www.ladepeche.fr | 50.81M | 1,026.77M |

Table 3: List of French sources used for the Occitan MT corpora. All three sources are generalist newspapers. Only a selection of correctly translated sentences are included in the MT corpora.

four LMs with Kaldi. LMs derived from the digital Occitan written corpora have been labelled natural models and those created by expanding the corpus with the MT corpus, synthetic models. We consider the natural models to be our baselines. Table 4 summarises the characteristics of each corpus. The natural LMs only use the natural corpora and the synthetic ones the sum of the natural and MT corpora. We also fine-tuned a Whisper model for both Occitan variants and compared its results against the classical Kaldi systems.

| Corpus | Hours | Sentences | Words |
|---|---|---|---|
| *Gascon* | | | |
| Speech | 126.6h | 91k | 960k |
| Nat. text | – | 289k | 3.5M |
| MT text | – | 25.3M | 521.4M |
| *Languedocian* | | | |
| Speech | 112.3h | 77k | 825k |
| Nat. text | – | 353k | 7M |
| MT text | – | 25.3M | 502.9M |

Table 4: Size of the Occitan corpus for each variant. The natural texts are a collection of texts originally written in its corresponding Occitan variant, whereas the MT texts are translated from a compilation of news and articles from French newspapers.

### 4.1. Kaldi Systems

We used a TDNN-LSTM system with a HMM-GMM based model. Following the usual pipeline, we extracted the MFCC acoustic features and employed the cepstral mean and variance normalization (CMVN) before performing Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) followed by feature space Maximum Likelihood Linear Regression (fMLLR). 100-dimensional i-vectors were used for the speaker adaptation part.

Some data augmentation techniques were employed, such as a 3-way speech velocity perturbation with 0.9 and 1.1 rates added, volume perturbation and audio distortion by means of introducing various out-of-speech signals and artefacts, i.e., music, background noise, babble and reverberation, which were obtained from the MUSAN corpus (Snyder et al. (2015)) and the Room Impulse Response and Noise Database (Ko et al. (2017)). Lastly, some telephone codecs (OPUS, AMR NB, GSM Full Rate, g722, g726, g723.1 and g711) were introduced. This should help reduce the negative effect of audio imperfections, out-of-speech segments and artefacts that real life audios tend to have.

The LMs were created using Kaldi's SRILM toolkit and the models are trigram based ones.

As previously mentioned, due to the intra-dialectal variety of Occitan (specially in Gascon), we expect the phonemic mapping of the lexicon not to always match the pronunciation given by every speaker. The fact that a standard pronunciation was used to simplify the phonetiser will likely evince that. Thus, some limitation is to be expected in the pronunciation lexicon and the AM's phone discrimination capacity, at least with the amount of speech data available to us. Besides, the homophonic nature of Occitan can accentuate the phonemic mapping problem by inserting statistical uncertainty (7.2% and 7.6% of the natural lexicon and 13% and 15% of the hybrid natural-synthetic lexicon, for Gascon and Languedocian, respectively, are homophones).

### 4.2. Whisper Systems

Whisper is an end-to-end robust multilingual ASR system trained on a vast amount of weakly supervised data. Compared to previous end-to-end models such as Wav2Vec (Schneider et al., 2019), it leverages significantly more data, specifically an order of magnitude more for English, and it is trained in a weakly supervised fashion in contrast to the fully unsupervised pre-training of Wav2Vec. It recognizes 96 languages off-the-shelf and via multitask learning it can also translate to English all those languages in an end-to-end manner. Since its release in 2022, it has pushed the ASR field a step further obtaining zero-shot state-of-the-art results in many standard benchmarks without any adaptation. Nevertheless, the authors claim that fine-tuning Whisper models to specific domains or data could further improve the results. Thus, we fine-tuned a Whisper model by feeding it with the Gascon and Languedocian speech corpora so that we can compare it to Kaldi's performance. Due to

resource constraints, we chose the small Whisper model (12 layers, width of 768, 12 heads and 244M parameters) for the fine-tuning process.

## 5. Results and Discussion

We first evaluated the Kaldi systems and compared the two types of LMs for both dialects. The natural LM (baseline) for Gascon achieved a WER of 20.86% and the synthetic one 23.14%. The natural Languedocian LM (baseline) obtained a 13.64% while the synthetic one slightly improved it with 13.52%. There is an evident gap between the performance of the system for one dialect and the other. We suspect that the main reason we obtained a higher overall WER in Gascon is related to its intra-dialectal variability. We believe that the less consistent pronunciation habits of Gascon speakers render the Gascon acoustic model less effective. The LMs may have also suffered from the same problem, as variants of words are not taken into account because of the standardised writing system that was selected. Finally, the fact that the Gascon natural LM uses a corpus half the size of the Languedocian one may have caused the dialect not to be properly modelled.

In any case, the experiments showed that a decent WER can be obtained with around 100h of speech corpus and a small-sized text corpus. However, the results suggest that a larger synthetic corpus does not improve the WER significantly, it even deteriorates it. We report a 10.9% relative increase for Gascon and a relative decrease of just 0.9% for Languedocian. This phenomenon has been reported in very under-resourced scenarios. Sikasote and Anastasopoulos (2021) have shown that a larger LM has a negative impact on the WER with 1.6% to 1.8% relative increase for the Bemba language when using a very small LM (123k token vs. 5.8M token text corpus). Similarly, Liu et al. (2023) found WER deterioration in some low-resource languages, concluding that larger LMs do not always yield lower WERs. They obtained relative increases of 1.42% to 3.16% (and a 37.73% deterioration in the worst case). Even if our scenario may not be considered a severely under-resourced one anymore, the fact that Gascon is phonetically varied could effectively be compared to an AM created with a smaller speech corpus, explaining why there are more errors in the transcriptions the synthetic system provides, even though it uses a bigger LM. Another hypothesis is that the larger corpus does not provide valuable information to the model because the size of the natural corpus is large enough for it to include enough everyday sentences that are contained in the test audios and hence, it does not make the LM more robust and it does not lead to a better modelling of the dialect. Yet another explanation is that the synthetic corpus produced with the MT method might not be very varied lexically and grammatically, resulting in a LM that favors specific word sequences too much. We also consider the possibility that the Gascon Occitan annotated speech corpus may be somewhat of an inferior quality compared to the Languedocian corpus, once again, probably due to its dialectal variety.

These results point in the direction that at least the Gascon corpus needs further refinement or an expansion in order to produce a superior system. In addition, we leave for future research the impact of an increasing synthetic corpus upon the WER.

On the other hand, Whisper outperforms Kaldi's baseline with an absolute gain of 4.14% (21.5% relative WER reduction) in Gascon and with a 1.9% (13.9% relative reduction) in Languedocian. Whisper takes advantage of a far larger speech corpus and the possibility of fine-tuning the default model so it does not come as a surprise that it outputs superior transcripts than Kaldi. Nonetheless, we observe a comparable performance in Languedocian. Taking into account that Whisper falls behind in Gascon in comparison to Languedocian too, it strengthens the hypothesis that Gascon's variability is, in fact, the cause of the poorer capacity of the ASR systems in this particular dialect. Still, we do not rule out the chance that the annotations of the Gascon speech corpus are poorer.

| Variant | System | WER | ΔWER |
|---|---|---|---|
| Gascon | $LM_{nat}$ (bs) | 20.86 | – |
| | $LM_{synth}$ | 23.14 | 10.9% |
| | Whisper | **16.37** | -21.5% |
| Languedocian | $LM_{nat}$ (bs) | 13.64 | – |
| | $LM_{synth}$ | 13.52 | -0.9% |
| | Whisper | **11.74** | -13.9% |

Table 5: Comparison of word error rates for each of the ASR systems and the relative WER change upon the baselines (bs). The best results are highlighted in bold.

After the quantitative analysis, we carried out a qualitative evaluation so that we can have a clearer view of the problems that need be worked out in future works. We took a new selection of 100 audio sentences in each dialect, in read speech style, from four different speakers, 50% male-female. It was confirmed that the natural model was, if not better in quality, at least equivalent to the synthetic model: we found a 79% success rate in Gascon transcriptions in contrast to a 83% in Languedocian. The synthetic models achieved, respectively, 76% and a 83% marks for Gascon and Languedocian. The results of the qualitative evaluation are

displayed in Table 6.

The results are consistent with what the quantitative evaluation shows. Overall, more errors are produced by the Gascon systems. However, linguistic errors, such as spelling errors and erroneous diacritics, are more frequent in the Languedocian systems. We have detected that some of these faulty transcriptions are rooted in the normalisation algorithms of the normalisation-phonetisation module: since the phonetiser was first developed, normalisation became secondary, so some phonemic approximation tricks were used. Thus, for correctly phonetising some word sequences, a number of letters are changed or eliminated to account for word contractions and the pronunciation of some words depending on the beginning of the next one. This phenomenon was mainly observed in Languedocian, which is why we counted more linguistic errors in spite of producing better overall transcriptions. In the contrary, the Gascon systems produce overall poorer transcriptions because they tend to fail the recognition of a higher number of regular words. This is in line with the explanation that the Gascon AM is weaker. We found out that several proper names and surnames were incorrectly transcribed too. This is related, in part, to different diacritical marks used in the same Occitan and French names, which the AMs cannot discriminate very well. The other contribution to these errors seem to be caused by the subpar corpora. If corrections were made in the normalisation-phonetisation module and we gathered more original texts in each dialect, we would likely be able to overcome some of these problems. Some examples of the ASR outputs are shown in Table 7.

## 6. Conclusions

We present, to our knowledge, the first semi-classical ASR system based on Kaldi and another E2E one based on Whisper for two of the main dialects of the Occitan language: Gascon Occitan and Languedocian Occitan. After an arduous effort, we have been able to gather a substantial amount of annotated speech and written corpora, part of which is publicly available and future agreements will hopefully make a portion of the other part available too. We have created a lengthy synthetic corpus obtained using MT for each variant and the pre-existing tool for phonetising Occitan text has been completed so that it also normalises text. One AM was produced for each dialect and two LMs for each system: a natural one and a synthetic one, the latter being produced with a text corpus augmentation approach based on MT. Promising WER results were reported for both dialects, with Languedocian having an overall better performance. This is probably due to a more robust annotated data and phonetic mapping, as Languedocian is a more consistent dialect compared to Gascon, both phonetically and lexically. Although the evaluation of the synthetic method showed minor improvements in Languedocian, it caused performance deterioration to the Gascon system. A WER of 20.86% was obtained for Gascon and 13.52% for Languedocian with Kaldi. The fine-tuned small Whisper model, which substantially improved Kaldi's results, obtained a WER of 16.37% for Gascon and 11.74% for Languedocian (21.5% and 13.9% relative improvements over the baseline Kaldi systems in Gascon and Languedocian, respectively). This project has set the path for future research, in which we could collect more recordings or we could augment the speech corpus by leveraging a TTS system so as to improve the acoustic models. We could also refine the quality of the MT texts, as well as the normalisation-phonetisation module for a more rigorous phonetic mapping in the pronunciation lexicon, in order to create superior LMs. These prototypes are ready for production and should contribute in reducing the time taken in other transcriptions that will, in turn, help feed the ASR systems with more data for future training processes as much as other speech related systems, such as TTS ones.

## 7. Acknowledgements

This project has been partially funded by the EGTC Euroregion New Aquitaine Euskadi Navarra, the Région Nouvelle-Aquitaine, the Région Occitanie and the Département des Pyrénées-Atlantiques. We thank the institutions, publishers and media that have contributed to the compilation of the corpora, as well as all the volunteers accross Occitania, without whose voice this project would have come to nothing. *Mercés hèra! Mercés plan!*

| Variant | System | Correct words | Total errors | Linguistic errors | Other errors |
|---|---|---|---|---|---|
| Gascon | $LM_{nat}$ | 79% | 252 | 35% | 65% |
| | $LM_{synth}$ | 76% | 289 | 40% | 60% |
| Languedocian | $LM_{nat}$ | 83% | 202 | 85% | 15% |
| | $LM_{synth}$ | 83% | 207 | 82% | 18% |

Table 6: Evaluation of the Occitan speech recognition prototypes, in percentage compared to a human transcription and nature of the errors compared to their total.

| | |
|---|---|
| Gascon | |
| Reference | qu'ei la darrèra annada abans de qui prenossi la retirada que la m'aparí de l'aver en classa |
| Hypothesis | que la darrèra annada abans de qui prenossi la retirada e qui l'a aparelh ne l'avem classa |
| English | *it's the last year before my retirement that there have been times that I had her in my class* |
| Reference | un dia maria qu'anà tau camp portar lo disnar au pair |
| Hypothesis | un dia maria qu'an atau camp portar lo disnar au pair |
| English | *one day maria went to the camp to bring lunch to her father* |
| Languedocian | |
| Reference | a son entorn dins l'ombra de punts fosforescents coma d'uòlhs se desplaçan rapidament |
| Hypothesis | a son entorn dins l'ombra de pus fosforescents podiái s se desplaçan rapidament |
| English | *around her in the shadows phosphorescent spots like eyes move quickly* |
| Reference | pauc a cha pauc sos uòlhs s'acostuman a l'escuresina |
| Hypothesis | pauc a chad pau sos uòlhs s'acostuman a l'escuresina |
| English | *little by little his eyes got used to the darkness* |

Table 7: ASR output examples for Gascon and Languedocian.

## 8. Bibliographical References


Solomon Teferra Abate, Martha Yifiru Tachbelie, and Tanja Schultz. 2020. Multilingual acoustic and language modeling for ethio-semitic languages. In *Interspeech*, pages 1047--1051.

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2015. Deep speech 2: End-to-end speech recognition in english and mandarin.

Apertium. 2016. Apertium translation pair for Occitan and French.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *arXiv preprint arXiv:2305.10951*.

P. Bec. 1986. *La langue occitane*. Que sais-je? Presses universitaires de France, Paris, 5th edition.

Myriam Bras and Marianne Vergez-Couret. 2008. BaTelÒc: A Text Base for the Occitan Language. *Language Documentation & Conservation*, Special Publication No. 9(Language Documentation and Conservation in Europe):133--149.

Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benaset Dazéas. 2017. Loflòc, lexic obèrt flechit occitan. In *Proceedings of the XIIn Congrès de l'Associacion internacionala d'estudis occitans*, Albi, France.

Lo Congrès. 2018a. Occitan gascon pack for AnySoftKeyboard.

Lo Congrès. 2018b. Occitan lengadocian pack for AnySoftKeyboard.

Ander Corral, Igor Leturia, Aure Séguier, Michäel Barret, Benaset Dazéas, Philippe Boula de Mareüil, and Nicolas Quint. 2020. Neural text-to-speech synthesis for an under-resourced language in a diglossic environment: the case of Gascon Occitan. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced*

*Languages (CCURL)*, pages 53--60, Marseille, France. European Language Resources association.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of cook islands māori.

Horia Cucu, Laurent Besacier, Corneliu Burileanu, and Andi Buzo. 2011. Enhancing automatic speech recognition for romanian by using machine translated and web-based text corpora. *SPECOM 2011*, pages 81--88.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. Synthasr: Unlocking synthetic data for speech recognition.

Mikel Forcada. 2006. Open-source machine translation: an opportunity for minor languages. 6:1--6.

Mikel Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127--144.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1764--1772, Bejing, China. PMLR.

Antton Gurrutxaga and Igor Leturia. 2014. Diagnostic et feuille de route pour le développement numérique de la langue occitane : 2015-2019. Technical report, Elhuyar Foundation, Media.kom.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735--1780.

Wenxin Hou, Han Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki. 2021. Exploiting adapters for cross-lingual low-resource speech recognition.

Zhuangqun Huang, Gil Keren, Ziran Jiang, Shashank Jain, David Goss-Grubbs, Nelson Cheng, Farnaz Abtahi, Duc Le, David Zhang, Antony D'Avirro, Ethan Campbell-Taylor, Jessie Salas, Irina-Elena Veliche, and Xi Chen. 2023. Text generation with speech synthesis for asr data augmentation.

Arnar Jensson, Koji Iwano, and Sadaoki Furui. 2008. Development of a speech recognition system for icelandic using machine translated text. In *Spoken Languages Technologies for Under-Resourced Languages*.

ArnarThor Jensson, Koji Iwano, and Sadaoki Furui. 2009. Language model adaptation using machine-translated text for resource-deficient languages. *EURASIP Journal on Audio, Speech, and Music Processing*, 2008:1--7.

Shreya Khare, Ashish R Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529--1533.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Michael L. Seltzer. 2021. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion.

Zoey Liu, Justin Spence, and Emily Prudhommeaux. 2023. Studying the impact of language model size for low-resource asr. In *COMPUTEL*.

Hideharu Nakajima, Hirofumi Yamamoto, and Taro Watanabe. 2002. Language model adaptation with additional text generated by machine translation. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

OPLO. 2020. Résultats de l'enquête sociolinguistique relative à la pratique et aux représentations de la langue occitane en nouvelle-aquitaine, en occitanie et au val d'aran. Technical report, Office Public de la Langue Occitane.

Joris Pelemans, Tom Vanallemeersch, Kris Demuynck, Lyan Verwimp, Patrick Wambacq, et al. 2016. Language model adaptation for asr of spoken translations using phrase-based translation models and named entity models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5985--5989. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian,

Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey.

Surabhi Punjabi, Harish Arsikere, and Sri Garimella. 2019. Language model bootstrapping using neural machine translation for conversational speech recognition. *CoRR*, abs/1912.00958.

Nicolas Quint. 2014. *L'occitan*. Assimil.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492--28518. PMLR.

Gökay Ramazan and Hülya Yalçın. 2019. Improving low resource turkish speech recognition with data augmentation and tts.

M. Romieu, J.C. Rixte, B. Molin, F. Marcouyre, V. Rivière, D. Escarpit, and A. Séguier. 2014. Tableaux de relations graphie-phonie. Lo Congrès permanent de la lenga occitana. Online: https://locongres.org/fr/competences/normes/graphie-phonie/prononciation.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533--536.

Luise Valentin Rygaard. 2015. Using synthesized speech to improve speech recognition for lowresource languages. *vol*, 8:2018.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition.

Claytone Sikasote and Antonios Anastasopoulos. 2021. Bembaspeech: A speech recognition corpus for the bemba language.

Tien-Ping Tan, Laurent Besacier, and Benjamin Lecouteux. 2014. Acoustic model merging using acoustic models from multilingual speakers for automatic speech recognition. In *2014 International Conference on Asian Language Processing (IALP)*, pages 42--45. IEEE.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université Toulouse 2 Le Mirail, Toulouse, France.

Emre Yılmaz, Henk van den Heuvel, and David A van Leeuwen. 2018. Acoustic and textual data augmentation for improved asr of code-switching speech. *arXiv preprint arXiv:1807.10945*.

## 9. Language Resource References

### Language Resources

hectoralos and Capsot and unuaiga-congres and ftyers and mr-martian and TinoDidriksen and xavivars and sushain97. 2022. *Apertium-ocifra (1.0.0) [Software]*. https://github.com/locongres/phonetizer-basics.

Ko, Tom and Peddinti, Vijayaditya and Povey, Daniel and Seltzer, Michael L. and Khudanpur, Sanjeev. 2017. *A study on data augmentation of reverberant speech for robust speech recognition*.

Lo Congrès. 2021. *Phonetizer basics*. Lo Congrès permanent de la lenga occitana. https://github.com/locongres/phonetizer-basics.

David Snyder and Guoguo Chen and Daniel Povey. 2015. *MUSAN: A Music, Speech, and Noise Corpus*. ArXiv:1510.08484v1.

Séguier Aure. 2023a. *Occitan Corpus from Lo Congrès news (1.0) [Data set]*. Lo Congrès permanent de la lenga occitana. Zenodo. https://doi.org/10.5281/zenodo.8411197.

Séguier Aure. 2023b. *SoftwaresOccitanTranslations corpus (1.0) [Data set]*. Lo Congrès permanent de la lenga occitana. Zenodo. https://doi.org/10.5281/zenodo.8411351.

Wikimédia. 2016. *Lingua Libre [Data set]*. Wikimédia France. Wikimédia France. https://lingualibre.org/LanguagesGallery.