orai
NLP TEKNOLOGIAK

Automatic Speech Recognition for Gascon and Languedocian Variants of Occitan

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

LREC-COLING 2024

Outline

1. Introduction
2. Motivation
3. Resources for an Occitan ASR
4. Experimental Setup
5. Results
6. Conclusions

orai
NLP TEKNOLOGIAK

Introduction

INTRODUCTION

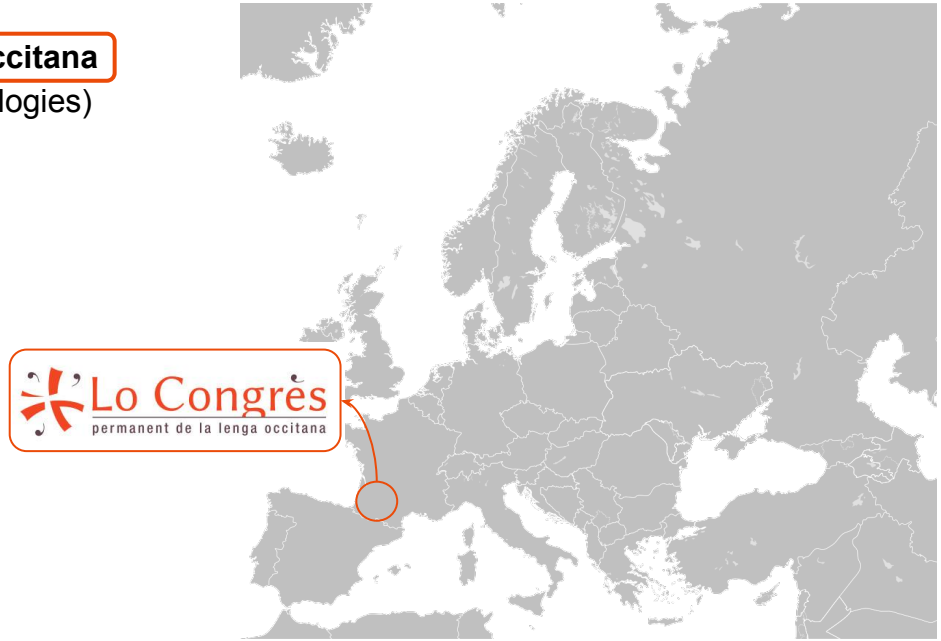
Involved Institutions

- Collaboration between
  - **Lo Congrès permanent de la lenga occitana**
  - **Elhuyar Foundation** (Orai NLP Technologies)

INTRODUCTION

Involved Institutions

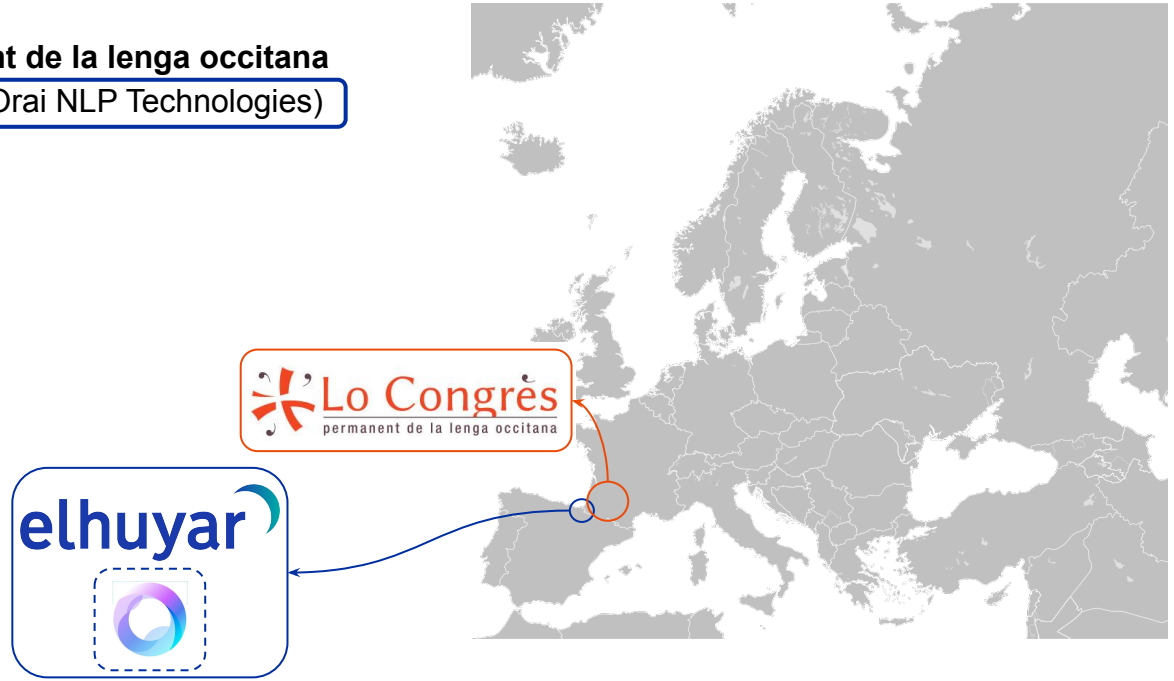
- Collaboration between
  - **Lo Congrès permanent de la lenga occitana**
  - **Elhuyar Foundation** (Orai NLP Technologies)



INTRODUCTION

Involved Institutions

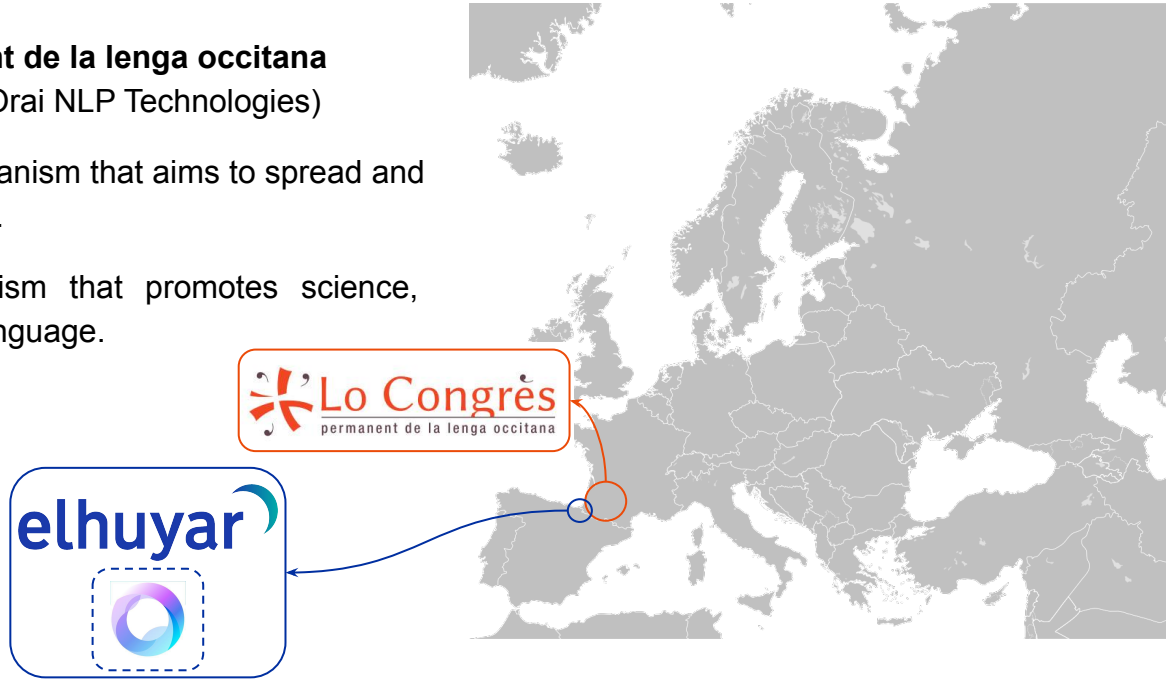
- Collaboration between
  - **Lo Congrès permanent de la lenga occitana**
  - **Elhuyar Foundation** (Orai NLP Technologies)



INTRODUCTION

Involved Institutions

- Collaboration between
  - **Lo Congrès permanent de la lenga occitana**
  - **Elhuyar Foundation** (Orai NLP Technologies)
- **Lo Congrès:** not-for-profit organism that aims to spread and regulate the Occitan language.
- **Elhuyar:** not-for-profit organism that promotes science, technology and the Basque language.



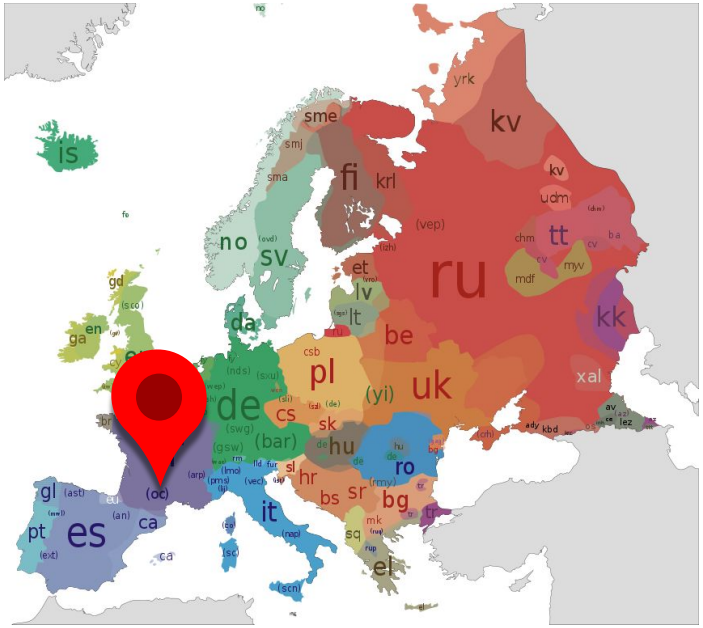
INTRODUCTION

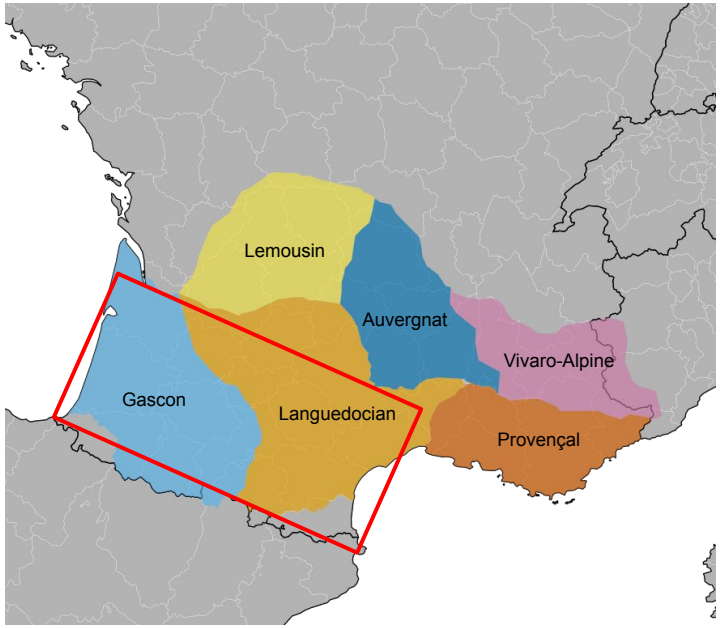
Occitan

- Roman language spoken in France, Spain and Italy by only 0.7%–3.6% of Occitans.
  - 110,000–580,000 people speak it (OPLO, 2020).
- Not official in France nor in Italy, a variant (Aranese) is co-official in a Spanish region (Val d'Aran).
- No official standard form.
- Definitely/Severely endangered (UNESCO).
- Generally divided into 6 dialects (Bec, 1986; Quint, 2014):
  - Auvergnat
  - Gascon
  - Languedocian
  - Limousin
  - Provençal
  - Vivaro-Alpine
- Great variability.

INTRODUCTION

Occitan Variants

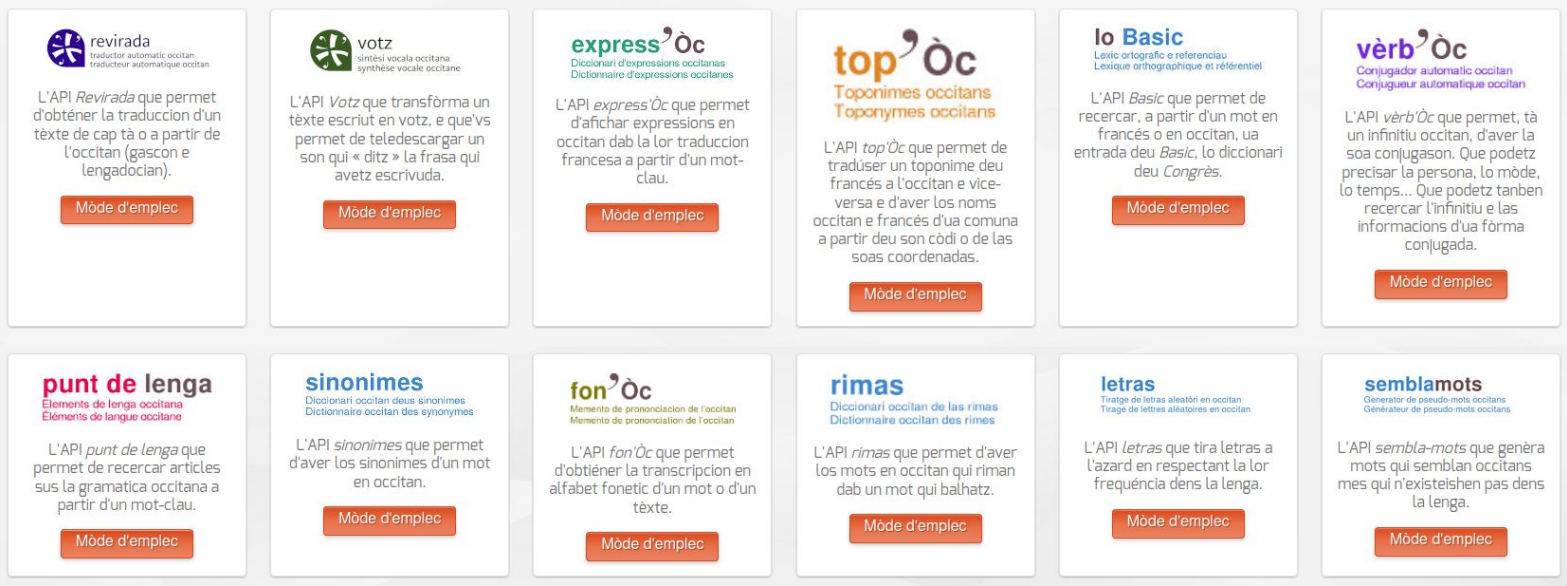




Motivation

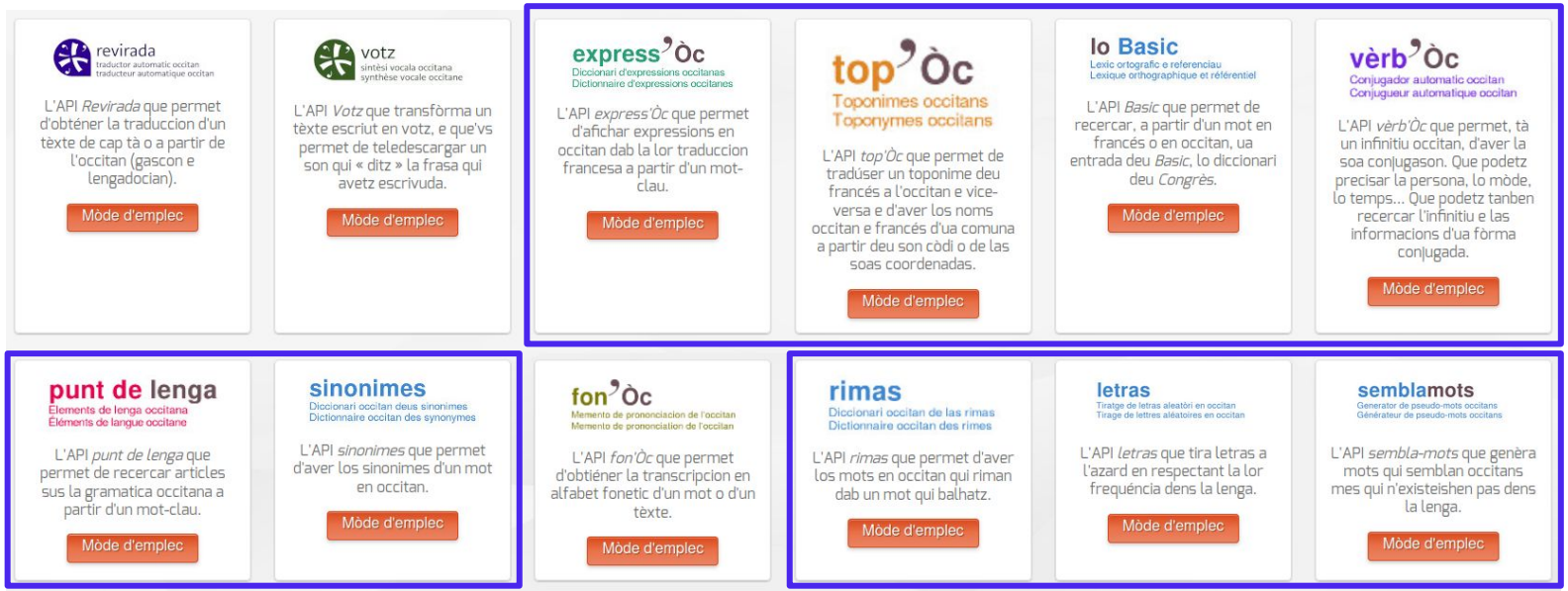
MOTIVATION

Occitan Language Tool Ecosystem



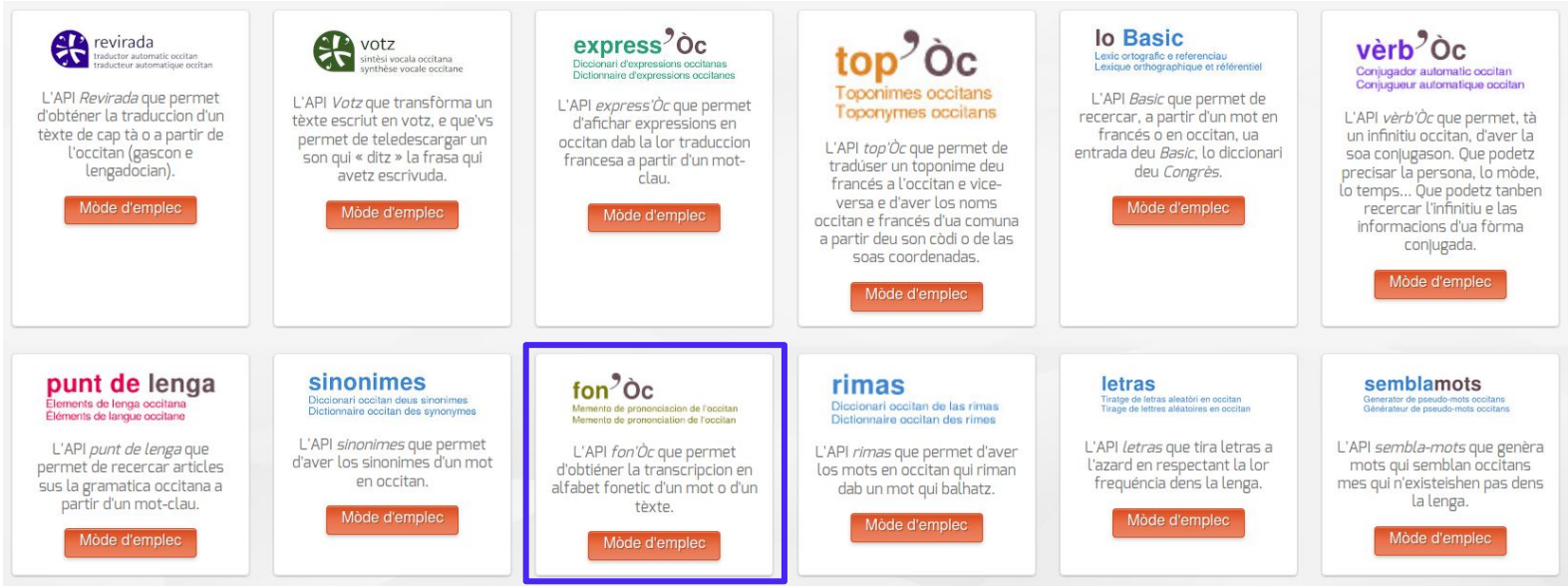
MOTIVATION

Occitan Language Tool Ecosystem: Dictionaries & Corpora



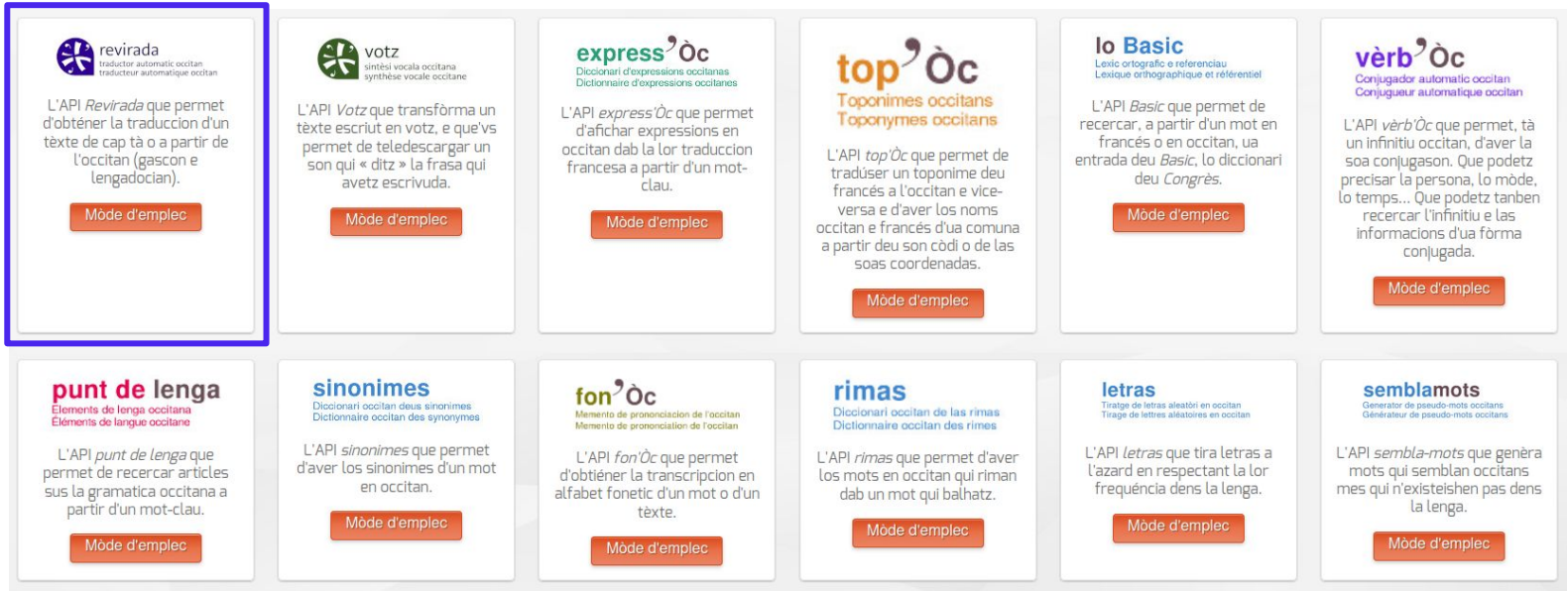
MOTIVATION

Occitan Language Tool Ecosystem: Phonetiser



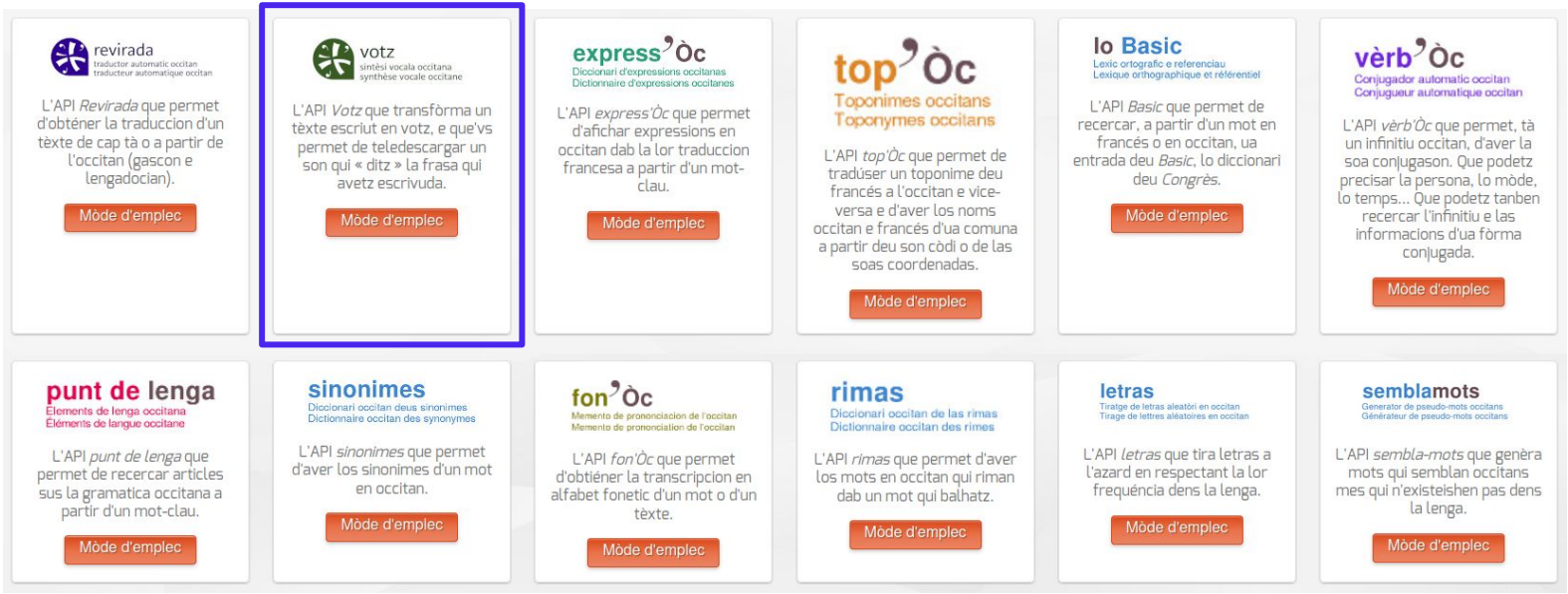
MOTIVATION

Occitan Language Tool Ecosystem: Machine Translation



MOTIVATION

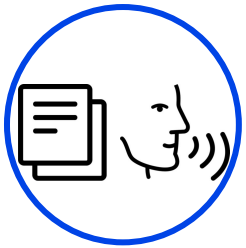
Occitan Language Tool Ecosystem: Text-to-speech



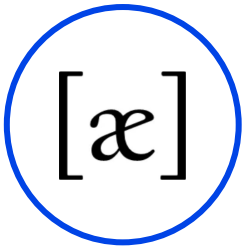
MOTIVATION

Occitan Language Tool Ecosystem

Dictionaries



Text and Speech Corpora



Phonetiser

Verb Conjugator

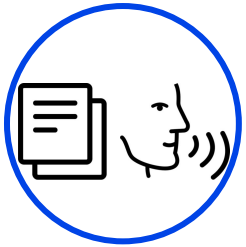
Machine Translation

Text-to-speech

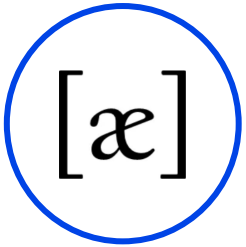
MOTIVATION

Occitan Language Tool Ecosystem

Dictionaries



Text and Speech Corpora



Phonetiser

Corpora far too small

Verb Conjugator

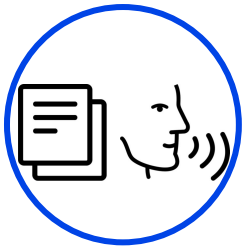
Machine Translation

Text-to-speech

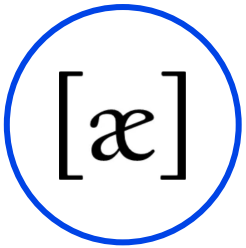
MOTIVATION

Occitan Language Tool Ecosystem

Dictionaries



Text and Speech Corpora



Phonetiser

Verb Conjugator

Machine Translation

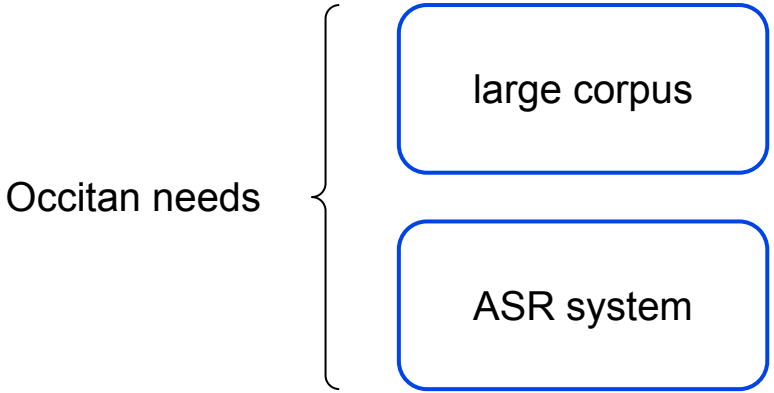
Text-to-speech

Corpora far too small

No ASR system

MOTIVATION

What Occitan needs



Resources for an Occitan ASR

orai
NLP TEKNOLOGIAK

RESOURCES FOR AN OCCITAN ASR

Needed Resources

Annotated speech corpus

Text corpus

Normalisation + phonetisation tool

} ASR system

RESOURCES FOR AN OCCITAN ASR

Obtaining the Speech Corpus

- TV, radio, podcasts, audiobooks, Lingua Libre (Wikimédia, 2016).
  - ~40h each dialect
- 10 people recruited (5h of recordings each).
  - 25h each dialect
- Recordings for TTS (Corral et al., 2020).
  - 7h Gascon
  - 12h Languedocian
- ReVOc (online contribution platform).
  - 51h Gascon
  - 37h Languedocian

RESOURCES FOR AN OCCITAN ASR

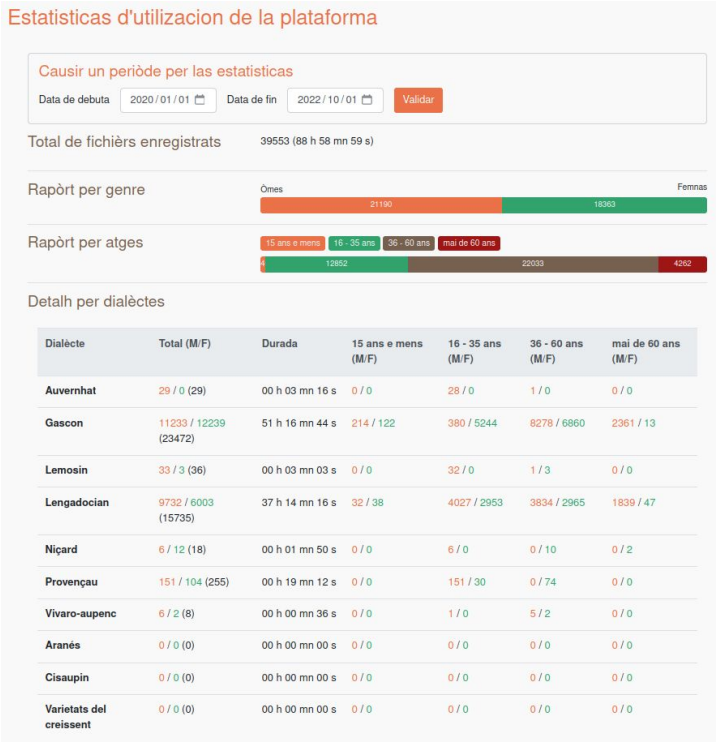
Obtaining the Speech Corpus

- TV, radio, podcasts, audiobooks, Lingua Libre (Wikimédia, 2016).
  - ~40h each dialect
- 10 people recruited (5h of recordings each).
  - 25h each dialect
- Recordings for TTS (Corral et al., 2020).
  - 7h Gascon
  - 12h Languedocian
- ReVOc (online contribution platform).
  - 51h Gascon
  - 37h Languedocian

→ Most of it is copyrighted

RESOURCES FOR AN OCCITAN ASR

ReVOc: Online Speech Contribution Platform

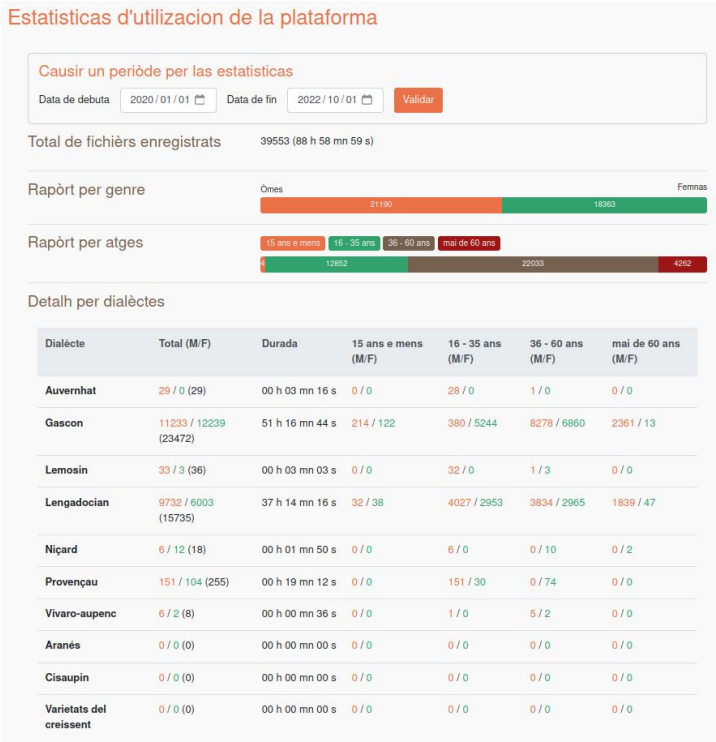


Variant	#recordings	#hours	#users	Gender split		Recordings by age			
				Men	Women	age ≤ 15	15 < age ≤ 35	35 < age ≤ 60	age > 60
Gascon	23.5k	51h	364	48%	52%	1.4%	24.0%	64.5%	10.1%
Languedocian	15.7k	37h		62%	38%	0.4%	44.4%	43.2%	12.0%

Table 1: Participation data at ReVOc (2020/01/01–2022/10/01 period).
Source: https://contribuir.locongres.com/revoc/stats.php

RESOURCES FOR AN OCCITAN ASR

ReVOc: Online Speech Contribution Platform

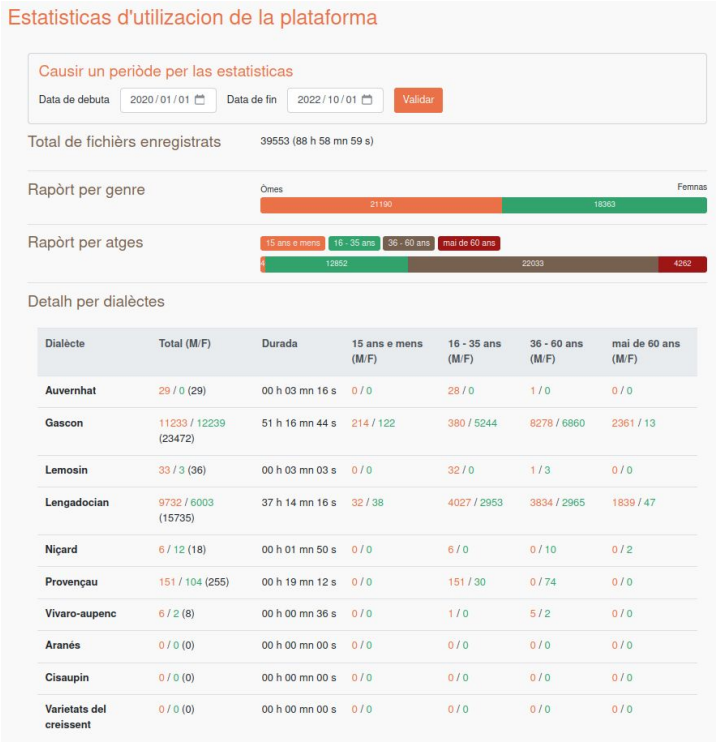


Variant	#recordings	#hours	#users	Gender split		Recordings by age			
				Men	Women	age ≤ 15	15 < age ≤ 35	35 < age ≤ 60	age > 60
Gascon	23.5k	51h	364	48%	52%	1.4%	24.0%	64.5%	10.1%
Languedocian	15.7k	37h		62%	38%	0.4%	44.4%	43.2%	12.0%

Table 1: Participation data at ReVOc (2020/01/01–2022/10/01 period).
Source: https://contribuir.locongres.com/revoc/stats.php

RESOURCES FOR AN OCCITAN ASR

ReVOc: Online Speech Contribution Platform



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

Table 1: Participation data at ReVOc (2020/01/01–2022/10/01 period).
Source: https://contribuir.locongres.com/revoc/stats.php

RESOURCES FOR AN OCCITAN ASR

Obtaining the Text Corpus

- Local newspapers
- Books
- Blogs
- Open source (Séguier, A. 2023a; Séguier, A. 2023b)
- Others

RESOURCES FOR AN OCCITAN ASR

Obtaining the Text Corpus

- Local newspapers
- Books
- Blogs
- Open source (Séguier, A. 2023a; Séguier, A. 2023b)
- Others

→ 3.5M words Gascon

→ 7M words Languedocian

RESOURCES FOR AN OCCITAN ASR

Machine Translation: Augmenting the Text Corpus

- Machine translation used to generate synthetic Occitan text.
- Apertium (Forcada et al., 2011) was used (hectoralos et al., 2022).
- fr–$oc_{Gasc.}$ (521M words)
- fr–$oc_{Lang.}$ (503M words)

Newspaper	URL	Words
La République des Pyrénées	https://www.larepubliquedespyrenees.fr	35.75M
Sud Ouest	https://www.sudouest.fr	171.81M
La Dépêche	https://www.ladepeche.fr	1026.77M

Table 2: 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.

RESOURCES FOR AN OCCITAN ASR

Occitan Corpus

	Corpus	Hours	#sentences	#words
Gascon				
	Speech	126.6h	91k	960k
	Natural text	–	289k	3.5M
	MT text	–	25.3M	521.4M
Languedocian				
	Speech	112.3h	77k	825k
	Natural text	–	353k	7M
	MT text	–	25.3	502.9M

Table 3: Speech and written Occitan corpus for Gascon and Languedocian.

RESOURCES FOR AN OCCITAN ASR

Occitan Corpus

Corpus	Hours	#sentences	#words
Gascon			
Speech	126.6h	91k	960k
Natural text	–	289k	3.5M
MT text	–	25.3M	521.4M
Languedocian			
Speech	112.3h	77k	825k
Natural text	–	353k	7M
MT text	–	25.3	502.9M

Table 3: Speech and written Occitan corpus for Gascon and Languedocian.

RESOURCES FOR AN OCCITAN ASR

Occitan Corpus

	Corpus	Hours	#sentences	#words
Gascon				
	Speech	126.6h	91k	960k
	Natural text	–	289k	3.5M
	MT text	–	25.3M	521.4M
Languedocian				
	Speech	112.3h	77k	825k
	Natural text	–	353k	7M
	MT text	–	25.3	502.9M

Table 3: Speech and written Occitan corpus for Gascon and Languedocian.

RESOURCES FOR AN OCCITAN ASR

Normalisation and Phonetisation

- Phonetisation tool by Lo Congrès (2021).
- Normalisation derived from the phonetisation tool.
  - Some cases not properly dealt with.
  - Problems are expected to arise.

Experimental Setup

EXPERIMENTAL SETUP

ASR Systems

- Kaldi
  - Acoustic Model: HMM-GMM + TDNN-LSTM
  - Language Model: 3-gram
    - LM_{nat} natural text
    - LM_{syn} natural text + MT text
- Whisper
  - Small (12 layers, width of 768, 12 heads, 244M params.)
  - Fine-tuned

EXPERIMENTAL SETUP

Data Augmentation

MUSAN (Snyder et al., 2015) and RIRs (Ko et al., 2017).

- Music (SNR 15:10:8:5)
- Background noise (SNR 15:10:5:0)
- Babble (SNR 20:17:15:13)
- Reverberation (small / medium / large room)
- Codecs (OPUS, AMR NB, GSM Full Rate, g722, g726, g723.1, g711)
- Speed perturbation (0.9 and 1.1)

Results

RESULTS

Results

Variant	System	WER	ΔWER
Gascon	LM_{nat}	20.86	–
	LM_{syn}	23.14	10.9%
	Whisper	**16.37**	-21.5%
Languedocian	LM_{nat}	13.64	–
	LM_{syn}	13.52	-0.9%
	Whisper	**11.74**	-13.9%

Table 4: Comparison of word error rates for each of the ASR systems and the relative WER change.

Variant	Transcription	
Gascon	ref	un dia maria qu'anà tau camp portar lo disnar au pair
	hyp	un dia maria qu'an atau camp portar lo disnar au pair
	eng	*one day maria went to the camp to bring lunch to her father*
Languedocian	ref	pauc a cha pauc sos uòlhs s'acostuman a l'escuresina
	hyp	pauc a chad pau sos uòlhs s'acostuman a l'escuresina
	eng	*little by little his eyes got used to the darkness*

Table 5: ASR output examples for Gascon and Languedocian.

RESULTS

Results

Variant	System	WER	ΔWER
Gascon	LM_{nat}	20.86	–
	LM_{syn}	23.14	10.9%
	Whisper	**16.37**	-21.5%
Languedocian	LM_{nat}	13.64	–
	LM_{syn}	13.52	-0.9%
	Whisper	**11.74**	-13.9%

Table 4: Comparison of word error rates for each of the ASR systems and the relative WER change.

Variant	Transcription	
Gascon	ref	un dia maria qu'anà tau camp portar lo disnar au pair
	hyp	un dia maria qu'an atau camp portar lo disnar au pair
	eng	*one day maria went to the camp to bring lunch to her father*
Languedocian	ref	pauc a cha pauc sos uòlhs s'acostuman a l'escuresina
	hyp	pauc a chad pau sos uòlhs s'acostuman a l'escuresina
	eng	*little by little his eyes got used to the darkness*

Table 5: ASR output examples for Gascon and Languedocian.

RESULTS

Results

Variant	System	WER	ΔWER
Gascon	LM_{nat}	20.86	–
	LM_{syn}	23.14	10.9%
	Whisper	**16.37**	-21.5%
Languedocian	LM_{nat}	13.64	–
	LM_{syn}	13.52	-0.9%
	Whisper	**11.74**	-13.9%

Table 4: Comparison of word error rates for each of the ASR systems and the relative WER change.

Variant	Transcription	
Gascon	ref	un dia maria qu'anà tau camp portar lo disnar au pair
	hyp	un dia maria qu'an atau camp portar lo disnar au pair
	eng	*one day maria went to the camp to bring lunch to her father*
Languedocian	ref	pauc a cha pauc sos uòlhs s'acostuman a l'escuresina
	hyp	pauc a chad pau sos uòlhs s'acostuman a l'escuresina
	eng	*little by little his eyes got used to the darkness*

Table 5: ASR output examples for Gascon and Languedocian.

RESULTS

Results

Variant	System	WER	ΔWER
Gascon	LM_{nat}	20.86	–
	LM_{syn}	23.14	10.9%
	Whisper	**16.37**	-21.5%
Languedocian	LM_{nat}	13.64	–
	LM_{syn}	13.52	-0.9%
	Whisper	**11.74**	-13.9%

Table 4: Comparison of word error rates for each of the ASR systems and the relative WER change.

Variant		Transcription
Gascon	ref	un dia maria qu'anà tau camp portar lo disnar au pair
	hyp	un dia maria qu'an atau camp portar lo disnar au pair
	eng	*one day maria went to the camp to bring lunch to her father*
Languedocian	ref	pauc a cha pauc sos uòlhs s'acostuman a l'escuresina
	hyp	pauc a chad pau sos uòlhs s'acostuman a l'escuresina
	eng	*little by little his eyes got used to the darkness*

Table 5: ASR output examples for Gascon and Languedocian.

RESULTS

Results

Variant	System	WER	ΔWER
Gascon	LM_{nat}	20.86	–
	LM_{syn}	23.14	10.9%
	Whisper	**16.37**	-21.5%
Languedocian	LM_{nat}	13.64	–
	LM_{syn}	13.52	-0.9%
	Whisper	**11.74**	-13.9%

Table 4: Comparison of word error rates for each of the ASR systems and the relative WER change.

Variant	Transcription	
Gascon	ref	un dia maria qu'anà tau camp portar lo disnar au pair
	hyp	un dia maria qu'an atau camp portar lo disnar au pair
	eng	*one day maria went to the camp to bring lunch to her father*
Languedocian	ref	pauc a cha pauc sos uòlhs s'acostuman a l'escuresina
	hyp	pauc a chad pau sos uòlhs s'acostuman a l'escuresina
	eng	*little by little his eyes got used to the darkness*

Table 5: ASR output examples for Gascon and Languedocian.

RESULTS

Results

- Kaldi.
  - The synthetic (MT) approach does not yield (much) better results.
  - Should have used an increasing synthetic corpus instead of going overkill.
  - Qualitative analysis showed that some errors are rooted in the normalisation module.
- Whisper.
  - Yields better results.
- Evident gap between Gascon and Languedocian results.
  - Gascon systems perform worse. Possible reasons:
    - Variability of the dialect: lack of pronunciation consistency.
    - Questionable quality of the annotated dataset.

Conclusions

CONCLUSIONS

Conclusions and Future Work

- First Occitan ASR systems for Gascon and Languedocian.
- For each dialect we evaluated a Kaldi system using a LM created with a corpus originally written in Occitan and a synthetic one obtained using MT.
- We fine-tuned a Whisper model for each dialect.
- Decent WER results obtained:
  - Kaldi: WER = 20.86 Gascon, WER = 13.52 Languedocian.
  - Whisper: WER = 16.37 Gascon, WER = 11.74 Languedocian.
- Whisper yields superior results.
- Kaldi: synthetic LM barely improves the WER for Languedocian, it deteriorates it for Gascon.

CONCLUSIONS

Conclusions and Future Work

- Gascon systems perform worse than Languedocian systems.
  - Due to Gascon dialectal variability?
  - Lower data quality?
- Great effort in collecting a large Occitan corpus.
  - Sadly, a big part of it is copyrighted.
  - Parts of it may end up being publicly available (stay tuned at https://locongres.org).
  - (There is a project for creating the largest open source Occitan corpus).
- Phonetiser is available.
- Future work:
  - Refine MT system and method.
  - Refine normalisation and phonetisation modules.
  - Augment speech corpus with TTS system.

References

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.

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

Nicolas Quint. 2014. L'occitan. Assimil.

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

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.

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.

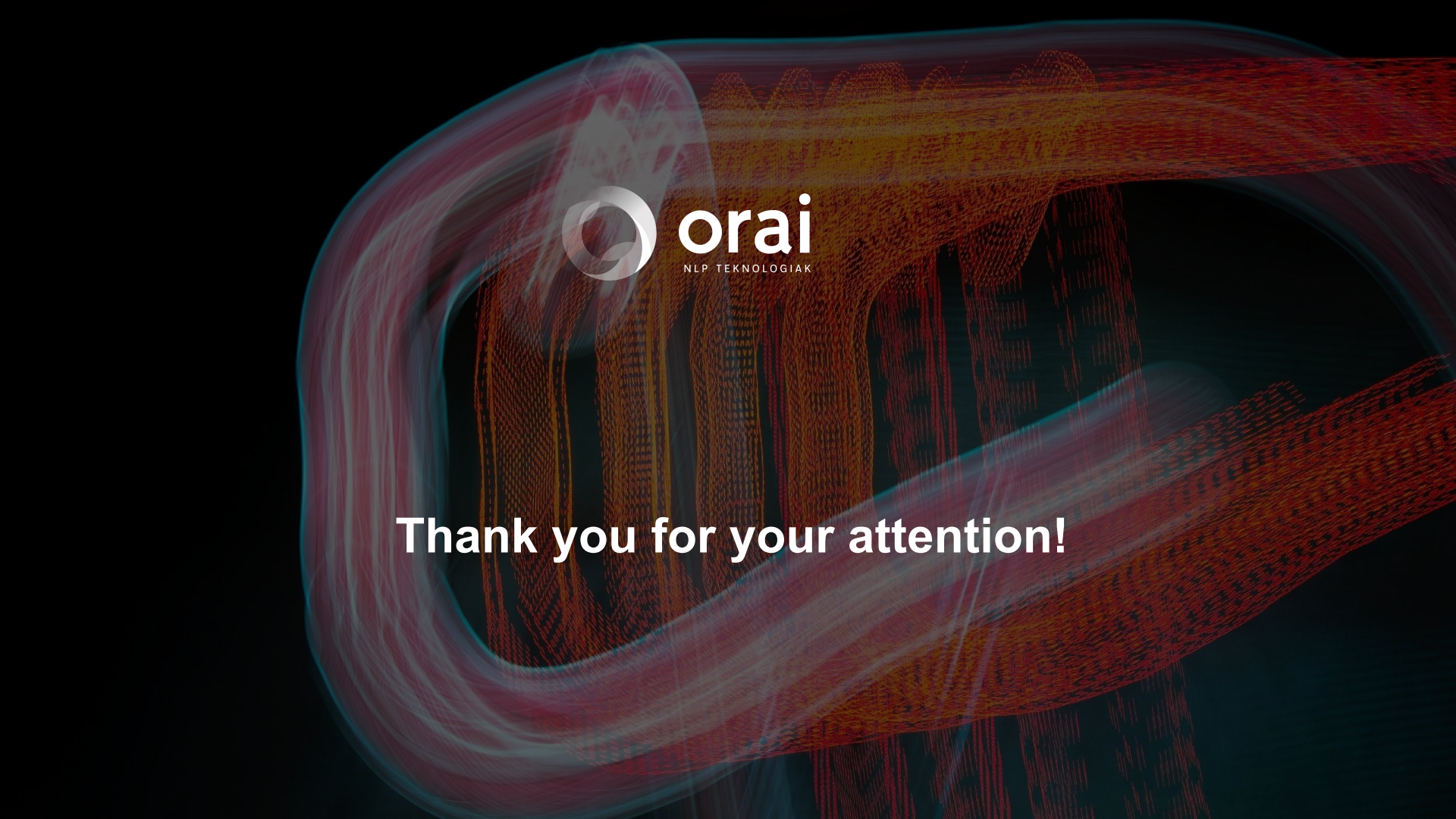
Mikel Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis Tyers. 2011. Apertium: A free/open-source platform for rulebased machine translation. Machine Translation, 25:127–144.

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.

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.

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.


orai
NLP TEKNOLOGIAK
Thank you for your attention!