



# 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



# Outline

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



# Introduction



## INTRODUCTION

# Involved Institutions

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

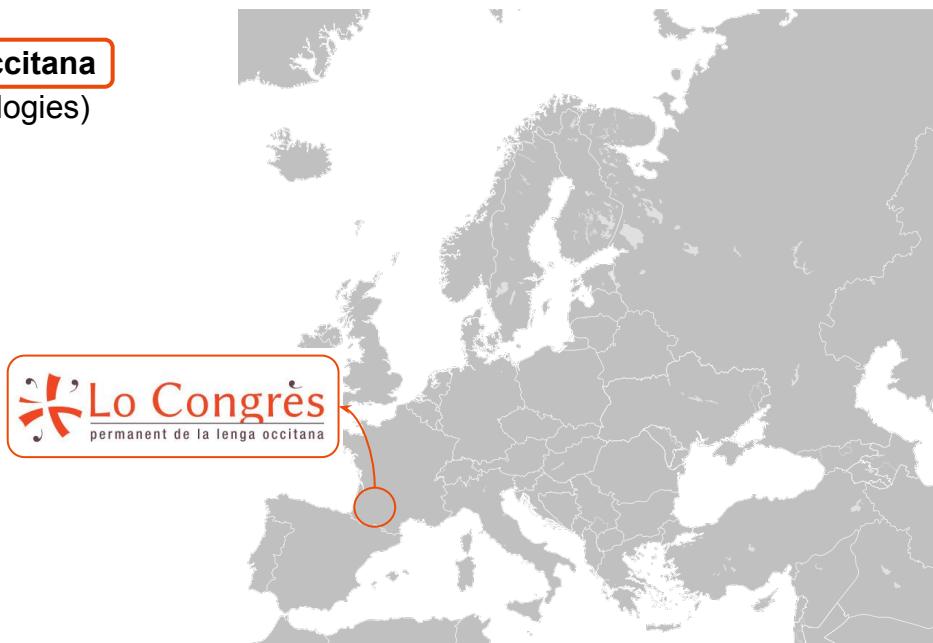




## INTRODUCTION

# Involved Institutions

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

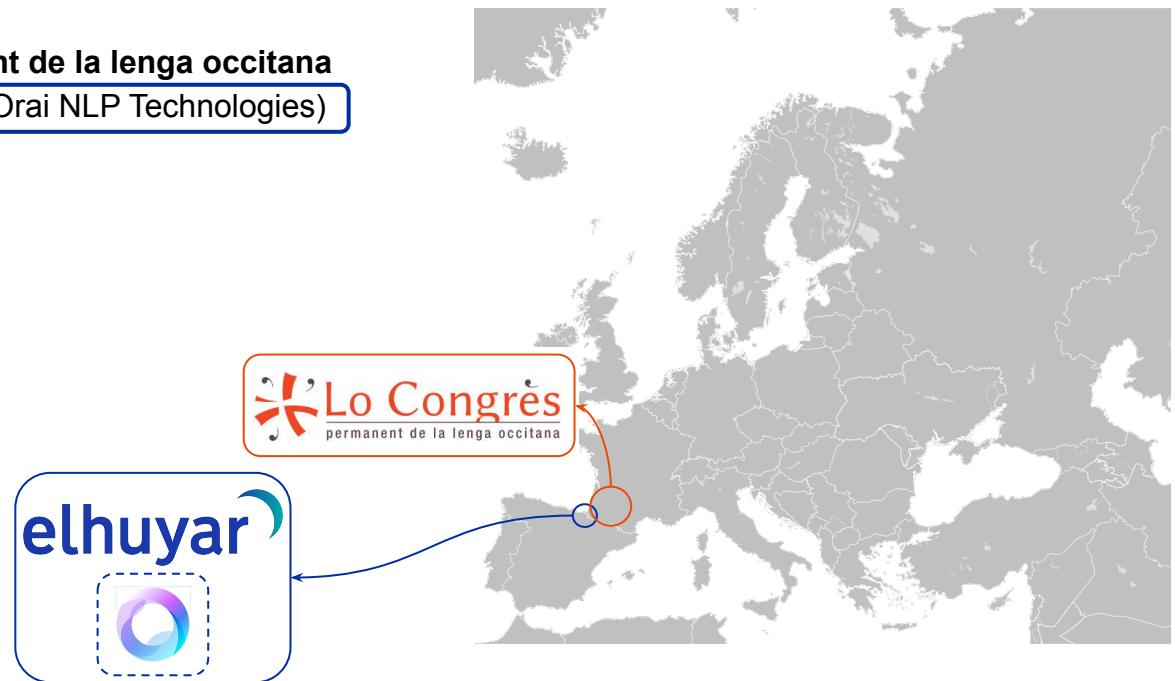




## INTRODUCTION

# 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.





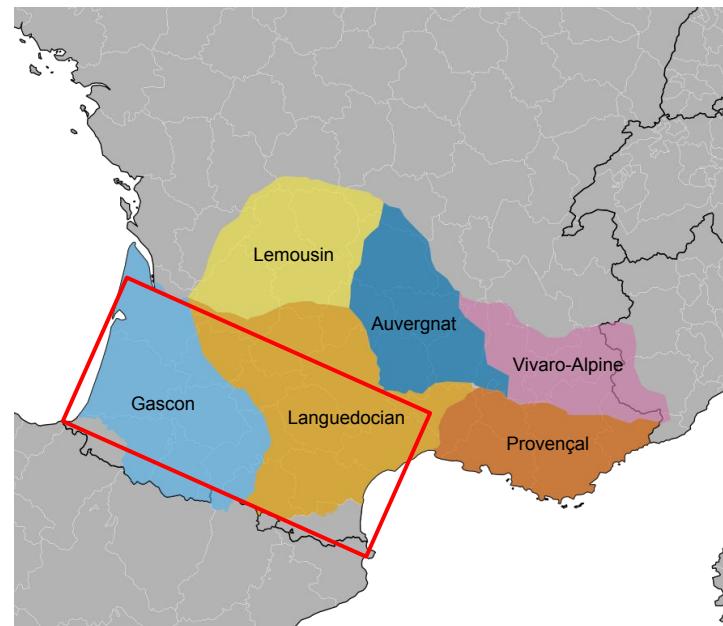
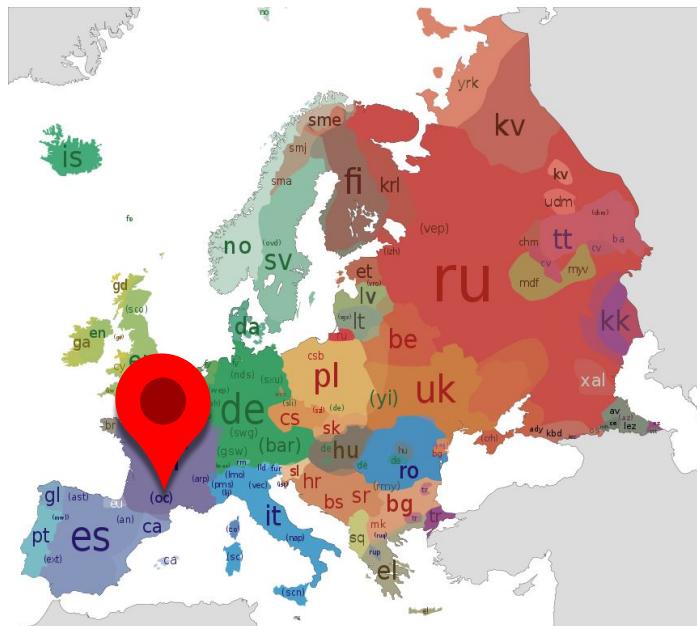
## INTRODUCTION

# 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
  - Limousin
  - Gascon
  - Provençal
  - Languedocian
  - Vivaro-Alpine
- Great variability.

## INTRODUCTION

# Occitan Variants



# Motivation

## MOTIVATION

# Occitan Language Tool Ecosystem



L'API *Revirada* que permet d'obtenir la traducción d'un tèxte de cap tò a partir de l'occitan (gascon e lengadocian).

[Mode d'emploi](#)

L'API *Votz* que transforma un tèxte escrit en votz, e que'ns permet de teledescargar un son qui « ditz » la frase qui avetz escriptura.

[Mode d'emploi](#)

L'API *express'òc* que permet d'afichar expressions en occitan dab la lor traducción francesa a partir d'un mot-clau.

[Mode d'emploi](#)

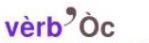
**top'òc**  
Topònimes occitans  
Toponymes occitans

L'API *top'òc* que permet de tradúser un topónime deu francés a l'occitan e vice-versa e d'aver los noms occitan e francés d'ua comuna a partir deu son còdi o de las soas coordenadas.

[Mode d'emploi](#)

**lo Basic**  
Lexic ortografic e referenciat  
Lexique orthographique et référentiel

L'API *Basic* que permet de recercar, a partir d'un mot en francés o en occitan, ua entrada deu *Basic*, lo diccionari deu *Congrès*.

[Mode d'emploi](#)

**vèrb'òc**  
Conjugador automatic occitan  
Conjugeur automatique occitan

L'API *vèrb'òc* que permet, tò un infinitiu occitan, d'aver la soa conjugason. Que podetz precisar la persona, lo mode, lo temps... Que podetz tanben recercar l'infinitiu e las informacions d'ua fòrma conjugada.

[Mode d'emploi](#)

Elements de lenga occitana  
Eléments de langue occitane

L'API *punt de lenga* que permet de recercar articles sus la gramatica occitana a partir d'un mot-clau.

[Mode d'emploi](#)

Diccionari occitan deus sinònimes  
Dictionnaire occitan des synonymes

L'API *sinonimes* que permet d'aver los sinònimes d'un mot en occitan.

[Mode d'emploi](#)

Memento de prononciacion de l'occitan  
Mémento de prononciation de l'occitan

L'API *fon'òc* que permet d'obtenir la transcripció en alfabet fonetic d'un mot o d'un tèxte.

[Mode d'emploi](#)

Diccionari occitan de las rimas  
Dictionnaire occitan des rimes

L'API *rimas* que permet d'aver los mots en occitan qui riman dab un mot qui balhatz.

[Mode d'emploi](#)

Tirage de letras aleatòri en occitan  
Tirage de lettres aléatoires en occitan

L'API *letras* que tira letras a l'azard en respectant la lor freqüencia dens la lenga.

[Mode d'emploi](#)

Generator de pseudo-mots occitans  
Générateur de pseudo-mots occitans

L'API *sembla-mots* que genera mots qui semblan occitans mes qui n'existeishen pas dens la lenga.

[Mode d'emploi](#)

## MOTIVATION

# Occitan Language Tool Ecosystem: Dictionaries & Corpora

 <p><b>revirada</b> traductor automàtic occitan traducteur automatique occitan</p> <p>L'API <i>Revirada</i> que permet d'obténer la traducció d'un tèxte de cap tò o a partir d'un occitan (gascon e lengadocian).</p> <p><a href="#">Mode d'emplec</a></p>	 <p><b>votz</b> sintèsi vocala occitana synthèse vocale occitane</p> <p>L'API <i>Votz</i> que transforma un tèxte escrit en votz, e que'vs permet de teledescargar un son qui « ditz » la frase qui avetz escrivuda.</p> <p><a href="#">Mode d'emplec</a></p>	 <p><b>express'Occ</b> Diccionari d'expressions occitanas Dictionnaire d'expressions occitanes</p> <p>L'API <i>express'Occ</i> que permet d'afichar expressions en occitan dab la lor traducción francesa a partir d'un mot-clau.</p> <p><a href="#">Mode d'emplec</a></p>	 <p><b>top'Occ</b> Topònimes occitans Toponymes occitans</p> <p>L'API <i>top'Occ</i> que permet de traduir un topónime deu francés a l'occitan e vice-versa e d'aver los noms occitan e francés d'ua comuna a partir deu son còdi o de las soas coordenadas.</p> <p><a href="#">Mode d'emplec</a></p>	 <p><b>lo Basic</b> Lexic ortografic e referencial Lexique orthographique et référentiel</p> <p>L'API <i>Basic</i> que permet de recercar, a partir d'un mot en francés o en occitan, ua entrada deu <i>Basic</i>, lo diccionari deu <i>Congrès</i>.</p> <p><a href="#">Mode d'emplec</a></p>	 <p><b>vèrb'Occ</b> Conjugador automàtic occitan Conjugueur automatique occitan</p> <p>L'API <i>vèrb'Occ</i> que permet, tè un infinitiu occitan, d'aver la soa conjugason. Que podetz precisar la persona, lo mode, lo temps... Que podetz tanben recercar l'infinitiu e las informacions d'ua fòrma conjugada.</p> <p><a href="#">Mode d'emplec</a></p>
--	--	---	--	--	--

## MOTIVATION

# Occitan Language Tool Ecosystem: Phonetiser

<b>revirada</b> traductor automàtic occità traducteur automatique occitan	<b>votz</b> sintesi vocala occitana synthèse vocale occitane	<b>express'òc</b> Diccionari d'expressions occitanas Dictionnaire d'expressions occitanes	<b>top'òc</b> Topònimes occitans Toponymes occitans	<b>lo Basic</b> Lexic ortografic e referenciat Lexique orthographique et référentiel	<b>vèrb'òc</b> Conjugador automàtic occità Conjugeur automatique occitan
L'API <i>Revirada</i> que permet d'obtenir la traducció d'un tèxte de cap tò a partir de l'occitan (gascon e lengadocian).  <a href="#">Mode d'emplec</a>	L'API <i>Votz</i> que transforma un tèxte escrit en <i>votz</i> , e que'vs permet de teledescargar un son qui « ditz » la frase qui avetz escriptura.  <a href="#">Mode d'emplec</a>	L'API <i>express'òc</i> que permet d'afichar expressions en occitan dab la lor traducció francesa a partir d'un mot-clau.  <a href="#">Mode d'emplec</a>	L'API <i>top'òc</i> que permet de tradúser un topònim deu francès a l'occitan e vice-versa e d'aver los noms occitan e francés d'ua comuna a partir deu son còdi o de las soas coordenadas.  <a href="#">Mode d'emplec</a>	L'API <i>Basic</i> que permet de recercar, a partir d'un mot en francès o en occitan, ua entrada deu <i>Basic</i> , lo diccionari deu <i>Congrès</i> .  <a href="#">Mode d'emplec</a>	L'API <i>vèrb'òc</i> que permet, tò un infinitiu occitan, d'aver la soa conjugason. Que podetz precisar la persona, lo mode, lo temps... Que podetz tanben recercar l'infinitiu e las informacions d'ua fòrma conjugada.  <a href="#">Mode d'emplec</a>
<b>punt de lenga</b> Elements de lenga occitana Eléments de langue occitane	<b>sinonimes</b> Diccionari occitan deus sinonimes Dictionnaire occitan des synonymes	<b>fon'òc</b> Memento de prononciacion de l'occitan Mémento de prononciation de l'occitan	<b>rimas</b> Diccionari occitan de las rimas Dictionnaire occitan des rimes	<b>letras</b> Tirage de letras aleatòri en occitan Tirage de lettres aléatoires en occitan	<b>semblamots</b> Generador de pseudo-mots occitans Générateur de pseudo-mots occitans
L'API <i>punt de lenga</i> que permet de recercar articles sus la gramatica occitana a partir d'un mot-clau.  <a href="#">Mode d'emplec</a>	L'API <i>sinonimes</i> que permet d'aver los sinonimes d'un mot en occitan.  <a href="#">Mode d'emplec</a>	L'API <i>fon'òc</i> que permet d'obtenir la transcripció en alfabet fonetic d'un mot o d'un tèxte.  <a href="#">Mode d'emplec</a>	L'API <i>rimas</i> que permet d'aver los mots en occitan qui riman dab un mot qui balhatz.  <a href="#">Mode d'emplec</a>	L'API <i>letras</i> que tira letras a l'azard en respectant la lor freqüència dens la lenga.  <a href="#">Mode d'emplec</a>	L'API <i>sembla-mots</i> que genera mots qui semblan occitans mes qui n'existeishen pas dens la lenga.  <a href="#">Mode d'emplec</a>

## MOTIVATION

# Occitan Language Tool Ecosystem: Machine Translation

 **revirada**  
traductor automàtic occitàn  
traducteur automatique occitan

L'API *Revirada* que permet d'obtenir la traducción d'un tèxte de cap tò a partir de l'occitan (gascon e lengadocian).

[Mode d'emploi](#)

 **votz**  
sintesi vocala occitana  
synthèse vocale occitane

L'API *Votz* que transforma un tèxte escrit en votz, e que'ns permet de teledescargar un son qui « ditz » la frase qui avetz escriptura.

[Mode d'emploi](#)

 **express'òc**  
Diccionari d'expressions occitanas  
Dictionnaire d'expressions occitanes

L'API *express'òc* que permet d'afichar expressions en occitan dab la lor traducción francesa a partir d'un mot-clau.

[Mode d'emploi](#)

 **top'òc**  
Topònimes occitans  
Toponymes occitans

L'API *top'òc* que permet de tradúser un topònime deu francès a l'occitan e vice-versa e d'aver los noms occitan e francés d'ua comuna a partir deu son còdi o de las soas coordenadas.

[Mode d'emploi](#)

 **lo Basic**  
Lexic ortografic e referenciat  
Lexique orthographique et référentiel

L'API *Basic* que permet de recercar, a partir d'un mot en francès o en occitan, ua entrada deu *Basic*, lo diccionari deu *Congrès*.

[Mode d'emploi](#)

 **verb'òc**  
Conjugador automatic occitan  
Conjugeur automatique occitan

L'API *verb'òc* que permet, tò un infinitiu occitan, d'aver la soa conjugason. Que podetz precisar la persona, lo mode, lo temps... Que podetz tanben recercar l'infinitiu e las informacions d'ua fòrma conjugada.

[Mode d'emploi](#)

 **punt de lenga**  
Elements de lenga occitana  
Eléments de langue occitane

L'API *punt de lenga* que permet de recercar articles sus la gramatica occitana a partir d'un mot-clau.

[Mode d'emploi](#)

 **sinonimes**  
Diccionari occitan deus sinonimes  
Dictionnaire occitan des synonymes

L'API *sinonimes* que permet d'aver los sinonimes d'un mot en occitan.

[Mode d'emploi](#)

 **fon'òc**  
Memento de prononciacion de l'occitan  
Mémento de prononciation de l'occitan

L'API *fon'òc* que permet d'obtenir la transcripciona en alfabet fonetic d'un mot o d'un tèxte.

[Mode d'emploi](#)

 **rimas**  
Diccionari occitan de las rimas  
Dictionnaire occitan des rimes

L'API *rimas* que permet d'aver los mots en occitan qui riman dab un mot qui balhatz.

[Mode d'emploi](#)

 **letras**  
Tirage de letras aleatori en occitan  
Tirage de lettres aléatoires en occitan

L'API *letras* que tira letras a l'azard en respectant la lor freqüencia dens la lenga.

[Mode d'emploi](#)

 **semblamots**  
Generador de pseudo-mots occitans  
Générateur de pseudo-mots occitans

L'API *sembla-mots* que genera mots qui semblan occitans mes qui n'existeishen pas dens la lenga.

[Mode d'emploi](#)

## MOTIVATION

# Occitan Language Tool Ecosystem: Text-to-speech

**revirada**  
traductor automàtic occità  
traducteur automatique occitan

L'API *Revirada* que permet d'obtenir la traducció d'un tèxte de cap tò a partir de l'occitan (gascon e lengadocian).

[Mode d'emplec](#)

**votz**  
sintesi vocala occitana  
synthèse vocale occitane

L'API *Votz* que transforma un tèxte escrit en votz, e que'ns permet de teledescargar un son qui « ditz » la frase qui avetz escriptada.

[Mode d'emplec](#)

**express'Òc**  
Diccionari d'expressions occitanas  
Dictionnaire d'expressions occitanes

L'API *express'Òc* que permet d'afichar expressions en occitan dab la lor traducció francesa a partir d'un mot-clau.

[Mode d'emplec](#)

**top'Òc**  
Topònimes occitans  
Toponymes occitans

L'API *top'Òc* que permet de tradúser un topònim deu francès a l'occitan e vice-versa e d'aver los noms occitan e francés d'ua comuna a partir deu son còdi o de las soas coordenadas.

[Mode d'emplec](#)

**lo Basic**  
Lexic ortografic e referencial  
Lexique orthographique et référentiel

L'API *Basic* que permet de recercar, a partir d'un mot en francès o en occitan, ua entrada deu *Basic*, lo diccionari deu *Congrès*.

[Mode d'emplec](#)

**verb'Òc**  
Conjugador automàtic occità  
Conjugateur automatique occitan

L'API *verb'Òc* que permet, tò un infinitiu occitan, d'aver la soa conjugacion. Que podetz precisar la persona, lo mode, lo temps... Que podetz tanben recercar l'infinitiu e las informacions d'ua fòrma conjugada.

[Mode d'emplec](#)

**punt de lenga**  
Elements de lenga occitana  
Eléments de langue occitane

L'API *punt de lenga* que permet de recercar articles sus la gramatica occitana a partir d'un mot-clau.

[Mode d'emplec](#)

**sinonimes**  
Diccionari occitan deus sinònimes  
Dictionnaire occitan des synonymes

L'API *sinonimes* que permet d'aver los sinònimes d'un mot en occitan.

[Mode d'emplec](#)

**fon'Òc**  
Memento de prononciacion de l'occitan  
Mémento de prononciation de l'occitan

L'API *fon'Òc* que permet d'obtenir la transcripció en alfabet fonetic d'un mot o d'un tèxte.

[Mode d'emplec](#)

**rimas**  
Diccionari occitan de las rimas  
Dictionnaire occitan des rimes

L'API *rimas* que permet d'aver los mots en occitan qui riman dab un mot qui balhatz.

[Mode d'emplec](#)

**letras**  
Tirage de letras aleatòri en occitan  
Tirage de lettres aléatoires en occitan

L'API *letras* que tira letras a l'azard en respectant la lor freqüència dens la lenga.

[Mode d'emplec](#)

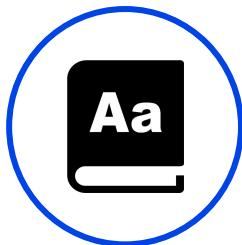
**semblamots**  
Generador de pseudo-mots occitans  
Générateur de pseudo-mots occitans

L'API *sembla-mots* que genera mots qui semblan occitans mes qui n'existeishen pas dens la lenga.

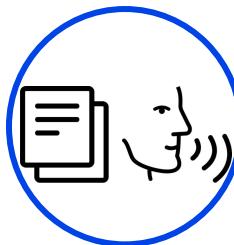
[Mode d'emplec](#)

## MOTIVATION

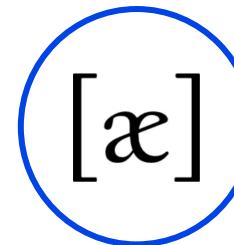
# Occitan Language Tool Ecosystem



Dictionaries



Text and Speech Corpora



Phonetiser



Verb Conjugator



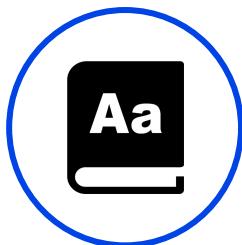
Machine Translation



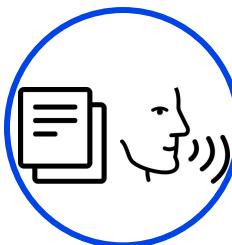
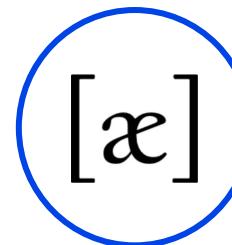
Text-to-speech

## MOTIVATION

# Occitan Language Tool Ecosystem



Dictionaries

Text and Speech  
Corpora

Phonetiser



Corpora far too small



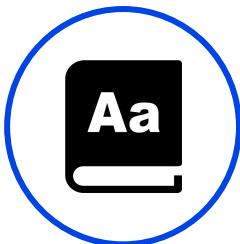
Verb Conjugator

Machine  
Translation

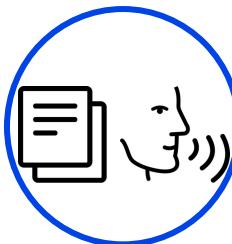
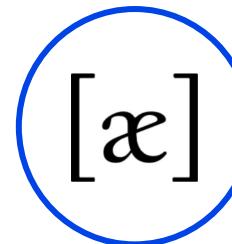
Text-to-speech

## MOTIVATION

# Occitan Language Tool Ecosystem



Dictionaries

Text and Speech  
Corpora

Phonetiser



Verb Conjugator

Machine  
Translation

Text-to-speech



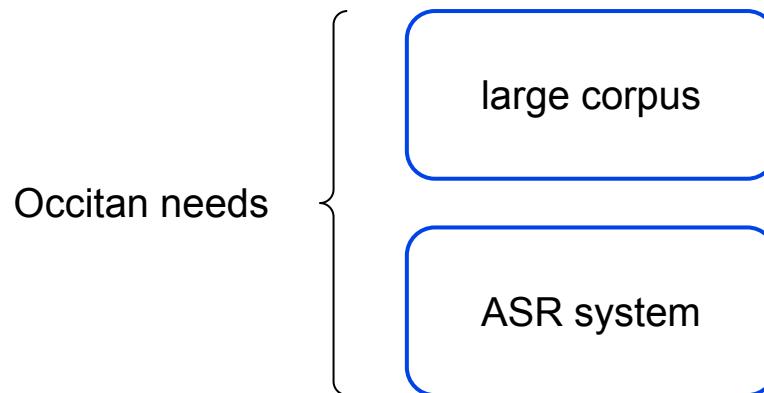
Corpora far too small



No ASR system

## MOTIVATION

# What Occitan needs





# Resources for an Occitan ASR



## 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

## Estatísticas d'utilització de la plataforma

Causir un període per les estadístiques				
Data de debutà	2020/01/01	Data de fin	2022/10/01	Validar
Total de ficheris enregistrats		39553 (88 h 58 min 59 s)		
Rapòrt per gènre				
Rapòrt per atges				
Detall per dialèctes				
Dialècte	Total (M/F)	Durada	15 ans e mens (M/F)	16 - 35 ans (M/F)
Auvernhat	29 / 0 (29)	00 h 03 mn 16 s	0 / 0	28 / 0
Gascon	11233 / 12239 (23472)	51 h 16 mn 44 s	214 / 122	380 / 5244
Lemosin	33 / 3 (36)	00 h 03 mn 03 s	0 / 0	32 / 0
Lengadocien	9732 / 6003 (15735)	37 h 14 mn 16 s	32 / 38	4027 / 2953
Niçard	6 / 12 (18)	00 h 01 mn 50 s	0 / 0	6 / 0
Provençau	151 / 104 (255)	00 h 19 mn 12 s	0 / 0	151 / 30
Vivaro-aupenc	6 / 2 (8)	00 h 00 mn 36 s	0 / 0	1 / 0
Aranés	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0
Cisaupin	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0
Varietats del creissent	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0

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%	
Languedocien	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

## Estatísticas d'utilització de la plataforma

Causir un període per les estadístiques						
Data de debutà	2020/01/01	Data de fin	2022/10/01	Validar		
Total de ficheris enregistrats						39553 (88 h 58 min 59 s)
Rapòrt per gènre						Omes Femnas
						21190 18363
Rapòrt per atges						15 ans e menys 16 - 35 anys 36 - 60 anys mai de 60 anys
						12852 22033 4262
Detall per dialèctes						
Dialècte	Total (M/F)	Durada	15 ans e mens (M/F)	16 - 35 ans (M/F)	36 - 60 ans (M/F)	mai de 60 ans (M/F)
Auvernhat	29 / 0 (29)	00 h 03 mn 16 s	0 / 0	28 / 0	1 / 0	0 / 0
Gascon	11233 / 12239 (23472)	51 h 16 mn 44 s	214 / 122	380 / 5244	8278 / 6960	2361 / 13
Lemosin	33 / 3 (36)	00 h 03 mn 03 s	0 / 0	32 / 0	1 / 3	0 / 0
Lengadocien	9732 / 6003 (15735)	37 h 14 mn 16 s	32 / 38	4027 / 2953	3834 / 2965	1839 / 47
Niçard	6 / 12 (18)	00 h 01 mn 50 s	0 / 0	6 / 0	0 / 10	0 / 2
Provençau	151 / 104 (255)	00 h 19 mn 12 s	0 / 0	151 / 30	0 / 74	0 / 0
Vivaro-aupenc	6 / 2 (8)	00 h 00 mn 36 s	0 / 0	1 / 0	5 / 2	0 / 0
Aranés	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0	0 / 0	0 / 0
Cisaupin	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0	0 / 0	0 / 0
Varietats del creissent	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0	0 / 0	0 / 0

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%	
Languedocien	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

## Estatísticas d'utilització de la plataforma

Causir un període per les estadístiques						
Data de debutà	2020/01/01	Data de fin	2022/10/01	Validar		
Total de ficheris enregistrats						39553 (88 h 58 min 59 s)
Rapòrt per gènre						Omes Femnas
						21190 18363
Rapòrt per atges						15 ans e mens 16 - 35 ans 36 - 60 ans mai de 60 ans
						12852 22033 4262
Detall per dialèctes						
Dialècte	Total (M/F)	Durada	15 ans e mens (M/F)	16 - 35 ans (M/F)	36 - 60 ans (M/F)	mai de 60 ans (M/F)
Auvernhat	29 / 0 (29)	00 h 03 mn 16 s	0 / 0	28 / 0	1 / 0	0 / 0
Gascon	11233 / 12239 (23472)	51 h 16 mn 44 s	214 / 122	380 / 5244	8278 / 6960	2361 / 13
Lemosin	33 / 3 (36)	00 h 03 mn 03 s	0 / 0	32 / 0	1 / 3	0 / 0
Lengadocien	9732 / 6003 (15735)	37 h 14 mn 16 s	32 / 38	4027 / 2953	3834 / 2965	1839 / 47
Niçard	6 / 12 (18)	00 h 01 mn 50 s	0 / 0	6 / 0	0 / 10	0 / 2
Provençau	151 / 104 (255)	00 h 19 mn 12 s	0 / 0	151 / 30	0 / 74	0 / 0
Vivaro-aupenc	6 / 2 (8)	00 h 00 mn 36 s	0 / 0	1 / 0	5 / 2	0 / 0
Aranés	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0	0 / 0	0 / 0
Cisaupin	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0	0 / 0	0 / 0
Varietats del creissent	0 / 0 (0)	00 h 00 mn 00 s	0 / 0	0 / 0	0 / 0	0 / 0

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%	
Languedocien	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<sub>Gasc.</sub> (521M words)
- fr-oc<sub>Lang.</sub> (503M words)

Newspaper	URL	Words
La République des Pyrénées	<a href="https://www.larepubliquedespyrenees.fr">https://www.larepubliquedespyrenees.fr</a>	35.75M
Sud Ouest	<a href="https://www.sudouest.fr">https://www.sudouest.fr</a>	171.81M
La Dépêche	<a href="https://www.ladepeche.fr">https://www.ladepeche.fr</a>	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
<b>Gascon</b>			
Speech	126.6h	91k	960k
Natural text	—	289k	3.5M
MT text	—	25.3M	521.4M
<b>Languedocian</b>			
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
<b>Gascon</b>			
Speech	126.6h	91k	960k
Natural text	—	289k	3.5M
MT text	—	25.3M	521.4M
<b>Languedocian</b>			
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
<b>Gascon</b>			
Speech	126.6h	91k	960k
Natural text	—	289k	3.5M
MT text	—	25.3M	521.4M
<b>Languedocian</b>			
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	$\Delta$ WER
Gascon	LM <sub>nat</sub>	20.86	—
	LM <sub>syn</sub>	23.14	10.9%
	<b>Whisper</b>	<b>16.37</b>	-21.5%
Languedocian	LM <sub>nat</sub>	13.64	—
	LM <sub>syn</sub>	13.52	-0.9%
	<b>Whisper</b>	<b>11.74</b>	-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	<i>one day maria went to the camp to bring lunch to her father</i>	
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	<i>little by little his eyes got used to the darkness</i>	

Table 5: ASR output examples for Gascon and Languedocian.

## RESULTS

# Results

Variant	System	WER	$\Delta$ WER
Gascon	LM <sub>nat</sub>	20.86	–
	LM <sub>syn</sub>	23.14	10.9%
	Whisper	<b>16.37</b>	-21.5%
Languedocian	LM <sub>nat</sub>	13.64	–
	LM <sub>syn</sub>	13.52	-0.9%
	Whisper	<b>11.74</b>	-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	<i>one day maria went to the camp to bring lunch to her father</i>	
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	<i>little by little his eyes got used to the darkness</i>	

Table 5: ASR output examples for Gascon and Languedocian.

## RESULTS

# Results

Variant	System	WER	$\Delta$ WER
Gascon	LM <sub>nat</sub>	20.86	—
	LM <sub>syn</sub>	23.14	10.9%
	<b>Whisper</b>	<b>16.37</b>	-21.5%
Languedocian	LM <sub>nat</sub>	13.64	—
	LM <sub>syn</sub>	13.52	-0.9%
	<b>Whisper</b>	<b>11.74</b>	-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	<i>one day maria went to the camp to bring lunch to her father</i>	
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	<i>little by little his eyes got used to the darkness</i>	

Table 5: ASR output examples for Gascon and Languedocian.

## RESULTS

# Results

Variant	System	WER	$\Delta$ WER
Gascon	LM <sub>nat</sub>	20.86	—
	LM <sub>syn</sub>	23.14	10.9%
	<b>Whisper</b>	<b>16.37</b>	-21.5%
Languedocian	LM <sub>nat</sub>	13.64	—
	LM <sub>syn</sub>	13.52	-0.9%
	<b>Whisper</b>	<b>11.74</b>	-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	<i>one day maria went to the camp to bring lunch to her father</i>	
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	<i>little by little his eyes got used to the darkness</i>	

Table 5: ASR output examples for Gascon and Languedocian.

## RESULTS

# Results

Variant	System	WER	$\Delta$ WER
Gascon	LM <sub>nat</sub>	20.86	—
	LM <sub>syn</sub>	23.14	10.9%
	<b>Whisper</b>	<b>16.37</b>	-21.5%
Languedocian	LM <sub>nat</sub>	13.64	—
	LM <sub>syn</sub>	13.52	-0.9%
	<b>Whisper</b>	<b>11.74</b>	-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	<i>one day maria went to the camp to bring lunch to her father</i>	
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	<i>little by little his eyes got used to the darkness</i>	

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, Michæl 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 unaiga-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!**