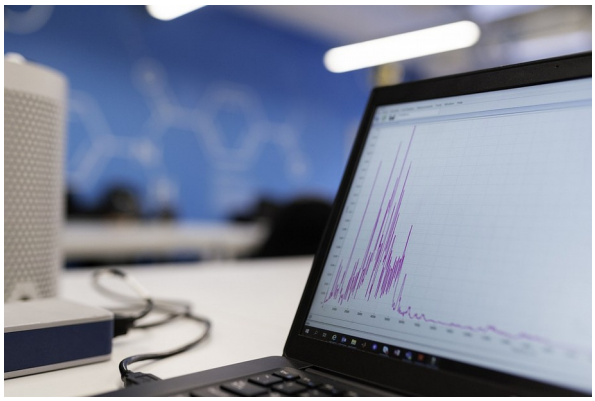
Lo Congrès
permanent de la lenga occitana

COMUNICAT

ReVoc

Programa de desvolopament de la **reconeishença vocau en occitan**



La reconeishença vocau qu'ei l'utís qui **analisa la votz** e qui la transcriu dab la fòrma d'un tèxte escriut. Que hè partida de las tecnologias de tractament de la paraula qui permeten aus umans d'**escambiar oraument dab las maquinas**, mercés a interfàcias vocaus.

La reconeishença vocau qu'ei indispensabla tà realizar utís com lo **sostitolatge automatic** de video, las aplicacions de **dictada vocau** o los **assistents personaus intelligents**.

Lo Congrès permanent de la lenga occitana que participa a un programa **transfronterèr** triennau dab lo prètzhèit de **dotar l'occitan** (tà las soas varietats gascona e lengadociana) d'aquera tecnologia.

Que tribalha en partenariat dab la Rolde de Estudios Aragoneses (qui desvolòpa la medisha tecnologia entà la lenga aragonesa), la fondacion basca Elhuyar (en carga de la partida tecnica deu programa) e mei d'ua estructura qui produseishen contienguts multimèdias en occitan.

Lo Congrès permanent de la lenga occitana

Rolde de Estudios Aragoneses

elhuyar

Région Nouvelle-Aquitaine

La Région Occitanie Pyrénées · Méditerranée

Pyrenees Atlantiques Le Département

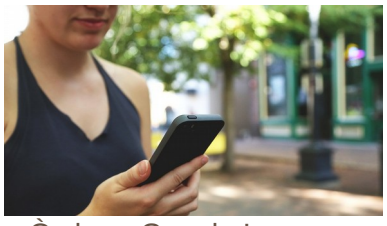
/ Perqué la reconeishença vocau en occitan ?

Las tecnologias de la lenga – reconeishença vocau, sintèsi vocau, traduccion automatica o enqüèra analisi semantic – que son un enjòc vitau entà las lengas minorizadas. Entà projectà's cap a ua societat mei anar mei numerizada, aquestas qu'an de dispausar de ressorsas e apèrs de qui cau entà que los locutors escàmbien en la lor lenga pròpia peu mejan d'interfacis. Mei d'un programa qu'estón realizats en aqueth sens entà la lenga occitana : Linguatec (traduccion automatica e sintèsi vocau), BaTelOc (basa textuau occitana), ROLF (clavèrs predictius).

La reconeishença vocau que permet la transcripcion de la votz en tèxte, ua tecnologia qui ei d'ara enlà difusada en abonde en aplicacions gran public, notadament peus assistents personaus (Siri d'Apple, Google Home o enqüèra Alexa d'Amazon entaus mei coneishuts) e tau sostitolatge automatic de video.

/ Exemples d'utilizacion de la reconeishença vocau

Assistents personaus



« Òc ben, Google ! »
Lo desvolopament de la reconeishença vocau que permeterà de har los assistents personals en occitan !

Sostitolatge automatic de videos

Un programa de reconeishença vocau que va permetre lo sostitolatge automatic de videos dins mai d'una lenga.

Transcripcion automatica de collectages

Un module de transcripcion automatica basat sus la reconeishença vocala qu'ajudarà lo tribalh de lingüistas e enquestaires.

- Demostracions de la reconeishença vocau basca e castelhana d'Elhuyar

/ La plataforma de contribucion

Entà atraçar ua quantitat bèra d'enregistraments transcriuts, e qui sian representatius de la diversitat deus locutors de l'occitan, Lo Congrès qu'a desvolopat un utís de contribucion tà la comunautat. Sus aquera plataforma, cadun que pòt enregistrar frasas qui seràn ajustadas au còrpus bastit dab los partenaris.

- Véder la video de presentacion
- Anar tà la plataforma

/ La reconeishença vocau, com fonciona ?

La reconeishença vocau qu'utiliza **l'intelligéncia artificiau** (los hialats neuronaus) tà transcríver automaticament la votz en tèxte escriut.

Abans d'ac poder har, que hè besonh d'entrainar l'IA dab frasas audio dejà transcriutas. Que hè doncas besonh un **bèth còrpus audio transcriut**, qu'ei a díser ua quantitat grana de tèxtes dab los enregistraments audios correspondents.

Que cau tanben « noirir » la maquina dab **còrpus grans de tèxte** e sonque. Atau que pòt apréner quaus fòrmas e son frequentas, quau mot apareish sovent a costat de tau aute...

Enfin, **desvolopar que cau mei d'un programa** :

- Un tà passar en letras los nombres, los simbèus, las abreviacions, las unitats de mesura... abans de balhar un tèxte a la maquina.
- Un « abreviator » qui hè lo contra, tà har mei legeders los tèxtes prepausats aus utilizators.
- Un fonetizaire tà obtiéner la prononciacion en alfabet fonetic internacionau d'un mot.
- Un programa tà aver tots los mots qui corresponen a ua prononciacion.

/ Las etapas de l'entrainament e de la transcripcion

Entrainament deu module de reconeishença deus sons

Enregistraments audio + transcripcions

⇓

chifras en letras

⇓

Enregistraments audio + transcripcions espandidas

⇓

alinhament de las frasas

⇓

Enregistraments audios + transcripcions alinhats

⇓

fonetizacion de las frasas

⇓

Enregistraments + transcripcions + fonetizacions

⇓

entrainament IA

⇓

Basa de correspondéncias sons-fonèmas

Entrainament deu module de reconeishença deus mots

Còrpus textuau

⇓

chifras en letras

⇓

Còrpus textuau espandit

⇓

entrainament IA

⇓

Basa de frequéncia deus mots en contèxte

Transcripcion de tèxte

Fichièr audio

⇓

seleccion deu fonèma mei probable peu module de reconeishença deus sons

⇓

Seguida de fonèmas (transcripcion fonetica de l'audio)

⇓

seleccion deus mots mei probables peu module de reconeishença deus mots

⇓

Transcripcion espandida

⇓

abreviator

⇓

Transcripcion finau

/ Lo calendari

// 2020 : Definicion de las exigéncias, especificacions foncionaus e constitucion deu còrpus

Que's haràn en ua purmèra etapa las exigéncias tecnicas atau com las especificacions foncionaus.

D'un punt de vista tecnic, los desvolopaments entà l'occitan que seràn realizats en l'estat de l'art, a saber per l'utilizacion deus hialats neuronaus (intelligéncia artificiau). Ad aquesta tecnologia de tria que'u hè totun besonh un nombre hòrt important de dadas. Sonque un còrpus ric, voluminós e variat que guarantirà un resultat de qualitat en fin de cadena.

Entad aquò har, lo Congrès qu'a engatjat un partenariat dab mantuns productors de contienguts textuaus multimèdias en occitan : institucions, mèdias, editors, productors de contienguts audiovisuaus...

Qu'ei pr'amor d'aquò aquesta purmèra fasa que serà essenciaument consacrada a un tribalh de collècta, tractament (alinhament tèxte/son) e enterpausatge de còrpus textuaus e acostics per l'occitan. Que s'estiman a 200 òras haut o baish lo besonh de transcripcions e a 500 milions de mots lo còrpus textuau necessari per cada varietat. Pr'amor d'estar l'occitan ua lenga enqüèra tròp chic dotada, que compensaram per l'utilizacion de còrpus gigants deu francés e de l'espanhòu per obtiéner, mercés a la traduccion automatica, còrpus textuaus occitans importants.

// 2021 : Finalizacion e desvolopament tecnologic

Ua part grana deu projècte que's harà ad aqueth moment : acabar la collècta de las dadas necessàrias, realizar tres deus quate lòts de tribalh mei tecnics entà arribar a ua version avançada deu desvolopament. Concrètament, que prevedem au mensh ua mesa en òbra avançada deus modules seguents :

- Creacion deu modèle lingüistic.
- Creacion deu modèle acostic.
- Desvolopament deu transcriptor.

// 2022 : Desvolopament finau e validacion

Dens la purmèra partida d'aquesta fasa darrèra, tots los desvolopaments tecnologics deu projècte que seràn acabats. La fasa de construccion deus transcriptors que serà tanben acabada. Integradas totas los compausantas tecnologicas, que seràn sosmetudas a ua seria de tèsts intensius d'avaloracion.

/ Los actors

// Sòcis e sostiens

ReVOc qu'ei un programa navèth de desvolopament de la reconeishença vocau en occitan (varietats gascona e lengadociana) engatjat peu Congrès permanent de la lenga occitana. Aqueth programa triennau (2020-2022) que's debana dens l'encastre d'un partenariat transfronterèr qui assòcia l'institucion aragonesa Rolde de Estudios Aragoneses (qui desvoloparà la medisha tecnologia entà la lenga aragonesa) e la fondacion basca Elhuyar (en carga de la partida tecnica deu programa). Qu'a lo sostien financèr de la Region Novèla Aquitània, de la Region Occitània e deu Departament deus Pirenèus Atlantics (aperets a projèctes transfronterèrs).

// Lo partenariat tà la constitucion deus còrpus

Tà entrainar l'intelligéncia artificiau, que cau quantitas granas de dadas. L'occitan, lenga dita « pauc dotada », n'a pas generaument aqueths ensembles de dadas. Tà constituir lo còrpus audio e lo còrpus textuau qui hèn mestièr au desvolopament de la reconeishença vocau, Lo Congrès non podè har-s'i solet.

Que's bastí doncas un partenariat d'ua pagèra inedita tà çò de l'occitan, tà constituir ua basa audio e textuau a la quau mei d'ua estructura e vienón portar la lor contribucion.

E tanben :

- Miquèu Baris
- Danís Chapduèlh
- David Grosclaude
- Lo Blòg Hadiu
- Patric Lavaud
- Canau Youtube Puta de mòrt

/ Lo Congrès permanent de la lenga occitana

Lo Congrès permanent de la lenga occitana qu'ei l'organisme interregionau de regulacion de l'occitan. Qu'òbra dens los domenis de la lingüistica e deu TAL (tractament automatic de la lenga).

Que produseish utís lingüistics numerics de referéncia (diccionaris, conjugators, correctors ortografics...), aplicacions tau TAL (sintèsi vocau, traduccion automatica...) e aplicacions taus telefonets (clavèrs predictius...).

Qu'a tanben missions de regulacion lingüistica e de recèrca scientifica aplicada.

Qu'ei l'editor d'un multidiccionari occitan (dicod'Òc) qui a cada an mei d'un milion de visitas.

/ Tà mei d'informacions :

Lo Congrès permanent de la lenga occitana
Castèth d'Este, Avenue de la Pléiade
64140 Billère

premsa.locongres.com/revoc

info@locongres.org

05 59 13 06 40