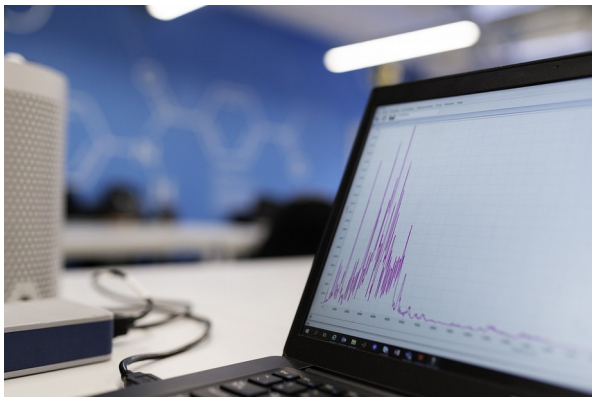
Lo Congrès
permanent de la lenga occitana

COMMUNIQUÉ

ReVoc

Programme de développement de la **reconnaissance vocale en occitan**



La reconnaissance vocale est l'outil qui **analyse la voix** et qui la transcrit sous la forme d'un texte écrit. Elle fait partie des technologies de traitement de la parole qui permettent aux humains d'**échanger oralement avec les machines**, grâce aux interfaces vocales.

La reconnaissance vocale est indispensable pour réaliser des outils comme le **sous-titrage automatique de vidéos**, les applications de **dictée vocale** ou les **assistants personnels intelligents**.

Le Congrès permanent de la langue occitane participe à un programme **transfrontalier** triennal avec l'objectif de **doter l'occitan** (pour ses variétés gasconne et languedocienne) de cette technologie.

Il travaille en partenariat avec la Rolde de Estudios Aragoneses (qui développe la même technologie pour la langue aragonaise), la fondation basque Elhuyar (en charge de la partie technique du programme) et plusieurs structures qui produisent des contenus multimédias en occitan.

Lo Congrès permanent de la lenga occitana

Rolde de estudios aragoneses

elhuyar

Région Nouvelle-Aquitaine

La Région Occitanie Pyrénées · Méditerranée

Pyrénées Atlantiques Le Département

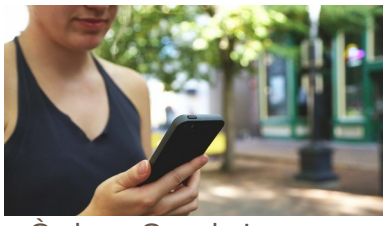
/ Pourquoi la reconnaissance vocale en occitan ?

Les technologies de la langue – reconnaissance vocale, synthèse vocale, traduction automatique ou encore analyse sémantique – sont un enjeu vital pour les langues minorisées. Pour se projeter vers une société de plus en plus numérisée, elles doivent disposer des ressources et des outils nécessaires pour que les locuteurs échangent dans leur propre langue à travers des interfaces. Plusieurs programmes ont été réalisés en ce sens pour la langue occitane : Linguatec (traduction automatique et synthèse vocale), BaTelOc (base textuelle occitane), ROLF (claviers prédictifs).

La reconnaissance vocale permet la transcription de la voix en texte, une technologie qui est aujourd'hui largement diffusée dans des applications grand public, notamment par les assistants personnels (Siri d'Apple, Google Home ou encore Alexa d'Amazon pour les plus connus) et pour le sous-titrage automatique de vidéos.

/ Exemples d'utilisation de la reconnaissance vocale

Assistants personnels

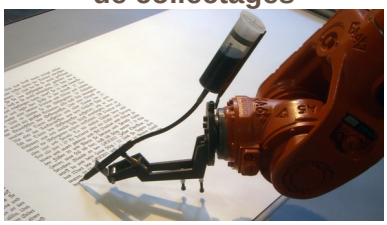


« Òc ben, Google ! »
Le développement de la reconnaissance vocale permettra de passer les assistants personnels en occitan !

Sous-titrage automatique de vidéos

Un programme de reconnaissance vocale permettra le sous-titrage automatique de vidéos dans plusieurs langues.

Transcription automatique de collectages



Un module de transcription automatique basé sur la reconnaissance vocale aidera le travail des linguistes et enquêteurs.

- Démonstrations de la reconnaissance vocale basque et espagnole d'Elhuyar

/ La plateforme de contribution

Pour récolter une grande quantité d'enregistrement transcrits, et qui soient représentatifs de la diversité des locuteurs de l'occitan, Le Congrès a développé un outil de contribution pour la communauté. Sur cette plateforme, chacun peut enregistrer des phrases qui seront ajoutées au corpus construit avec les partenaires.

- Voir la vidéo de présentation
- Aller à la plateforme

/ La reconnaissance vocale, comment ça marche ?

La reconnaissance vocale utilise l'**intelligence artificielle** (les réseaux neuronaux) pour transcrire automatiquement la voix en texte écrit.

Avant de pouvoir le faire, il faut entraîner l'IA avec des phrases audios déjà transcrites. Il y a donc besoin d'un **grand corpus audio transcrit**, c'est à dire une grande quantité de texte avec les enregistrements audios correspondants.

Il faut également « nourrir » la machine avec de **grands corpus de texte seul**. Ainsi, elle peut apprendre quelles formes sont fréquentes, quel mot apparaît souvent à côté de tel autre...

Enfin, il faut **développer des programmes** :

- Un pour passer en lettres les nombres, les symboles, les abréviations, les unités de mesure... avant de donner un texte à la machine.
- Un « abréviateur » qui fait l'inverse, pour rendre plus lisibles les textes proposés aux utilisateurs.
- Un phonétiseur pour obtenir la prononciation en alphabet phonétique international d'un mot.
- Un programme pour avoir tous les mots qui correspondent à une prononciation.

/ Les étapes de l'entraînement et de la transcription

Entraînement du module de reconnaissance des sons

Enregistrements audio + transcriptions

⇓ chiffres en lettres ⇓

Enregistrements audio + transcriptions étendues

⇓ alignement des phrases ⇓

Enregistrements audios + transcriptions alignés

⇓ phonétisation des phrases ⇓

Enregistrements + transcriptions + phonétisations

⇓ entraînement IA ⇓

Base de correspondances sons-phonèmes

Entraînement du module de reconnaissance des mots

Corpus textuel

⇓ chiffres en lettres ⇓

Corpus textuel étendu

⇓ entraînement IA ⇓

Base de fréquence des mots en contexte

Transcription de texte

Fichier audio

⇓ sélection du phonème le plus probable par le module de reconnaissance des sons ⇓

Suite de phonèmes (transcription phonétique de l'audio)

⇓ sélection des mots les plus probables par le module de reconnaissance des mots ⇓

Transcription étendue

⇓ abréviateur ⇓

Transcription finale

/ Le calendrier

// 2020 : Définition des exigences, spécifications fonctionnelles et constitution du corpus

Une première étape consistera à décrire les exigences techniques, ainsi que les spécifications fonctionnelles.

D'un point de vue technique, les développements pour l'occitan seront réalisés dans l'état de l'art, à savoir par l'utilisation de réseaux neuronaux (Intelligence Artificielle). Mais cette technologie de pointe nécessite un nombre très important de données. Seul un corpus riche, volumineux et varié garantira un résultat de qualité en fin de chaîne.

Pour ce faire, le Congrès a engagé un partenariat avec plusieurs producteurs de contenus textuels multimédias en occitan : institutions, médias, éditeurs, producteurs de contenus audiovisuels...

C'est pour cela que cette première phase sera essentiellement consacrée à un travail de collecte, de traitement (alignement texte/son) et de stockage de corpus textuels et audios pour l'occitan. On estime à 200 heures environ le besoin de transcriptions et à 500 millions de mots le corpus textuel pour chaque variété. L'occitan étant une langue encore trop peu dotée, nous compenserons par l'utilisation de corpus géants du français et de l'espagnol en obtenant, grâce à la traduction automatique, des corpus textuels occitans importants.

// 2021 : Finalisation et développement technologique

Une grande partie du projet sera réalisée pendant cette phase : terminer la collecte des données nécessaires, réaliser trois des quatre lots de travail plus techniques pour arriver à une version avancée du développement. Concrètement, nous prévoyons au moins une mise en oeuvre avancée des modules suivants :

- Création du modèle linguistique.
- Création du modèle acoustique.
- Développement du transcripteur.

// 2022 : Développement final et validation

Dans la première partie de cette dernière phase, tous les développements technologiques du projet seront terminés. La phase de construction des transcripteurs sera également finie. Une fois intégrées toutes les composantes technologiques, elles seront soumises à une série de tests intensifs d'évaluation.

/ Les acteurs

// Membres et soutiens

ReVOc est un nouveau programme de développement de la reconnaissance vocale en occitan (variétés gasconne et languedocienne) engagé par le Congrès permanent de la langue occitane. Ce programme triennal (2020-2022) se déroule dans le cadre d'un partenariat transfrontalier qui associe l'institution aragonaise Rolde de Estudios Aragoneses (qui développera la même technologie pour la langue aragonaise) et la fondation basque Elhuyar (en charge de la partie technique du programme). Il est soutenu financièrement par la Région Nouvelle-Aquitaine, la Région Occitanie et le Département des Pyrénées-Atlantiques (appels à projets transfrontaliers).

// Le partenariat pour la constitution des corpus

Pour entraîner l'intelligence artificielle, il faut de grandes quantités de données. L'occitan, langue dite « peu dotée », n'a généralement pas ces grands ensembles de données. Pour constituer le corpus audio et le corpus textuel nécessaires au développement de la reconnaissance vocale, Le Congrès ne pouvait pas œuvrer seul.

Un partenariat, d'une envergure inédite en ce qui concerne l'occitan, a donc été construit, pour constituer une base audio et textuelle à laquelle plusieurs structures sont venues apporter leur contribution.

Et aussi :

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

/ Le Congrès permanent de la langue occitane

Le Congrès permanent de la langue occitane est l'organisme interrégional de régulation de l'occitan. Il œuvre dans les domaines de la linguistique et du TAL (traitement automatique de la langue).

Il produit des outils linguistiques numériques de référence (dictionnaires, conjugueurs, correcteurs orthographiques...), des applications pour le TAL (synthèse vocale, traduction automatique...) et des applications pour les mobiles (claviers prédictifs...).

Il a également des missions de régulation linguistique et de recherche scientifique appliquée.

Il est l'éditeur d'un multidictionnaire occitan (dicod'Òc) qui a chaque année plus d'un million de visites.

/ Pour plus d'informations :

Le Congrès permanent de la langue occitane
Château d'Este, Avenue de la Pléiade
64140 Billère

premsa.locongres.com/revoc

info@locongres.org

05 59 13 06 40