Ouvrir aux linguistes « de terrain » un accès à la transcription automatique

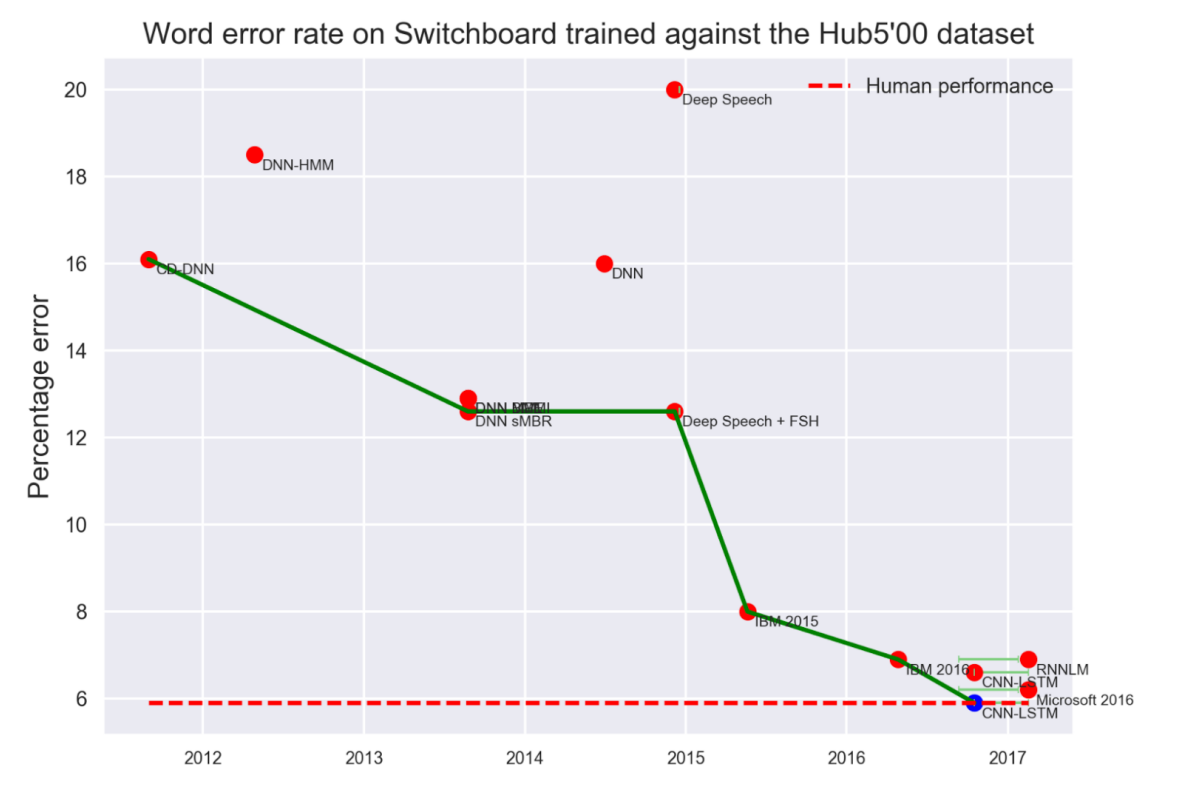
Journées LIFT 2020

Guillaume Wisniewski, Alexis Michaud, Benjamin Galliot, Laurent Besacier, Séverine Guillaume, Katya Aplonova et Guillaume Jacques

Les besoins des linguistes de terrain

- matériaux de base : enregistrement + transcription
- coût de la transcription élevé : 1h d'enregistrement $\leftrightarrow$ 50h à 100h de travail
- transcrire tout le corpus enregistré = premier défi pour les linguistes de terrain

Les promesses de la reconnaissance de la parole (1)



La « révolution » de l'apprentissage profond en TAL :

$\hookrightarrow$ puissance de calcul $\oplus$ algorithmes d'apprentissage $\oplus$ données

$\hookrightarrow$ performances « remarquables » (reco, trad, ...)

Les promesses de la reconnaissance de la parole (2)

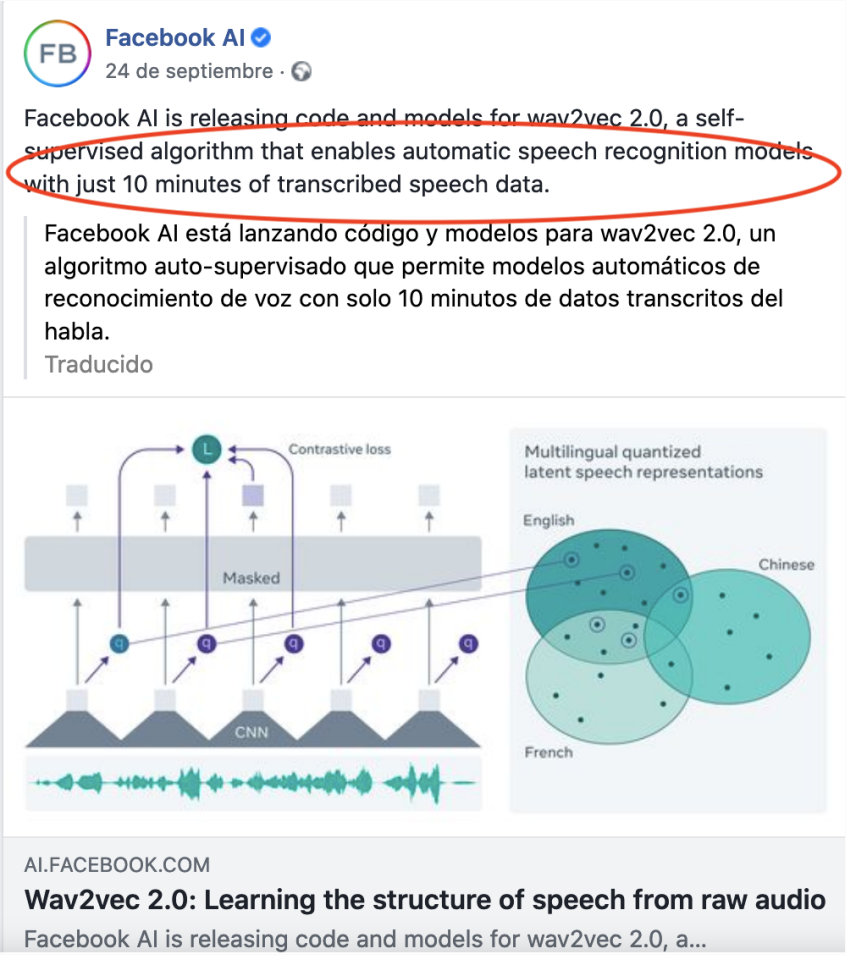
The next frontier will be unsupervised learning Y. LeCun

↪ de plus en plus de techniques pour réduire la quantité de données annotées requise

↪ TAL aujourd'hui : beaucoup de données non annotées (textes & vidéos issus du web) + « quelques » données annotées

Ces méthodes peuvent-elles faciliter le travail des linguistes de terrain ?

Les promesses de la reconnaissance de la parole (2)



The next frontier will be unsupervised learning Y. LeCun

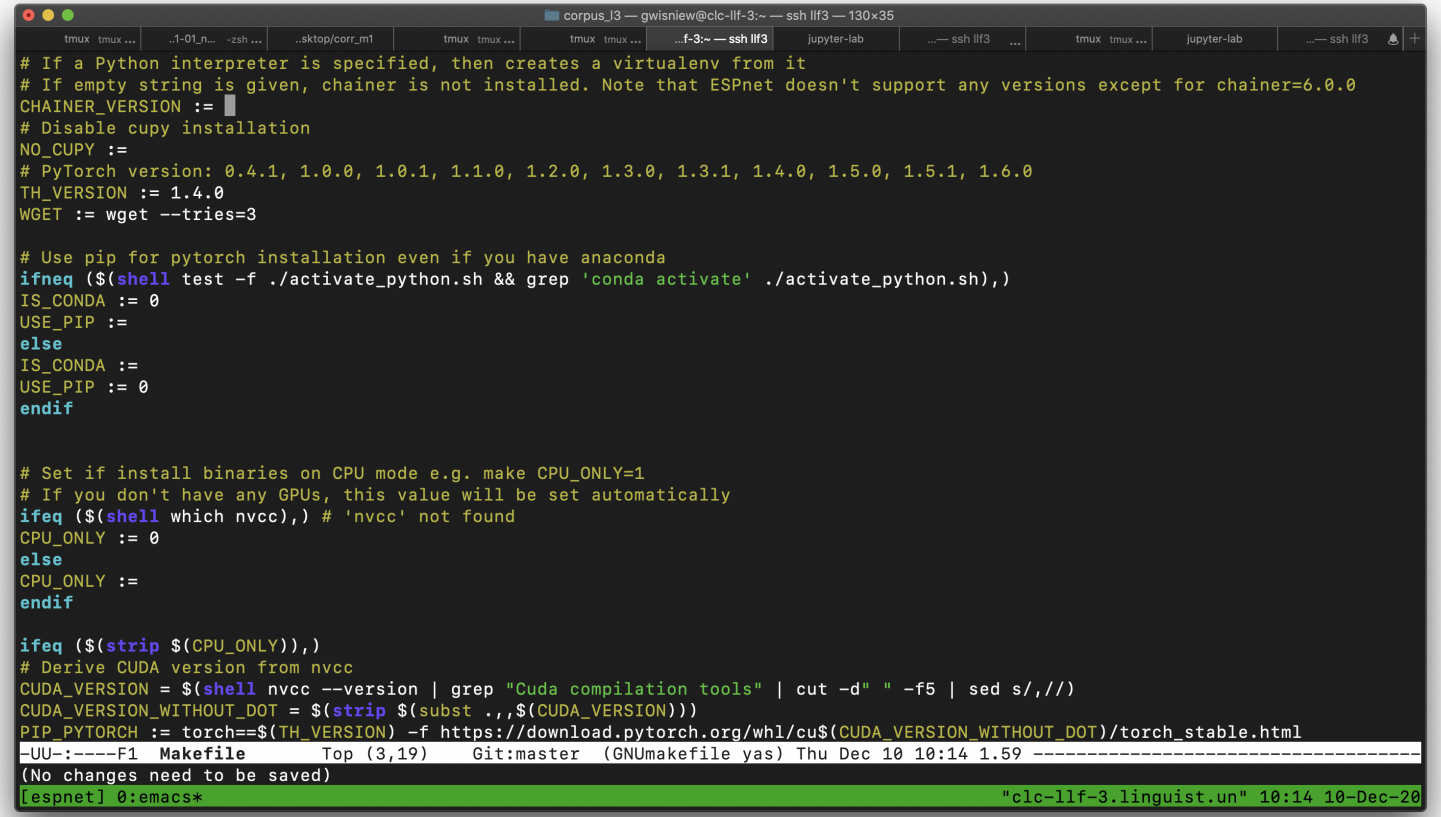
↪ de plus en plus de techniques pour réduire la quantité de données annotées requise

↪ TAL aujourd'hui : beaucoup de données non annotées (textes & vidéos issus du web) + « quelques » données annotées

Ces méthodes peuvent-elles faciliter le travail des linguistes de terrain ?

Une première difficulté

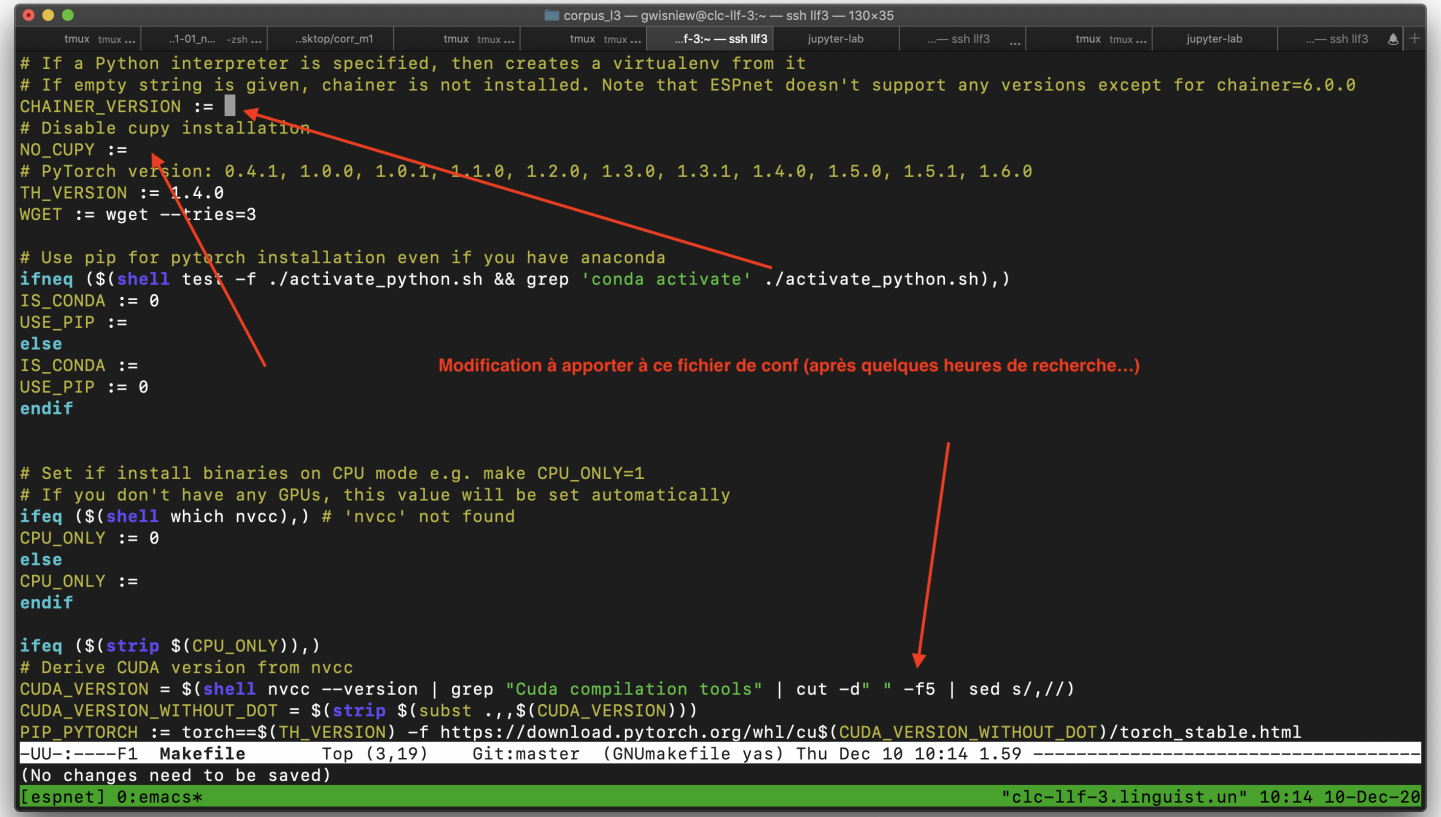
Installation ESPnet



Utiliser des outils de TAL nécessite une (bonne) maîtrise de la ligne de commande (euphémisme)

Une première difficulté

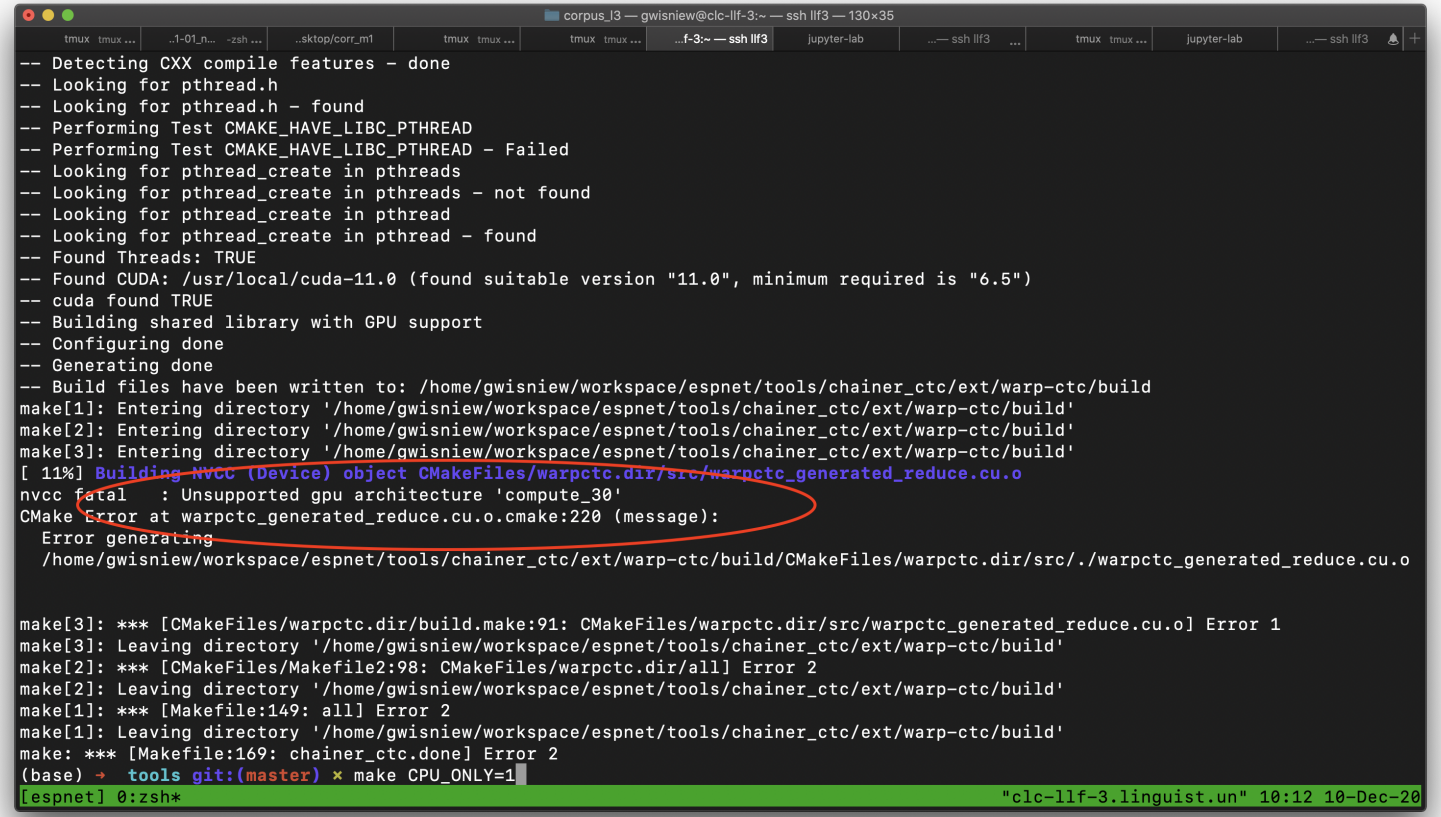
Installation ESPnet



Utiliser des outils de TAL nécessite une (bonne) maîtrise de la ligne de commande (euphémisme)

Une première difficulté

Installation ESPnet



Utiliser des outils de TAL nécessite une (bonne) maîtrise de la ligne de commande (euphémisme)

Une première difficulté

Installation ESPnet

Utiliser des outils de TAL nécessite une (bonne) maîtrise de la ligne de commande (euphémisme)

Le projet Elpis

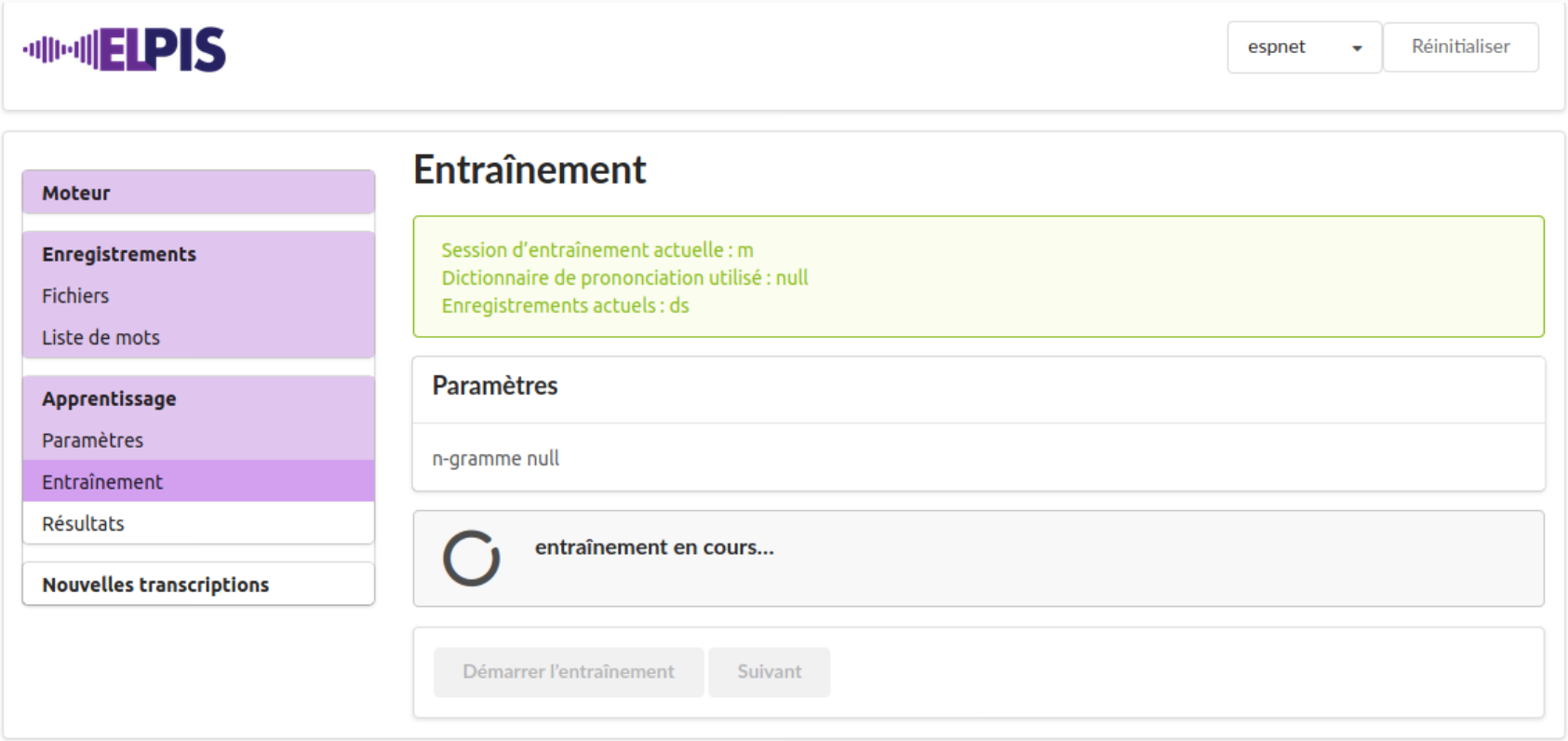
Le projet Elpis `https://github.com/CoEDL/elpis`

- fournir aux linguistes de terrain un accès aux techniques de reco
- à l'aide d'une interface « utilisable »

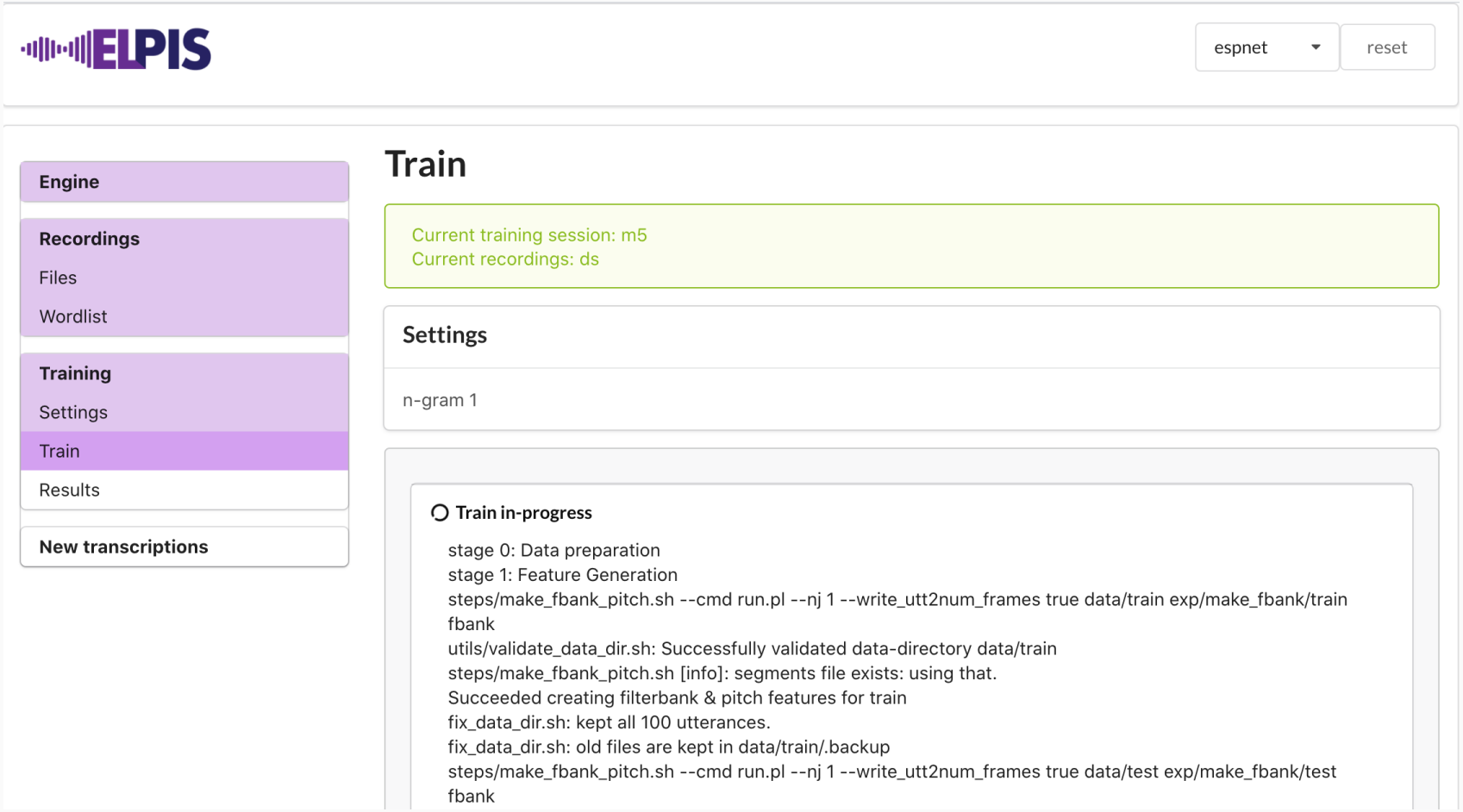
Dans le cadre de ce travail :

- intégration du moteur de reconnaissance ESPnet
- facilitation de l'installation (image Docker)
- chaîne de conversion : Pangloss $\leftrightarrow$ Elan $\rightarrow$ Elpis
- traduction de l'interface
- expériences sur de nouvelles langues

Au final :

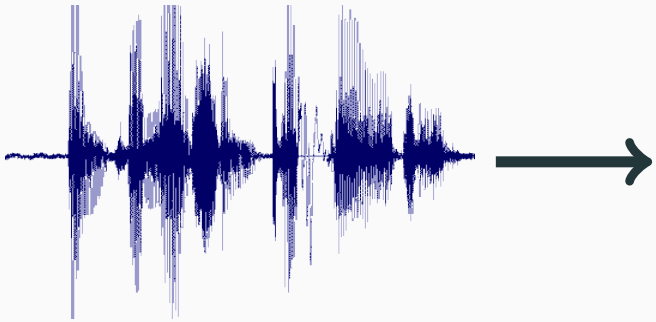

ELPIS
espnet
Réinitialiser
Moteur
Enregistrements
Fichiers
Liste de mots
Apprentissage
Paramètres
Entraînement
Résultats
Nouvelles transcriptions
Entraînement
Session d'entraînement actuelle : m
Dictionnaire de prononciation utilisé : null
Enregistrements actuels : ds
Paramètres
n-gramme null
entraînement en cours...
Démarrer l'entraînement
Suivant

Au final :


ELPIS
espnet
reset
Engine
Recordings
Files
Wordlist
Training
Settings
Train
Results
New transcriptions
Train
Current training session: m5
Current recordings: ds
Settings
n-gram 1
Train in-progress
stage 0: Data preparation
stage 1: Feature Generation
steps/make_fbank_pitch.sh --cmd run.pl --nj 1 --write_utt2num_frames true data/train exp/make_fbank/train fbank
utils/validate_data_dir.sh: Successfully validated data-directory data/train
steps/make_fbank_pitch.sh [info]: segments file exists: using that.
Succeeded creating filterbank & pitch features for train
fix_data_dir.sh: kept all 100 utterances.
fix_data_dir.sh: old files are kept in data/train/.backup
steps/make_fbank_pitch.sh --cmd run.pl --nj 1 --write_utt2num_frames true data/test exp/make_fbank/test fbank

Résultats expérimentaux (1)

Reconnaissance phonémique



→ ˈkɯɕɯŋgɯ kɯɕɯŋgɯ tɕe, tɤru ʁnɯs pjɤ-tú-ndʑi.

↪ sortie = séquence de caractères (y compris espaces)

↪ pour quatre langues

Résultats expérimentaux (2)

Résultats

Langue	Nb locuteurs	Type	Taille (mn)	CER (%)
Na	1	*Récits spontanés*	273	14.5
Na	1	*Expressions élicitées*	188	4.7
Chatino	1	*Parole lue*	81	23.5
Japhug	1	*Récits spontanés*	170	12.8
Bashkir	36	*Récits spontanés*	273	33

↪ Résultats dépendent fortement du corpus : nombre de locuteurs, quantité de données, conditions d'enregistrement, genres

↪ loin des performances de la reco « traditionnelle »...

↪ mais d'une qualité suffisante pour être utile ?

Intérêt pour les TAListes

We are not coding monkey ! !

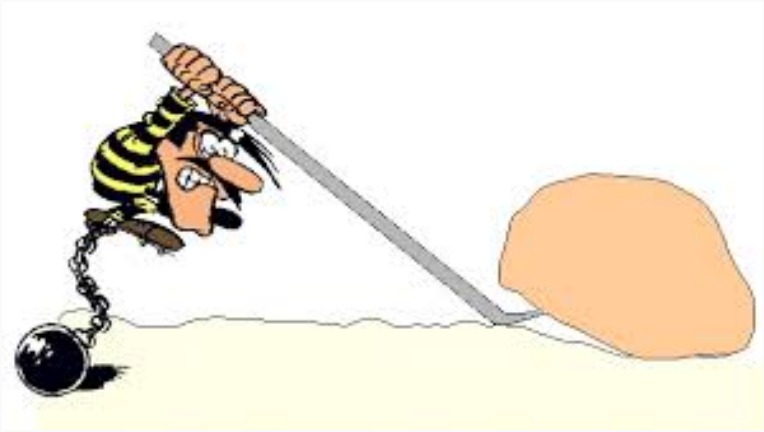
↪ nombreux défis techniques

↪ encore plus de défis scientifiques

Les défis qui nous attendent

1. *one model to rule them all*
    - comment adapter les modèles aux spécificités de chaque langue ?

Les défis qui nous attendent



1. *one model to rule them all*
2. tirer avantage des données non annotées

Les défis qui nous attendent

1. *one model to rule them all*
2. tirer avantage des données non annotées
3. réduire la quantité de données nécessaires
   - $\hookrightarrow$ annotées
   - $\hookrightarrow$ non annotées

   $\Leftarrow$ partage de paramètres entre langues, pré-entrainement, self-training, ...

Les défis qui nous attendent

1. *one model to rule them all*
2. tirer avantage des données non annotées
3. réduire la quantité de données nécessaires
4. interactions avec les linguistes de terrain
   - ↪ a-t-on assez de données ?
   - ↪ les performances sont-elles suffisantes ?

Merci pour votre attention

Remerciements

Merci à Oliver Adams (qui a identifié ESPnet comme un outil prometteur, et a réalisé son intégration dans Elpis), et à l'équipe du projet Elpis : Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles. Nous remercions également Christopher Cox, qui a réalisé un greffon *(plugin)* de transcription automatique pour le logiciel Elan ; Nick Evans, Nick Thieberger, Steven Morey, qui jouent un rôle important dans la coordination internationale du projet Elpis et sa diffusion auprès des linguistes ; et Hilaria Cruz, Martine Adda, Graham Neubig et Nathan Hill pour leur appui.

Nous remercions l'Institut des langues rares (ILARA) de l'École Pratique des Hautes Études, l'Université du Queensland et l'*Australian Research Council Centre of Excellence for the Dynamics of Language* pour le soutien financier apporté au développement d'outils de transcription automatique pour la documentation linguistique. Le présent travail est en outre une contribution au projet Labex « Fondements empiriques de la linguistique » (ANR-10-LABX-0083) ainsi qu'au projet « La documentation computationnelle des langues à l'horizon 2025 » (ANR-19-CE38-0015-04).