Transcription automatisée et progression vers l'open source

Max Beligné – Ingénieur de recherche
Plateforme Universitaire de Données Grenoble Alpes (PUD-GA)
max.beligne@univ-grenoble-alpes.fr

OSDays
Novembre 2025

Éléments pour comprendre d'où je parle ?

- **Ingénieur de recherche** à la *Plateforme Universitaire de Données - Grenoble Alpes (PUD-GA)*
  (*Maison des Sciences Humaines Alpes*, relai de l'*IR* Progedo*)
  Appui à la recherche en SHS

- **Participation et développement de la plateforme TADDDAM :**
  (**T**ransformations, **A**nalyses et **D**éveloppements de **D**onnées, **D**ocuments et **A**rchives **M**ultimédias)
  service porté par la *PUD-GA* et le laboratoire *PACTE* avec le soutien de *GRICAD*
  connu à l'*UGA* pour son **service de transcription** (https://tadddam.univ-grenoble-alpes.fr)

- **Animation d'un groupe de travail TIPS-IA** (https://mate-shs.cnrs.fr/les-groupes/groupes-thematiques/tips-ia/)
  (**T**ranscription, **I**nterface, **P**ipeline et **S**ynergie IA)
  au sein du **réseau national *MATE-SHS* pour mutualiser nos réflexions et actions**
  avec plusieurs financements : *Progedo*, *Huma-Num* et *URFIST*

Introduction

- Transcription : transformation de l'audio en texte (« speech to text »)
- Champ de recherche et d'application bien antérieur au deep learning
- Un logiciel historique qui à été très utilisé :

- Très utilisé des années 2000 jusqu'à milieu des années 2010
- 60 ou 250 euros (version standard ou pro)
- Préapprentissage sur la voix de l'utilisateur
- Rapidité équivalente à un dactylo moyen/avancé
  (environ 6,7h de travail
  pour une transcription simple d'un entretien d'1h)

Suite du « previously » …

Au début des années 2020, les technologies évoluent avec la progression de l'IAG :

- Wav2Vec
- Wav2Vec 2.0, HuBert

➡ Apparition de nouveaux services comme Noota :

- Pour l'utilisateur,
  pas de pré-entrainement avec sa voix
  pas besoin de dicter
- 15 euros / mois dans la formule de base
- Gain de temps important

https://www.logiciels.pro/logiciel-saas/noota/

Le moment « Whisper »

- Modèles (du tiny 39 millions de paramètres au large 1550 millions) lancés par OpenAI fin septembre 2022 : https://github.com/openai/whisper
- A été entrainé sur 680 000 heures d'audio (corpus non diffusé)
- Architecture encoder-decoder : **facilité d'usage**
- **Amélioration nette des performances**
  (Avec Whisper large, environ 50 % d'erreurs en moins par rapport à Wav2Vec 2.0 sur plusieurs benchmarks)
- **Robustesse** : avec de l'audio bruité, arrive à certains résultats
  (il y a une perte de performance mais ne s'écroule pas)
- **Diversité linguistique** : 96 langues différentes
  (avec des performances inégales mais tranche avec les modèles spécialisés préexistants)

Impacts macro

- Démocratisation de la retranscription automatisée
- Gain de temps mais aussi financier
  (prestataires externes ou vacataires souvent payés pour ce travail)
- Impacts négatifs aussi :

  - RGPD (sans information, beaucoup passent par ChatGpt)

  - Des usages inquiétants :

    https://pulitzercenter.org/stories/researchers-say-ai-powered-transcription-tool-used-hospitals-invents-things-no-one-ever

Plan

Par rapport à cette introduction très macro
l'objectif est de rentrer dans une compréhension plus fine et micro

I) Une adoption de la transcription automatisée à la fois lente et rapide

II) Quelques impacts de cette adoption dans le cadre d'enquête par entretiens en SHS

III) Transition vers des modèles plus petits et totalement open-source

I) 1) une adoption « lente »

- Whisper a surement été éclipsé par la sortie ChatGpt deux mois après.
- Beaucoup de personnes potentiellement intéressées
  mais au départ pas au courant
- Apparait dans le scope des sciences sociales cinq mois après le lancement
  https://www.css.cnrs.fr/using-whisper-to-transcribe-oral-interviews/
  février 2024, Yacine Chitour et Julien Boelaert

➡ Ce qui est mis en avant dans cet article de blog
c'est un exemple sur google colab
et en effet, c'était la solution la plus simple à l'époque

Des premières expérimentations… à l'idée d'un service en ligne

- 1er essai sur le JupyterHub de l'*UGA* :
  ok avec les plus petits modèles (whisper tiny, base…)
  mais ne marche pas avec les modèles les plus gros qui sont les plus performants.

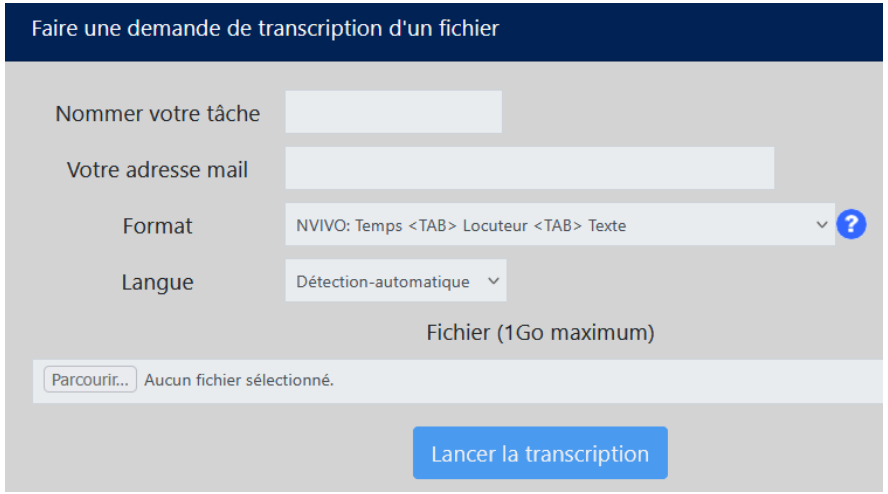
- Passage et apprentissage sur l'infrastructure de calcul intensif de l'*UGA* (*Gricad*)

- Idée du service en ligne après discussion avec Patrick Juen (*PACTE* / *Gricad*)
  car disposait des connaissances pour le faire

Formalisation du service

Fichier
audio ou vidéo

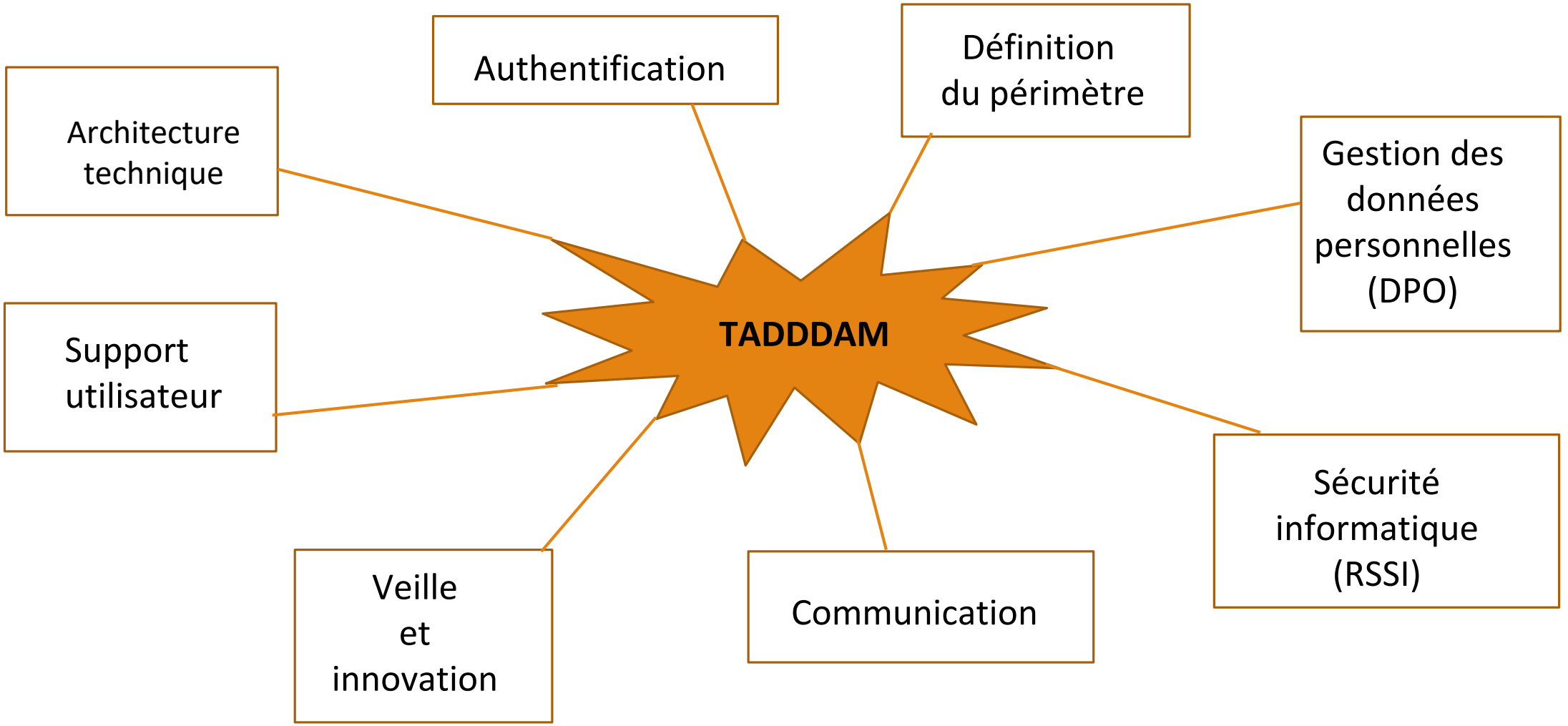
Interface
pour déposer son fichier
et choisir paramètres



Traitement
sur les serveurs
de l'UGA (GRICAD)

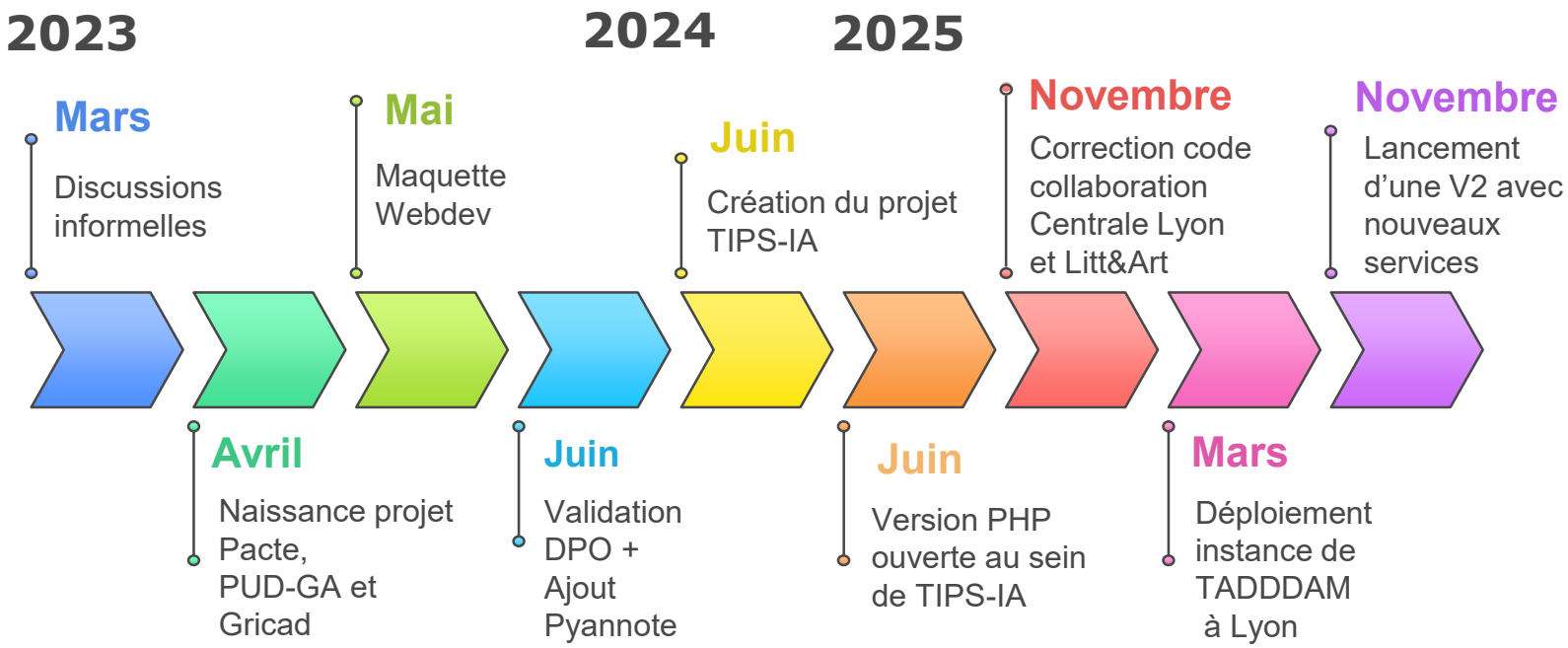
Récupération
du résultat :
fichier texte

Complexité d'un tel service de transcription



I)2) Une adoption rapide !

- Lancement officiel du service TADDDAM en novembre 2023
  En 2 ans, **850 utilisateurs différents et 15 000 transcriptions effectuées**



- **Au niveau national, croissance forte aussi**
  Exemple du service Huma-Num via Sharedoc
  Investissement 55k€ dans nouveaux GPU pour ce service de transcription

Au niveau technologique, beaucoup de dynamiques autour de Whisper

- Optimisation : temps / performance

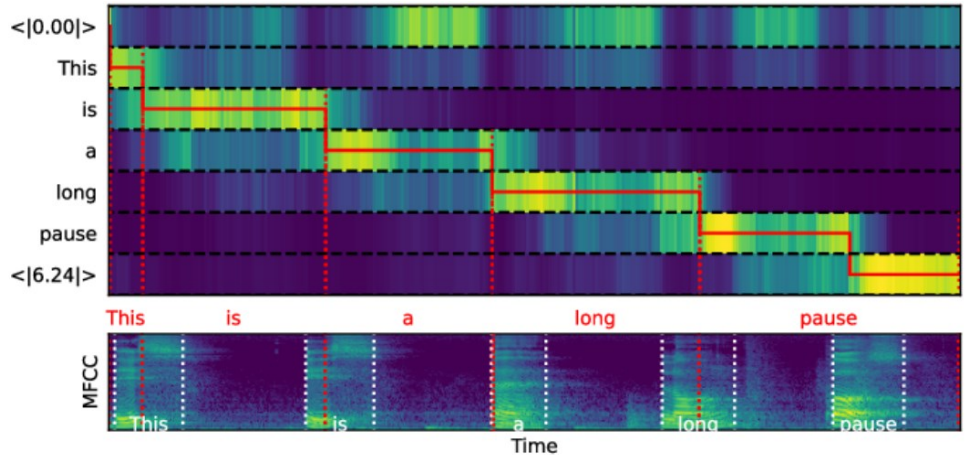
  faster-whisper : perte d'un peu de qualité
  mais possible de diviser le temps de traitement par 4
  insanely-fast-whisper : encore plus rapide

- Fine-tuning : amélioration par exemple pour des cas spécifiques

  pour des langues moins représentées, pour des domaines spécialisés…

- Ajout technologique

  exemple DTW (Dynamic Time Warping)
  pour avoir les timestamps au niveau du mot



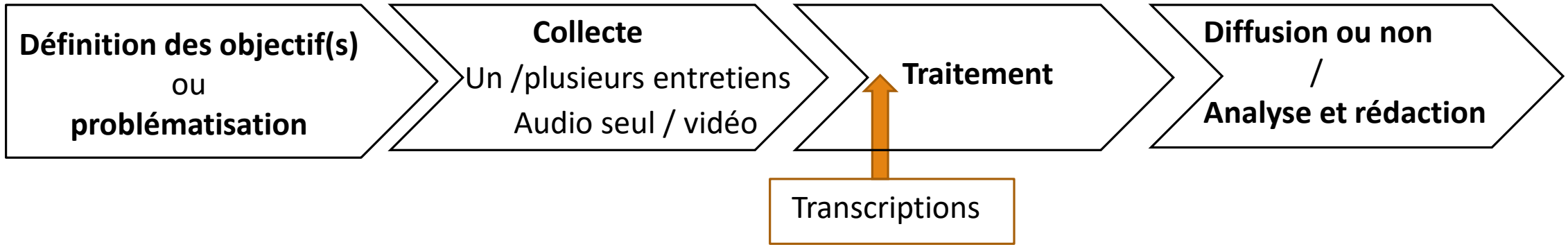
CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions

Partie 2 :

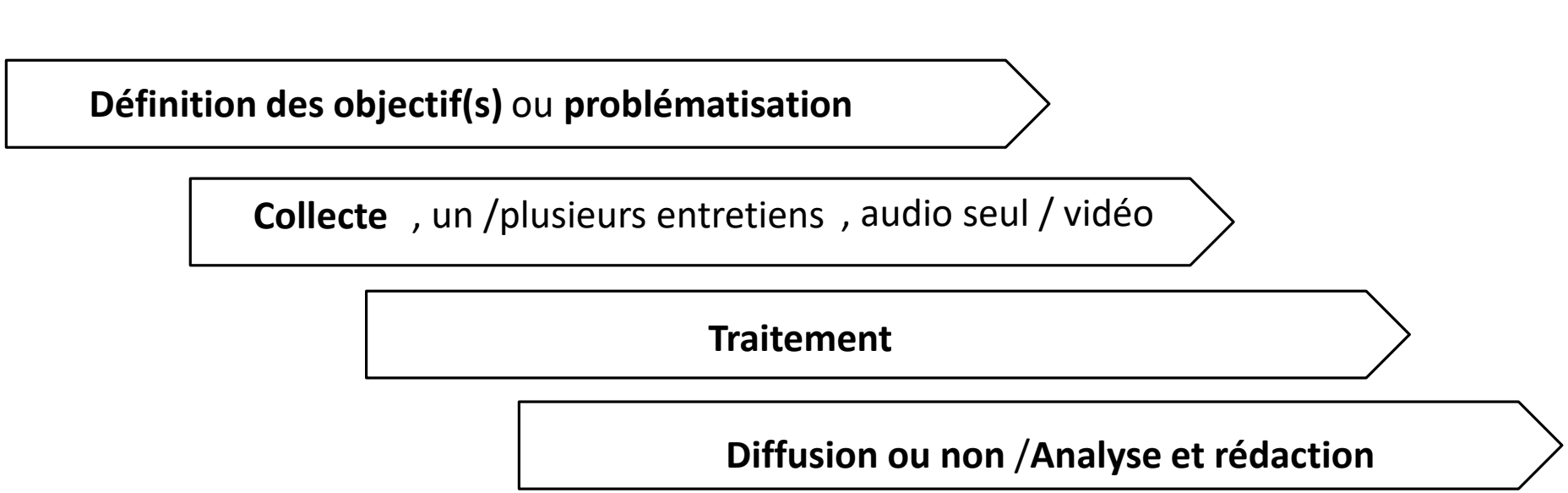
Quelques impacts de cette adoption dans le cadre d'enquête par entretiens en SHS

Rappel : Une étape dans une chaîne

Approche linéaire



Approche parallèle



Multiples questions dans ce moment de changement

- Par rapport à l'augmentation du nombre d'entretiens possible
  → est-ce que ça **change les méthodologies** qui auraient tendance à devenir **plus quantitatives** ?

- Par rapport à la diminution du temps de transcription (qui est aussi un premier moment d'analyse)
  → est-ce que ça **change le rapport au matériau** de recherche qui serait **plus distancié** ?

- Est-ce qu'il y des **différences disciplinaires dans les évolutions** ?
  car globalement, on ne retranscrit pas de la même manière par exemple en sociologie et en linguistique.

- Est-ce qu'au fur et à mesure que l'usage se généralise,
  les utilisateurs **deviennent moins critiques par rapport aux impacts** de la transcription automatisée ?

➡ Nécessité d'enquêter …

Projet « Chuchotons » à Rennes (https://eso.cnrs.fr/fr/node/projet-de-recherche/22818/15000/chuchotons)
+ relai au niveau national avec TIPS-IA

Un élargissement progressif de l'intérêt dans TIPS-IA

Grâce à dynamiques collectives + financements + refonte ou nouvel outil lié à l'IA

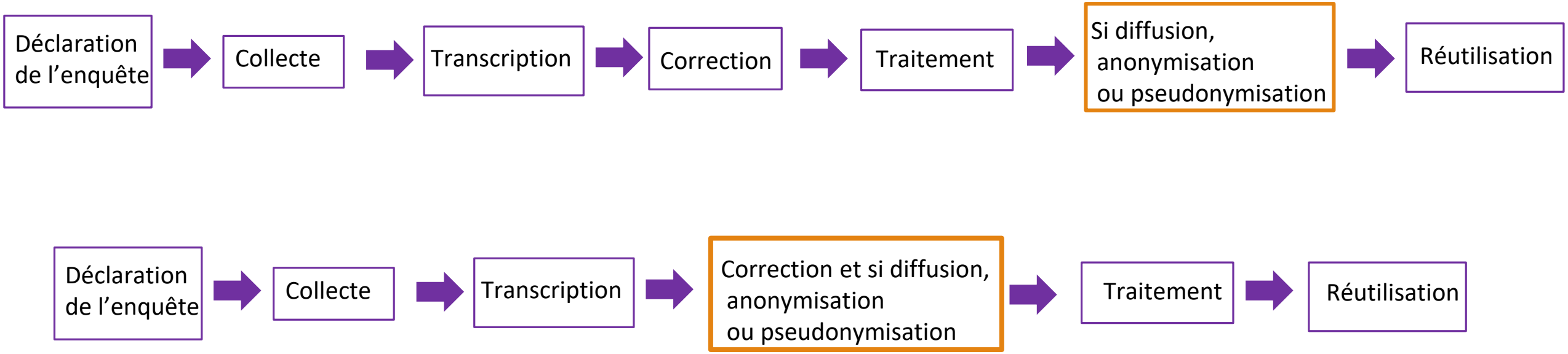
Transcription

Transcription → Correction

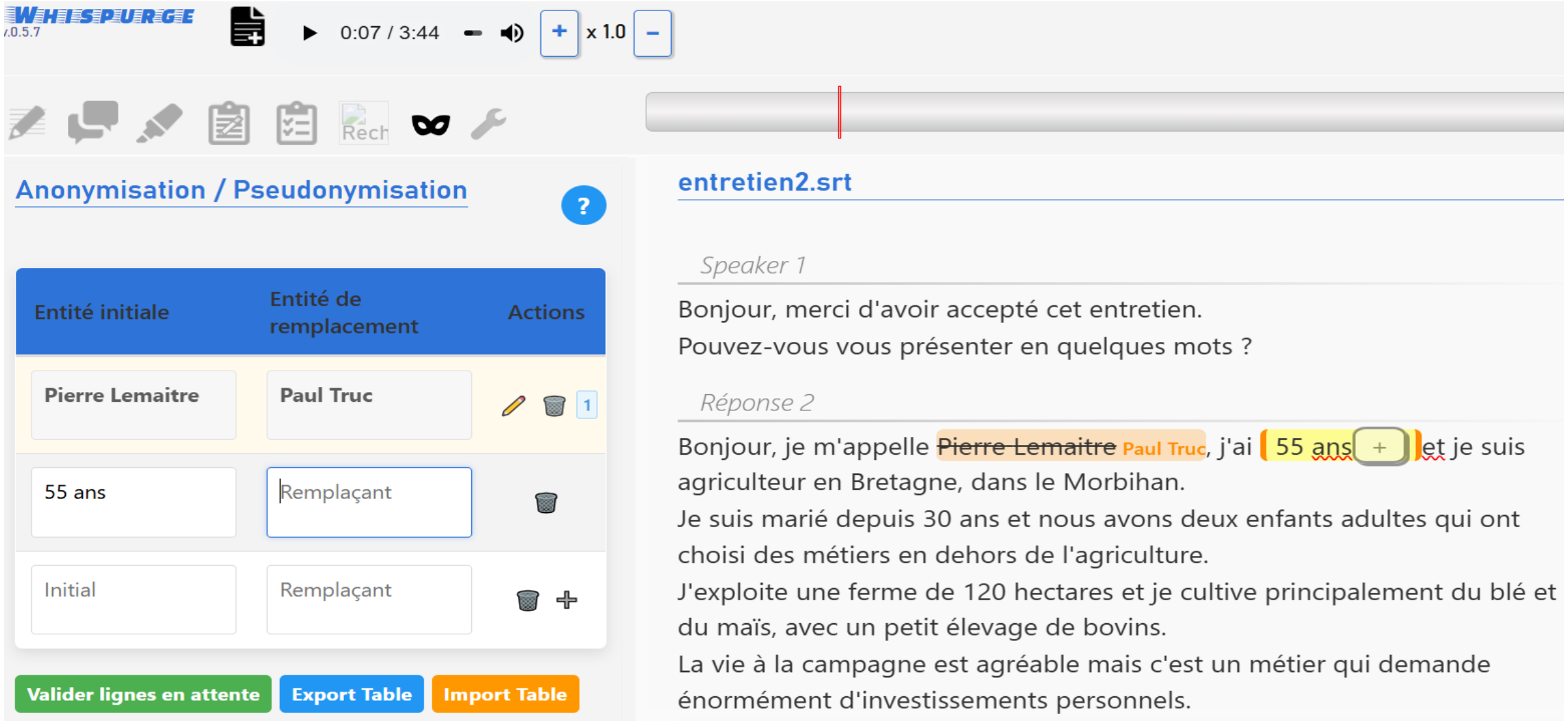
Transcription → Correction → Traitement → Si diffusion, anonymisation ou pseudonymisation

Déclaration de l'enquête → Collecte → Transcription → Correction → Traitement → Si diffusion, anonymisation ou pseudonymisation → Réutilisation

Une volonté de modifier la chaîne



Travail en cours sur Whispurge / Sonal



Partie 3 :

Transition vers des modèles plus petits et totalement open-source

Un espace de référence (https://huggingface.co/spaces/hf-audio/open_asr_leaderboard)

The Open ASR Leaderboard ranks and evaluates speech recognition models on the Hugging Face Hub.
We report the Average WER (lower the better) and RTFx (higher the better). Models are ranked based on their Average WER, from lowest to highest. Check the Metrics tab to understand how the models are evaluated.
If you want results for a model that is not listed here, you can submit a request for it to be included.
The leaderboard includes both English ASR evaluation and multilingual benchmarks across the top European languages.

Leaderboard | Multilingual | Long-form | Metrics | Request a model here! | About

model	Average WER	RTFx	License	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeech	Tedlium	Voxpopuli
nvidia/canary-qwen-2.5b	5.63	418.28	Open	10.1	10.45	9.43	1.61	3.1	1.9	2.71	5.66
ibm-granite/granite-speech-3.3-8b	5.74	145.42	Open	8.98	9.42	10.19	1.43	2.86	3.91	3.4	5.72
ibm-granite/granite-speech-3.3-2b	6	270.57	Open	8.9	10.25	10.69	1.53	3.26	3.87	3.57	5.93
microsoft/Phi-4-multimodal-instruct	6.02	151.1	Open	11.6	10.16	9.33	1.69	3.82	3.06	2.94	6.04
nvidia/parakeet-tdt-0.6b-v2	6.05	3386.02	Open	11.1	11.15	9.74	1.69	3.19	2.17	3.38	5.95
aquavoice/avalon-v1-en	6.24	NA	Proprietar	11.5	11.37	9.52	1.68	3.28	2.1	3.02	7.33
nvidia/parakeet-tdt-0.6b-v3	6.32	3332.74	Open	11.3	11.19	9.57	1.92	3.59	3.98	2.8	6.09
nvidia/canary-1b-flash	6.35	1045.75	Open	13.1	12.77	9.85	1.48	2.87	1.95	3.12	5.63
kyutai/stt-2.6b-en	6.4	88.37	Open	12.1	10.99	9.81	1.7	4.32	2.03	3.35	6.79
nvidia/canary-1b	6.5	235.34	Open	13.9	12.19	10.12	1.48	2.93	2.06	3.56	5.79
nyrahealth/CrisperWhisper	6.67	84.05	Open	8.71	12.89	10.24	1.82	4	2.7	3.2	9.82

De gros enjeux sur l'efficacité des petits modèles : la transcription synchrone (stream…)

- Change en profondeur notre manière d'interagir avec tous les outils numériques
- Pipeline classique "Voice Agent" : Speech To Text → NLP / LLM → Text to Speech
- En combinaison avec des agents IA capable de réaliser des actions.
- Possible aussi que pour tout une partie, ça ne passe plus par du texte,
  et donc saute l'étape de transcription
  modèles « speech to speech » ou « speech to agent » (moins de latence)

Des étapes récentes et importantes pour l'open-source

2 modèles de Nvidia :
- https://huggingface.co/nvidia/canary-1b-v2
- https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3

Leur dataset d'entrainement : https://huggingface.co/datasets/nvidia/Granary

1 millions d'heures d'audio dans 25 langues européennes
avec leur transcription et traduction si différent de l'anglais

Le pipeline pour créer de nouvelle données : https://github.com/NVIDIA/NeMo-speech-data-processor/tree/main/dataset_configs/multilingual/granary

Autre initiative intéressante : OLMoASR (détaille toutes les étapes)

https://github.com/allenai/OLMoASR

Vers des architectures plus complexes

Un des travaux de l'année prochaine :

En fonction de la requête de l'utilisateur, utiliser le modèle le plus adapté

Exemple fictif car dépendra de l'évolution technologique :

Whisper-large-V3 pour les langues non européennes
Possible Parakeet-V4 pour les langues européennes
Un autre modèle pour ceux qui sont intéressés par les disfluences (hésitation, répétitions,…)

Idem pour la diarisation :

Sortformer2 si pas plus de 4 locuteurs
pyannote dans les autres cas

Conclusion

- Aventure qui m’a conduit beaucoup plus loin que tout ce qui j’aurais pu imaginer
- Importance du travail collectif
- Information et pédagogie dès le travail de Master
  car service de transcription souvent inconnu
  et utilisation fréquente de ChatGPT
- Pédagogie aussi ensuite pour éviter que ce service soit vecteur de mauvaises pratiques
- Utilité de garder la main sur cet outillage