Journée finale du projet ANR ANTRACT
Sophia-Antipolis, mardi 14 juin 2022

Une relecture de l'expérience d'accompagnement de l'utilisation de TXM au sein du projet ANTRACT : fécondité et limites de l'interdisciplinarité

Bénédicte PINCEMIN
IHRIM, université de Lyon, CNRS

Itinéraire

- Point sur le chemin parcouru
  - Historique des versions pour les deux principaux corpus, NOTICES et VOIX-OFF
  - Mémo des avancées récentes
  - Quelques apports du projet pour la textométrie et TXM
- Les richesses et limites de l’expérience de recherche en Histoire
  - Niveau/ampleur d’usage vs qualité/pertinence scientifique
  - Rendez-vous manqué sur les documents collaboratifs
  - Limites rencontrées sur l’usage textométrique
  - Qualité humaine et scientifique de l’équipe

Petit historique des deux principaux corpus
(AF-NOTICES et AF-VOIXOFF)

Quand	***Quoi***	***Descriptif***
Jan. 2018	NOTICES-V0	Import des 23 114 notices documentaires → premier aperçu des possibilités textométriques
Fév. 2018	NOTICES-V1	Structuration des descripteurs et de la date
Oct. 2018	NOTICES-V2	Amélioration des métadonnées. Formation.
Oct. 2018	VOIXOFF-V0	Import des 12 298 transcriptions automatiques, 1er aperçu (ASR3 fait en allant lire le passage vidéo associé à chaque sujet)
Avr.2019	VOIXOFF-V1	Ajout des métadonnées. Retour à la vidéo, prototype. (ASR3b sur 10 000 vidéos découpées et nommées par sujet)

Les (quasi)-doublons d'AF-VOIXOFF-V1

AFVOIXOFFV01/<foule>

Query foule

ref	Left context	Pivot	Right context
AFE86004649, S55, 0:06:31	j'un état martin etan il y a	foule	était de pour encourager les rescapés de ce tour d
AFE85003165, S8, 0:02:33	non quinzième étape martha état il y a	foule	des belges pour encourager les rescapés de ce tou
AFE86000584, S8, 0:00:20	de s'ouvrir à paris il y a	foule	depuis quelque temps au magique le saint des sai
AFE86004035, S3, 0:00:33	de s'ouvrir à paris il y a	foule	depuis quelques temps où magie club le saint des
AFE86000447, S11, 0:01:13	dimanche matin place louis quatorze il y a	foule	à l'église notre-dame des victoires riquier deux mi
AFE86004010, S51, 0:01:24	dimanche matin place louis quatorze il y a	foule	l'église notre-dame des victoires riquier deux milli
AFE85008648, S24, 0:00:30	le président de gaulle au milieu des acclamations	foule	généralement réservé jusqu'à buckingham palace
AFE86003718, S78, 0:08:36	le président de gaulle au milieu des acclamations	foule	généralement réservé jusqu'à buckingham palace
AFE86004447, S74, 0:08:18	taire c'était en dentelle débile une autre	foule	attendait le geste symbolique bon c'est oui et on r
AFE86003239, S40, 0:02:45	de devant l'hôtel de ville une autre	foule	attendait le geste symbolique oui et non et en rem
AFE86004608, S1, 0:00:06	découverte la terre dans la connaître d'autres	foule	et n'aura -t -on montre à travers paris centaines de
AFE85008338, S9, 0:00:05	il y avait	foule	à reims un jour de l'été mille neuf cents m le

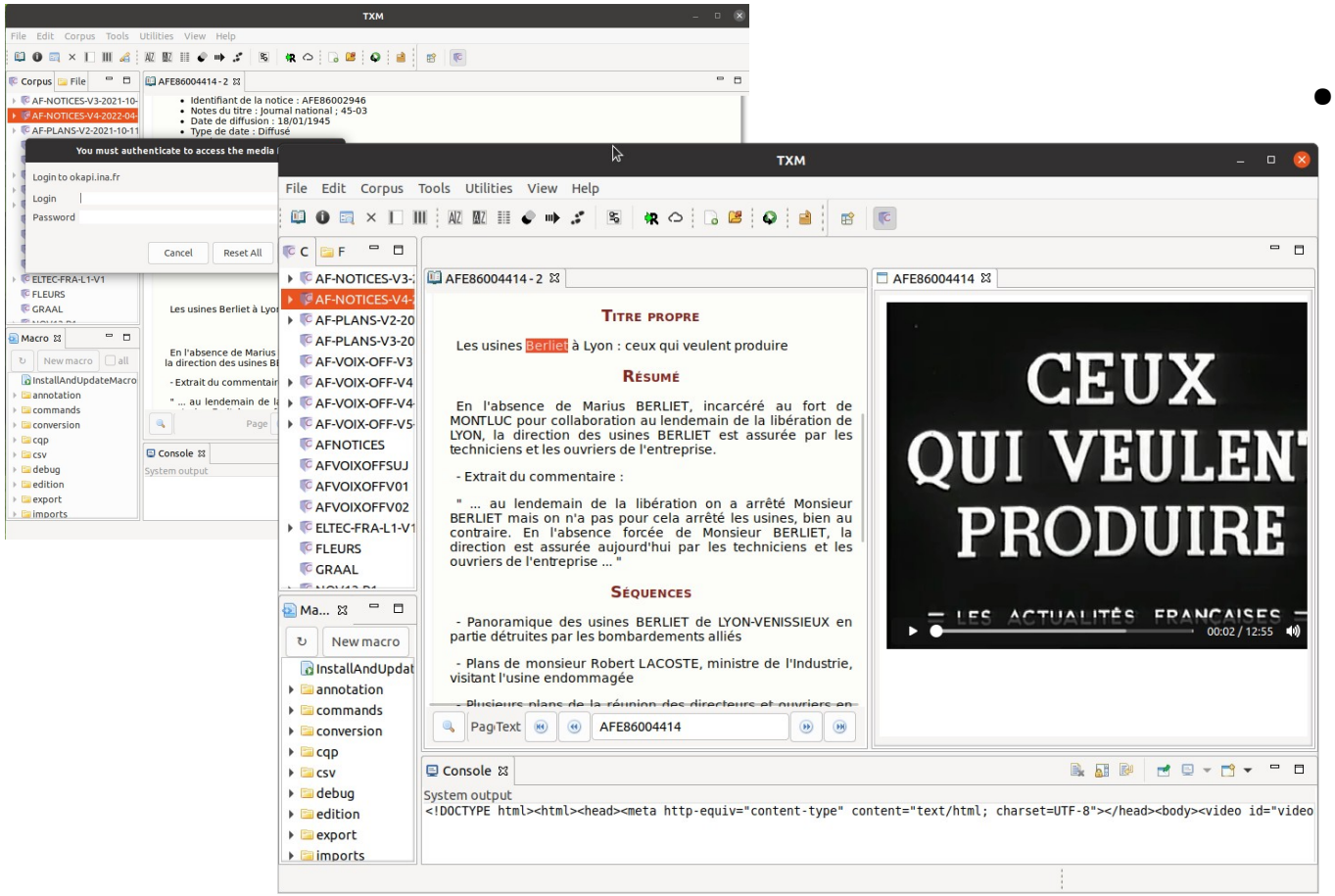
1 - 100 / 978

Rq. : on a aussi affaire ici aux premières transcriptions automatiques, avant que l'outil soit adapté/entraîné pour ces données d'archives (il est mis en difficulté par la qualité ± dégradée de l'enregistrement, la façon de parler qui a beaucoup changé, l'évolution du vocabulaire, etc.)

Petit historique des deux principaux corpus
(AF-NOTICES et AF-VOIXOFF), suite

Quand	***Quoi***	***Descriptif***
Nov. 2019	VOIXOFF-V2	Élimination des doublons/recouvrements en structurant par émission (uniquement édition métropolitaine). Retour à la vidéo en ligne. (ASR4 sur les vidéos d'émissions)
Déc. 2020	VOIXOFF-V3	Ergonomie (documentation intégrée, noms de propriétés). (ASR5 et calcul de synchronisations de sujets manquantes)
Mai 2021	VOIXOFF-V4	(ASR6)
Oct. 2021	NOTICES-V3	Restructuration par émission (comme VOIXOFF), traitement des lacunes → *Mais disparition des sujets régionaux*
Avr. 2022	VOIXOFF-V5	Sujets tous synchronisés, synchronisation au mot (*vs* tour de parole) (complétion manuelle des synchronisations manquantes) → *Mais absence de sujets sélectionnés par ailleurs ? (inclus)*
Avr. 2022	NOTICES-V4	Sujets ordonnés, dates sur les sujets, lien aux vidéos par sujet.

Mémo des avancées récentes



- Pour **AF-NOTICES-V4-2022-04-27**
  - Le retour à la vidéo
  - L’ordre des sujets dans l’émission

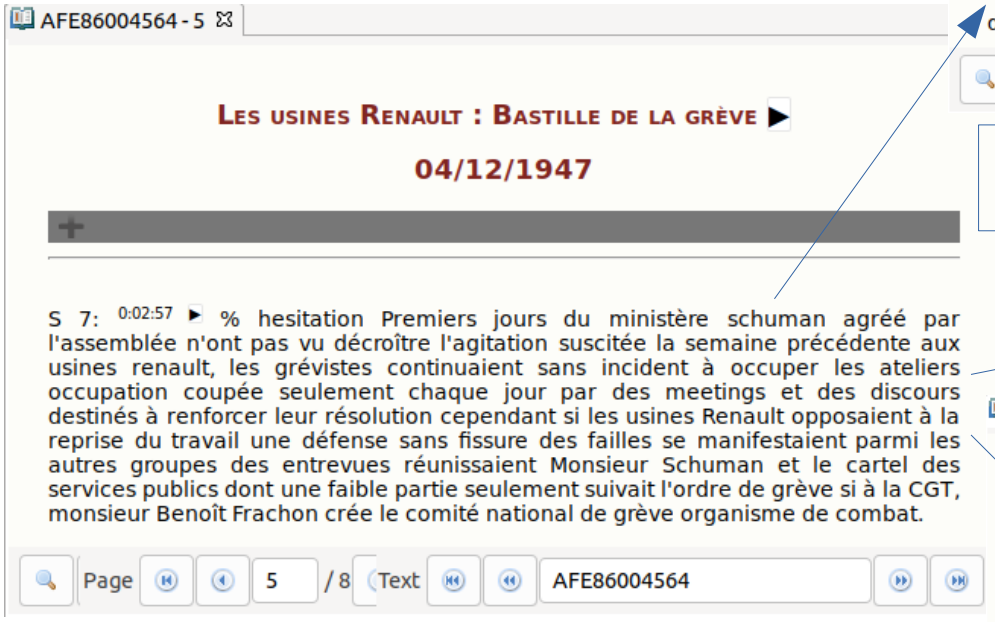
Mémo des avancées récentes

- Pour **AF-VOIX-OFF-V5-2022-04-27**
  - La résorption des sujets non-synchronisés
  - Le calage des sujets au mot près (vs au tour de parole près)

- Sujets inclus
- Vidéos muettes
- Vidéos manquantes
- Sujets supprimés dans la màj de la base INA

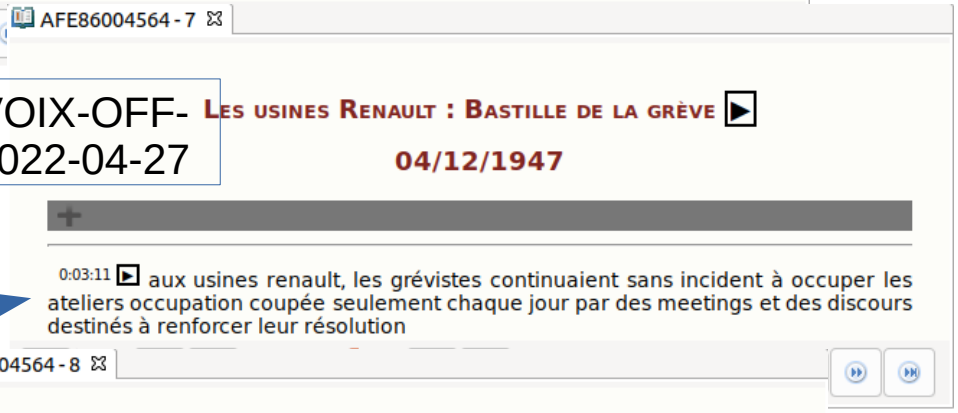
Sujets identifiés...	… **présents** dans **V4**	… **absents** dans **V4**	***Total*** :
… **présents** dans **V5**	9 372	1 311	10 683
… **absents** dans **V5**	12	81	93
Total :	9 384	1 392	10 776

Exemple de redécoupage de tour de parole



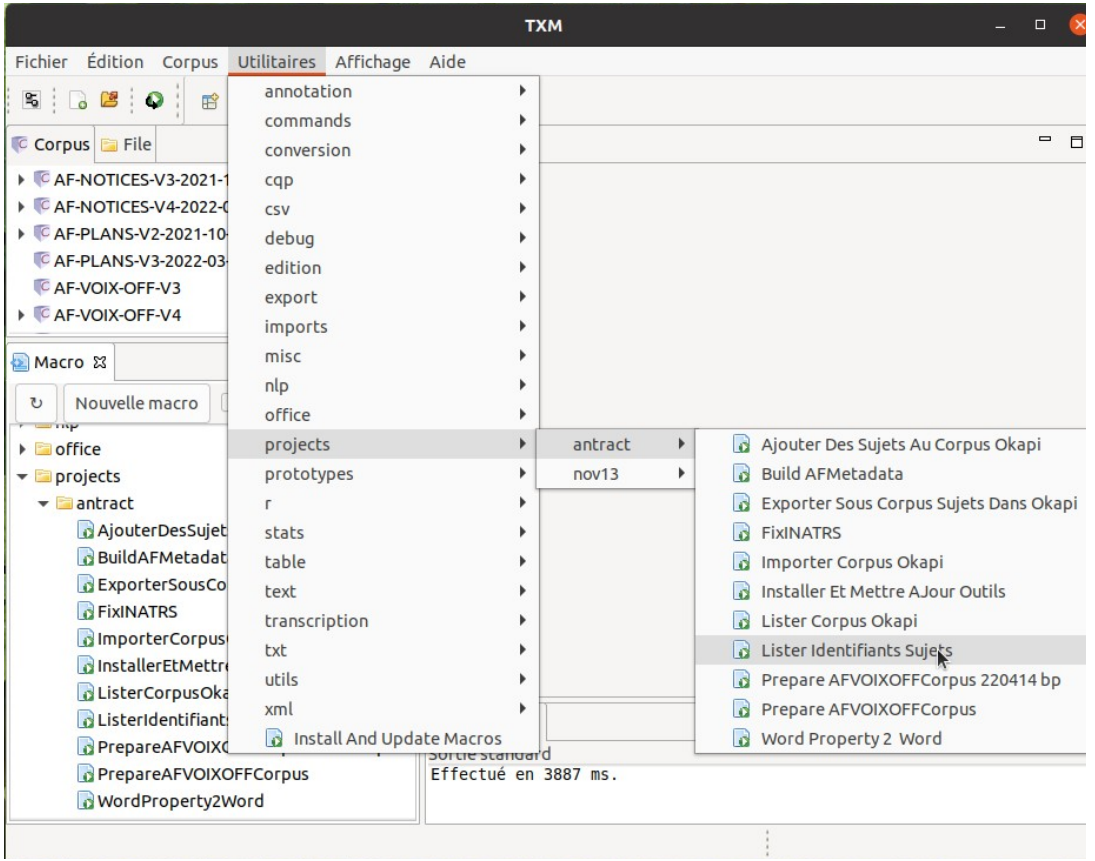
AF-VOIX-OFF-V4-2021-05-19

AF-VOIX-OFF-V5-2022-04-27



Mémo des avancées récentes

- Pour TXM
  - La facilité de manier des sous-corpus de sujets avec les utilitaires
    - Communication entre Okapi et TXM, dans les deux sens
    - Et aussi ListerIdentifiantsSujets pour « déménager » un sous-corpus entre corpus au sein de TXM



Quelques apports du projet pour la textométrie

- Questionnement scientifique intéressant sur la notion d'unité textuelle d'analyse (un peu évoqué dans DHQ 2021)
  - Relation sujet / émission (comme article / journal ? Etc.)
  - ≠ bobine, ≠ page de tapuscrit...
- Des développements dans le logiciel TXM utiles à une large communauté :
  - **MediaPlayer** : consolidation, accès aux vidéos à distance... (JADT 2020)
  - accès direct à un texte dans l'**ÉDITION**
  - **opérateurs de sous-corpus**
  - nouvelles possibilités pour l'**import** de tableaux (à partir de l'expérience d'AF-NOTICES)
  - expérience d'**annotation/projection** pour la recherche de Franck sur la grammaire cinématographique (JADT 2022)

Les richesses et limites de la collaboration de recherche en Histoire

Niveau/ampleur d'usage *vs* qualité/pertinence scientifique

- Pas de problème avec une éventuelle sous-exploitation des outils en soi :
  - le but n'est pas d'utiliser toutes les fonctionnalités ou d'avoir un usage avancé,
  - ni d'avoir un grand nombre d'utilisateurs
- Plutôt : enjeu d'avoir des explorations pertinentes par rapport à un questionnement de recherche
  - comprendre et valoriser les possibilités de l’outil pour la recherche
  - ne pas passer à côté d’une fonctionnalité disponible et adaptée

=> importance des échanges entre utilisateurs et équipe TXM

Rendez-vous manqué sur les documents collaboratifs de travail sur les corpus ?

- On a peu eu les échanges espérés avec les gdocs, en particulier ceux dédiés à la mise au point d'une analyse.
- Expliciter / garder trace de la démarche d'analyse avait pour moi deux intérêts :
  - 1. pouvoir **travailler ensemble** sur la meilleure façon d'ajuster le questionnement historien et les possibilités du corpus et de l'outil
  - 2. capitaliser l'information et se doter des moyens de reproduire l'analyse (**reproductibilité**), soi-même plus tard (**capitaliser**) ou pour suivre les progrès des corpus.

Limites rencontrées du point de vue de l'usage textométrique

- Le côté un peu frustrant de souvent ne pas pouvoir vous accompagner dans la mise au point de votre parcours d'analyse, pour un usage éclairé des outils
  - crainte de trop solliciter ou déranger, le manque de temps
  - avancée en autonomie sauf sur les points bloquants
  - quelquefois perte de temps et d'énergie dans des fausses pistes
- Difficulté à être efficaces dans notre documentation et dans nos supports d'aide
  - réel besoin de documents-supports car pratique intermittente/fragmentée
  - récurrence de certaines difficultés
  - problème de communication sur la façon de gérer l'évolution des données (investir dans les procédures plutôt que dans les données)

Qualité humaine et scientifique de l'équipe

- Équilibre et complémentarité
  - articulation entre les partenaires,
  - ...sans avoir épuisé le potentiel dans le cadre du projet (analyse d'image & TXM, tapuscrits,…)
- La motivation et l'ouverture d'esprit des historiens
  - Curiosité, expérience concrète et approche critique des Humanités numériques
  - malgré un certains nombre d'aspects déstabilisants
    - l'évolution des corpus,
    - la prise en main des outils pas complètement intuitive ni très efficacement documentée
- La perspective que l'investissement pourra profiter au-delà du projet
  - accès aux données sources et enrichies
  - la communication de l'expérience rassemblée dans un ouvrage rédigé