



Documentation du corpus **AF-VOIX-OFF-V5**

contributor, editor Bénédicte PINCEMIN
contributor Serge HEIDEN
contributor Matthieu DECORDE

Copyright © avril 2022 ENS de Lyon

<http://textometrie.org>



Ce document est publié sous licence Creative Commons Attribution-NonCommercial-ShareAlike 4.0
<https://creativecommons.org/licenses/by-nc-sa/4.0>

Contenu et version du corpus

Ce corpus représente la collection complète des journaux métropolitains hebdomadaires des *Actualités françaises*, de janvier 1945 à février 1969, soit 1261 journaux.¹

L'import de ce corpus a été réalisé par l'équipe TXM en avril 2022. Ses sources sont la transcription automatique du 8 mai 2021, dite ASR6 ou v12 (pour le contenu textuel du corpus), et le tableau des notices INA du 27 avril 2022 (pour les métadonnées) bénéficiant des synchronisations manuelles de Matthieu Frey et Bénédicte Pincemin. Le corpus TXM intègre des liens aux vidéos en ligne sur Okapi (il faut avoir installé l'extension Media Player de TXM, et une authentification INA est requise pour l'accès aux vidéos). Il compte près de 1,5 millions de mots.

Schéma simplifié

- **<text>** = un journal hebdomadaire
 - **id** = identifiant INA
 - **date-de-diffusion-tri** = AAAA-MM-JJ
 - **date-de-diffusion-année** = AAAA
 - ...
- **<div>** = un sujet
 - **id** = identifiant INA
 - **date-de-diffusion-emission-tri** = AAAA-MM-JJ
 - **date-de-diffusion-annee** = AAAA
 - **date-de-diffusion** = JJ/MM/AAAA (propre au sujet, potentiellement différente de celle de l'émission)
 - **n** = position du sujet dans le journal (1, 2 etc.)
 - **synchronized** = "true" ou "false"
 - **titre-propre**
 - **resume**
 - **sequences**
 - **descripteurs-aff-lig**
 - **generique-aff-lig**
 - ...
- **<sp>** = un segment de parole
 - **time** = H:MM:SS
 - ...
 - **[mot]**
 - **word**
 - **frlemma**
 - **frpos**
 - **ne**
 - ...

¹ Il y a en fait 1260 journaux dans ce corpus, car la vidéo pour l'émission du 18 septembre 1963 étant muette, sa transcription est vide, ce qui ne permet pas à l'émission correspondante d'être présente.

Structuration et caractéristiques du contenu

Les textes (structure <text>) du corpus sont les **1260 journaux** hebdomadaires.

L'ordre des textes est **chronologique** (ce qui ordonne les pages des textes affichées par TXM -commande Édition- et donne sens à la commande Progression sur l'ensemble du corpus).

Le texte de chaque journal est la transcription automatique du commentaire audio, synchronisée à la vidéo.

Chaque journal est composé d'une succession de **sujets** (structure <div>) (tout mot transcrit relève d'**un et au plus un seul sujet**) :

- Il reste quelques rares cas où pour un segment vidéo avec du contenu transcrit (quelquefois au début ou à la fin de la vidéo, ou entre deux sujets), il n'y a **pas de sujet** décrit dans la base INA. Ces passages non identifiés sont notés comme sujets (div) non-synchronisés (propriété `synchronized="false"`). Il reste 8 passages de ce type d'ampleur pouvant correspondre à un sujet, dont 3 doublons (passages déjà décrits ailleurs), donc seuls 5 sujets manquants (présents dans la transcription mais non identifiés) (cf. détail en Annexe 1).
- Inversement, il arrive que la description documentaire INA détaille quelques (sous-)sujets à l'intérieur d'un sujet. Mais dans le modèle actuel de TXM pour les transcriptions il n'est pas possible d'emboîter des sujets, les sujets sont nécessairement successifs. Le corpus est donc **privé des sujets inclus** : plus exactement, de leur description issue des notices INA qui permet de les repérer, mais pas de leur contenu qui est bien transcrit comme le reste (cf. section *Comment prendre en compte les sujets inclus* plus bas).

Au total on a donc **10 688 sujets** dans AF-VOIX-OFF-V5-2022-04-27², dont maintenant seulement **5 non identifiés**.

Un travail important a été fait dans la préparation de cette version v5 pour ajuster automatiquement les tours de parole créés par la reconnaissance automatique de la parole, et les sujets définis dans la base INA. En effet, les tours n'étaient pas naturellement inclus dans les emplans temporels attribués aux sujets. Le traitement a cherché à faire correspondre le mieux possible les sujets et les passages de transcription correspondants, au mot près tout en gardant un peu de souplesse.

Le gain apporté par les sujets synchronisés à la main et par l'ajustement plus fin des tours de parole aux sujets a permis d'identifier **1 311 sujets supplémentaires par rapport à AF-VOIX-OFF-V4** (soit 14 % de sujets identifiés en plus). Ces sujets sont regroupés dans le sous-corpus `sujets-non-synchronises-dans-v4`.

Le contenu des sujets se présente comme une succession de segments de parole (structure <sp>). Ces segments de parole sont visualisés dans l'édition TXM comme des paragraphes.

Chaque segment de parole est enfin composé de mots.³

Remarque : dans le cadre de TXM nous utilisons des éléments et attributs XML de la TEI (*Text encoding initiative*) les plus proches du sens que l'on souhaite véhiculer dans l'encodage, pour avoir une terminologie uniforme et standardisée. Le sens effectif de données calculées automatiquement provient des traitements qui les ont produites.

Métadonnées des journaux hebdomadaires

Chaque journal des Actualités françaises correspond à une structure `text`, dotée de propriétés dont :

- `id` : l'identifiant de la notice INA (ex. AFE86004415)
- `date-de-diffusion` : notée JJ/MM/AAAA (ex. : 25/01/1945 pour le 25 janvier 1945).
- `date-de-diffusion-tri` : notée AAAA-MM-JJ (ex. : 1945-01-25 pour le 25 janvier 1945).

2 Calcul : 10 884 (structures <div>) - 193 (passages non-synchronisés ne correspondant pas à des sujets) - 3 (doublons) = 10 688. Voir Annexe 1 pour l'obtention des chiffres utilisés.

3 Plus exactement, chaque segment de parole <sp> contient un énoncé (<u>, *utterance*), qui contient les mots. Cependant dans ce corpus le découpage en énoncés correspond à celui en segments de parole, il n'apporte pas d'information supplémentaire utile, si bien que nous pouvons en pratiquer l'ignorer.

On notera aussi la disponibilité de variantes d'expression de la date de diffusion, potentiellement utiles pour certaines interrogations (construction de partitions, affichage, etc.) : `date-de-diffusion-annee`, `date-de-diffusion-mois`, `date-de-diffusion-jour-semaine`, `date-de-diffusion-jour`.

Si besoin, un inventaire complet des métadonnées des journaux est consultable via la commande Propriétés, onglet Détails, section Propriétés des structures, point text. Ces propriétés sont :

- soit directement issues des bases de données INA, en reprenant le nom du champ (en minuscules, sans accents, sans parenthèses, avec des tirets entre les mots : `duree`, `notes-du-titre`...);
- soit produites par le projet Antract (nom commençant par "antract-");
- soit ajoutées par l'import TXM, comme les variantes d'écriture de date ci-dessus, mais aussi d'autres propriétés pour des considérations techniques (ex. `text-order`, ou `title` -qui est équivalent à `titre-propre`).

Métadonnées des sujets

Les propriétés des sujets (structure `div`) indiquent d'abord :

- `synchronized` : si le sujet est identifié ("`true`", on a alors les informations documentaires associées) ou si c'est un passage non reconnu et attribué ("`false`", sans informations documentaires).
- `id` : l'identifiant de la notice INA (vide en cas de sujet non-synchronisé).
- `n` : le numéro d'ordre du sujet dans le journal (1 pour le premier, 2 pour le deuxième, etc.).

Les `date-de-diffusion` des sujets ne sont pas toujours égales à celle du journal (elles peuvent être quelques jours antérieures, ou manquer pour les sujets non-synchronisés), c'est pourquoi pour les dates on s'appuiera généralement sur celle du journal qui est disponible dans la propriété `date-de-diffusion-emission-tri`.

Les propriétés des sujets apportent des informations des notices sujets INA, en particulier les informations "textuelles" du corpus des Notices :

- `titre-propre` : le titre du sujet
- `resume` : le "Résumé" du sujet dans la base INA
- `sequences` : la description plan à plan dite "Séquences" de la base INA.
- `descripteurs-aff-lig` : liste des descripteurs INA, les descripteurs sont précédés d'un code qui indique leur type (DET : thème, DEL : lieu, DEI : ce qui apparaît à l'image, DSO : ce qu'on entend dans la bande son ex. telle chanson, tel slogan), ex. DET: `Seconde Guerre mondiale` ; DEL: `France` ; DEL: `Alsace` ;
- `generique-aff-lig` : liste de personnes impliquées, avec indication de leur rôle par un code (ex. PAR participant, OPV opérateur de prise de vue, etc.), ex. OPV `Persin, René` ; OPS `Remoué, Francis` ; PAR `Lacoste, Robert` ;

Deux remarques sur ces propriétés "textuelles" :

1. Dans AF-VOIX-OFF-V5 (comme dans les versions antérieures) ces contenus textuels sont codés comme une (seule et grande) chaîne de caractères, en valeur d'une propriété, ce ne sont pas les mots du corpus (les mots du corpus AF-VOIX-OFF-V5 sont ceux des transcriptions). Cette représentation est moins précise (on n'a pas les frontières des mots) et moins riche (on n'a pas d'informations linguistiques : lemmes, catégories grammaticales) que dans AF-NOTICES. **Pour les interrogations se limitant aux notices documentaires, on préférera donc toujours le corpus AF-NOTICES** ; en revanche, le corpus **AF-VOIX-OFF-V5 sera utile pour des recherches croisées** entre notices et transcriptions.

2. Les Séquences ne sont pas toujours renseignées ; le contenu du Résumé prend parfois la forme d'une description de type Séquences ; en l'état actuel des données il vaut donc mieux généralement **interroger ensemble `resume` et `sequences`**.

Les propriétés des sujets reprennent également d'autres informations des notices INA (`genre`, `nature-de-production`, `producteurs-aff`, `duree`, etc.) et des informations construites dans le projet avec un nom commençant par `antract-` (notamment la position temporelle dans l'émission avec les `time-codes` `antract-debut`, `antract-fin`). Si besoin, un inventaire complet des propriétés disponibles pour les sujets est consultable via la commande Propriétés, onglet Détails, section Propriétés des structures, point div.

Propriétés des segments de parole et des mots

Les segments de parole (<sp>) sont notamment dotés d'une propriété donnant leur repérage temporel dans la vidéo, le time-code du début du segment (`time`). Ils portent également une information de locuteur (`who`) tel qu'a pu le détecter le traitement automatique ASR, mais cette information est d'intérêt limité dans le cas de notre corpus, essentiellement monologal.

Pour les mots, les principales propriétés sont :

- `word` : forme graphique, telle qu'elle se présente dans la transcription ASR fournie ;
- `fr lemma` : lemme, c'est-à-dire l'entrée de dictionnaire correspondante (singulier, masculin, infinitif, etc.)
- `fr pos` : catégorie grammaticale (liste disponible dans le manuel TXM) (`pos`= part-of-speech = partie-du-discours)

Les lemmes et les catégories grammaticales ont été affectés automatiquement par le logiciel TreeTagger avec le principal modèle de langue pour le français. Cela enrichit les interrogations possibles mais il peut y avoir des erreurs.

Les mots portent également toutes les informations que leur a attribuées l'ASR (automatic speech recognition), par exemple un degré de fiabilité de la reconnaissance (`conf`) et une autre indication de catégorie grammaticale (`pos`) exprimée dans un autre jeu de catégories. Les entités nommées sont codées avec la propriété `ne` (comme *named entities*) selon le système de codage IOB ([https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_\(tagging\)](https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging))). Les entités sont de 8 types (`amount`, `event`, `func`, `loc`, `org`, `pers`, `prod`, `time`). Le premier mot d'une mention a une catégorie préfixée par `B-` (comme *beginning*), et les éventuels mots suivants constituant la même désignation ont leur catégorie préfixée par `I-` (comme *in*). Donc par exemple pour trouver les mentions de type `event` on peut utiliser la requête CQL :

```
[ne="B-event"][ne="I-event"]*[:ne!="I.*":]
```

Glose de la requête : on cherche le premier mot d'une entité de type `event`, suivi d'éventuels autres mots composant la suite de la désignation de l'entité, jusqu'à ce qu'on trouve un mot n'étant plus dans la suite d'une désignation d'entité, ce dernier mot n'étant pas considéré dans le résultat de la requête (effet des deux points à l'intérieur des crochets).

Si besoin, les inventaires complets des propriétés sont consultables via la commande Propriétés, onglet Détails, section Propriétés des unités lexicales (pour les mots), et section Propriétés des structures, point `sp` pour les segments de parole.

Comment récupérer et mettre dans AF-VOIX-OFF-V5 un sous-corpus (sélection de sujets) déjà constitué

- Si le sous-corpus à récupérer est dans Okapi : utilitaire `ImporterCorpusOkapi`.
- Si le sous-corpus à récupérer est dans un autre corpus TXM : utilitaires `ListerIdentifiantsSujets` puis `ImporterCorpusOkapi`.

L'installation et la mise en œuvre des utilitaires sont décrites dans cette documentation en ligne : https://groupes.renater.fr/wiki/txm-info/public/chantier_antract#tutoriel_d_utilisation_des_outils_d_echanges_entre_txm_et_okapi

Noter également que le sous-corpus `sujets-non-synchronises-dans-v4` peut être utilisé pour chercher d'éventuels sujets pouvant compléter un sous-corpus qui aurait été initialement **construit** par des requêtes dans AF-VOIX-OFF-V4: on lance les mêmes requêtes sur ce sous-ensemble complémentaire pour ne pas avoir à retraiter des sujets déjà vus.

Comment prendre en compte les sujets inclus

Comme le corpus AF-VOIX-OFF-V5 n'a quasiment plus de sujets non identifiés (non synchronisés), les sujets qu'on peut ne pas trouver sont essentiellement des sujets inclus. Pour chaque identifiant non trouvé dans le corpus AF-VOIX-OFF-V5, on peut donc généralement le remplacer par celui du sujet englobant. On s'appuiera sur le corpus AF-NOTICES qui contient tous les sujets (englobants et inclus) :

- dans AF-NOTICES, chercher le sujet inclus, avec une concordance sur : `<div>[_div_id="AFEXXXXXXXX"]`
- double-cliquer sur la ligne pour ouvrir l'édition du corpus à la page de la notice du sujet ;
- feuilleter les notices et regarder les notices voisines (notamment celles d'identifiant voisin) dans la même émission, consulter les indications de time-code (antract-debut, antract-fin) au début de chaque notice, et en déduire quelle est la notice englobante. Noter son identifiant pour l'utiliser dans AF-VOIX-OFF à la place de celui du sujet inclus.

Annexe 1 : Observations des sujets non-synchronisés restants

Synthèse

En pratique, tous les sujets INA des Actualités françaises (tels que rassemblés dans le corpus AF-NOTICES-V4) sont maintenant synchronisés dans le corpus AF-VOIX-OFF, sachant cependant que les sujets inclus sont couverts par leur sujet englobant. Il ne reste que 5 passages vidéo de type sujet pour lesquels nous ne disposons pas de notice INA correspondante, et 3 passages vidéo en doublon pour lesquels le sujet correspondant est déjà synchronisé sur un passage identique.

Détail

En effet, le corpus AF-VOIX-OFF-V5-2022-04-27 compte 201 passages transcrits non rattachés à un sujet, on peut les consulter avec une concordance sur :

```
<div>[_div_synchronized="false"]
```

Mais dans la très grande majorité des cas (192 / 201), il s'agit de transcriptions très courtes, associées à des mots qui techniquement montrent que la reconnaissance automatique de la parole était en difficulté, typiquement lorsqu'elle essaye de transcrire des passages non parlés (musique, bruitage). L'INDEX de

```
<div>[_div_synchronized="false"]{1,3}</div>
```

trouve 193 passages de cette sorte qui se détaillent comme suit :

<i>Contenu textuel du passage transcrit non synchronisé</i>	<i>Nombre de segments (div) avec ce contenu</i>
(%hesitation)	150
Hum .	22
Oui .	15
(%hesitation) (%hesitation)	1
(%hesitation) Hum .	1
(%hesitation) Oui .	1
de Strasbourg ,	1
Hum . (%hesitation)	1
Non	1
Total :	193

Seul le passage « de Strasbourg, » correspond aux derniers mots d'un sujet qui a été time-codé un peu court (et l'algorithme de positionnement des tours de parole dans les sujets a ici été dans une situation où il a coupé le tour exactement au temps indiqué dans la base INA).

Considérons à présent les 8 passages plus longs, correspondant de fait réellement à des contenus de sujets. On peut les parcourir avec une concordance sur :

<div>[_div_synchronized="false"]{4,} within div

L'analyse de ces 8 passages permet de conclure que tous ces passages correspondent soit à des sujets effectivement absents de la base INA (telle que représentée dans le tableau du 17 mars 2022 et ses versions ultérieures) ou bien à des passages vidéos en doublon, le sujet est bien présent mais est associé à un autre segment vidéo de contenu identique. Dans le détail,

- 5 sujets non-synchronisés correspondent à un sujet manquant :
 - 1947-01-23, AFE86004519, 0:02:38
 - 1949-09-29, AFE86004659, 0:06:28
 - 1949-10-13, AFE86004661, 0:06:48
 - 1952-10-02, AFE86004816, 0:06:00 (sujet AFE85004768 manquant ?)
 - 1962-12-05, AFE86003857, 0:00:18 (sujet AFE85009742 manquant ?)

(les identifiants des sujets manquants sont suggérés par les identifiants des sujets voisins, en observant que globalement des sujets successifs reçoivent des identifiants successifs.)

- 3 sujets non-synchronisés correspondent à des passages vidéo en doublon, un seul des deux passages est associé au sujet correspondant :
 - 1968-05-15, AFE86004141 : la vidéo de l'émission complète est composée de deux fois la même vidéo mais le 2^e sujet est muet dans la première moitié. Il est time-codé sur la seconde moitié, alors que les deux autres sujets sont time-codés sur la première. Les deux sujets non-synchronisés sont les 1^{er} et 3^e sujets de la deuxième moitié (début à 0:10:58 et 0:12:39).
 - 1967-04-18, AFE86004085, 0:00:18 : ce passage est un doublon de AFE86000896, situé en dernier dans l'émission précédente (1967-04-11, AFE86004084, 0:09:54).

Annexe 2 : Comparaison des sujets présents dans AF-VOIX-OFF-V4, AF-VOIX-OFF-V5, et AF-NOTICES

On considère AF-VOIX-OFF-V4-2021-05-19 et AF-VOIX-OFF-V5-2022-04-27. Regardons quantitativement le nombre de sujets identifiés (c'est-à-dire avec un identifiant INA) présents dans les deux corpus :

<i>Sujets identifiés...</i>	<i>... présents dans V4</i>	<i>... absents dans V4</i>	<i>Total :</i>
<i>... présents dans V5</i>	9 372	1 311	10 683
<i>... absents dans V5</i>	12	81	93
<i>Total :</i>	9 384	1 392	10 776

Le nombre total de sujets dans la base INA pour les Actualités françaises (représentée par le tableau ANTRACT_Notices_220427abp.xlsx) est de 10 776, c'est le nombre de sujets du corpus AF-NOTICES-V4-2022-04-27 qui contient l'intégralité des sujets disponibles.

Les sujets identifiés absents dans V5 sont les sujets inclus et les sujets pour lesquels il n'y a pas de transcription (vidéo sans son, ou vidéo manquante). Par exemple voici ce qu'il en est exactement pour les 12 sujets de v4 ne sont plus présents dans v5. La disparition de ces 12 sujets s'explique ainsi :

- soit il s'agit de sujets inclus :
 - AFE85003777 → inclus dans AFE85003779 (contenant AFE8500376, AFE85003777, AFE85003778 et AFE85003781).
 - AFE85003778 → inclus dans AFE85003779 (contenant AFE8500376, AFE85003777, AFE85003778 et AFE85003781).
 - AFE85004547 → inclus dans AFE85004546 (contenant AFE85004547 et AFE85004548)
 - AFE85004548 → inclus dans AFE85004546 (contenant AFE85004547 et AFE85004548)
 - AFE85005746 → inclus dans AFE85005747 (contenant AFE85005746, AFE85005748, AFE85005749)
 - AFE85009352 → inclus dans AFE85009350 (contenant AFE85009351 et AFE85009352)

- AFE86002998 → inclus dans AFE86002997 (contenant AFE86002998 et AFE86002999)
- AFE86002999 → inclus dans AFE86002997 (contenant AFE86002998 et AFE86002999)
- soit leur présence était due à des erreurs de time-code (vidéo sans parole ou absence de vidéo correspondante) :
 - AFE85002929 → sans paroles donc n'a pas de texte dans AF-VOIX-OFF
 - AFE85005841 → vidéo manquante
 - AFE85005842 → vidéo manquante
- soit ils ne font plus parti de la base INA dans la dernière version des données :
 - AFE85002783 → a disparu des versions plus récentes du tableau INA, cette notice a dû être supprimée lors d'une mise à jour de la base.