

**RAPPORT PROJET OCR**



* **PENPYTEXT -**

**SOMMAIRE**

1- Description du projet

2 - Source du dataset

3 - Structure et éléments du dataset

4- Existence de biais

5 - Représentation des données d'entraînement (X, y)

6- Harmonisation des données d'entraînement

**1- Description du projet**

Le projet consiste à utiliser des algorithmes de deep learning pour reconnaître les caractères manuscrits dans un document numérique (pdf ou png)

**2- Source du dataset**

Pour réaliser ce projet, nous utiliserons le dataset **IAM Handwriting Database 3.0** disponible [ici](https://fki.tic.heia-fr.ch/databases/iam-handwriting-database#LOBcorpus).

La base de données contient des formes de **texte anglais** écrites à la main sans contrainte, qui ont été scannées à une résolution de 300 dpi et enregistrées sous forme d'images PNG.

Caractéristiques de la IAM Handwriting Database 3.0:

▪ 657 personnes ont participé à la rédaction des fichiers manuscrits

▪ 1 539 pages de texte scannées

▪ 5 685 phrases isolées et labellisées

▪ 13 353 lignes de texte isolées et labellisées

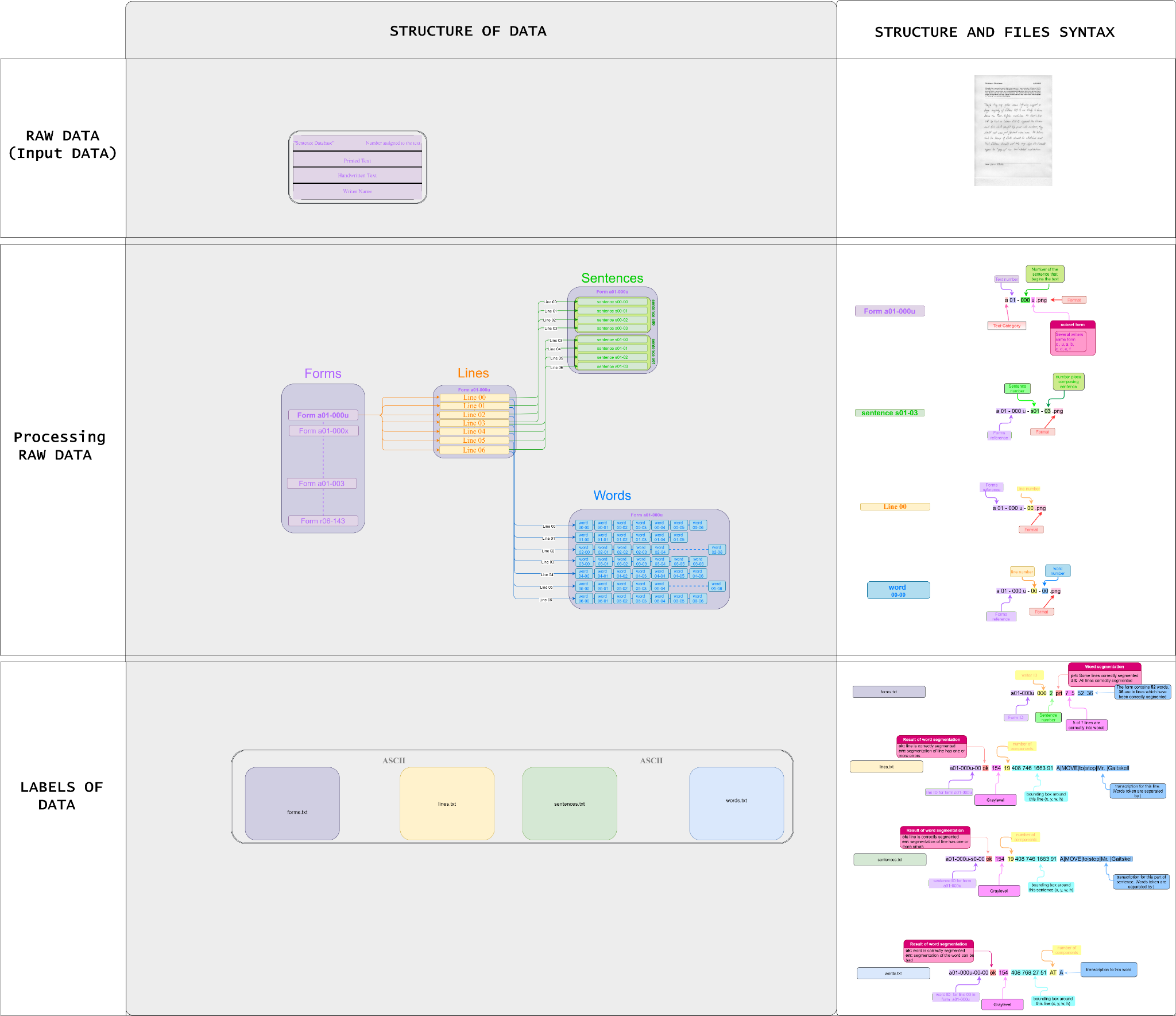
▪ 115 320 mots isolés et labellisés

**3 - Structure et éléments du dataset**

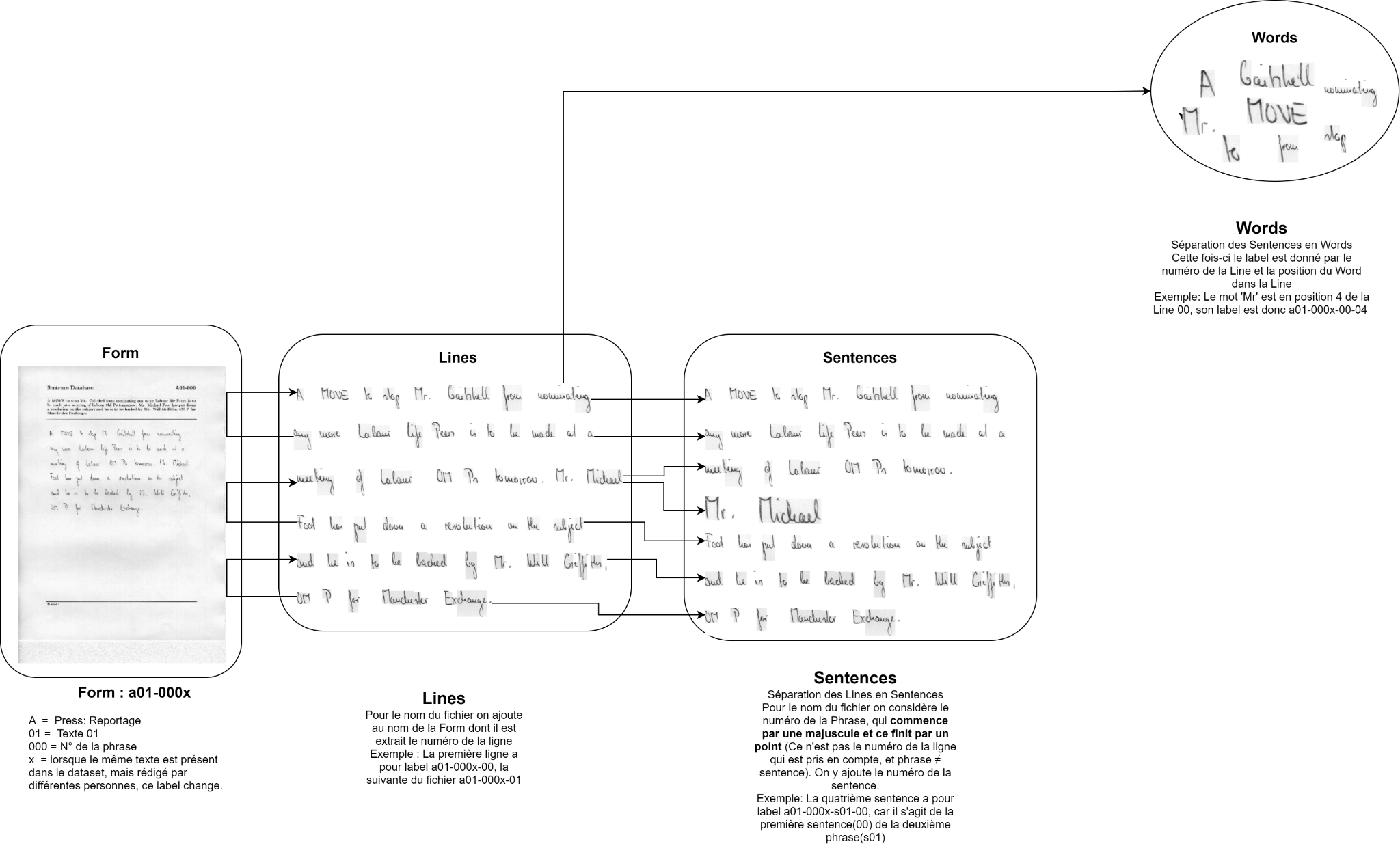
Le dataset IAM est constitué de deux types de fichier structurés hiérarchiquement en “Forms”, “Lines”, “Sentences” et “Words” :

* Les fichiers **.png** correspondent aux images issues des différentes opérations de segmentation réalisées sur les “forms”.
* Les fichiers **.txt** contiennent la transcription des différentes segmentations.

Ci-dessous, vous trouverez la structure globale des données.



*Fig.1 Structure globale des données*



*Fig.2 Exemple de représentation de la structure des données*

**4- Existence de biais**

La base de données contient des formes de **texte anglais** écrites à la main sans contrainte sous forme de différents types de formulaire (journalistique, vie,..).657 personnes ont participé à la rédaction de ces formulaires. Vu la diversité des formulaires et des rédacteurs, nous avons pris l’hypothèse de ne pas tenir en compte des biais.

**5 - Représentation des données (X, y)**

La première étape de la modélisation du problème est de formaliser la structure des données. Les fichiers ont été divisés en plusieurs dossiers (voir 3.Structure et éléments du dataset).

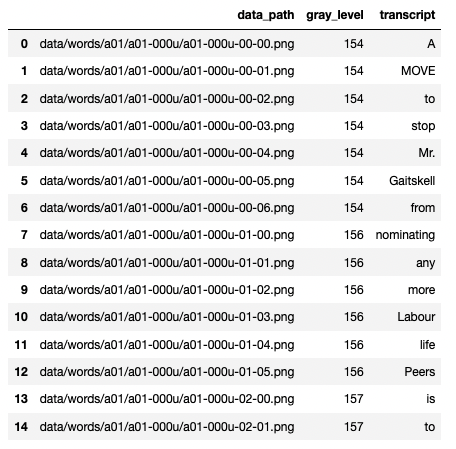
Les fichiers du jeu de données dans un dossier nommé *data*. Pour plus de lisibilité et afin que les données soient plus facilement exploitables, nous avons associé dans un fichier ‘words\_data.csv’ les informations suivantes :

- *data\_path* : chaîne de caractères qui représente le chemin d’accès à l’image d’un mot.

- *gray\_level* : entier compris dans l’intervalle [0,255] qui représente le niveau de gris de notre texte dans l’image. Cet entier est destiné au nettoyage des images.

- *transcript* : représente le label de l’image mot, il s’agit de notre cible.

Extrait du fichier ‘words\_data.csv’



Dans un second temps, nous procéderons de la même manière pour les données concernant les phrases

**6- Harmonisation des données d'entraînement**

Les fichiers .png qui constituent notre base de données, sont représentés sous la forme d’une matrice de pixels. Elle a deux dimensions qui définissent la taille de l’image et une troisième pour représenter la couleur.

Lors de la modélisation, l’harmonisation du dataset permettra d’améliorer l'apprentissage. Il consistera en la standardisation de la taille de l’image, d’une part. L’image sera ainsi positionnée à différentes coordonnées dans un rectangle blanc de taille standard, dont les dimensions seront celles de la plus grande image. Cela entraînera le modèle à lire les images à différents endroits et permettra d’améliorer les performances.

D’autre part, lors de l’exploration du jeu de données nous avons observé la présence de bruit sur les images. Une autre étape de l’harmonisation consistera en la suppression de ce dernier afin de rendre les images plus nettes, et permettre une meilleure reconnaissance des mots.

Une dernière étape sera de s’assurer que la résolution des données soit uniformisée.