

# ANALYSE DE DESCRIPTIONS ARCHITECTURALES PAR NLP

*Extraction d'informations sur les bâtiments historiques à partir de textes*



**S6 – Licence d'Excellence en Intelligence Artificielle**

**Réalisé par :**

Mouad Radouani  
Abdelkarim Narjiss

**Encadré par :**

Pr. EL HABIB Ben Lahmar  
Pr. Oussama Kaich

**Année universitaire : 2024–2025**

## Table des Matières

<b>Introduction Générale .....</b>	<b>2</b>
<b>Partie I : Présentation Générale du Projet .....</b>	<b>3</b>
1. Contexte générale : .....	3
2. Problématique .....	3
4. Objectifs Spécifiques : .....	4
5. Méthodologie Globale : .....	4
<b>Partie II : Les Données .....</b>	<b>5</b>
1. Problématique des Données .....	5
2. Génération manuelle de données synthétiques .....	5
3. Annotation manuelle ciblée .....	5
<b>Partie III : Entraînement du Modèle NER .....</b>	<b>6</b>
1. Préparation des données annotées .....	6
2. Construction d'un modèle linguistique sur mesure .....	6
3. Apprentissage du modèle.....	6
4. Sauvegarde du modèle.....	7
<b>Partie IV : Déploiement d'une Interface d'Analyse Interactive avec Streamlit .....</b>	<b>7</b>
1. Introduction.....	7
2. Structure Générale de l'Application .....	7
3. Présentation Visuelle de l'Application Streamlit.....	8
<b>Conclusion Générale.....</b>	<b>10</b>

## Liste des Figures

Figure 1 : Interface au Lancement.....	8
Figure 2 : Résultat Après Analyse d'un Paragraphe .....	9

# Introduction Générale

Depuis des siècles, les bâtiments historiques témoignent des civilisations, des savoir-faire, et des styles artistiques d'époques passées. Que ce soit une cathédrale gothique, un palais baroque ou une mosquée ottomane, chaque monument raconte une histoire à travers son architecture, ses matériaux et son style. Ces constructions font partie intégrante du patrimoine culturel mondial, et les comprendre aide à mieux connaître l'histoire, l'art, et même la société de leur temps.

Avec le temps, beaucoup de ces bâtiments ont été documentés dans des textes descriptifs : brochures, articles, bases de données culturelles, etc. Ces textes sont souvent riches, mais restent difficilement exploitables de manière automatique, car ils sont écrits en langage naturel. Or, dans une époque où l'on génère et traite d'énormes quantités de données, structurer ces informations devient essentiel pour les rendre accessibles, analysables et réutilisables.

C'est là que le progrès scientifique et technologique entre en jeu. Ces dernières années, les techniques d'intelligence artificielle, et plus particulièrement le traitement automatique du langage naturel (NLP), ont permis d'extraire automatiquement du sens à partir de textes non structurés. Des modèles sont aujourd'hui capables de lire, d'identifier des entités, de comprendre des relations et d'organiser tout cela sous forme exploitable.

C'est dans ce contexte que s'inscrit notre projet. L'idée est simple : concevoir une application capable d'analyser des paragraphes décrivant des bâtiments historiques, et d'en extraire automatiquement trois types d'informations clés :

- les caractéristiques architecturales (comme colonnes, coupoles, arcs...).
- les matériaux de construction (pierre, marbre, bois...).
- les styles architecturaux (gothique, renaissance, baroque...).

# Partie I : Présentation Générale du Projet

## 1. Contexte générale :

Le patrimoine architectural constitue un héritage précieux, reflet de l'histoire, de l'art et du savoir-faire des civilisations passées. Chaque monument, qu'il soit religieux, civil ou militaire, porte une identité propre et témoigne d'une époque et d'un style de construction. Ces édifices ne sont pas de simples structures, mais des témoins vivants de notre mémoire collective.

Depuis des siècles, ils sont décrits dans divers textes — fiches patrimoniales, brochures, articles ou ouvrages scientifiques. Ces descriptions, bien que riches en contenu (styles, matériaux, éléments architecturaux), sont souvent peu structurées, rédigées en langage naturel, ce qui rend leur exploitation automatique complexe.

À l'ère du numérique, valoriser ces textes devient un enjeu essentiel. Grâce aux avancées en intelligence artificielle, notamment en traitement automatique du langage (NLP), il devient possible d'en extraire le sens, de rendre ces informations accessibles et analysables à grande échelle.

Notre projet s'inscrit dans cette démarche : exploiter des outils simples mais puissants pour redonner vie aux textes du patrimoine, de façon interactive, lisible et intelligente.

## 2. Problématique

Les descriptions de bâtiments historiques, bien qu'informées et riches, sont souvent rédigées en langage naturel, sans structure claire. Elles mêlent formulations variées, synonymes, et tournures littéraires, ce qui rend leur traitement automatique difficile.

**La question centrale est donc la suivante :**

**Comment un système peut-il reconnaître et extraire automatiquement les éléments clés (styles, matériaux, caractéristiques) à partir d'un texte descriptif non structuré sur un bâtiment historique ?**

Le défi est de concevoir une solution légère, entraînée sur peu de données, mais suffisamment intelligente pour comprendre la diversité du langage architectural et restituer l'information de façon exploitable.

## 3. Objectif Général :

L'objectif général de ce projet est de **concevoir un outil capable d'analyser automatiquement des paragraphes décrivant des bâtiments historiques afin d'en extraire les informations clés**. Il s'agit de transformer du texte non structuré en données organisées, compréhensibles et exploitables.

Concrètement, l'outil vise à identifier trois types d'éléments dans les descriptions :

- **les caractéristiques architecturales** (ex. : arc, colonne, coupole).
- **les matériaux de construction** (ex. : pierre, bois, marbre).
- **les styles architecturaux** (ex. : gothique, baroque, renaissance).

Ce processus permettrait d'automatiser l'analyse de textes patrimoniaux, et donc de faciliter la **classification, la recherche, ou encore la mise en valeur de monuments historiques** sur des plateformes numériques. Un tel système pourrait également servir de base pour d'autres applications, dans les domaines de la culture, de l'éducation, du tourisme ou de la recherche scientifique.

À travers une approche simple mais fonctionnelle, ce projet cherche donc à démontrer qu'il est possible d'utiliser des techniques de **traitement du langage naturel (NLP)**, même avec un volume limité de données, pour rendre accessibles et exploitables des informations enfouies dans les textes.

## 4. Objectifs Spécifiques :

- Créer un jeu de données adapté au domaine de l'architecture historique.
- Annoter manuellement les entités pertinentes : styles, matériaux, caractéristiques.
- Entraîner un modèle de reconnaissance d'entités nommées (NER) avec spaCy.
- Créer une interface utilisateur simple et fonctionnelle avec Streamlit.
- Permettre une visualisation claire des résultats extraits à partir d'un texte donné.

## 5. Méthodologie Globale :

La méthodologie adoptée dans ce projet suit une logique **progressive et orientée vers la mise en pratique**. Elle repose sur des étapes simples mais cohérentes, de la création des données jusqu'au déploiement de l'application finale.

### 1. Génération et annotation des données

En l'absence de corpus annoté existant, des paragraphes décrivant des bâtiments ont été générés avec des outils d'IA, puis annotés manuellement pour identifier les entités (caractéristiques, matériaux, styles).

### 2. Préparation des données d'entraînement

Les annotations ont été converties dans un format compatible avec spaCy, permettant l'entraînement d'un modèle NER personnalisé.

### 3. Entraînement du modèle NLP

Un modèle de reconnaissance d'entités a été entraîné avec spaCy à partir des données annotées, en veillant à obtenir un bon équilibre entre simplicité et performance.

### 4. Développement de l'interface avec Streamlit

Une application Web légère a été créée pour permettre l'analyse en ligne d'un texte, avec surlignage coloré et tableau récapitulatif.

### 5. Test et validation fonctionnelle

Le système a été testé avec plusieurs exemples pour vérifier la qualité des extractions et la fluidité de l'interface.

## Partie II : Les Données

### 1. Problématique des Données

L'un des premiers obstacles rencontrés dans ce projet a été l'absence de jeux de données publics spécifiquement conçus pour l'extraction d'informations architecturales à partir de textes descriptifs. Aucune base annotée ne permettait de relier automatiquement les éléments comme les styles, les matériaux ou les formes structurelles dans les paragraphes sur les bâtiments historiques.

Une alternative aurait pu être de **collecter des textes existants** via le **web scraping**. Cependant, même si cette méthode permet d'obtenir un grand volume de descriptions, elle pose un problème majeur : **l'annotation manuelle à grande échelle**. Annoter correctement plusieurs milliers de paragraphes est une tâche extrêmement longue, fastidieuse, et sujette à erreurs. Il serait par exemple **inimaginable d'annoter 10 000 descriptions manuellement**, sans outils automatisés, surtout dans un cadre académique ou individuel.

Face à ces contraintes, une approche alternative s'est imposée : **générer manuellement un petit corpus synthétique**, maîtrisé et annoté avec précision.

### 2. Génération manuelle de données synthétiques

Pour surmonter ce manque, nous avons décidé de générer nous-mêmes les données. À l'aide d'outils de génération de texte (comme des modèles d'IA), nous avons conçu une série de paragraphes imitant des descriptions réalistes de monuments historiques. Ces textes incluent volontairement des mentions de styles (ex. : gothique, roman), de matériaux (ex. : pierre, bois), ainsi que d'éléments architecturaux (ex. : dôme, colonne, vitrail).

Cette approche nous a permis de produire rapidement un corpus varié, tout en maîtrisant le contenu pour s'assurer que les entités d'intérêt étaient bien représentées.

### 3. Annotation manuelle ciblée

Une fois les paragraphes générés, chaque texte a été **annoté manuellement**. Nous avons identifié et marqué trois types d'entités :

- **FEATURE** : éléments architecturaux (ex. : arc, coupole, pilier)
- **MATERIAL** : matériaux de construction (ex. : marbre, pierre, bois)
- **STYLE** : styles architecturaux (ex. : baroque, roman, Renaissance)

Les annotations ont été saisies dans un fichier CSV structuré, dans lequel chaque ligne contenait le paragraphe ainsi que les entités correspondantes, regroupées par catégorie.

## Partie III : Entraînement du Modèle NER

### 1. Préparation des données annotées

Après la génération et l'annotation manuelle d'un corpus synthétique, la première étape a été de transformer ces données en un format utilisable par l'outil d'apprentissage de spaCy. Chaque ligne du fichier CSV contient un paragraphe descriptif accompagné d'une liste d'éléments correspondant à trois catégories d'intérêt : les caractéristiques architecturales (FEATURE), les matériaux (MATERIAL), et les styles architecturaux (STYLE).

Le script Python développé parcourt chaque ligne du fichier, repère les expressions associées à ces catégories, et calcule leur position exacte dans le texte (indices de début et de fin). Ces informations sont ensuite regroupées sous la forme d'exemples d'entraînement standardisés pour spaCy, qui apprend à localiser et à catégoriser ces éléments dans un paragraphe.

### 2. Construction d'un modèle linguistique sur mesure

Dans une optique de légèreté et de simplicité, un modèle vierge (`spacy.blank("en")`) a été utilisé. Cela permet de bâtir un système spécifiquement adapté à la tâche, sans être influencé par des connaissances générales préalablement intégrées.

Un composant de reconnaissance d'entités nommés (NER) est alors ajouté au pipeline de traitement. On y introduit explicitement les trois étiquettes d'entités à apprendre (FEATURE, MATERIAL, STYLE). Ce choix permet au modèle de concentrer toute son attention sur les types d'éléments réellement pertinents dans le contexte des descriptions architecturales.

### 3. Apprentissage du modèle

L'entraînement est effectué via une boucle personnalisée sur 30 itérations. À chaque époque, les exemples sont mélangés de manière aléatoire afin de renforcer la robustesse du modèle. Chaque exemple est transformé en un objet `Example` propre à spaCy, à partir duquel le système ajuste ses prédictions internes grâce à une fonction de perte.

Un taux d'abandon (dropout) de 0.3 est appliqué à chaque mise à jour pour éviter le surapprentissage. Cela signifie qu'une partie des connexions internes du modèle est volontairement ignorée à chaque étape, ce qui améliore sa capacité à généraliser sur des textes qu'il n'a jamais vus.

Tout au long du processus, les pertes sont affichées à chaque itération, permettant d'évaluer la progression de l'apprentissage.

## 4. Sauvegarde du modèle

À la fin de l'entraînement, le modèle est sauvegardé dans un dossier local (my\_custom\_model). Il peut ensuite être chargé et utilisé dans n'importe quel environnement Python pour analyser automatiquement de nouveaux paragraphes décrivant des bâtiments historiques, et en extraire les éléments d'intérêt.

Ce modèle constitue le cœur du système intelligent conçu dans le cadre de ce projet : il permet de transformer un texte libre en informations structurées exploitables par des systèmes numériques.

# Partie IV : Déploiement d'une Interface d'Analyse Interactive avec Streamlit

## 1. Introduction

Une fois le modèle d'apprentissage entraîné et validé, l'étape suivante consiste à le rendre accessible via une interface conviviale. Pour cela, nous avons opté pour Streamlit, un framework léger et intuitif en Python permettant de créer des applications web interactives avec un minimum de code. Cette interface permet aux utilisateurs, même non techniques, de tester et exploiter le modèle directement à partir de leur navigateur.

Le but est de démocratiser l'accès à l'analyse automatique des textes descriptifs d'édifices historiques, en facilitant la saisie, la visualisation des entités extraites, et la compréhension des résultats.

## 2. Structure Générale de l'Application

L'application se divise en plusieurs modules principaux :

- Chargement du modèle entraîné (my\_custom\_model)
- Saisie du texte par l'utilisateur
- Extraction automatique des entités via le modèle NLP
- Mise en évidence des entités détectées dans le paragraphe
- Affichage d'un tableau synthétique des résultats

Cette architecture assure un flux clair et logique, de l'entrée à l'interprétation des données.

### 3. Présentation Visuelle de l'Application Streamlit

Afin d'illustrer concrètement le fonctionnement de l'interface développée avec Streamlit, deux captures d'écran ont été réalisées, correspondant aux deux étapes clés de l'interaction utilisateur.

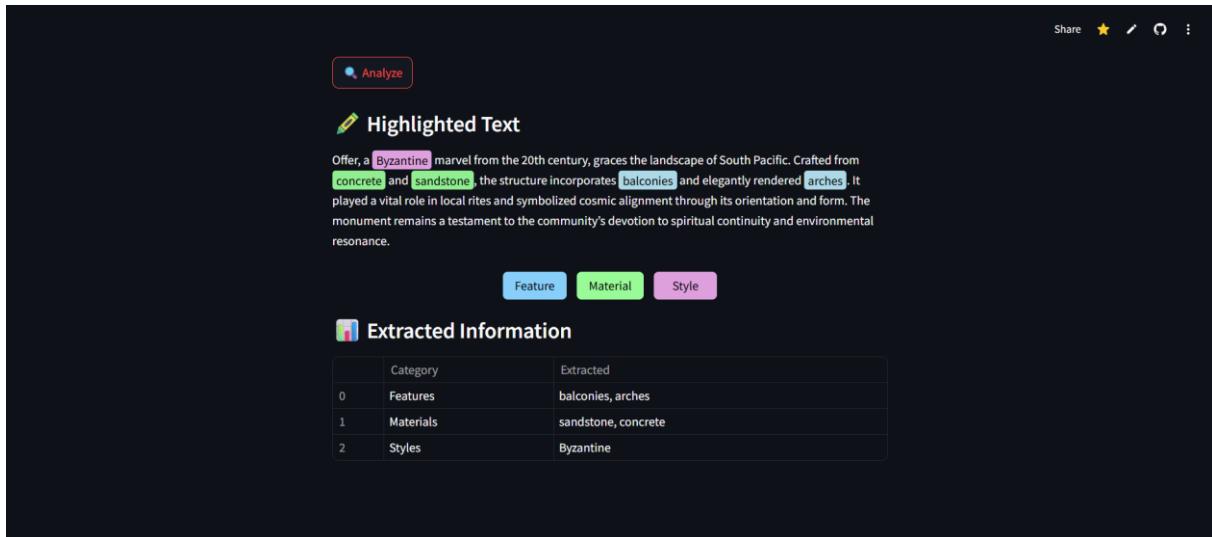


**Figure 1 : Interface au Lancement**

Dans cette première vue, l'utilisateur accède à une interface épurée composée de :

- Un titre principal (“ Historical Buildings Analyzer”) qui indique clairement l’objectif de l’outil.
- Une zone de saisie où l’utilisateur est invité à coller ou rédiger un paragraphe décrivant un bâtiment historique.
- Un bouton “ Analyze” permettant de déclencher l’analyse linguistique.

À ce stade, aucune sortie n'est encore affichée. L'objectif est de permettre une prise en main simple et immédiate.



**Figure 2 : Résultat Après Analyse d'un Paragraphe**

Une fois le paragraphe saisi et le bouton “Analyze” cliqué, l’interface génère automatiquement deux blocs d’informations :

#### ❖ Texte Surligné (Highlighted Text)

Le paragraphe est affiché avec les entités reconnues mises en évidence par un code couleur intuitif :

- Bleu clair (FEATURE) pour les éléments architecturaux (ex. : fenêtre, coupole, arc).
- Vert clair (MATERIAL) pour les matériaux (ex. : marbre, bois, pierre).
- Violet (STYLE) pour les styles (ex. : roman, gothique).

Chaque mot reconnu par le modèle est entouré d’une couleur correspondant à sa catégorie, ce qui permet une lecture immédiate et visuelle des éléments identifiés.

Un bloc de légende colorée placé juste en dessous rappelle la signification de chaque couleur pour faciliter l’interprétation.

#### ❖ Tableau d’Informations Extraites

Sous le texte, un tableau synthétique est généré automatiquement, regroupant :

- Les Features (éléments architecturaux)
- Les Materials (matériaux)
- Les Styles (styles architecturaux)

Chaque ligne du tableau présente la catégorie ainsi que les mots extraits qui y sont associés dans le paragraphe analysé. Ce tableau permet d’obtenir une vue condensée et exploitable des données textuelles.

Ces deux vues combinées offrent une expérience utilisateur claire, interactive, et pédagogique. Elles permettent à la fois une lecture visuelle intuitive et une interprétation analytique structurée du texte, valorisant ainsi le contenu patrimonial de manière intelligente.

## Conclusion Générale

Le patrimoine architectural constitue une richesse inestimable, témoin de l'histoire, des savoir-faire et des civilisations passées. Pourtant, malgré son importance, il reste encore difficilement accessible et exploitable à grande échelle, notamment en raison de la forme non structurée des textes qui le décrivent. Face à l'essor des technologies numériques et à la montée en puissance de l'intelligence artificielle, il devient essentiel de repenser la manière dont ces informations sont collectées, traitées et valorisées.

Ce projet s'inscrit dans cette démarche de modernisation et d'innovation. Il a permis de concevoir une solution complète qui repose sur le traitement automatique du langage naturel (NLP) pour extraire des informations pertinentes à partir de descriptions textuelles de bâtiments historiques. Grâce à un modèle personnalisé entraîné à partir de données annotées, il est désormais possible d'identifier automatiquement des éléments clés comme les styles architecturaux, les matériaux utilisés ou les caractéristiques structurelles d'un édifice.

L'intégration de cette solution dans une interface interactive, développée avec Streamlit, offre une expérience simple, visuelle et intuitive. Elle permet non seulement d'analyser un texte de manière instantanée, mais aussi de visualiser les entités reconnues grâce à un système de surlignage coloré, tout en les regroupant dans un tableau clair et synthétique.

Ce travail démontre qu'il est possible, avec des moyens relativement simples et bien ciblés, d'apporter une réelle valeur ajoutée à la compréhension et à l'exploitation du patrimoine écrit. Il ouvre également la voie à de nombreuses perspectives : enrichissement de bases de données patrimoniales, outils pédagogiques pour les étudiants en histoire ou architecture, ou encore valorisation touristique à travers des applications intelligentes.

Ce projet n'est donc qu'un point de départ. Il montre qu'en alliant technologie et culture, on peut faire émerger de nouveaux usages, plus accessibles, plus interactifs et surtout plus connectés à notre époque. L'avenir du patrimoine passera sans doute par ce type d'initiatives hybrides, qui mettent l'intelligence artificielle au service de la mémoire collective.