



Rapport

Module : Traitement automatique du langage naturel

Master 1 SII

Mini projet

**Réalisation d'un outil d'aide au
développement d'un dictionnaire de la
langue arabe sur des bases historiques**

- Réalisé par :

BENHADDAD Wissam

BOURAHLA Yasser

MOHAMEDI Haroune

LAHBIB Abdelghani

10-12-2018

Table des matières

Table des matières	2
Table des figures	3
1 Motivations et problématique	4
1.1 Introduction	4
1.2 Définitions	5
1.2.1 DataSet et Corpus	5
1.2.2 TALN	5
1.2.3 Dictionnaire historique	5
1.3 Conclusion	5
2 Conception du système	6
2.1 Introduction	6
2.2 Schéma global du système	6
2.3 Les modules du système	7
2.3.1 Aspirateur de sites web	7
2.3.2 Organisateur de corpus	7
2.3.3 Corpus reader	8
2.3.4 Base de données	9
2.3.5 Application (Front-end et Back-end)	10
2.4 Déploiement des modules dans le cloud	10
2.5 Conclusion	10
3 Réalisation de l'application	11
3.1 Introduction	11
3.2 Environnement de travail et outils utilisés	11
3.2.1 Python	11
3.2.2 JavaScript	11
3.2.3 NLTK	11
3.2.4 VueJS	12
3.2.5 Django	12
3.2.6 PostgreSQL	12
3.2.7 Google Cloud Platform	12
3.3 Présentation de l'application	13
3.3.1 Interface principale	13
3.3.2 Corpus	13
3.3.3 Corpus Browser	13
3.3.4 Add Entry	14
3.3.5 Dictionary	14
3.3.6 Historical dictionary	15
3.3.7 Graphs	15
3.4 Fonctions supplémentaires	15
3.5 Conclusion	15
4 Conclusion générale	16
4.1 Objectifs atteints	16
4.2 Limites du système	16
4.3 Perspectives futures	16
Bibliographie	17

Table des figures

2.1	Schéma global du système	6
2.2	Schéma de la base de données	9

1.1 Introduction

Depuis son apparition (au II^e siècle), la langue arabe n'a cessé d'évoluer, donnant naissance à de nouveaux mots, ou modifiant le sens de mots existants. Cette évolution a particulièrement enrichi le vocabulaire de la langue de l'islam, en conséquent et au fil du temps, plusieurs ouvrages destinés à recenser les différentes définitions et sens d'un mot ont vu le jour, chacun durant sa période. néanmoins, il est primordial de garder une trace des différents changements qui ont eu lieu sur ces mots, et cela depuis leur émergence. C'est avec cette idée en tête que les lexicographes des temps modernes en ont eu l'initiative d'entamer la construction de dictionnaires historiques afin de regrouper toutes les nuances des mots à travers les âges.

La création d'un tel ouvrage n'est pas chose facile, en effet elle demande d'une part une grande connaissance sur les différentes périodes historiques de la langue, ainsi que sur la langue en elle-même durant ces périodes. Chercher et regrouper des écrits, documents et ouvrages des différents auteurs durant ces périodes est une tâche qui est en elle-même très ardue, cela peut prendre plusieurs décennies pour créer une collection de documents assez représentative de chaque période. Analyser le contenu de tous ces documents est la phase qui dure le plus de temps, une vérification minutieuse de chaque information ajoutée au dictionnaire final doit être faite, puis soumise à l'approbation de plusieurs experts du domaine.

C'est avec l'avènement de l'informatique, de l'intelligence artificielle et plus récemment avec l'explosion du volume de données présentes sur internet, que l'idée d'utiliser ces technologies pour faciliter et accélérer le processus de création d'un dictionnaire historique de la langue arabe a émergé. En effet la grande quantité et diversité de documents présents sur internet pourraient être exploités par un lexicographe pour ne pas s'attarder sur la fastidieuse tâche de collecte des données, et cela en utilisant des techniques de traitement automatique du langage (TALN), de recherche d'information (RI) et d'intelligence artificielle.

Ce besoin d'un outillage informatique est la principale motivation derrière ce mini-projet, avec suffisamment de données et une bonne conception, la réalisation d'un tel outillage pourrait faire gagner énormément de temps aux lexicographes du monde arabe.

Le but de projet étant maintenant établi, nous allons maintenant passer à la schématisation de ce rapport. Nous commencerons d'abord par de petites définitions pour se situer dans la suite du rapport, nous enchaînerons ensuite sur la conception du système pour expliquer le travail réalisé, viendra ensuite la présentation de notre application, enfin nous finirons par une conclusion générale comportant un bilan du projet, des critiques sur notre système ainsi que les perspectives envisagées.

1.2 Définitions

1.2.1 DataSet et Corpus

Un jeu de données (DataSet) est un ensemble de données traité et organisé dans un schéma spécifique aux besoins d'un système, un dataset peut être une base de données relationnelle, un ensemble de fichiers texte, une banque d'images/videos ...

Dans notre cas nous nous intéresserons plus particulièrement à un type de dataset appelé Corpus, informellement un corpus est un dataset principalement utilisé dans le domaine du TALN, il est constitué d'un ensemble de fichier texte (annotés ou pas) qui représentent un domaine, une thématique, un(ou des) type(s) d'ouvrages ...

Un corpus est un composant essentiel pour la l'application des techniques de TALN, la taille et la qualité d'un corpus est donc un facteur primordial pour assurer une bonne performance d'un système.

1.2.2 TALN

Le Traitement Automatique du Langage Naturel (TALN) est un sous domaine de l'intelligence artificielle qui vise à analyser et à modéliser les composants du langage humain, que ce soit du point de vue syntaxique, sémantique ou pragmatique. l'aspect principal du TALN est le fait de permettre aux machine de traiter les séquence de texte non plus comme une simple suite de symboles, mais comme des entités informationnelles. Des connaissances sur la langue sont un prérequis essentiel pour le développement d'un système utilisant le TALN, ainsi que la disponibilité d'un grand ensemble de données pour faire de l'apprentissage automatique.

Le TALN est découpé en un ensemble de techniques et opérations à appliquer sur du texte, une multitude de domaine d'application existent pour l'utilisation de ces derniers. Dans ce projet nous nous intéresserons principalement aux techniques suivantes :

Lemmatisation

La lemmatisation est un terme désignant l'analyse lexicale d'un texte dans le but de regrouper les mots d'une même famille. Les mots d'une même famille sont donc réduits en une unique entité appelée « **lemme** ». Ainsi la lemmatisation consiste à regrouper les différentes flexions d'un mot unique.[1]

Segmentation

La segmentation d'un texte est l'opération de découpage de ce dernier en composantes linguistiques plus petites(des phrases, des groupes nominaux, des mots ...), c'est un processus non-trivial car chaque langue dispose de règles spécifiques en ce qui concerne les marqueurs de fin de phrases.

Étiquetage morphosyntaxique (PoS-Tagging)

il consiste à identifier pour chaque mot sa classe morphosyntaxique(Nom,Verbe,Nom pluriel, ...) à partir de son contexte dans un corpus ou texte ,ainsi que de connaissances lexicales de la langue.

1.2.3 Dictionnaire historique

Informellement, un dictionnaire historique est un ouvrage qui rassemble, sous forme d'un liste d'entrées, un ensemble de mot d'une langue donnée avec leurs définitions et/ou des exemples d'utilisation selon des périodes historiques prédéfinies (en rapport avec la langue ou pas).

1.3 Conclusion

Au terme de ce chapitre, nous avons une idée plus claire sur le travail qui doit être réaliser, nous allons donc attaquer l'aspect conceptualisation, il s'agira principalement de définir les composants de notre système.

2.1 Introduction

Comme mentionné précédemment, nous allons nous intéresser dans ce chapitre à la conception que nous avons réalisé, nous présenterons un schéma global du système, puis nous nous détaillerons le rôle de chaque composant, en donnant un exemple d'utilisation et/ou du flux de données qui entre/sort de ce dernier, nous parlerons ensuite du déploiement du système dans une plateforme serverless(dans le cloud), principalement car c'est un aspect important de l'expérience d'utilisation(UX).

2.2 Schéma global du système

Notre système se compose essentiellement de deux parties(elles même subdivisées en plusieurs modules) :

- **Récupération et pré-traitement des données** : principalement, c'est depuis des sites web que le système cherche des données, puis il se charge d'organiser les fichiers téléchargés dans un espace de stockage.
- **Exploitation et mise à jour des données récupérées** : c'est la partie où les données qui sont maintenant structurées et organisées seront utilisées par l'application, qui dans notre cas se trouve être une application web hébergé dans le cloud.

Le schémas suivant explicite un peu plus l'explication précédente :

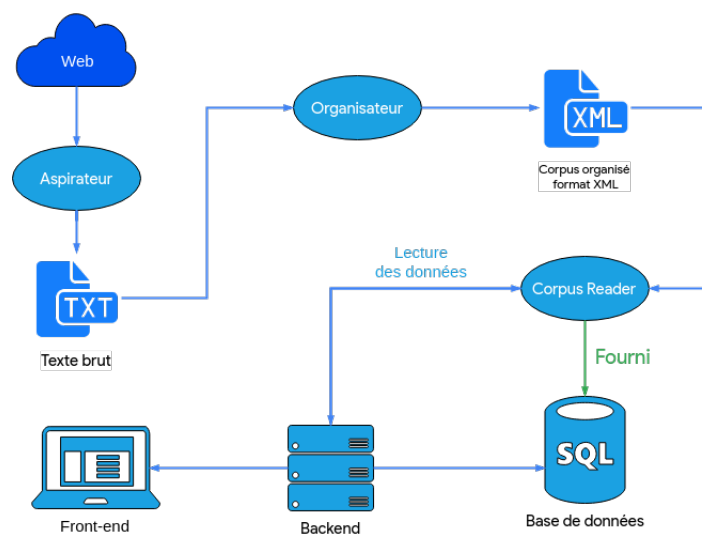


Figure 2.1 – Schéma global du système

2.3 Les modules du système

Cette section se consacre à la description en détails des différentes modules du système, avec éventuellement des exemples d'utilisation de chacun.

2.3.1 Aspirateur de sites web

Le module d'aspiration est composé de plusieurs petits scripts, le premier d'entre eux **initializer.py**, son rôle est d'initialiser certaines variables globales (comme les dates de début/fin des périodes historiques de la langue arabe (... العصر الجاهلي، عصر صدر الإسلام ...), ou encore les chemins d'accès des différents fichiers récupérés et des fichiers xml produits. Il permet aussi la création dynamique d'une arborescence de répertoire où placer chaque fichier aspiré selon son époque, son genre, son type ...

Les autres scripts sont destinés à aspirer un type de document en particulier, nous pouvons citer les différents "scrapeur"¹ utilisé :

- **books_scrape.py** Il a pour but d'aspirer le contenu du site alwaraq.net, qui est un recueil de livres arabes.
- **chi3r_scrape.py** Son but est d'aspirer le contenu du site aldiwan.net, qui est une collection de poèmes arabes organisés par auteurs, périodes, pays ...
- **islamicbook_scrape.py** Son but est d'aspirer le contenu du site islamicbook.ws, qui est un recueil d'ouvrages dédiés à l'islam.
- **news_scrape.py** Son but est de télécharger les corpus déjà organisés depuis le site aracorpus.e3rab.com qui est une banque de blogs rédigés en arabe.
- **quran_corpus_builder.py** par respect pour les verset coraniques, nous avons décidé de télécharger un corpus déjà organisé sous format xml et certifié comme étant authentique par le groupe **Tanzil Project** (voir tanzil.net), puis nous l'avons ré-organisé en utilisant le script en une structure **XML** (que nous définirons plus bas).

Il a noté que tout ces scripts utilisent des fonctions d'organisation communes, pour placer chaque fichier acquis dans son répertoire adéquat dans l'arborescence créée par **initializer.py**, cela a pour but de faciliter l'accès ultérieur aux données bruts, pour les nettoyer et construire les corpus au format **XML**.

2.3.2 Organisateur de corpus

Ce module est composé principalement d'un seul script qui effectue deux tâches, une fonction de nettoyage "**cleaner**", et un constructeur de fichiers **_cleanNotFound** pour générer le fichier **XML** associé à chaque texte brut extrait. Le format des fichiers XML que nous avons choisis est le suivant :

```
1 <root encoding="utf-8">
2   <metadata>
3     <book_name>ألا حي بالبردين دارا ولا أرى</book_name>
4     <era>Umayyad</era>
5     <author>
6       <name>جرير</name>
7       <birth>650</birth>
8       <death>728</death>
9     </author>
10    <id>116</id>
11    <type>شعر</type>
12    <size>66</size>
13  </metadata>
14  <doc>
15    <sentence>ألا حي بالبردين دارا ولا أرى</sentence>
16    ...
17    <sentence>كرتيا ولم تعلق عنانا يقيمها</sentence>
18  </doc>
19 </root>
```

1. Terme anglais pour désigner un aspirateur de site web

Remarque : Pour les textes coraniques, chaque **sura** (سورة) sera représenté par un fichier **XML**, dont les balises **<sentence>** contiendront ses versets (آيات السورة)

Chaque document du corpus sera donc divisé en deux ensembles de balise, une partie méta-données contenant une variété d'information sur cet élément du corpus (Nom de l'auteur, période historique, id , type ...), puis une partie données, qui contiendra plusieurs phrases qui composent le document.

Après avoir construit l'arborescence, l'initialisation du système sera presque complète (il faudra remplir les tables de la base de données, nous verrons le schéma relationnel plus tard), il s'agira maintenant de passer la main au module suivant pour l'exploitation de ce corpus.

2.3.3 Corpus reader

Ce module à joue le rôle d'une interface d'abstraction entre les fichiers XML et le serveur qui désire récupérer ces données, ce dernier devra instancier un objet de la classe **HistoricalCorpus**, cette classe hérite directement de la classe **nlk.corpus.XMLCorpusReader**, en redéfinissant les méthodes approprié, nous pourrons donc accéder au corpus et ses composants en invoquant des méthodes comme :

```
1 def sents(fileid=None,start=None,end=None,era=None,category=None):
```

Cette fonction retourne une liste de phrase du corpus tout entier, ou d'une portion selon les valeurs des paramètres qui lui sont passer (période,catégorie...).

```
1 def tagged_sents(fileid=None,start=None,end=None,era=None,category=None):
```

Retourne une liste de phrases avec leurs étiquetages morpho-syntaxique.

```
1 def lemma_sents(self,fileid=None,start=None,end=None,era=None,category=None):
```

Retourne une liste de phrases auxquelles une lemmatisation à été appliquée.

```
1 def words(fileid=None,start=None,end=None,era=None,category=None):
```

Retourne une liste de tout les mots du corpus (ou d'une partie du corpus).

```
1 def tagged_words(fileid=None,start=None,end=None,era=None,category=None):
```

Retourne une liste de tuples (mots,Étiquette-Morpho-Syntaxique) de tout les mots du corpus (ou d'une partie du corpus).

```
1 def lemma_words(self, fileid=None,start=None,end=None,era=None,category=None):
```

Retourne une liste de de tout les lemmes (forme infléchié des mots) du corpus (ou d'une partie du corpus)

Le CorpusReader est invoqué principalement durant l'initialisation du serveur (avant l'utilisation de l'application), et cela pour garantir une utilisation plus aisée de l'application après déploiement. Il reste donc à voir comment le serveur remplit la base de données.

2.3.4 Base de données

Le but d’avoir cette base de données et principalement de stocker les différentes méta-données sur les documents du corpus, ainsi que les relations entres les contenus des documents, cela facilitera la tâche au serveur de ne considérer que le niveau le plus abstrait des données, et laisser la tâche de récupérer et manipuler ses données au CorpusReader.

Nous avons choisis le schéma relationnel suivant pour la base de données :

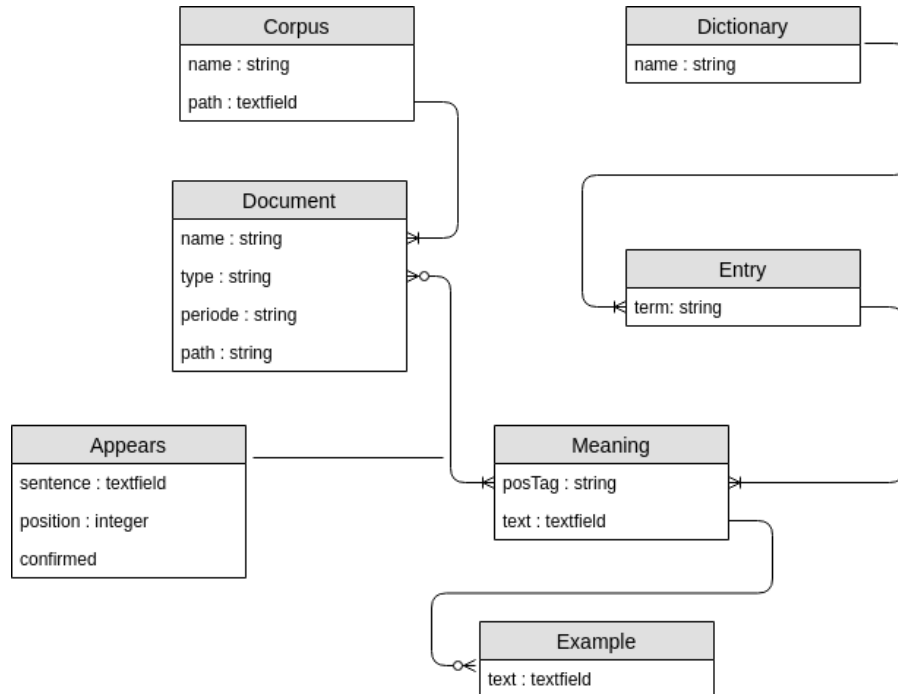


Figure 2.2 – Schéma de la base de données

La table **Dictionary** est rempli avant la fin de la phase d’initialisation, nous avons pour cela utilisé les données provenant de deux sources :

- المعجم الرائد
- المعجم الجامع
- المعجم الوسيط

Nous avons tout d’abord transformer les entrées du dictionnaire en un fichier **JSON** dont le format est le suivant :

```
1  {
2    "root" : {
3      "postag-1" : [
4        "def-1",
5        ...
6      ],
7
8      ...
9
10     "postag-n" : [
11       "def-1",
12       ...
13     ]
14   }, ...
15
16 }
```

2.3.5 Application (Front-end et Back-end)

Back-end : C'est une application web qui expose une api REST-ful , elle est accessible a partir de certains **end-points**² qui seront exploités avec des requêtes **http**, l'avantage d'une telle architecture est qu'elle est 100% inter-opérable (elle ne dépends ni du système, ni des technologies ni du langage du serveur ou des clients).

Front-end le front-end utilise ces **endpoints** pour récupérer de la donnée et l'afficher à travers des requêtes lancée selon les besoins de l'utilisateur (en manipulant l'interface par exemple). On peut aussi développer une application mobile ou desktop de la même façon.

2.4 Déploiement des modules dans le cloud

Nous avons choisi de déployer notre application suivant une architecture server-less dans le but de faciliter son utilisation, de la rendre publique et disponible à tout ceux qui voudraient la tester, donner un feedback ou proposer une amélioration en consultant le dépôt sur [Github](#)

2.5 Conclusion

Au terme de chapitre, nous avons donc fais le tour des principales composantes du système, sans trop rentrer dans les détails, nous introduirons les composantes non-citées dans le chapitre suivant lors de la présentation des fonctionnalités de l'application si besoin est.

2. Emplacement dans le serveur renvoyant une réponse à une requête

3.1 Introduction

Durant ce chapitre, nous allons présenter l'aspect pratique de notre système, à savoir l'application qui lui est dédiée, nous commencerons par une description des outils utilisés, suivit d'une présentation de l'interface d'utilisation pour enfin introduire les fonctionnalités supplémentaire non-apparente dans cette dernière.

3.2 Environnement de travail et outils utilisés

3.2.1 Python



Python est un langage de programmation scripté très puissant, de par sa simplicité d'utilisation et de sa syntaxe très intuitive, il est aussi un des langages préférés des adeptes du TALN, pour la raison qu'un grand nombre de bibliothèques y sont disponibles. Nous l'avons choisi principalement pour des traitement de TALN mais aussi dans la partie back-end du système.

3.2.2 JavaScript



Très populaire chez les développeur web, JavaScript(JS) est un outil très puissant pour un développement rapide d'application web rapide, puissantes et élégantes.

3.2.3 NLTK

Natural Language Toolkit (NLTK) est un framework (ensemble de librairies) implémentant un grand nombre d'algorithmes et modèles du TALN, écrite en python elle sera donc parfaitement compatible avec notre back-end qui lui sera développé avec **DJANGO** (voir suite).

3.2.4 VueJS



Vue.js est un framework écrit en JavaScript pour le développement d'interface web, basé sur le principe de **components**, Vue donne au développeurs une grande liberté et flexibilité pour la création d'interface dynamiques et **reponsive**.

3.2.5 Django



Django est un framework écrit en Python qui facilite le développement du back-end d'un site web, il s'occupe de gérer les couches basses de ce dernier (sessions, sécurité...) et peut même générer une interface d'administration automatiquement. L'objectif de Django est de proposer un développement plus efficace et plus rapide d'un site web dynamique tout en maintenant sa qualité.

3.2.6 PostgreSQL



PostgreSQL est un système de gestion de base de données relationnelle et objet (SGBDRO). C'est un outil libre disponible selon les termes d'une licence de type BSD.

3.2.7 Google Cloud Platform



Google Cloud Platform est une plateforme de cloud computing fournie par Google, proposant un hébergement sur la même infrastructure que celle que Google utilise en interne pour des produits tels que son moteur de recherche¹. Cloud Platform fournit aux développeurs des produits permettant de construire une gamme de programmes allant de simples sites web à des applications complexes.

3.3 Présentation de l'application

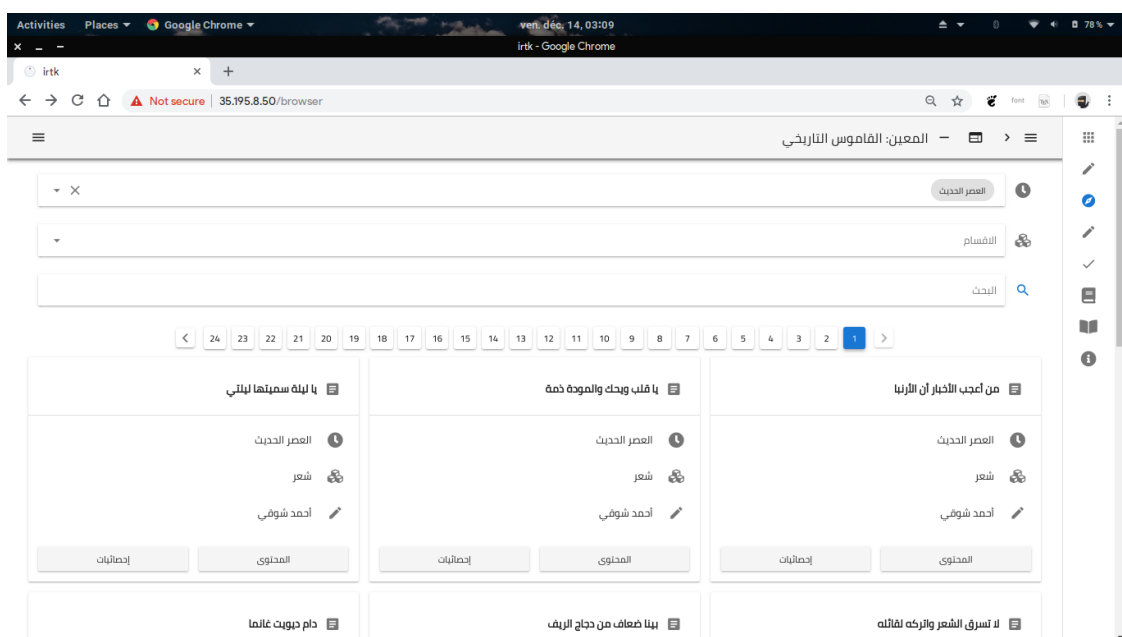
3.3.1 Interface principale

L'application se présente comme composé de trois parties, une barre d'outils en haut, un panneau d'onglet latéral permettant d'accéder aux différentes sections de l'application, et une zone d'affichage principale pour afficher le contenu de chaque section.

3.3.2 Corpus

Dans cette section, la main est donnée à l'utilisateur pour uploader un nouveau corpus sur le serveur principal, il suffit de donner un nom au corpus, une description pour faciliter son utilisation, et sélectionner les fichiers **XML** correspondant au corpus.

3.3.3 Corpus Browser



Dans cette section, l'utilisateur pourra parcourir chaque composant du corpus, le choix lui est donné pour filtrer sa recherche, que ce soit par période (الفترة الزمنية), par section (قسم) ou par mots clés, le résultat est une succession de pages dynamiquement ajoutées à la mémoire (grâce au corpus-reader), chaque page est composée de tuile contenant les informations relatives aux documents pertinents selon le filtre de recherche.

En cliquant sur le bouton **content** (المحتوى), une nouvelle section est affichée, elle permet à l'utilisateur de lire le fichier sélectionné. De plus s'il désire ajouter une nouvelle entrée au dictionnaire historique, il lui suffit de sélectionner la partie du texte qui l'intéresse, une liste déroulante s'affiche lui proposant d'ajouter un des mots précédemment sélectionnés au dictionnaire (nous verrons les détails de cette opération dans la section suivante).

3.3.4 Add Entry

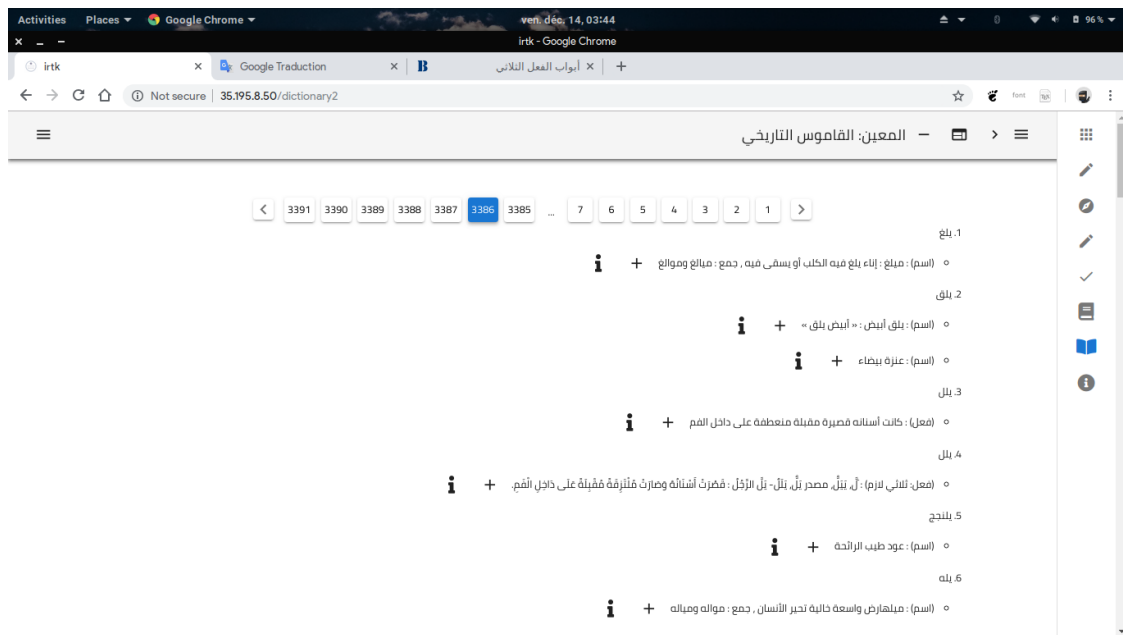
Cette section donne la main au lexicographe pour ajouter une nouvelle entrée dans le dictionnaire historique, pour ce faire il faut entrer le terme à ajouter (المصطلح), il faudra ensuite remplir un (ou plusieurs) champs selon le besoin de l'utilisateur :

- Une liste de lignes réservées aux sens des termes à ajouter, il faudra fournir une étiquette morpho-syntaxique parmi une liste déroulante de choix possible (النوع), suivit d'un extrait de texte où le terme apparaît (النص). L'utilisateur pourra répéter ce processus autant de fois qu'il désire ajouter de sens au terme.
- Une liste de lignes réservées aux informations chronologiques (الأمثلة التاريخية), il faudra fournir la période historique (الفترة الزمنية), la section visée (القسم), le document (الوثيقة) puis finalement un exemple d'utilisation (الجملة), un champ **confirmé** est à cocher pour valider les informations saisies, sinon l'ajout n'est pas autorisé.

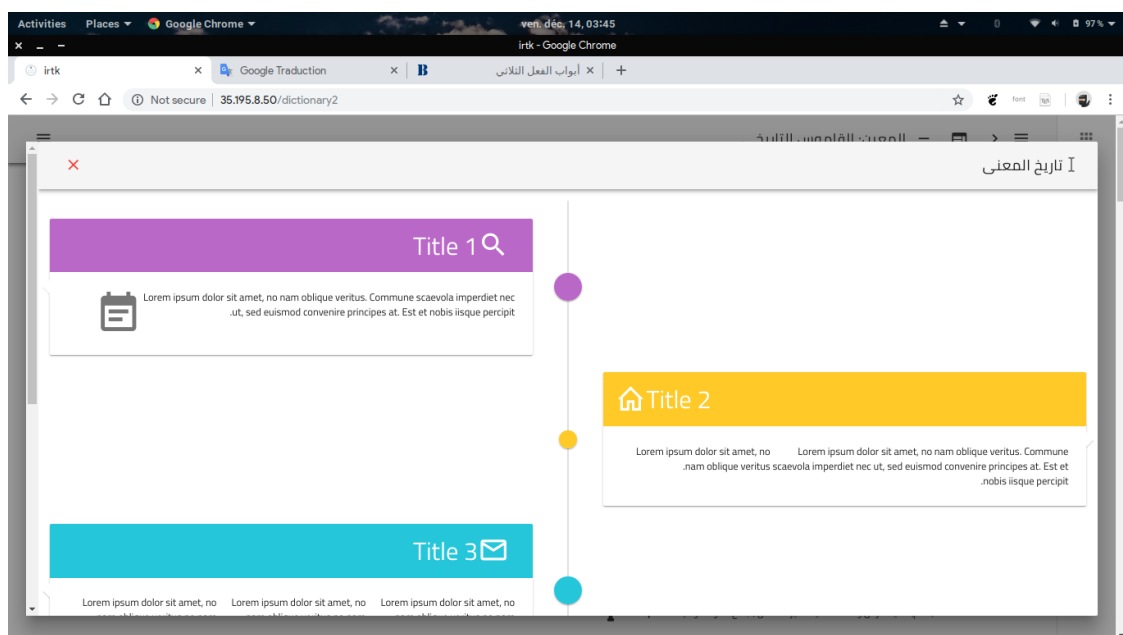
3.3.5 Dictionary

Tout simplement cette section récupère tous les mots du dictionnaire classiques, et affiche les différents sens liés au mot auquel on s'intéresse.

3.3.6 Historical dictionary



Dans cette section, nous trouverons notre dictionnaire historique, organisé comme un dictionnaire classique avec des entrées(des mots) et leurs définitions, mais avec l'option de consulter les différentes utilisation du mot lors des siècles en cliquant sur un bouton, une timeline s'affiche en donnant un récapitulatif chronologique du mot :



L'option + permet de modifier cette entrée, elle renvoie à la section **add-entry** mais avec le champs **terme** déjà rempli avec le terme courant à modifier.

3.3.7 Graphs

Display different statistics for some words

3.4 Fonctions supplémentaires

3.5 Conclusion

4.1 Objectifs atteints

4.2 Limites du système

4.3 Perspectives futures

BIBLIOGRAPHIE

- [1] "La lemmatisation : Optimiser le seo avec la lemmatisation." <https://www.yakaferci.com/lemmatisation-seo/>. (Accessed on 12/13/2018).