

**TUTRICE DE STAGE : RESPONSABLE DE STAGE :**

**Mme. MM.**

**Martine GAUTIER Bruno GUILLAUME**

**Yves LEPAGE**

**Jean LIEBER**

**Emmanuel NAUER**

**Hue-Nam LY**

**Rapport De Stage**

**Développer une infrastructure Web pour un système de correction de phrases en français à partir de cas**

**Du 9 avril au 6 juin 2018**

**Année Universitaire 2017-2018**

****



**Licence 3 informatique**

**SOMMAIRE**

**REMERCIEMENTS 3**

**INTRODUCTION 4**

**PRÉSENTATION DE L’ENTREPRISE 5**

1. **BASE DE DONNÉES 6**
2. Présentation de la base de donnes 6
3. Génération du script de la création et de l’insertion 7
4. **CONCEPTION**
5. Diagramme contexte statique 8
6. Diagramme de cas d’utilisation 8
7. Diagramme d’activité 9
8. Diagramme de séquences 11
9. **ENVIRONNEMENT DE DÉVELOPPEMENT**
10. MySQL et phpMyAdmin 13
11. Bootstrap 13
12. SublimeText 14
13. Balsamiq 14
14. Wampserver 15
15. GIT 15
16. **INTERFACE WEB**
17. Corrector un site multilingue 16
18. La maquette du site 17
19. Interface d’utilisateur 18
20. Interface d’administrateur 19

**CONCLUSION 20**

**SITOGRAPHIE 21**

**ANNEXE 21**

**REMERCIEMENTS**

Je tiens à présenter mes sincères remerciements à :

Messieurs Bruno GUILLAUME, Yves LEPAGE, Jean LIEBER, Emmanuelle NAUER, pour m’avoir accepté dans ce stage, ainsi que leurs explications et leurs conseils durant mes deux mois de stage.

Madame Martine GAUTHIER pour ses enseignements pendant mes années universitaires et son soutien pendant mon stage.

Mes collaborateurs : Messieurs André GIANG, Damien LEVY pour leur aide et leurs idées sur le design du site web.

**INTRODUCTION**

De nos jours, il existe plusieurs systèmes qui corrigent des fautes d’orthographe ou la grammaire automatiquement. Mais que se passe-t-il si on rentrait une phrase syntaxiquement incorrecte ? Est-ce que ces systèmes nous donnent une bonne ou mauvaise solution ?

Dans notre cas, on cherche à corriger ce genre de problème et fournir à l’utilisateur une correction acceptable en utilisant le raisonnement à partir de cas.

Le raisonnement à partir de cas (RàPC) : consiste donc à s’appuyer sur des couples (problème, solution) où le problème est une phrase incorrecte qui est corrigée en solution pour proposer une correction à une nouvelle phrase cible. Par exemple :

*Problème : Je aimer les pommes.*

*Solution : J’aime les pommes.*

La phrase cible sera proposée par les utilisateurs via une interface web. Cela déclenchera le système de remémoration puis d’adaptation et la solution sera affichée sur l’interface web. Par exemple :

*Cible : Je manger des tomates.*

*Solution : Je mange des tomates.*

Si le résultat est accepté par l’utilisateur il forme alors un couple (cible, solution) qui sera insérée dans la base de cas pour être utilisable ultérieurement en tant que nouveau cas.

La base de cas minimale est alimentée de manière manuelle par Mr. Levy. Ensuite elle sera complétée de manière semi-automatique à partir du corpus WiCoPaCo qui est un historique de modification de pages sous Wikipédia. Puis, elle sera améliorée au fur et à mesure de l’utilisation. La remémoration et l’adaptation sont des algorithmes développés par Mr. Giang en Python. En s’appuyant sur le RàPC pour corriger un problème et donner une solution.

Dans ce stage, je dois créer une base de données qui stocke les cas et développer une interface web multilingue qui donnera la possibilité aux utilisateurs de rentrer une phrase incorrecte en français ou anglais et affichera la solution aux utilisateurs.

Tout d’abord, je vais présenter la création d’une base de données, la conception, suivi de l’environnement de développement, et enfin le développement de l’interface web Corrector.

**PRESÉNTATION DE L’ENTREPRISE**

Le stage s’est déroulé au Loria[[1]](#endnote-1), Laboratoire lorrain de recherche en informatique et ses applications situé à Nancy. Ce laboratoire est une Unité Mixte de Recherche (UMR), il est commun à plusieurs établissements le CNRS, l’université de Lorraine et l’Inria. Créé depuis 1997 son domaine est la recherche fondamentale et appliquée en sciences informatiques.

Cet UMR possède un effectif total de plus de 400 personnes et est composée de 28 équipes répartis en 5 départements, dont 15 sont communes avec Inria. Le Loria est un des plus grands laboratoires de la région lorraine.

# Notre stage s’est déroulé au sein du département 4 TALC[[2]](#endnote-2) (traitement automatique des langues et des connaissances) qui comme son nom l’indique est spécialisé dans 3 domaines :

* Langue : concernent notamment la parole, la syntaxe, la sémantique et la représentation du discours.
* Les connaissances : la représentation des connaissances et la formalisation du raisonnement, la découverte de connaissances par l’utilisation de méthodes symboliques ou numériques de fouille de données, les méthodes d’apprentissage et la classification.
* Documents : la reconnaissance de formes, la reconnaissance de symboles et la modélisation de l’écriture manuscrite.

Nous étions dans l’équipe Orpailleur qui compte 12 membres.

1. **BASE DE DONNÉES**
2. Présentation de la base de données

Pour ce stage, le type de base de données n’était pas imposé. J’ai choisi mySQL, car c’est celui que j’ai utilisé pendant mon année universitaire.

La principale difficulté est que cette base doit être adaptée aux travaux de Mr. Giang et de Mr. Levy. Ainsi, après plusieurs réunions, on a décidé que cette base est de la forme présente en figure 1. Le MCD a été créé par un site web mocodo[[3]](#endnote-3).

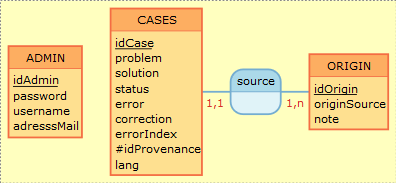


Fig.1 Base de données du système Corrector.

Notre base de données contient trois relations :

La relation CASES est composée d’une clé primaire idCase, ainsi que deux colonnes « problem » et « solution », ces deux dernières forment alors le couple (problème, solution) constituant un cas. La colonne « status » est de forme « correct, incorrect, en attente » pour indiquer si la solution est correcte. Les deux colonnes suivantes sont « erreur et correction » elles indiquent quel(s) le(s) mot(s) a(ont) été corrigé(s). « errorIndex » donne la position du mot erroné dans la phrase. « lang » pour l’instant notre projet ne comporte que du français et de l’anglais mais cela peut évoluer dans le futur.

La relation ORIGIN (ou provenance) stocke les informations de la provenance du couple (problème, solution), c’est-à-dire de quelle source Mr Levy a trouvé ce genre de cas. La clé primaire est « idOrigin » et « originSource » est le nom de la source (Par exemple : WiCoPaCo un site de correction et modification de la page wikipedia).

La relation ADMIN contient les informations nécessaires à l’administrateur.

1. Génération du script de création et d’insertion
2. Création

Le code de création d’une base est très standard (voir figure 2) :

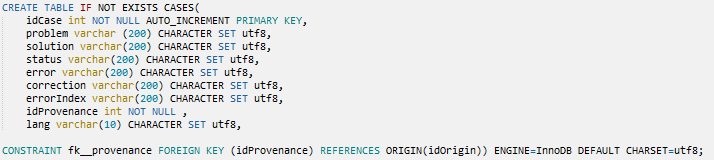


Fig.2 Code de création de la relation CASES.

Le seul problème que j’ai rencontré ici est la partie « ENGINE=InnoDB ». Il faut précise le type ENGINE si non le système va prendre celui par défaut : MYISM. Cette dernière ne prend pas en charge les clés étrangères.

1. Insertion

Les données du projet sont collectées par Mr Levy et stockées dans un fichier.csv. A partir de ce fichier on a deux choix :

1. On peut l’importer directement sur phpMyAdmin avec le système import fichier que fournit le site.
2. On développe un script php pour importer les données et les insérer dans la base.

J’utilise la deuxième méthode. Je développe deux fonctions : une qui stocke les informations du fichier.csv dans un « array », détaille du code dans la figure 3. Et la deuxième fonction insère les contenus de cette « array » dans la base de données (voir figure 4).



Fig.3 importCSVFile. Fig.4 insertData.

1. **LA CONCEPTION**
2. Diagramme de contexte statique

Le diagramme de contexte permet de visionner les interfaces entre le système sous enquête et le système des entités externes avec lesquels il est en interaction représenter par la figure 5. Dans notre cas, des entités externes sont des utilisateurs (USER) et le système sous enquête est le moteur de correction.

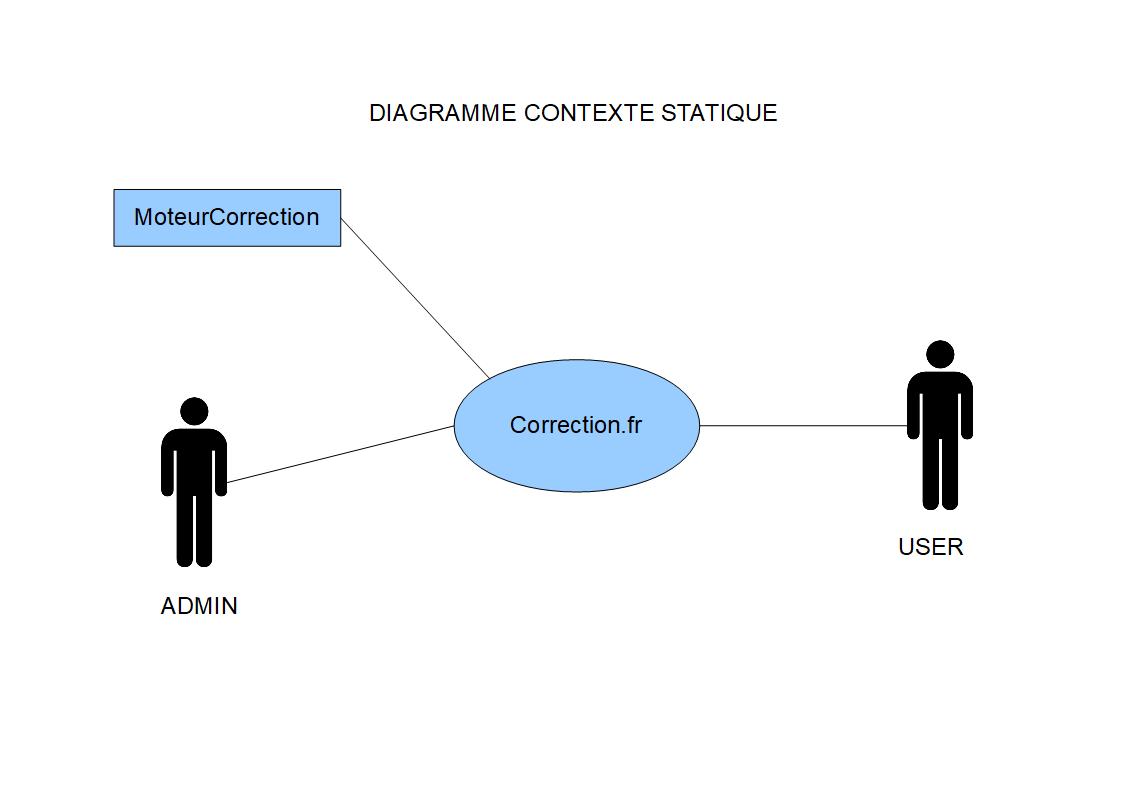
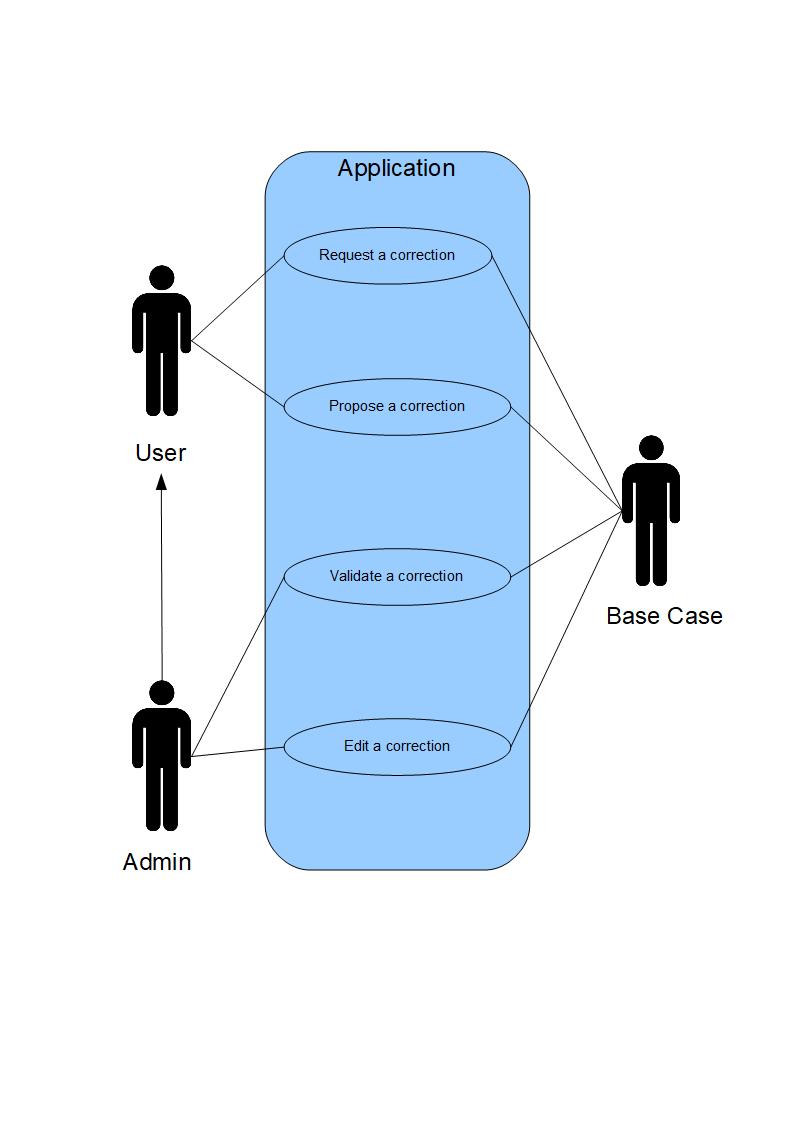


Fig.5 Diagramme de contexte statique.

1. Diagramme de cas d’utilisation

Le diagramme de cas d’utilisation représente les fonctionnalités nécessaires pour les utilisateurs (voir figure 6). Après les études des besoins des utilisateurs, on a trouvé quatre types de cas :

* Demander d’une correction (*request correction*).
* Proposer une correction (*propose correction*).
* Valider une correction (*validate a correction*).
* Modifier une base de cas (*Edit case base*).

Les deux derniers types sont réservés aux utilisateurs ayant un droit d’administrateur.

Fig.6 Diagramme de cas d’utilisation.

1. Diagramme d’Activité

Le diagramme d’activité représente le déroulement des actions, il est utilisé pour détailler les spécifications d’un cas d’utilisation et le déclenchement d’événements en fonction des états du système et la modélisation de ses comportements. Ainsi après, le diagramme de cas d’utilisation, on possède alors quatre diagrammes d’activité.

1. Demander une correction

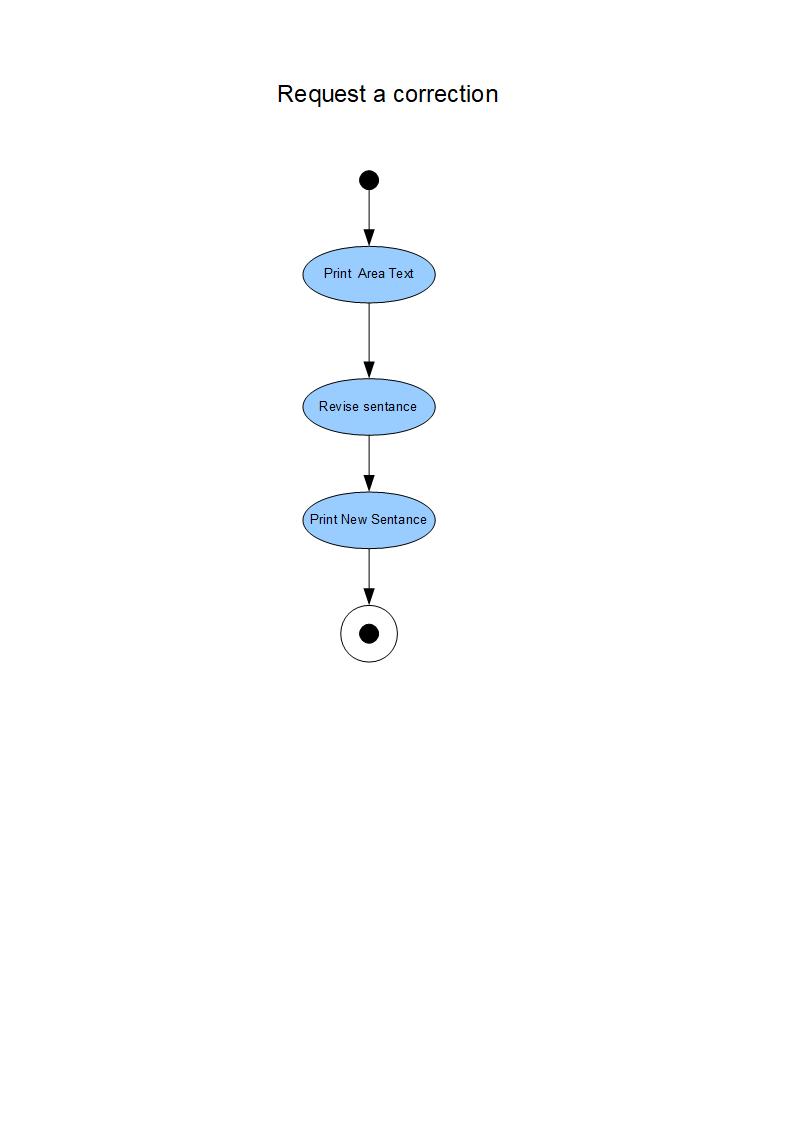
Le système affiche une zone de texte où un utilisateur peut entrer la phrase incorrecte. Ensuite il récupère la phrase et déclenche le moteur de correction (Revise Sentence), puis, il affiche le résultat à l’utilisateur (présenter dans la figure 7).

Fig.7 Demander une correction.

1. Proposer une correction

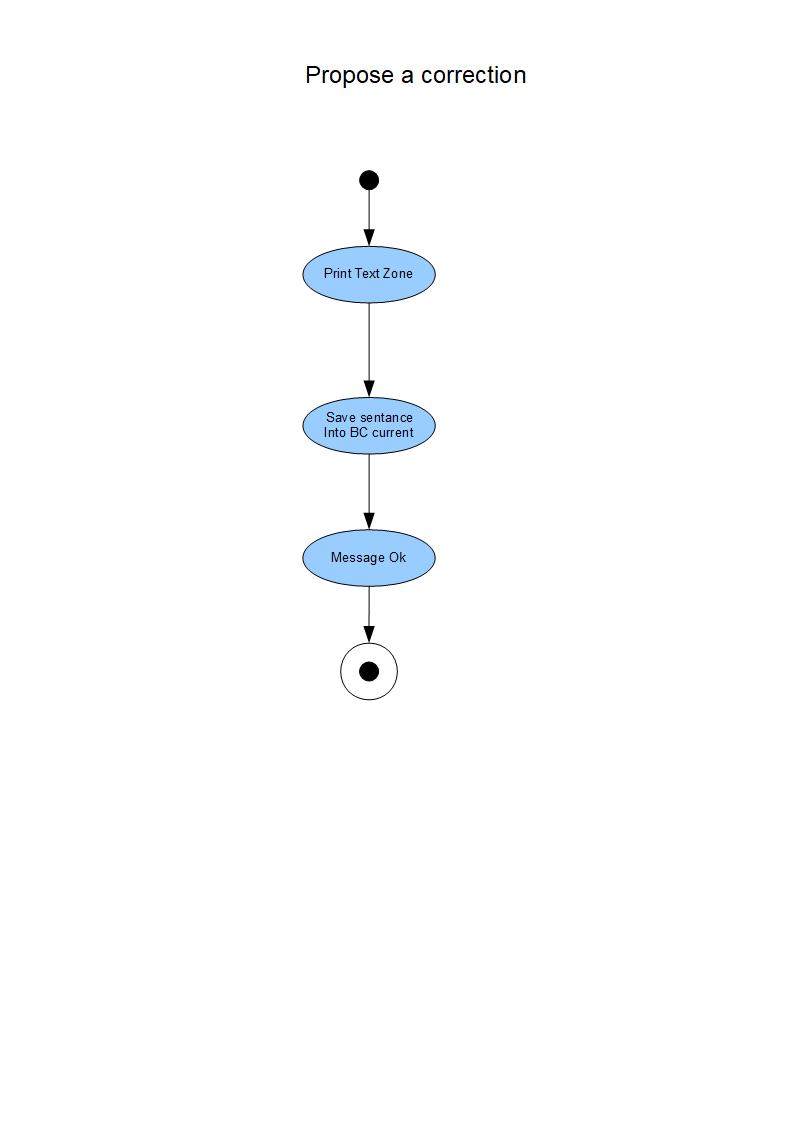
Le système a le même comportement que ‘’Demander une correction ‘’ (voir figure 8), mais au lieu de déclencher le moteur de correction, il sauvegarde la phrase proposée par l’utilisateur dans la base courante.

Fig.8 Proposer une correction.

1. Valider la phrase

Cette activité est réservée pour les administrateurs. Le système affiche la phrase proposée par un utilisateur, puis elle déclenche le moteur de correction et affiche le résultat. Il attend alors une confirmation de l’administrateur sur ce résultat pour savoir s’il doit stocker la phrase proposée dans la base de données ou l’enlever de la base de case (figure 9).

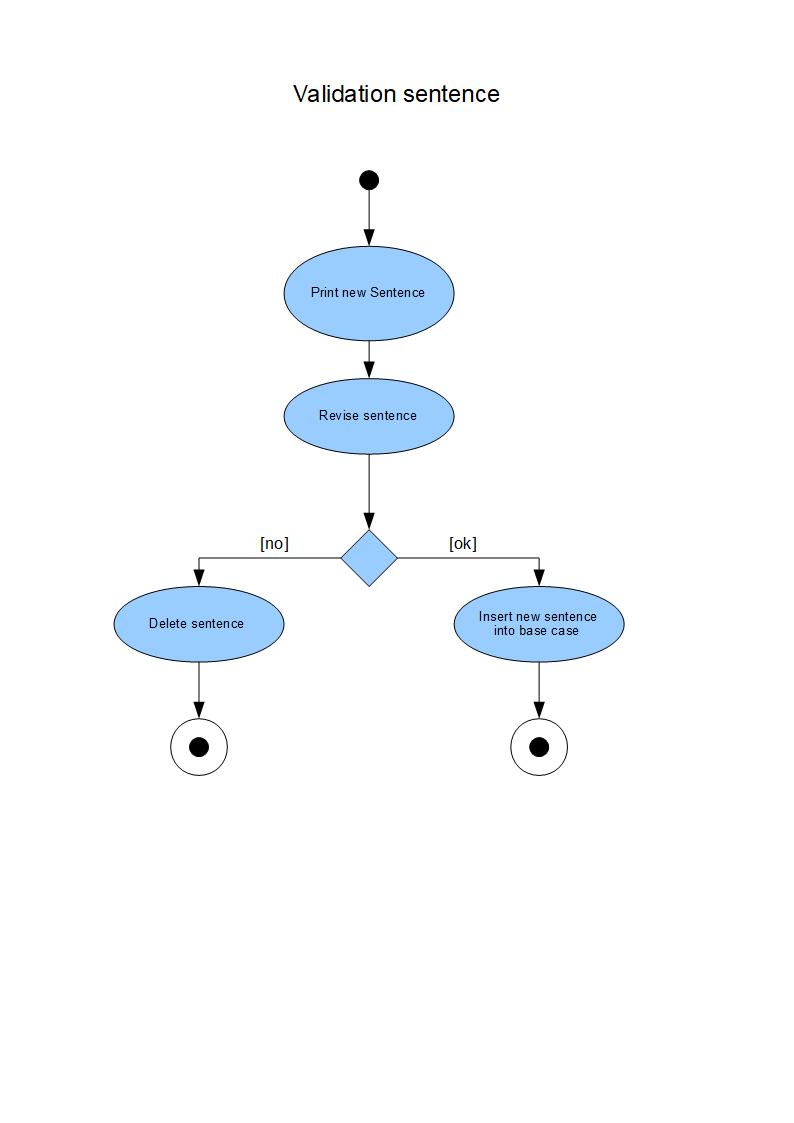


Fig.9 Validation une phrase par un administrateur.

1. Modification d’une phrase

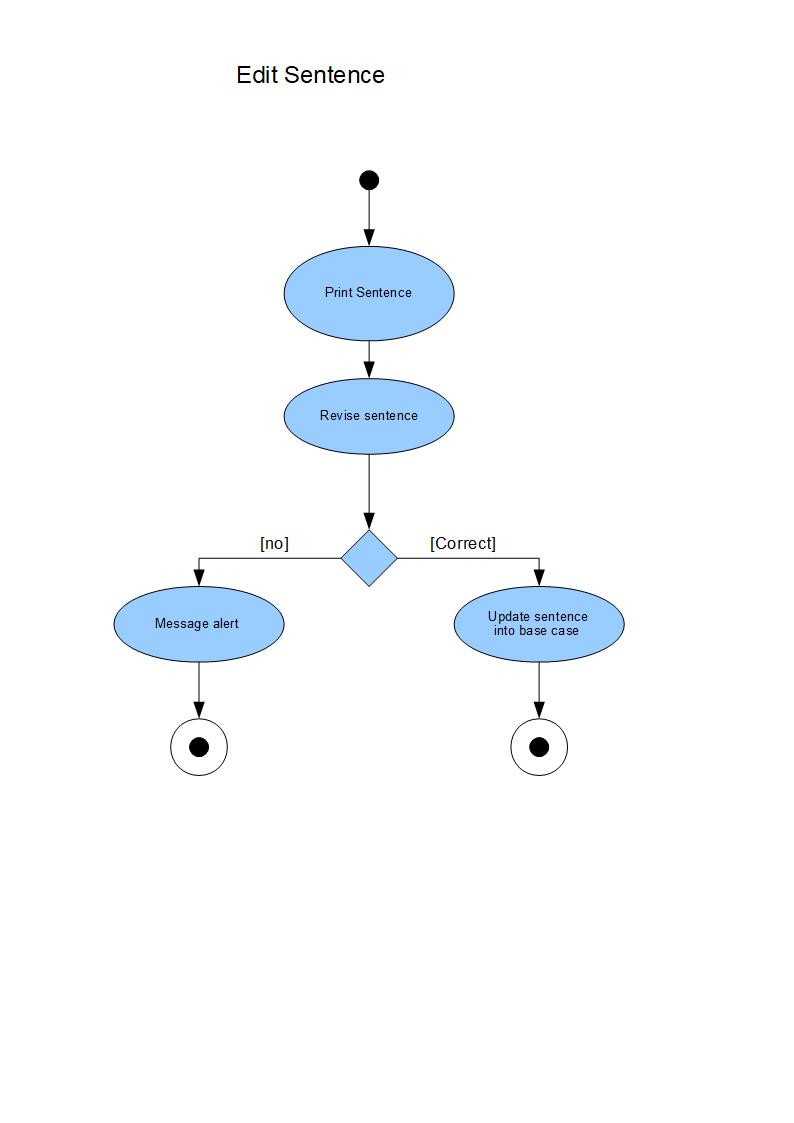
Cette activité est identique à la précèdente. Sauf qu’au lieu d’insérer dans la base courante, le système va modifier la phrase que l’administrateur pense être incorrecte (figure 10).

Fig.10 Modifications d’une phrase par un administrateur.

1. Diagramme de séquence

Ce type de diagramme permet de décrire les différents scénarios d’utilisation du système (figure 11). Il représente des interactions entre les acteurs et le système dans l’ordre chronologique. Pour générer ce genre de diagramme je me sers du site websequencediagrams[[4]](#endnote-4).

1. Demander une correction

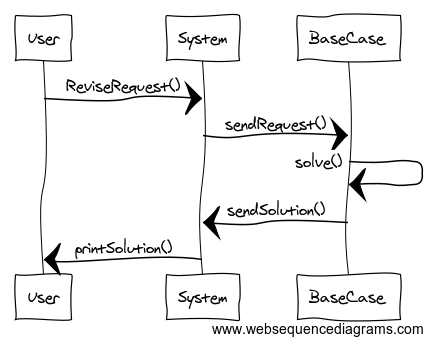


Fig.11 Demander d’une correction.

L’utilisateur demande une correction au système, le système envoie la requête au moteur de correction, qui va résoudre le problème et donner une solution au système, le système va afficher le résultat a l’utilisateur.

1. Proposer une correction

L’utilisateur propose une phrase, le système l’envoi à la base de cas, qui le stocke dans une relation CASES avec le statut « en attente » (figure 12).

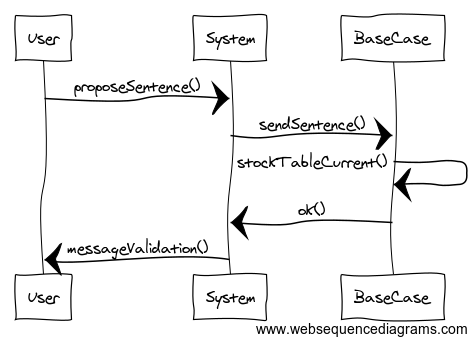


Fig.12 Proposer une Correction.

1. Validation une phrase

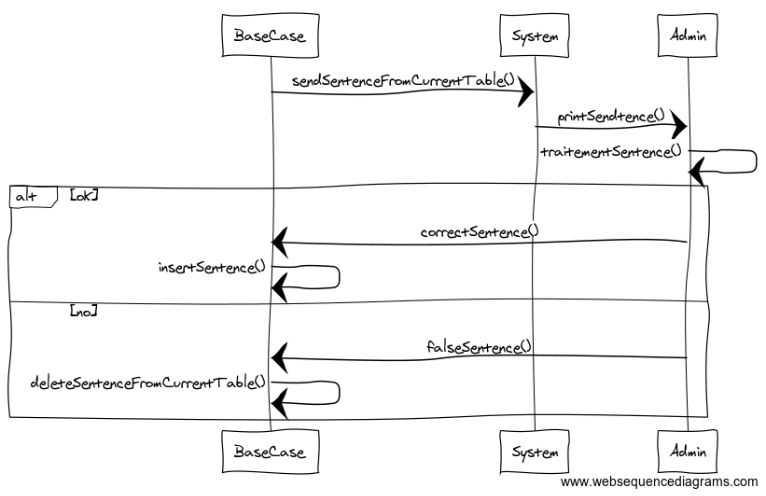


Fig.13 Validation une phrase.

Le système selectionne dans la relation CASES le cas avec le statut « faux ou en attente », et l’affiche à l’administrateurs. Ce dernier valide la phrase si le cas est correct, il sera inséré dans la base de cas, sinon il sera enlevé de la relation CASES (figure 13).

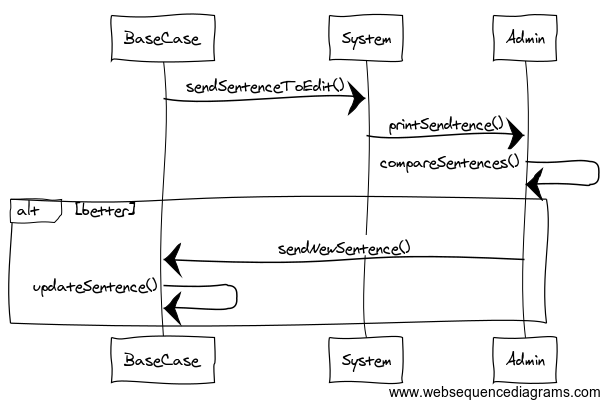
1.  Modifications d’une phrase

Fig.14 Modifications d’une phrase

Le système affiche la phrase à modifier, l’administrateur compare la nouvelle phrase par rapport à l’ancienne. Si elle est meilleure, la nouvelle phrase remplacera alors l’ancienne dans la base de cas (figure 14).

1. **ENVIRONNEMENT DE DÉVELOPPEMENT**
2. MySQL[[5]](#endnote-5) et phpMyAdmin

J’ai choisi MySQL comme gestionnaire de base de données pour notre site web. PhpMyAdmin (figure 15) est une application web pour la gestion de MySQL. Comme vous pouvez le constater, ces trois relations portent le même nom que dans le MCD (figure 1). Ici je choisis « Interclassement (dans figure 15) » est l’utf8[[6]](#endnote-6). Cette convention a été imposée par les encadrants.

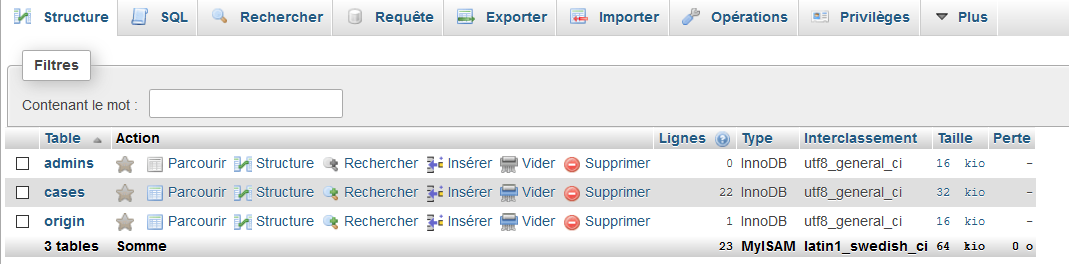


Fig.15 : Interface phpMyAdmin.

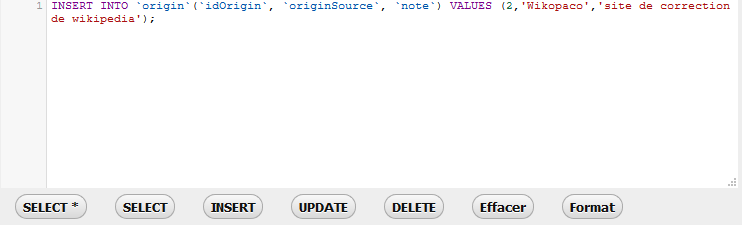
L’interface est simple à utiliser. On peut rentrer directement les requêtes SQL via les formulaires proposés par le site (fig16). J’utilise souvent ces formulaires lors de la création de la base de données et les tests d’insertion.

Fig.16 : Insertion d’une ligne dans la relation Origin.

1. Bootstrap[[7]](#endnote-7)

Bootstrap est une collection d'outils utile à la création du design (graphisme, animation et interactions avec la page dans le navigateur, etc.) de sites et d'applications web. C'est un ensemble qui contient des codes HTML et CSS, des formulaires, boutons, outils de navigation et autres éléments interactifs, ainsi que des extensions JavaScript en option. C’est la première fois que, j’utilise ce framework. Mais il est simple à utiliser et intuitif. En plus, il y a des tutoriaux sur l’Internet pour apprendre à l’utiliser. Ainsi après une dizaine d’heures, je commence à maîtriser ce framework.

1. SublimeText[[8]](#endnote-8)

SublimeText est un éditeur de texte code en C++ et Python. C’est une extension pour VIM. Il possède une interface de développement standard. Le côté droit (Fig17.A) est une arborescence de projet. La partie centrale (Fig17.B) est la zone de code.

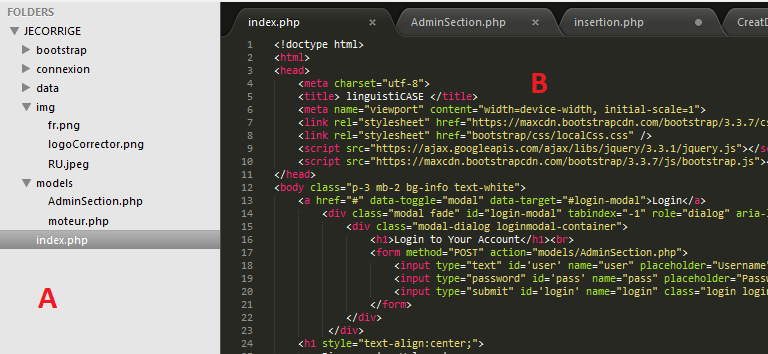


Fig18 : Interface de SublimeText.

Ce choix n’est pas imposé pour le stage. Au début, j’ai choisi phpStorm[[9]](#endnote-9) (c’est un environnement de développement, pour le développement PHP). Mais j’ai un problème avec l’installation et la performance du PC, j’ai décidé de prendre SublimeText pour ce projet. Certes, SublimeText est moins puissant que phpStorm car c’est un éditeur de texte et l’autre est un IDE. Mais SublimeText dispose d’assez de fonctionnalités pour mes besoins.

1. Balsamiq[[10]](#endnote-10)

Balsamiq est un logiciel payant pour faire les maquettes d’interface utilisateur. Je l’ai déjà utilisé pour un de mes projets universitaires donc, je connais et maîtrise certaines de ses fonctionnalités. À la demande du responsable, j’ai installé une version d’essai pour mes besoins, c’est très suffisant pour les éléments de notre interface web.

1. Wampserver[[11]](#endnote-11)

WampServer est une plateforme de développement Web de type WAMP, permettant de faire fonctionner localement (sans avoir à se connecter à un serveur externe) des scripts PHP. WampServer n'est pas un logiciel, mais un environnement qui contient trois serveurs (Apache, MySQL et MariaDB), un interpréteur de script (PHP), ainsi que phpMyAdmin pour l'administration Web des bases MySQL.

J’utilise souvent cette interface souvent pour accéder au site phpMyAdmin et localhost.

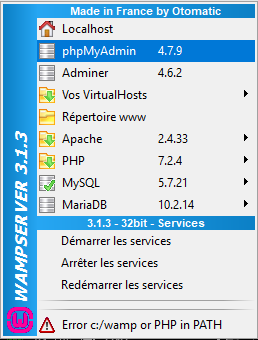


Fig18. Wampserver.

Parmi les options disponibles(Fig18) je n’utilise que :

Localhost : pour accéder au site web en cours de développement.

PhpMyAdmin : pour afficher l’interface web de la base de données.

1. GIT

Comme expliqué dans l’introduction, je suis en collaboration avec deux autres personnes pour ce projet. Ainsi pour mettre en commun notre code, on a décidé de tous utiliser un logiciel de gestion de version GIT, le dépôt est sur le site de GitHub[[12]](#endnote-12). On a choisi ce logiciel, parce qu’’il est libre, et qu’on le connait bien.

1. **INTERFACE WEB**
2. Corrector un site multilingue
3. Historique

Ce projet est un système de correction de phrases. Ainsi, on a dû trouver un nom et un logo pour le site web, le premier nom a été lingustiCase, mais après une certaine réunion, on a constaté que le nom était trop vague, il nous fallait un nom plus précis, on a donc trouvé « Corrector » et le nom n’est pas encore utilisé (Fig19). Pourquoi Corrector ? Cette idée vient de notre responsable Mr. Lieber, pendant une réunion, on n’avait pas d’idée, donc il a proposé Corrector (Figure 20).



Fig.19 Logo de Corrector.

1. Site multilingue

Ce projet est à la base un sujet de recherche changé en sujet de stage. Ainsi, pour présenter notre travail à l’internationale, c’est mieux d’être en anglais. Donc, le responsable préfère que le site disponible en français et en anglais. Pour représenter les langues, on a décidé de les représenter avec des drapeaux (Fig20).

Ainsi, l’interface du site change de langue si on clique sur le drapeau.





Fig.20 Site multilingue.

Pour développer ce site, j’utilise PHP[[13]](#endnote-13) et Javascript[[14]](#endnote-14), langages que j’ai déjà utilisé une ou deux fois pendant mes projets universitaires, j’ai aussi suivi des tutoriels sur l’Internet[[15]](#endnote-15) pour le débugage et pour chercher les informations pour mes besoins. Je n’ai pas rencontré de difficultés particulières pendant le développement. Le seul problème que j’ai eu étant au niveau du design du site.

1. La maquette du site

Avant de commencer à développer le site web, je dois d’abord faire la conception (PARTIE II) et une maquette à présenter à mes responsables. Ainsi j’ai décidé d’utiliser balsamiq pour faire la maquette du site (Fig.21). Après les analyses de besoins d’utilisateur. Il suffit d’une zone de texte (Fig.21A), un bouton pour valider(Fig.21B), une zone de réponse du serveur(Fig.21C), des boutons sur les retours d’utilisateurs (Fig.21D).

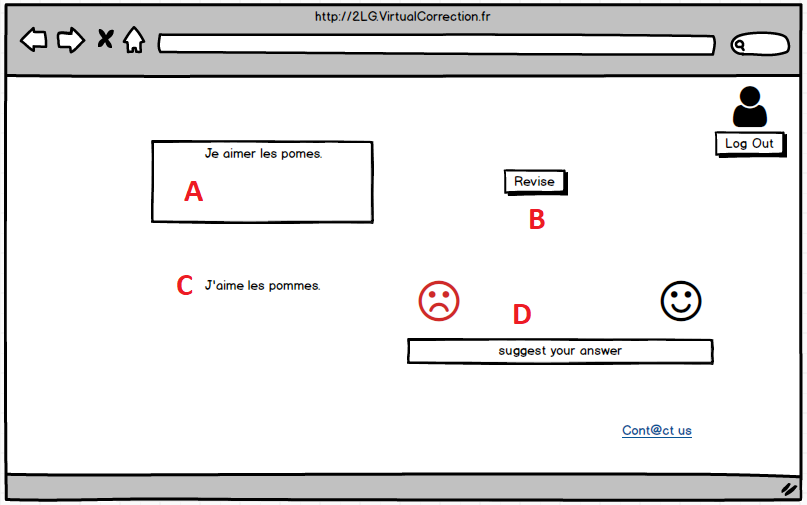


Fig.21 Une première maquette du site.

Cette première version a été jugée améliorable suite à une réunion avec mes encadrants : les éléments doivent être alignés, la section « identification » réservée pour l’administrateur est trop mise en valeur. ELle doit être plus petit et changer en LoginAdmin. De plus, les noms des boutons doivent être modifiés au fur et à mesure du projet. Si je me rappelle bien, j’ai changé le nom des boutons une dizaine de fois.

Pour la partie d’identification de l’administrateur (Fig.22), j’ai pensé à faire un pop-up standard qu’on rencontre souvent sur les sites internet, avec une identification, un mot de passe et un bouton « ok ».

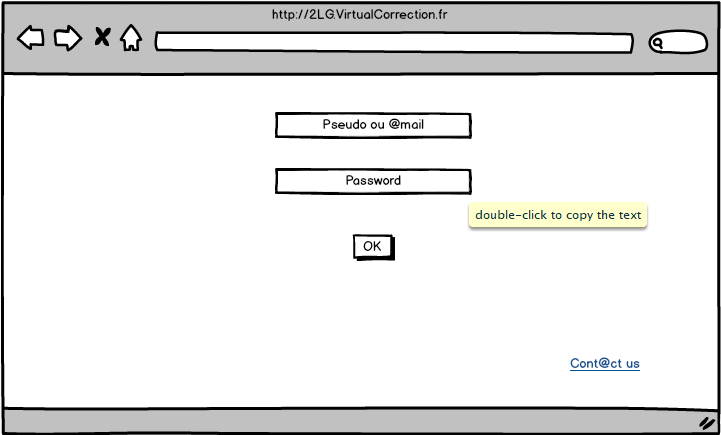


Fig.22 : Connexion à la section de l’administrateur.

1. Interface utilisateur

La première version du site(Fig.23) n’utilise pas Bootstrap. De plus lors de la présentation devant mes responsables, ils m’ont fait remarquer qu’il y a trop de changement de page. C’est-à-dire si je clique sur « Français » ou « English » cela redirige vers une autre page au lieu d’actualiser la page (Fig.24). En parallèle, le style du texte du site web n’était pas esthétique. Ainsi, après la réunion, Mr Nauer m’a conseillé d’utiliser Bootstrap, et au lieu de faire des redirections de page, il préfère que le déroulement se fasse en une seule page avec des balises qui s’affichent au fur et mesure (Fig25).



Fig.23 accueil. Fig.24 Page en français.

Lorsque j’utilise bootstrap, surtout avec des couleurs, j’ai appris que chaque couleur choisie par bootstrap correspond à une signification. Le bleu clair est pour l’information, le bleu standard pour valider, le vert est pour le succès, et le rouge c’est pour un « warning ».

L’utilisation de Corrector est très intuitive : l’utilisateur entre une phrase dans le champ « Entrez la phrase à corriger », il clique sur « valider », le site affiche la réponse avec deux boutons juste en dessous, ces deux boutons sont « Correct et Incorrect ».

Si l’utilisateur clique sur « Correct », le site demande si l’utilisateur veut insérer sa solution dans la base de données pour améliorer la performance du site.

Si l’utilisateur clique sur « Incorrect » le site affiche alors un autre formulaire pour demander à l’utilisateur de rentrer sa correction. Et « OK » est pour insérer dans la base de données le problème de l’utilisateur et sa correction.

Toutes les insertions des utilisateurs sont stockées dans la relation CASES mais avec un statut « EN ATTENTE ».



Fig.25 Corrector.

1. Interface administrateur

L’interface administrateur a moins d’éléments que l’interface utilisateur. Après avoir identifié l’administrateur, le site redirige vers l’interface réservé uniquement à l’administrateur (pas d’inscription).

L’interface administrateur contient seulement deux éléments (Fig.26).

« Restaurer la base de données » : elle a pour le but de remettre l’état de la base de données du site a l’état initial, c’est-à-dire à partir d’un fichier CSV que Mr Levy a créé. Cette partie est en cours de discussion avec les responsables, si on restaure, toutes les insertions des utilisateurs seront effacées.

« Valider les cas utilisateur » : si l’administrateur clique sur afficher, le site affiche dans un tableau les données de la relation CASES avec statut « en attente » avec un bouton « OK ». Le bouton OK permet de changer le statut « en attente » au statut « true » qui signifie que la phrase est validée et est un bon couple (problème, solution) à ajouter dans la relation CASES. Cette partie est en cours de réflexion, dois-je créer un bouton supplémentaire pour refuser une proposition ? Ainsi, la proposition sera supprimée de la relation CASES.

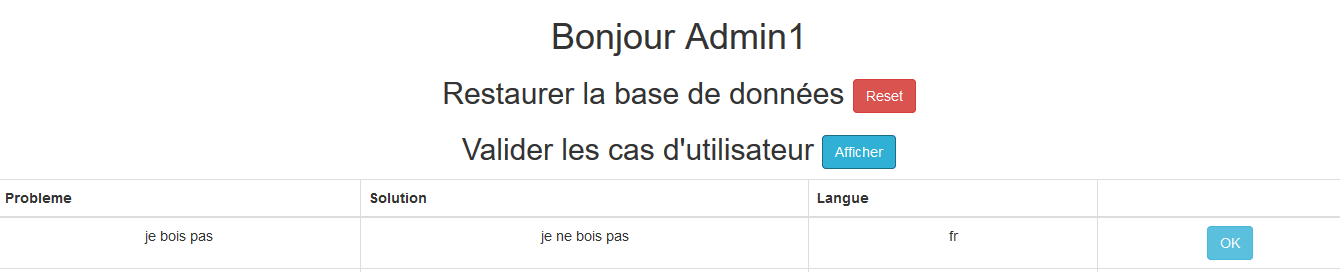


Fig.26 Interface d’Administrateur.

Ici on ne peut pas s’inscrire car les administrateurs sont les responsables du projet. Cette partie est en cours de développement.

**CONCLUSION**

Niveau professionnel, comme ce travaill s’integre dán un nouveau sujet de recherche, à part le fait qu’il faille respecter les consignes de l’encadrant. Je suis libre de mes décisions, que ce soit, au niveau du langage, du design du site web, du type de base de données choisie. Pendant les réunions, l’encadrant donne un objectif à atteindre et à faire pour la prochaine réunion. On fait une réunion toutes les deux semaines avec les encadrants. On a aussi fait des visioconférences avec Mr LEPAGE professeur d’université Waseda[[16]](#endnote-16) au Japon.

Niveau formation, les connaissances que j’ai acquises pendant mes années universitaires m’ont bien servi. Par exemple : la conception, les logiciels utilisés ou la façon de coder.

Niveau personnel, Corrector est le premier site web que j’ai développé tout seul. Je suis très content quand les encadrants sont satisfaits de mon travail. Au début, pour mettre en place c’est difficile. Soit, je n’ai pas d’idée sur la façon de comment à développer une fonction, soit il y en a trop et il faut choisir la meilleure. À partir de ce projet, maintenant, si j’utilise une phrase négative, je me rappelle qu’il y a « ne ou n’». De plus, si je disposais de plus de temps, je pourrais ajouter plus de fonctionnalités pour le Corrector, et surtout la gestion d’administrateur.

**Sitographie[[17]](#endnote-17) :**

1. <http://www.loria.fr/fr/presentation/> [↑](#endnote-ref-1)
2. <http://www.loria.fr/fr/la-recherche/departements/4-traitement-des-langues-et-des-connaissances/> [↑](#endnote-ref-2)
3. <http://mocodo.wingi.net/> [↑](#endnote-ref-3)
4. <https://www.websequencediagrams.com/> [↑](#endnote-ref-4)
5. <https://www.mysql.com/fr/> [↑](#endnote-ref-5)
6. <https://fr.wikipedia.org/wiki/UTF-8> [↑](#endnote-ref-6)
7. <https://www.w3schools.com/bootstrap/> [↑](#endnote-ref-7)
8. <https://www.sublimetext.com/> [↑](#endnote-ref-8)
9. <https://www.jetbrains.com/phpstorm/> [↑](#endnote-ref-9)
10. <https://balsamiq.com/> [↑](#endnote-ref-10)
11. <http://www.wampserver.com/> [↑](#endnote-ref-11)
12. <https://github.com/> [↑](#endnote-ref-12)
13. <https://openclassrooms.com/courses/concevez-votre-site-web-avec-php-et-mysql> [↑](#endnote-ref-13)
14. <https://openclassrooms.com/courses/dynamisez-vos-sites-web-avec-javascript> [↑](#endnote-ref-14)
15. <https://stackoverflow.com/> , <https://openclassrooms.com/forum/> et <https://fr.wikipedia.org/wiki/Wiki> [↑](#endnote-ref-15)
16. <https://www.waseda.jp/top/en> [↑](#endnote-ref-16)
17. <https://fr.wiktionary.org/wiki/sitographie>

    **Annexe :**

    Les codes sources sont disponibles sur le dépôt Github :

    https://github.com/Sle3pingForest/StageL3/tree/master/WEB [↑](#endnote-ref-17)