

# Projet Moteur de recherche en langage naturel



**Groupe :** *IMBERT OLIVIER*  
*SIMMA GRIDSADA*

*L3 Informatique*  
*Université méditerranée aix-marseille II*

**Responsable :** Marie-Hélène STEFANINI

**L'année universitaire :** 2009/2010

## Table de matière

I. Introduction .....	2
II. Qu'est-ce qu'un moteur de recherche ? .....	2
A. Moteur de recherche traditionnel (Exploration/Indexation/Recherche). ....	2
B. Moteur de recherche en langage naturel.....	3
III. Listes des moteurs de recherche. ....	5
C. Moteurs de recherche de type statistique. ....	5
D. Moteurs de recherche en langage naturel. ....	5
IV. Projet : Un système de moteur de recherche simple. ....	6
E. Problématique.....	6
F. Réalisation.....	6
G. Recherche Statistique vs Recherche en “Langage Naturel” .....	6
V. Conclusion .....	8
VI. Bibliographies.....	9
VII. Annexes .....	10

# I. Introduction

Dans une approche plutôt de découverte, nous avons choisi ce sujet basé sur les moteurs de recherche et plus spécifiquement les moteurs de recherche en langage naturel. Il est devenu quasiment impossible de trouver une information sans eux sur Internet. Leur but est de nous simplifier la vie.

Avant toute chose le “*Natural Language*” traduit de l'Anglais par langage naturel signifie :  
“ Possibilité de poser une question à un Outil de recherche sous la forme d'une phrase intelligible comme « *quel temps fait-il aujourd'hui à Marseille ?* »”.

C'est ce que réalise certains moteurs de recherche que nous vous présenterons au cours de ce rapide exposé. De plus, nous vous expliquerons le fonctionnement d'un moteur de recherche en langage naturel, de conception personnelle.



## II. Qu'est-ce qu'un moteur de recherche ?

### A. Moteur de recherche traditionnel (Exploration/Indexation/Recherche).

Un moteur de recherche est un logiciel capable d'orienter un utilisateur lors d'une recherche d'informations à travers une base documentaire, à partir d'une requête formulée dans une syntaxe qui lui est propre, au moyen d'opérateurs spécifiques, ou en « *langage naturel* ». Le moteur de recherche propose alors en réponse une liste de documents correspondant au mieux aux paramètres de la recherche.

Sur Internet, un moteur de recherche est accessible par sa page d'accueil. Celle-ci peut être présentée sous forme de portail d'informations, ou être d'un aspect plus épuré en ne faisant figurer que les éléments nécessaires à l'exécution d'une requête : une zone de saisie et un bouton de démarrage de la recherche.

A la différence d'un annuaire, pour lequel le référencement des pages s'effectue manuellement, le moteur de recherche doit « *aspirer* » le web afin d'enrichir sa base de connaissances pour proposer des réponses les plus pertinentes possibles. Cette étape est réalisée par les robots d'indexation dont la fonction est de parcourir les pages rencontrées en suivant récursivement les liens pointant vers d'autres pages et d'en indexer le contenu textuel.

Si toutes les pages présentes sur Internet étaient jugées de la même importance, il serait difficile de définir un ordre d'affichage des résultats, et d'autant plus fastidieux pour l'utilisateur d'analyser les centaines de pages qui lui seraient proposées. C'est en partie grâce à l'indicateur d'importance relative des pages que la disposition des pages pourra être définie lors de l'affichage.

L'importance relative d'une page est fonction de l'importance relative des pages web qui ont des liens pointant vers elle et du nombre de liens émis par ces mêmes pages. Le calcul pour une page nécessite donc un calcul itératif, puisque doter une page d'une importance relative modifie par ricochet l'importance relative de toutes les pages vers lesquelles pointe un lien de cette même page, et de toutes les pages reliées à ces pages.

L'indexation consiste en une analyse statistique, et non sémantique, des pages aspirées par les robots d'indexation. Elle est réalisée en deux étapes :

- Calcul de l'indice de densité des mots dans la page : c'est le rapport entre le nombre d'occurrence d'un mot et le nombre de mots total de la page. Cette opération consiste à trier et classer les mots trouvés, à l'exception des termes filtrés et des mots insignifiants.
- Classement de la page, suivant le poids des mots, dans l'index inversé, où chaque mot renvoie à la liste des pages sur lesquelles il est présent.

Par conséquent, le principe de fonctionnement d'un moteur de recherche est en partie basé sur sa capacité à « *s'auto-alimenter* ». D'autre part, nous avons vu que l'agencement des pages, lors de l'affichage du résultat d'une requête, dépendait à la fois de l'indicateur d'importance relative et du classement des pages dans l'index inversé. A aucun moment le contexte de la requête n'est pris en compte. Donc, rien ne garantit la pertinence des résultats par rapport à la thématique évoquée. C'est dans le but de proposer des réponses plus justes, que les sociétés s'orientent naturellement vers le développement d'un nouveau type de moteur de recherche : les moteurs de recherche contextuels. Mais pour ça il faut mieux connaître l'utilisateur...

En essayant d'en savoir plus sur nous, les moteurs de recherche ne risquent-ils pas, en fin de compte, d'en savoir trop ?

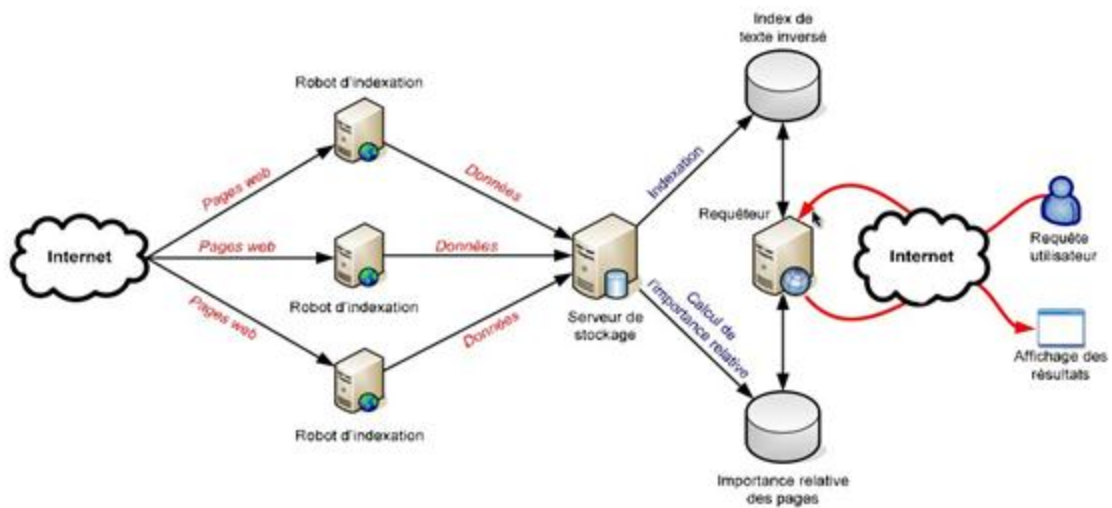
## **B. Moteur de recherche en langage naturel**

Les moteurs en langage naturel doivent théoriquement pouvoir comprendre n'importe quelle question posée avec des mots courants. Certains vont même jusqu'à donner une réponse précise en fonction de la phrase ou de la question posée. Ces moteurs de recherche font appel à des grammaires de structures de phrase interne (grammaires de type syntaxique et grammaires de type lexical), analysent la phrase ou la question posée en fonction des grammaires présentes, recherchent les informations présentes au niveau du web et en retournent l'information pure sans lien spécifique avec une quelconque pages web.

Par exemple, via le célèbre moteur de recherche en langage naturel Wolfram-Alpha (l'analyse

syntaxique ne s'opère uniquement qu'en anglais), si vous écrivez :

“How old was Mickael Jackson in 2006 ?”, le moteur de recherche vous répondra seulement et uniquement : “47 years”.



*Illustration 1: Principe de fonctionnement d'un moteur de recherche*

### **III. Listes des moteurs de recherche.**

#### **C. Moteurs de recherche de type statistique.**

Le principe de fonctionnement de ces moteurs de recherche est similaire à celui des moteurs de recherche les plus usuellement connus :

- Aguiip et AltaVista
- Ask
- Baidu
- Bing
- Duck Duck Go
- Ecogine
- Ecosia
- Ethicle
- Exalead
- Google
- Panguso
- Scroogle
- Seeks
- Yahoo!
- Yauba
- Orange
- Optimal Search
- Yandex
- Blekko

#### **D. Moteurs de recherche en langage naturel.**

- Powerset : Recherche contenu encyclopédique.
- True Knowledge : Recherche contenu général.
- Hakia : Recherche contenu général.
- Wolfram-Alpha : Recherche contenu général.

## IV. Projet : Un système de moteur de recherche simple.

### E. Problématique

Il s'agit de mettre en place un programme fenêtré en JAVA GUI (lancer avec Netbean ou Eclipse) qui a une facilité d'utilisation pour pouvoir rechercher des mots souhaités.

- Utiliser les outils libres ( open source) pour ce projet;
- Possibilité de choisir et ouvrir le fichier .txt;
- Possibilité de rechercher les fichiers souhaités avec des mots clés de manière rapide;
- Possibilité de rechercher des informations de mots clés souhaités qu'on a trouvé dans le texte sur google (dont images , vidéos);
- Possibilité de développer le programme dans l'avenir grâce au Java DOC pour ceux que cela intéresse;
- Possibilité d'implémentation d'un moteur de recherche de type sémantique en Java à l'avenir.

### F. Réalisation

La réalisation nous a pris un peu de temps, surtout au niveau de la conception même du programme et de la mise en place de nos propres algorithmes pour une recherche par la méthode du langage naturel en java sous Eclipse et NetBean.

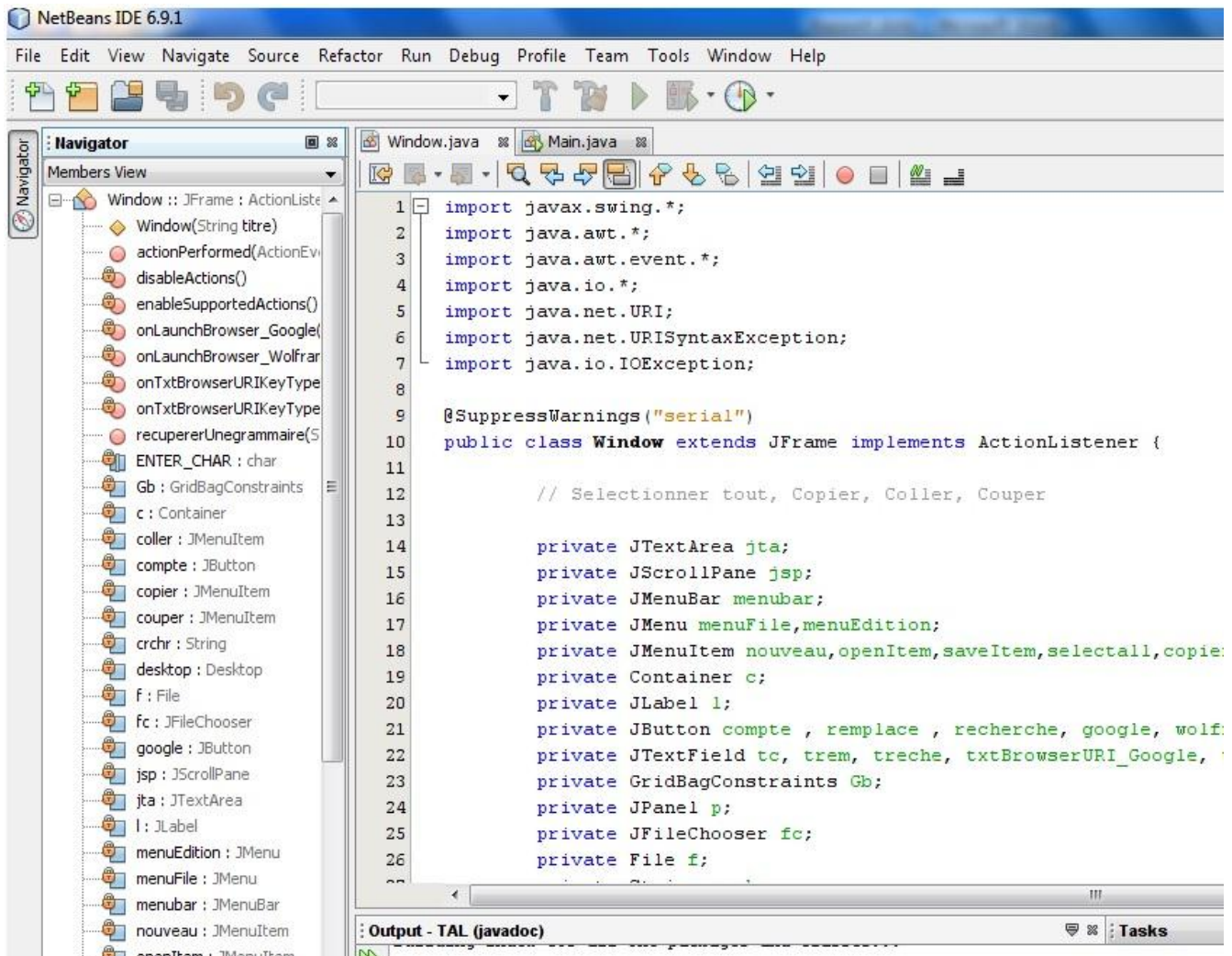
### G. Recherche Statistique vs Recherche en “Langage Naturel”

Lorsqu'on utilise un moteur de recherche utilisant une recherche internet de type statistique, on obtient des réponses sur toutes les pages web concernant le mot ou l'expression entré(e). Il faudra donc par la suite que l'utilisateur prenne son temps afin de récupérer les informations souhaitées. L'avantage est que l'on connaît la provenance des informations présentes au niveau d'internet.

Si on utilise un moteur de recherche basé sur le “Langage Naturel” tel qu'il est présent sur notre petite application, c'est-à-dire Wolfram-Alpha, on ne pourra pas sélectionner n'importe quelle requête sous n'importe quel langage. Le seul langage accepté est l'Anglais. De plus, on ne pourra pas sélectionner n'importe quelle sentence, car certaines grammaires ont été implémentées mais pas toutes.

Par exemple : “How old was << N'importe qui >> in << N'importe quelle date >> ?” fonctionne.  
“The President of << Le pays que vous souhaitez >>” fonctionne également.

Au cas où la phrase n'est pas comprise par le moteur de recherche en langage naturel (Wolfram-Alpha), il analysera alors le dernier mot de la phrase. Autre point négatif à ajouter à un moteur de recherche en langage naturel, c'est qu'il est assez long à trouver une réponse pour une phrase précise, si grammaticalement correcte. En effet celui-ci doit rechercher chaque information demandée sur chaque site internet dédié, d'où une complexité plus importante et cela demande de la ressource de la part du navigateur et de la machine plus généralement.





## V. Conclusion

De nombreux moteurs de recherche de type recherche statistique utilise aussi des parties en langage naturel. Le plus célèbre dans le monde étant « *Google* » utilise ce nouveau procédé pour effectuer des recherches, tout en gardant une approche visuelle au niveau des résultats trouvés de type statistique.

Une recherche scientifique plus avancée des moteurs actuels en langage naturel, serait de nature à faciliter les utilisateurs, en ce qu'il permettrait un gain de temps et par conséquent d'argent, notamment dans les secteurs de l'informatique et industriel en général (industriel - commerce - services).

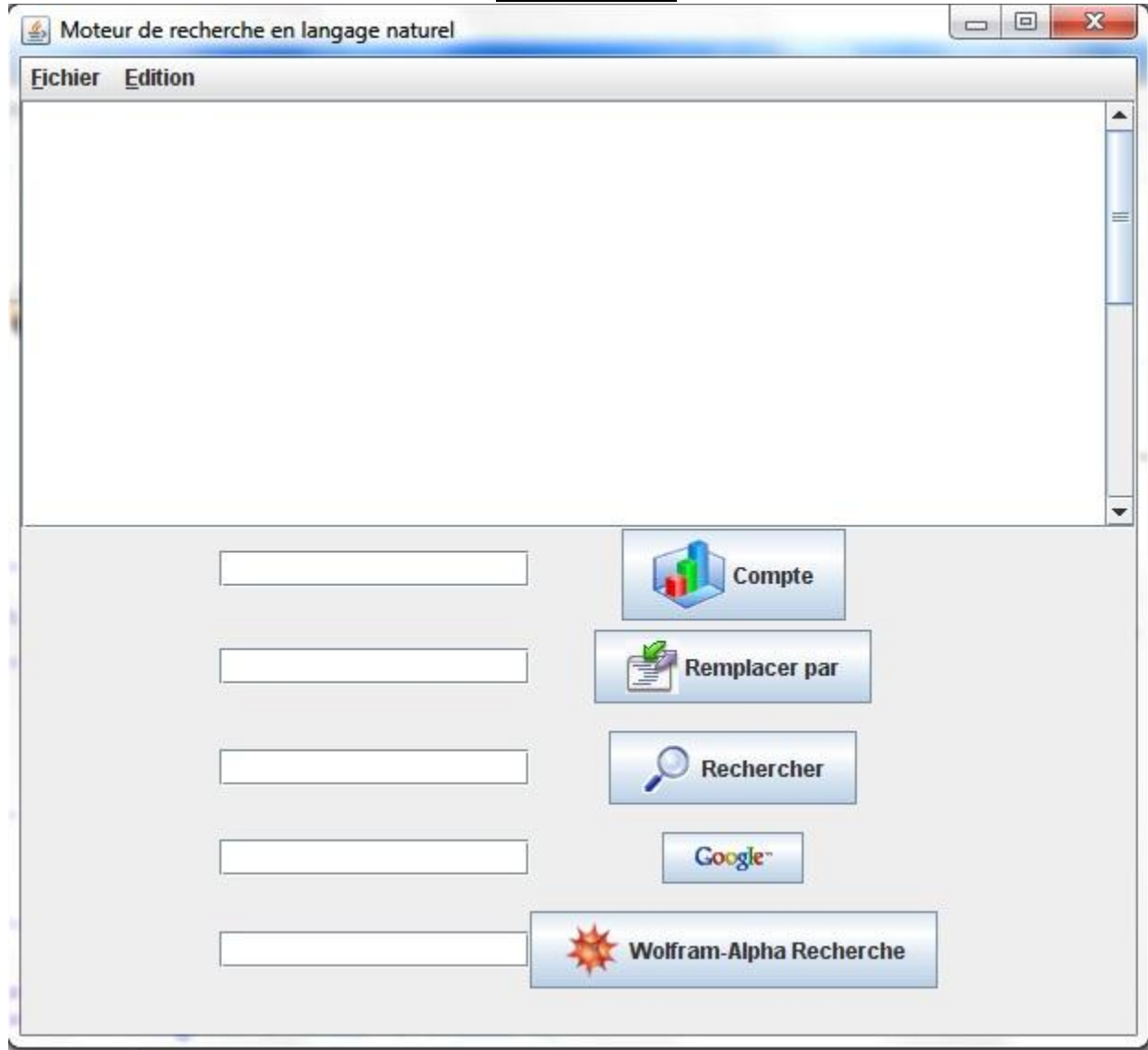


## VI. Bibliographies

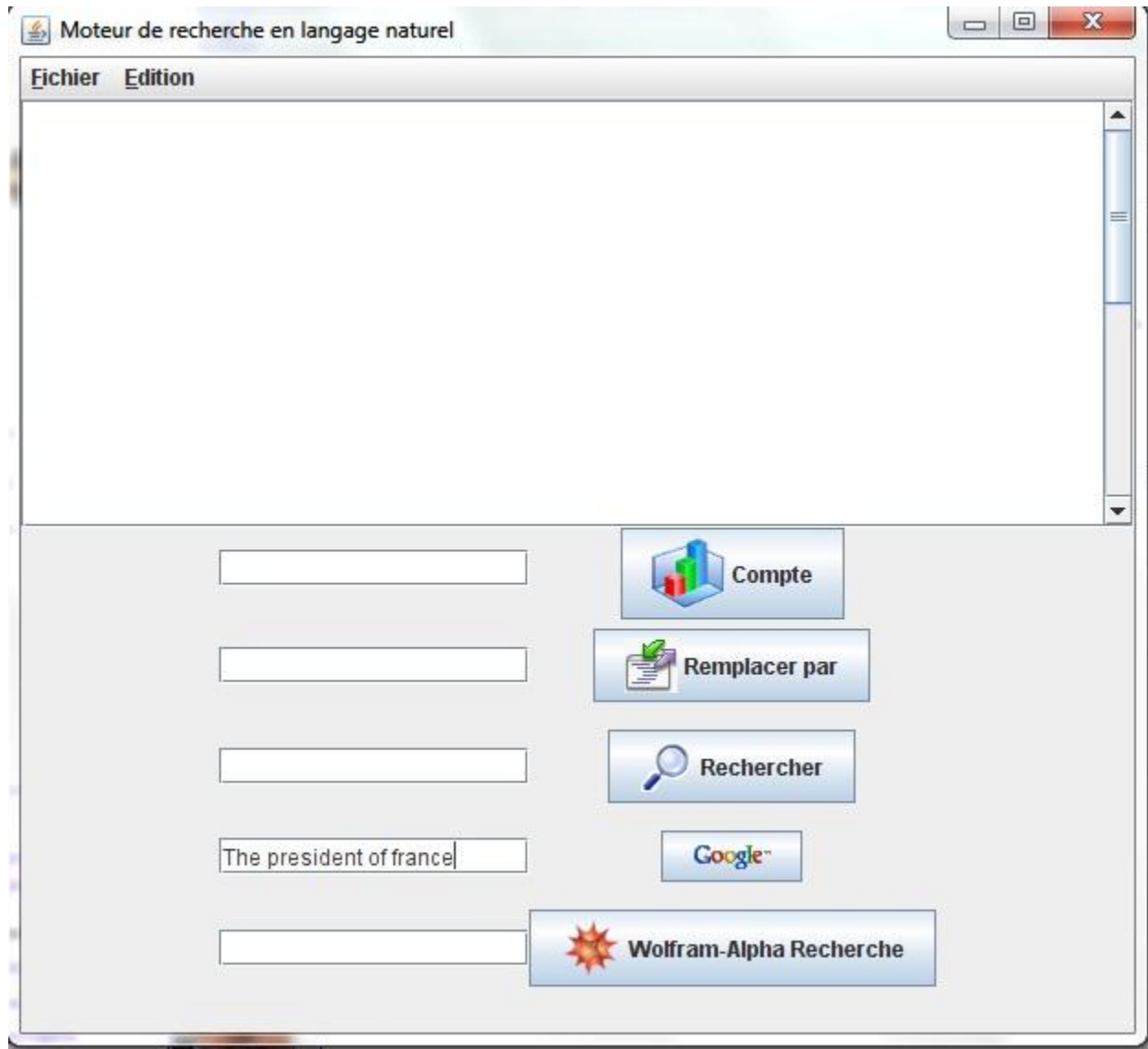
- ❖ [http://fr.wikipedia.org/wiki/Moteur\\_de\\_recherche](http://fr.wikipedia.org/wiki/Moteur_de_recherche)
- ❖ <http://fr.wikipedia.org/wiki/PageRank>
- ❖ <http://www.pagerank.fr/>
- ❖ <http://www.pagerank-direct.fr/>
- ❖ [http://www.prchecker.info/check\\_page\\_rank.php/](http://www.prchecker.info/check_page_rank.php/)
- ❖ [http://fr.wikipedia.org/wiki/Analyse\\_semantique\\_latente](http://fr.wikipedia.org/wiki/Analyse_semantique_latente)
- ❖ <http://blog.cytise.fr/moteurs-de-recherche/powerset-moteur-de-recherche-semantique/>
- ❖ <http://www.valhalla.fr/ressources/java/old/cours/texte/texte.pdf>
- ❖ <http://www.journaldunet.com/solutions/moteur-referencement/selection/5-moteurs-de-recherche-en-langage-naturel/5-moteurs-de-recherche-en-langage-naturel.shtml>
- ❖ [www.clubic.com/actualite-139538-wikipedia-moteur-recherche-langage-naturel.html](http://www.clubic.com/actualite-139538-wikipedia-moteur-recherche-langage-naturel.html)
- ❖ <http://www.linformaticien.com/actualites/id/6339/wolfram-alpha-moteur-de-recherche-en-langage-naturel.aspx>
- ❖ <http://www.dicodunet.com/definitions/moteurs-de-recherche/langage-naturel.html>
- ❖ <http://www.wolframalpha.com/>
- ❖ <http://www.google.fr/>

## VII. Annexes

### Notre projet :



**1) Projet testé par une recherche de type statistique « Google » :**



a) Résultat de la recherche par « Google » de « The president of France »

The screenshot shows a Firefox browser window with the Google search results for the query "the president of france". The address bar shows the URL "http://www.google.fr/search?q=the+president+of+france". The search bar contains the text "the president of france" and a "Rechercher" button. Below the search bar, it indicates "Environ 229 000 000 résultats (0,22 secondes)".

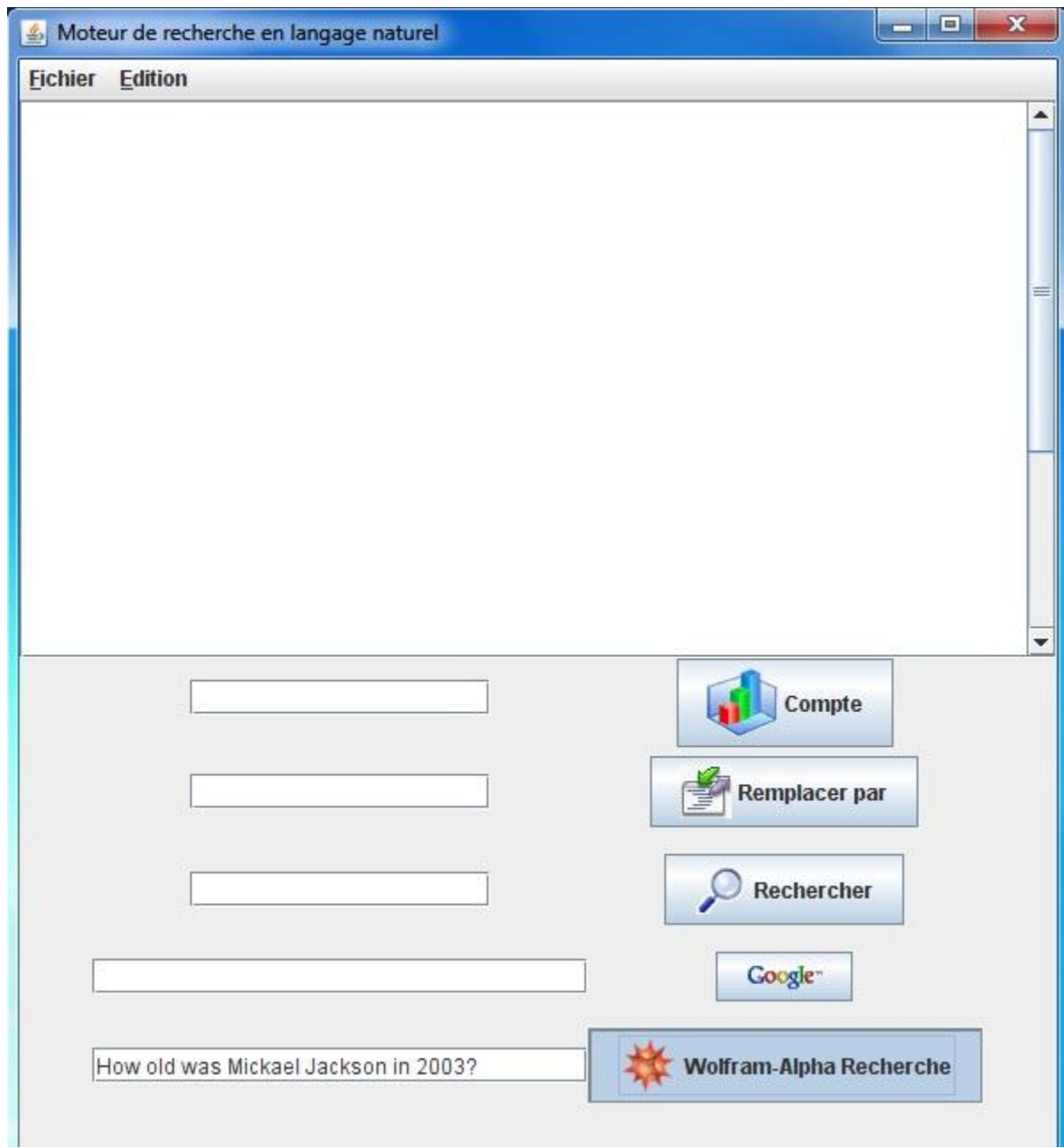
On the left side, there is a sidebar with navigation links: "Tout", "Images", "Vidéos", "Actualités", "Shopping", "En temps réel", and "Plus". Below these, there are sections for "Marseille" (with a link to "Changer le lieu"), "Le Web" (with links for "Pages en français", "Pays : France", and "Pages en langue étrangère traduites"), and "Date indifférente" (with a link for "Les plus récentes").

The main search results are listed on the right:

- Site officiel - Présidence de la République**: Voir tous les déplacements du **Président en France** ..... Politique maritime de la **France**: le **Président** se rend au port du Havre jeudi 21 avril ...  
Écrire au président - Agenda - Discours - Photos  
[www.elysee.fr/](http://www.elysee.fr/) - En cache - Pages similaires
- President of France - Wikipedia, the free encyclopedia** - [ Traduire cette page ]  
The President of the French Republic colloquially referred to in English as **the President of France**, is France's elected Head of State. ...  
Current presidential powers - Criminal responsibility and ... - Election  
[en.wikipedia.org/wiki/President\\_of\\_France](http://en.wikipedia.org/wiki/President_of_France) - En cache - Pages similaires
- Liste des présidents de la République française - Wikipédia**  
... chef de la **France** libre en exil depuis l'armistice de 1940, exerce à son ...  
Deuxième République - Troisième République - Période 1940 - 1947  
[fr.wikipedia.org/.../Liste\\_des\\_présidents\\_de\\_la\\_République\\_française](http://fr.wikipedia.org/.../Liste_des_présidents_de_la_République_française) - En cache - Pages similaires
- Président de la République française - Wikipédia**  
Depuis l'élection du **président** au suffrage universel direct en 1962, il s ...  
[fr.wikipedia.org/.../Président\\_de\\_la\\_République\\_française](http://fr.wikipedia.org/.../Président_de_la_République_française) - En cache - Pages similaires

Below the results, there is a link for "Actualités correspondant à the president of france".

**2) Projet testé par une recherche de type sémantique « Wolfram-Alpha » :**



a) Résultat de la recherche par « Wolfram-Alpha » de « How old was Mickael Jackson in 2003 ? » :

The screenshot shows the WolframAlpha website interface. At the top, the browser's address bar displays the URL: `http://www.wolframalpha.com/input/?i=How+old+was+Mickael+Jackson+in+2003?`. Below the address bar, the search query "How old was Mickael Jackson in 2003?" is entered into the main input field. The website's navigation menu includes links for HOME, EXAMPLES, PRODUCTS, BLOG, and ABOUT. A banner below the menu encourages downloading toolbars, gadgets, or add-ons. The WolframAlpha logo, featuring a red star and the text "WolframAlpha computational knowledge engine", is prominently displayed. Below the input field, a box shows the interpretation: "Interpreting 'mickael' as 'michael'". The "Input interpretation:" section displays the query as "age of Michael Jackson (singer) in 2003". The "Result for start of 2003:" section shows the answer "44 years" with a "Show details" button. At the bottom, it states "Computed by Wolfram Mathematica" and provides links for "Source information" and "Download as: PDF | Live Mathematica".

Firefox

<http://www.wolframalpha.com/input/?i=How+old+was+Mickael+Jackson+in+2003?>

How old was Mickael Jackson in 2003? - ...

HOME | EXAMPLES | PRODUCTS | BLOG | ABOUT

Download one of our [toolbars](#), [gadgets](#), or [add-ons](#) for easy access to Wolfram|Alpha »

**WolframAlpha**™ computational knowledge engine

How old was Mickael Jackson in 2003?

Interpreting "mickael" as "michael"

Input interpretation:

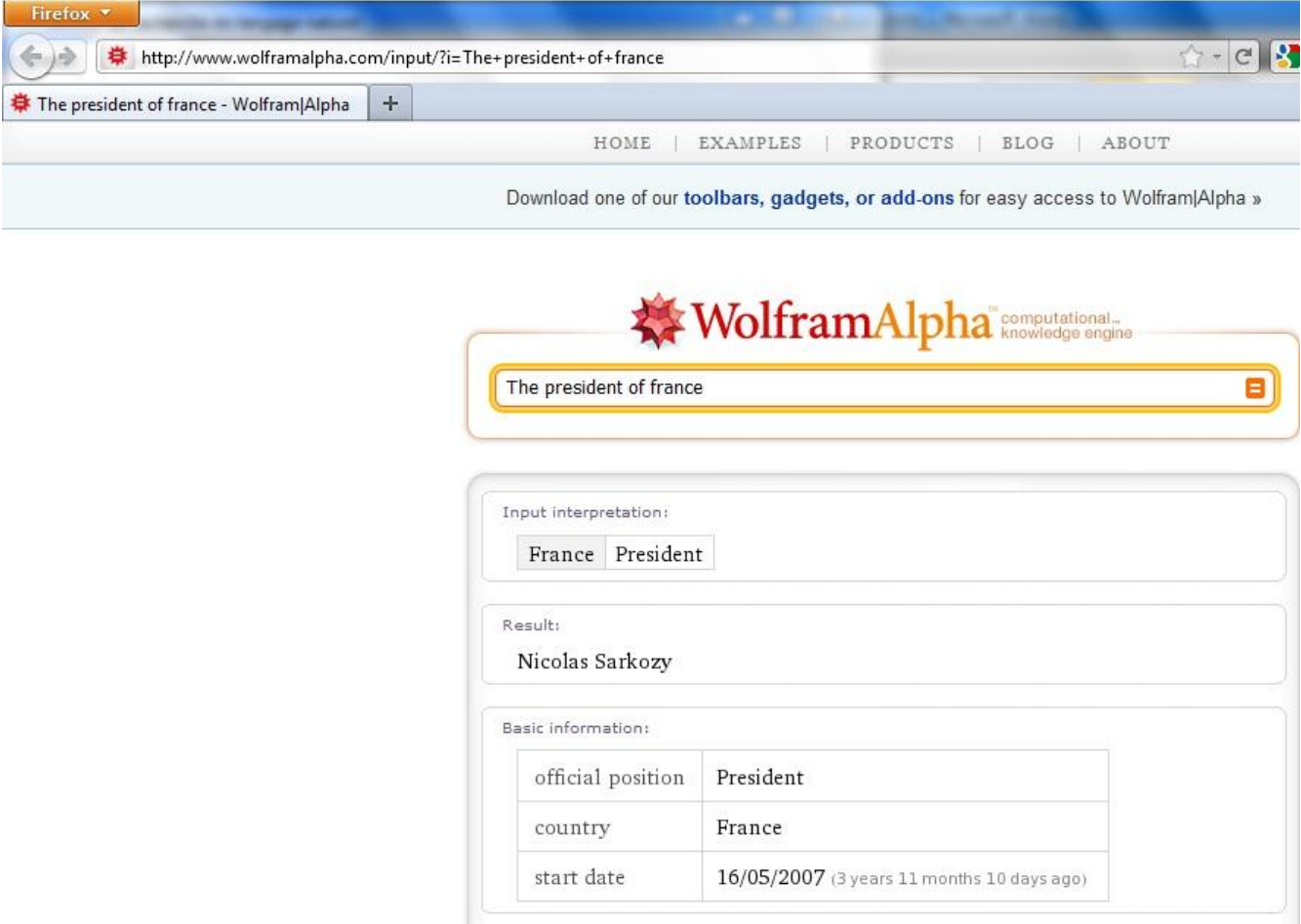
age of Michael Jackson (singer) in 2003

Result for start of 2003:

44 years [Show details](#)

Computed by [Wolfram Mathematica](#) [Source information »](#) Download as: [PDF](#) | [Live Mathematica](#)

**b) Résultat de la recherche par « Wolfram-Alpha » de « The president of France » :**




Firefox

http://www.wolframalpha.com/input/?i=The+president+of+france

The president of france - Wolfram|Alpha

HOME | EXAMPLES | PRODUCTS | BLOG | ABOUT

Download one of our [toolbars](#), [gadgets](#), or [add-ons](#) for easy access to Wolfram|Alpha »

 **WolframAlpha**™ computational knowledge engine

The president of france

Input interpretation:

France President

Result:

Nicolas Sarkozy

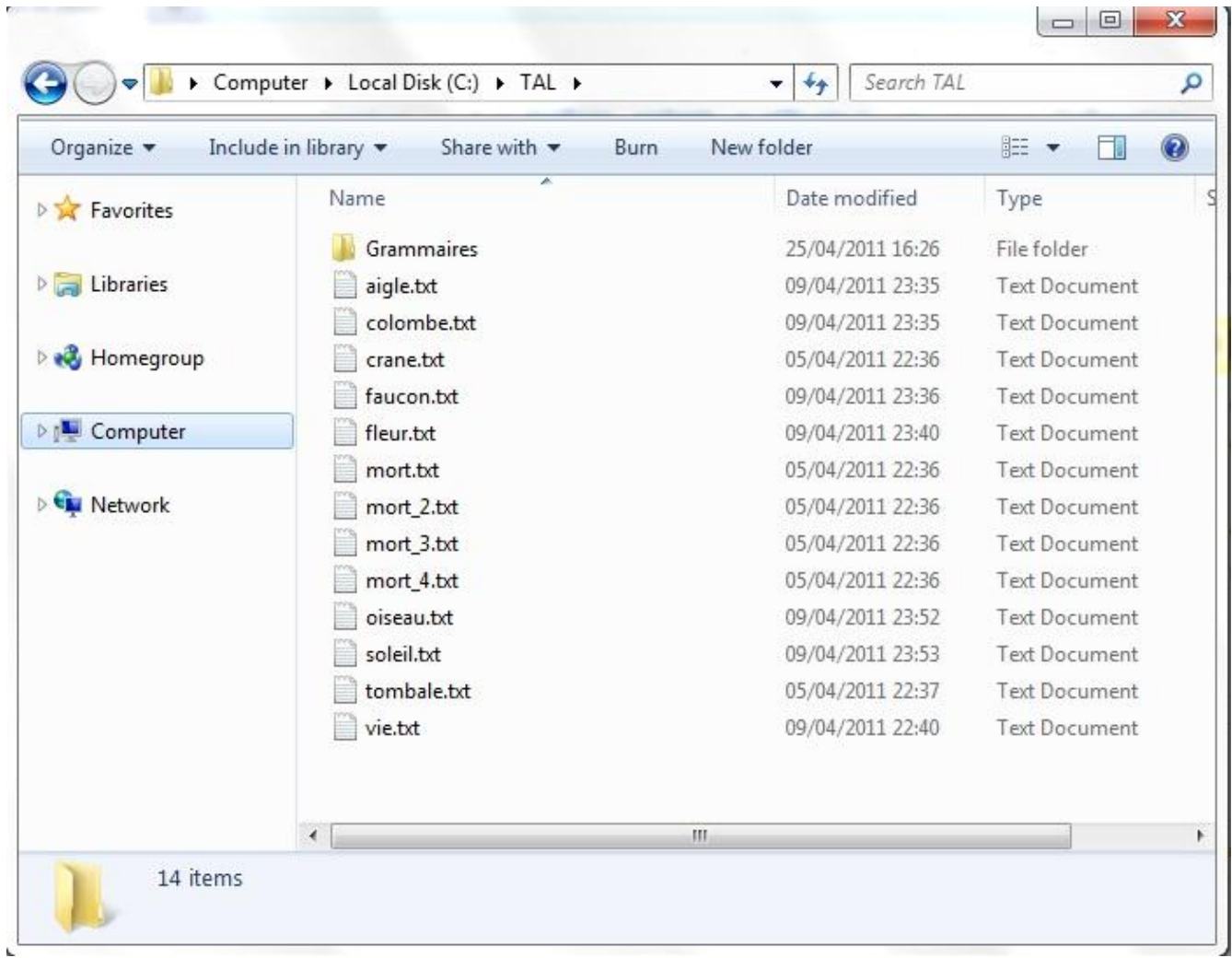
Basic information:

official position	President
country	France
start date	16/05/2007 (3 years 11 months 10 days ago)

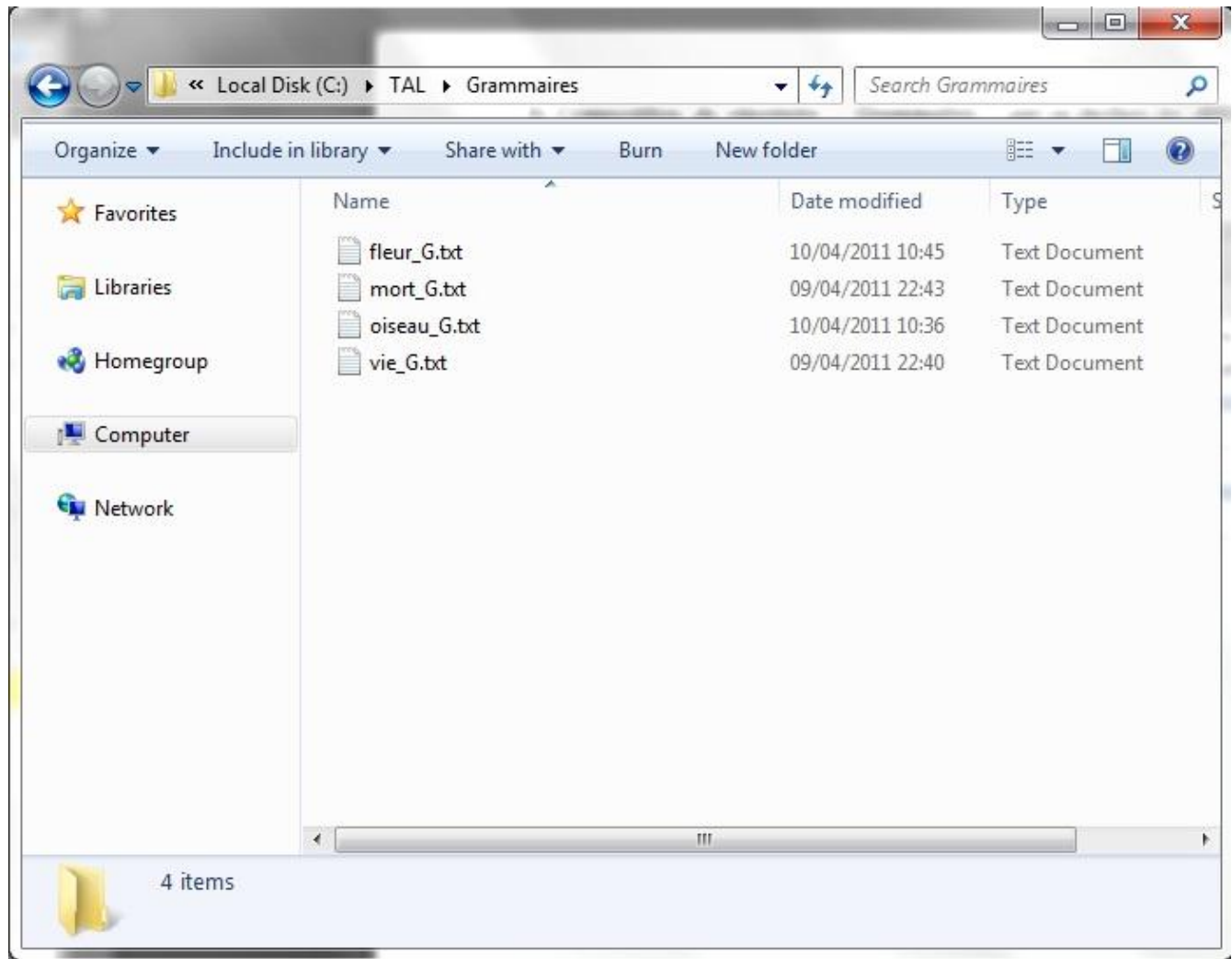


### 3) Recherche par notre moteur de recherche :

*a) Composition du répertoire « TAL » qui va composer notre répertoire de recherche de fichiers :*



***b) Composition du répertoire « Grammaires » qui va inclure les différentes grammaires afin de réaliser une recherche de type sémantique :***



c) Résultats de recherche par notre outil de moteur de recherche via les mots clés  
« mort », « vie », « crane », « oiseau » et « aigle » :

Annexe(tableaux 1 à 6)

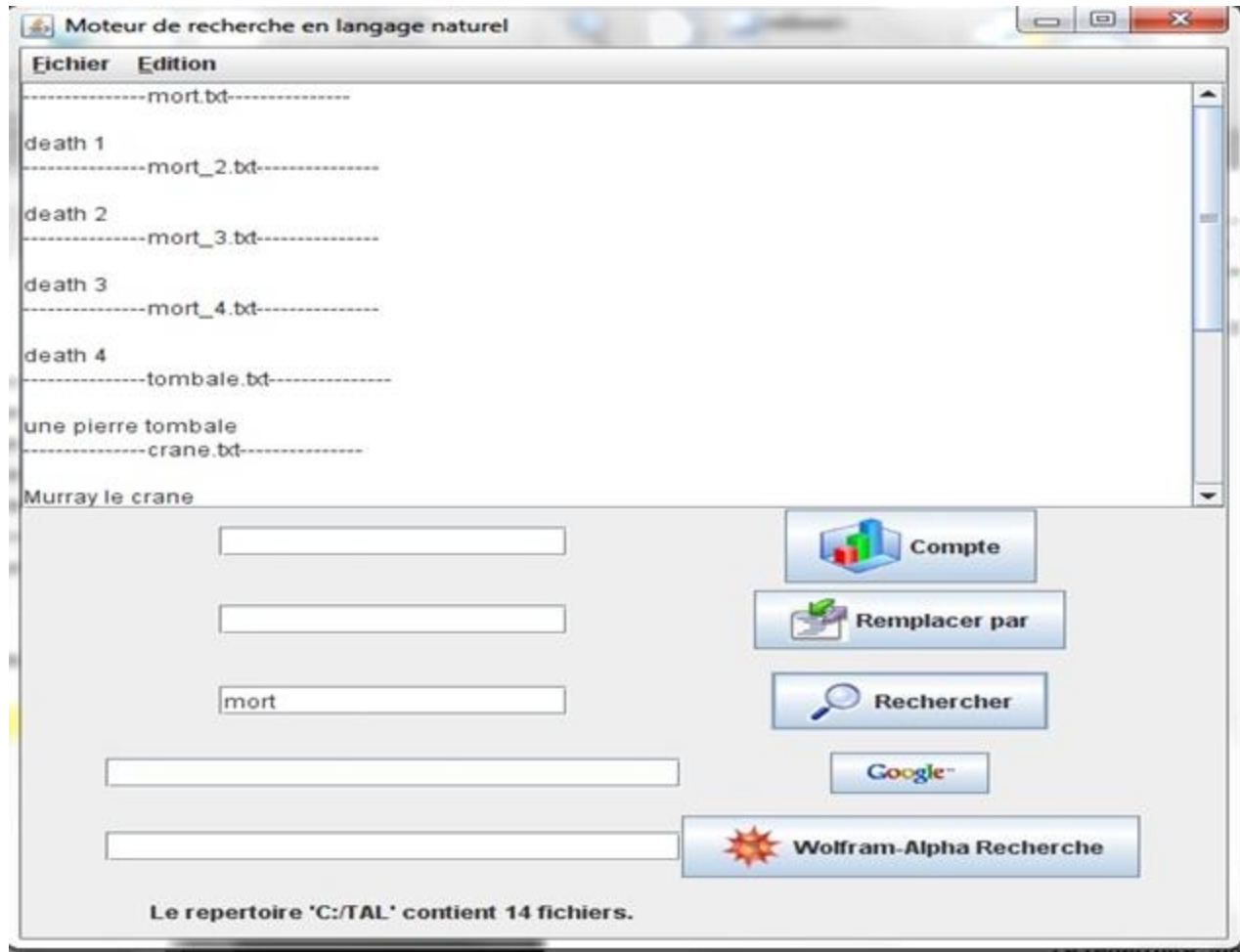
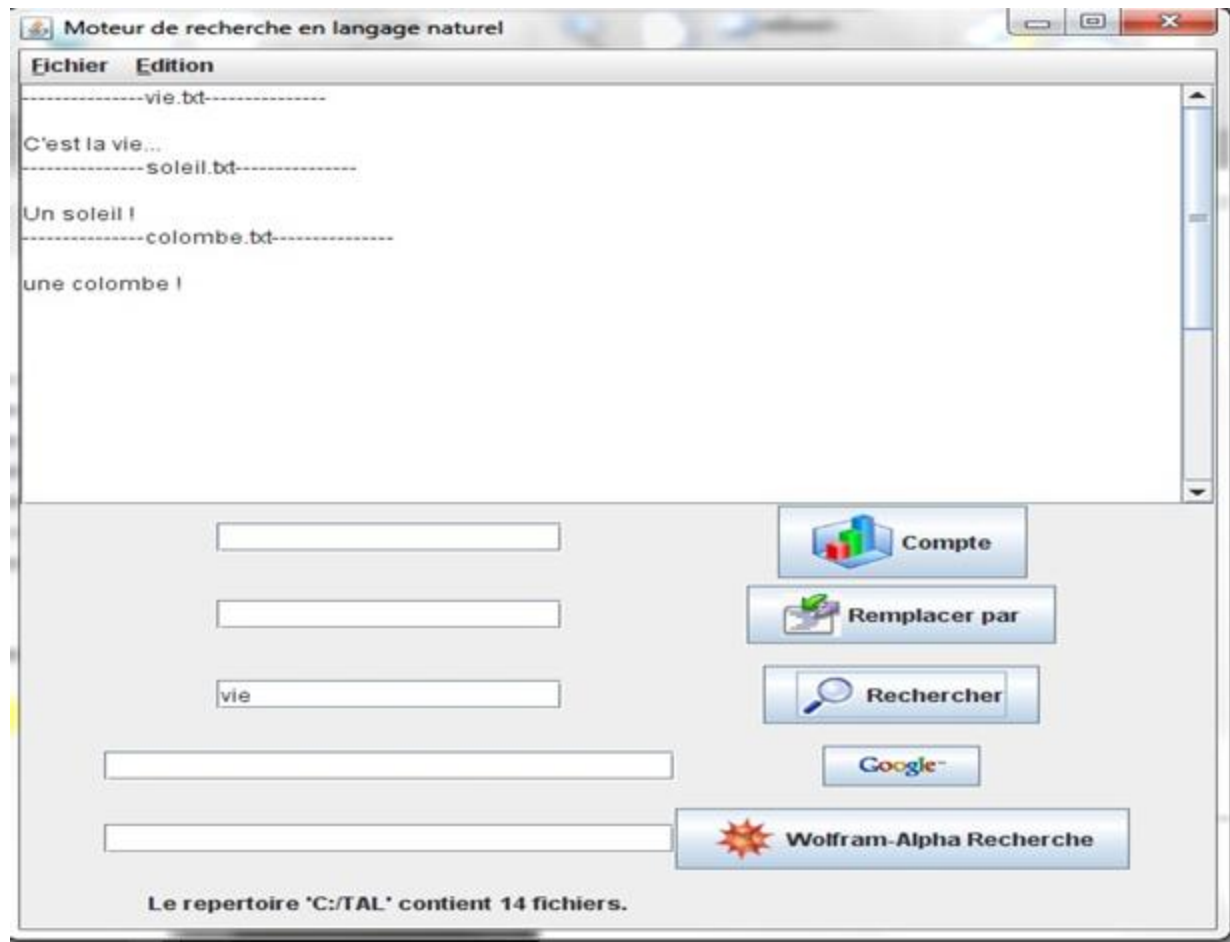
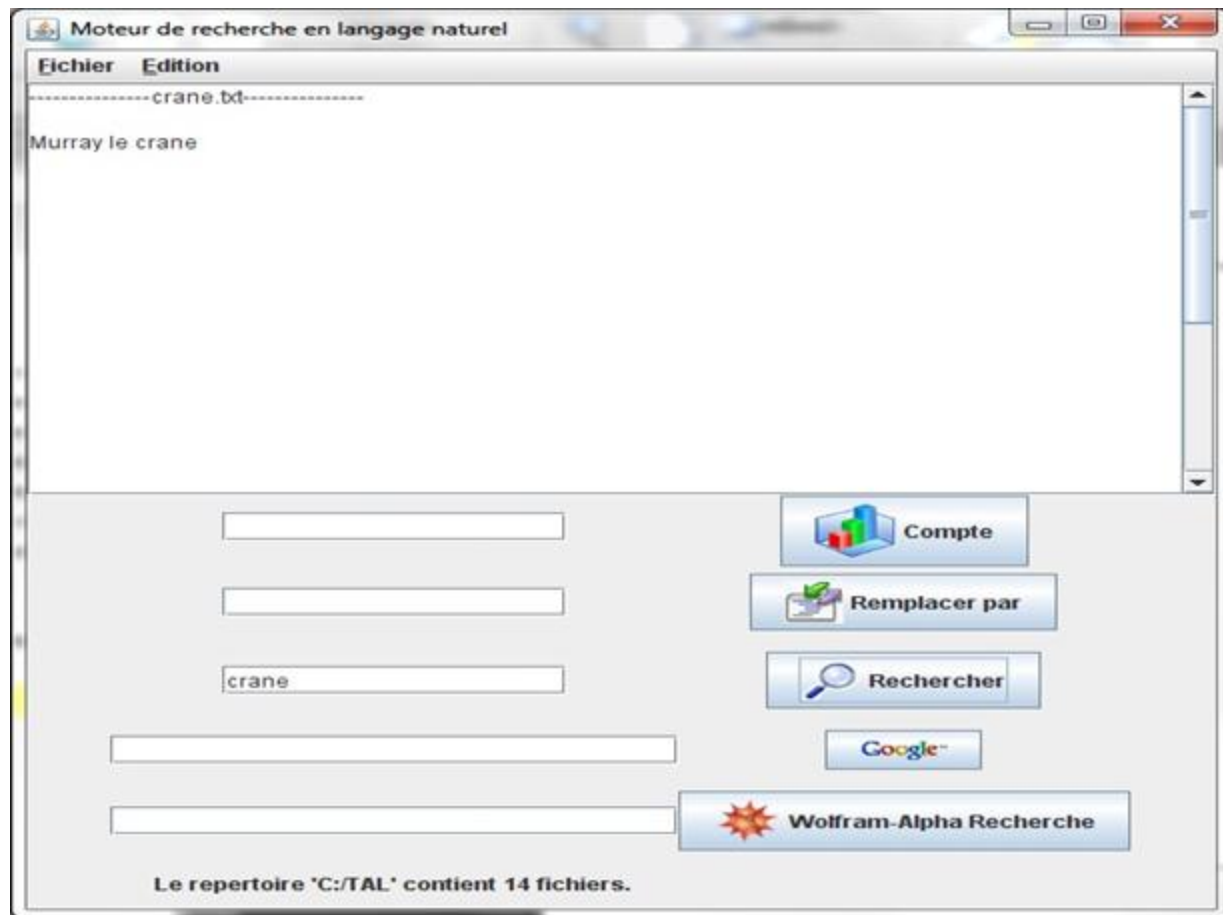


Illustration 2: tableau 1



*Illustration 3: tableau 2*



*Illustration 4: tableau 3*

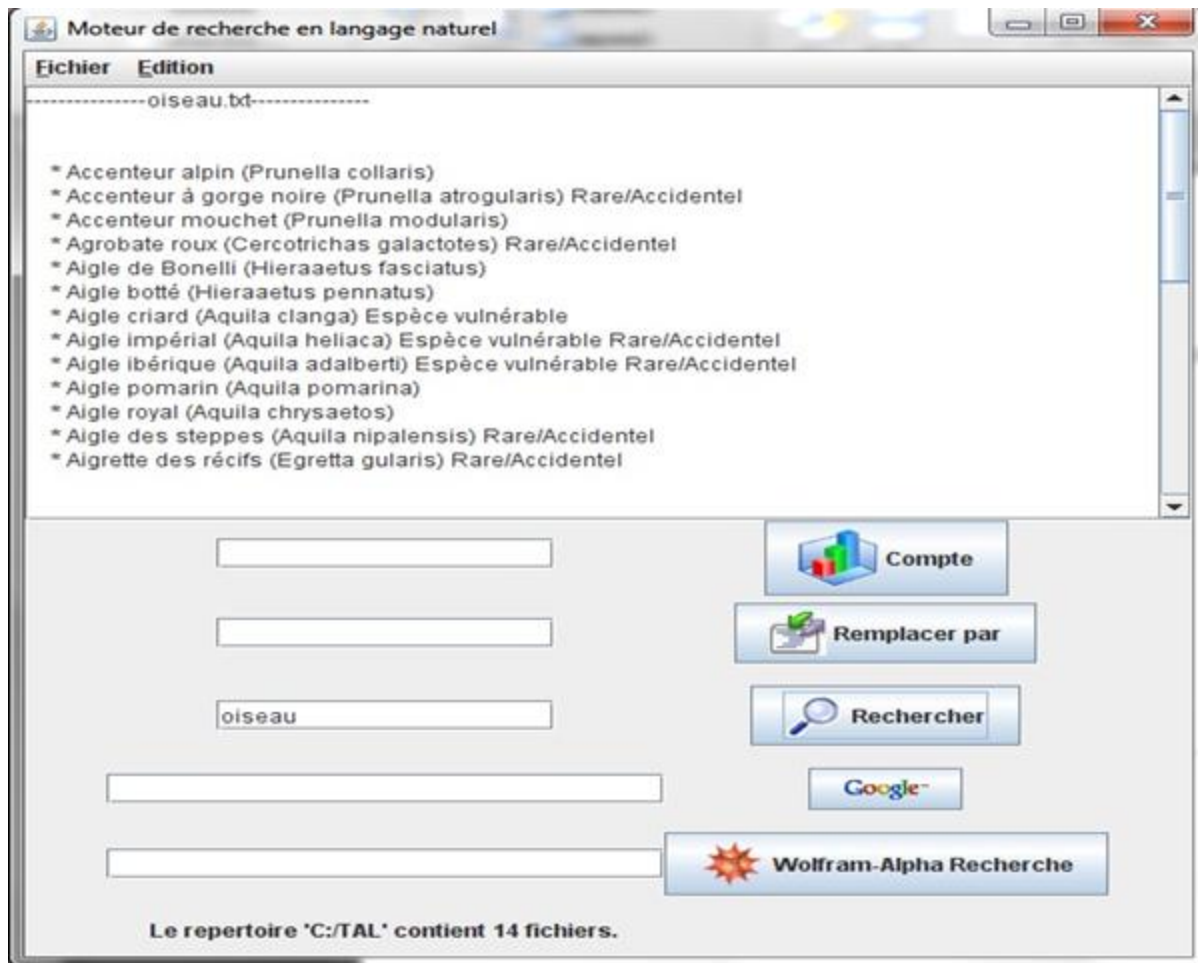


Illustration 5: tableau 4

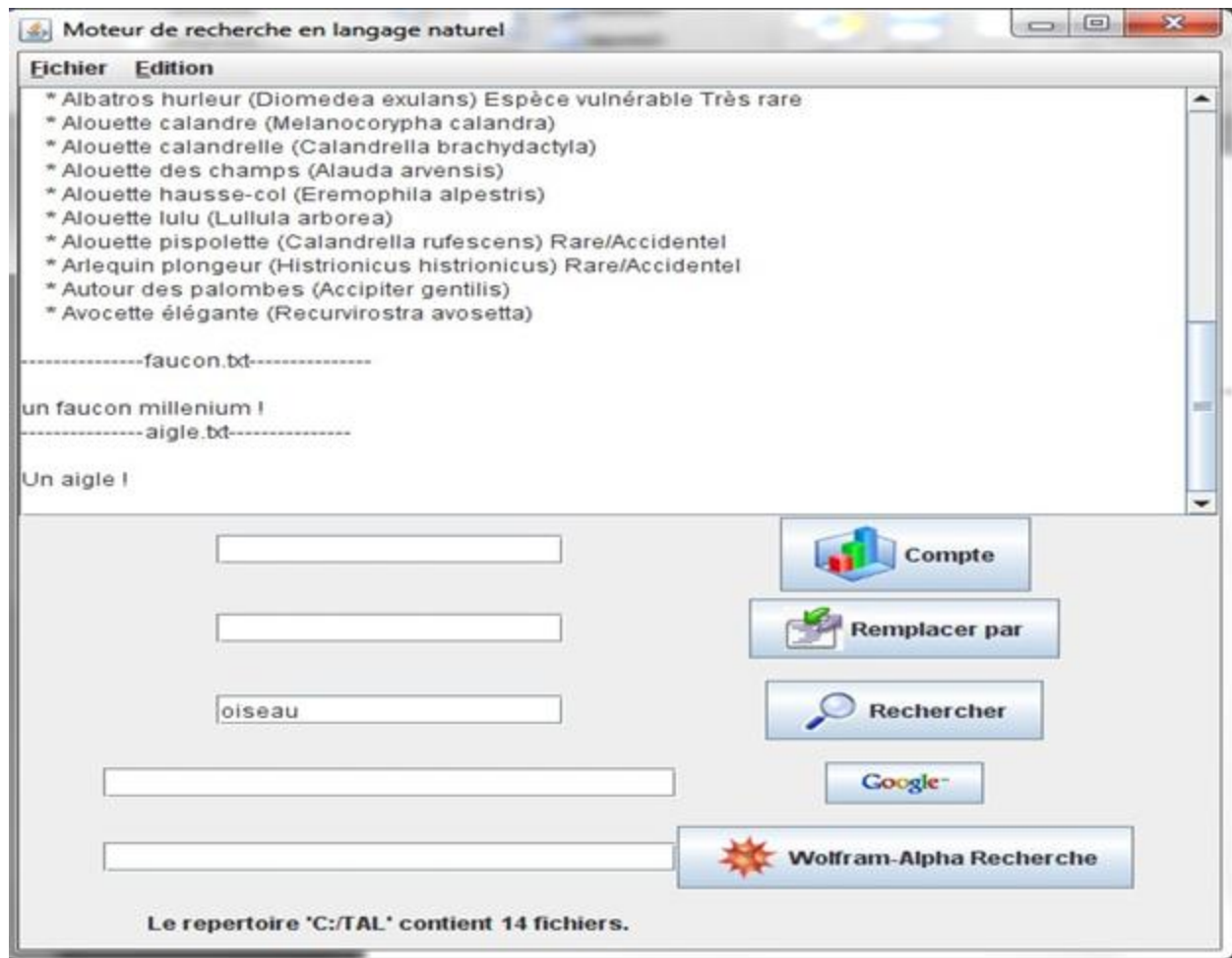


Illustration 6: tableau 5

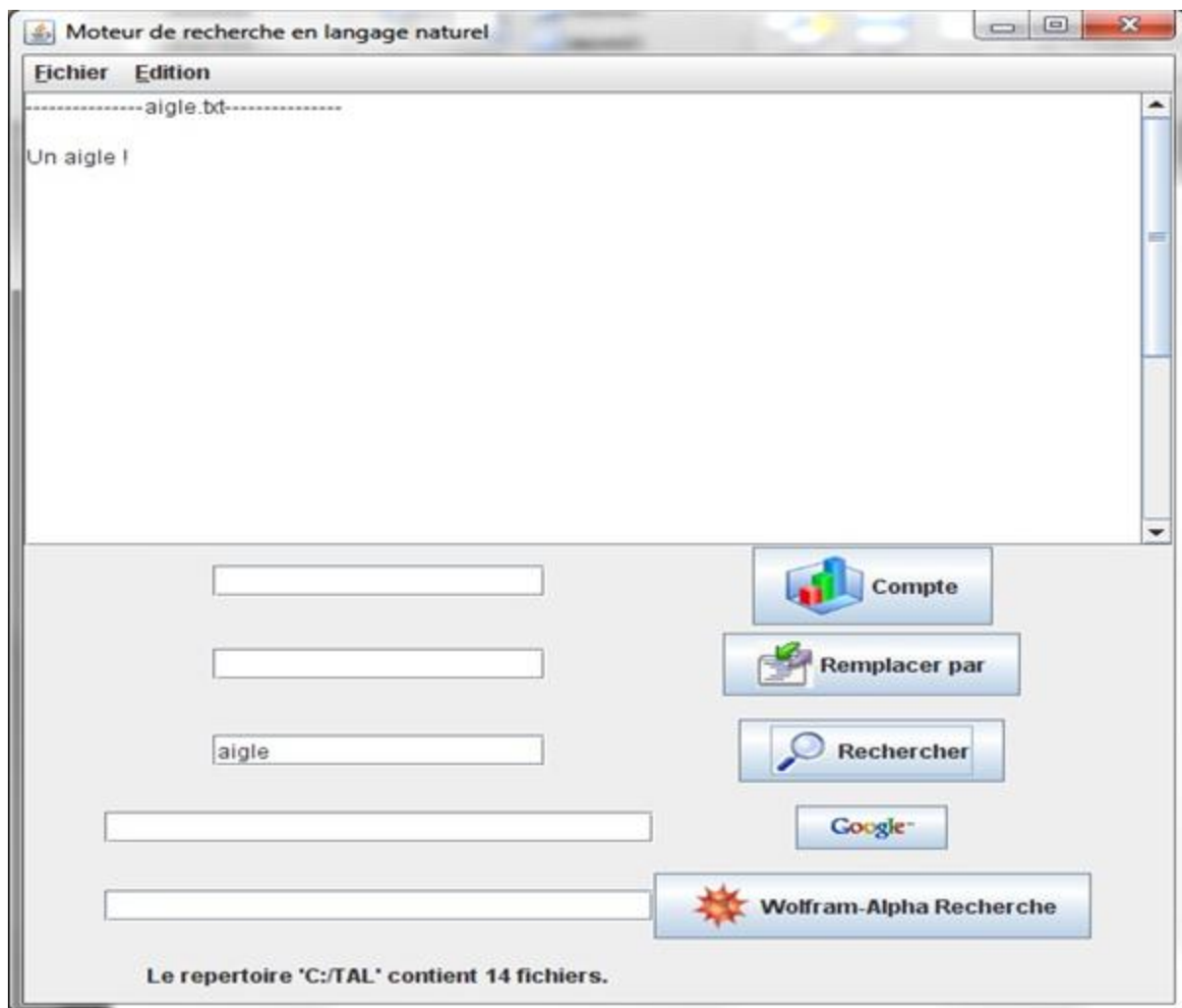


Illustration 7: tableau 6