

## Add a French Language Text to TAPoR

*This recipe takes a French language text and adds it to the [TAPoR](#) workspace for textual analysis.*



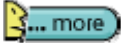
This recipe ensures that the fundamental task of loading text into a text analysis environment is accomplished correctly. For proper analysis, the text must be interpreted by the computer in the same way in which you enter it, including accented characters.

There are a variety of ways in which text can be [encoded](#) by operating systems and applications during text entry and storage.

This recipe will ensure that your text has been entered and encoded properly for analysis and that you can enter search terms and parameters from your browser to complete analytical tasks.



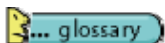
### Ingredients

---

- An electronic text in the French language
- A List Words Tool such as [TAPoR List Words Tool](#)
- A Concordance tool such as the [TAPoR Find Words - Concordance Tool](#)
- A Text Editor capable of converting between character encodings 

### Recipe Steps

---

1. **Prepare** Text in an external editor to ensure that it is [encoded](#) correctly ;  
or
2. **Confirm** that the French language web page that you wish to use is encoded properly ;
3. **Log in** to [TAPoR](#);
4. **Add** your encoded French language text file to **MyTexts** ;
5. **Generate** a word list (sorted by frequency) using the [TAPoR List Words Tool](#);
6. **Explore** an accented word individually using [Find Words - Concordance Tool](#)

### Discussion

---

## Text Editors

You may require a text editor to [encode](#) your text into [UTF-8](#) or [Latin-1](#) to maintain the accents and special characters in the textual language. On a Windows system, this can be done through the \*NotePad\* and under Macintosh OSX through \*TextEdit\*. On Unix-based systems, you will find a text editor installed as part of the standard system install. Word processors typically provide a much deeper tool set for formatting text and generally save documents in their native format which is not appropriate for importing into a text analysis environment. However, they too can be used to save a plain text file with appropriate encoding by following the appropriate steps.

- Instructions for saving as UTF-8 or Latin-1 using [NotePad](#)
- Instructions for saving as UTF-8 or Latin-1 using [TextEdit](#)
- Instructions for saving as UTF-8 or Latin-1 using [MicrosoftWord](#)

## Web Page Encoding

To verify that the web page that you wish to import into TAPoR is encoded in either [UTF-8](#) or [Latin-1](#), you need to check the browser settings. In Internet Explorer, simply got to the *View* Menu and select the Encoding Option. This should read Unicode (UTF-8). On Firefox, the option is *Character Encoding* under the *View* menu. This should also read Unicode (UTF-8). If this is not the case, then you can manually select the encoding you wish to use from this menu. On other web browsers, the process should be similar. Please consult their help files for specific instructions on character encoding. If you view the page source for your web page, it may contain the HTML line:

```
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
```

or

```
"<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1"/>"
```

This will indicate that it is encoded properly for text analysis.

## Exercise Steps

---

## Glossary

---

### Text Encoding

One of the most important aspects of the text input process is the encoding of the text which you are working with. It must be encoded as either UTF8 or Latin-1, which provides proper mapping of accented and other extended characters. See the links below for more background information on encoding processes. For

example, when properly encoded the character 'e' is differentiated from the character 'é' and 'é' is not seen as the character 'e' + some symbol.

### **UTF-8 (8-bit Unicode Character Encoding)**

Unicode character encoding is an evolution of the ASCII set to permit support of a greater number of alphanumeric characters including those with diacritical marks such as accents. More information on UTF-8 is available at: [Wikipedia](#)

### **Latin-1 (ISO 8859-1)**

Latin-1 character encoding is an evolution of the ASCII character set to permit support of a greater number of alphanumeric characters including those with diacritical marks such as accents. It is being supplanted by the [UTF-8 Character Encoding](#) More information on Latin-1 is available at: [Wikipedia](#)

### **MyTexts**

This is an area of the TAPoR in which you collect your private texts for analysis. It is also a portal to access publicly available texts which have been added by other users. In this area you can view the catalogue of texts available to you or add, edit, tag, and view the contents of specific texts.

### [A Complete Glossary](#)

<statement of origin, caveats, etc>

V1.0 20 May 2006