# Improving the Built-in Dictionary

*Gedit is a great text editor complete with a spellchecker, but many words spelled correctly are underlined in red as misspelled words. Why?*

This is not gedit's fault. Linux, and Ubuntu, contains a basic built-in dictionary located at **/usr/share/dict/words**that any program, including gedit, can tap into to check for misspelled words in a text file.

The problem is that the built-in dictionary is extremely basic, but it is easily improved by adding words to the dictionary's text file.

## Finding the Dictionary File

Words used for spellchecking are stored in a plain text dictionary file located at **/usr/share/dict/words**. Entering **cat usr/share/dict/words** will show the words in a terminal, but **words** is actually a soft link to the real dictionary file based upon the set locale. For an American Ubuntu locale, usr/share/dict/words points to the file**american-english**, which contains a small list of words specific to American English. If the locale is the United Kingdom, the words link points to the file **british-english**.

To keep things easy and access the dictionary of the current locale, access the dictionary file using usr/share/dict/words. The correct file will be edited.

## Gathering Wordlists

Collect as many word lists as possible. Wordlist files are plain text files containing one word per line without any extra formatting or markup. These can be lists of country names, cities, custom dictionary files, words often misspelled, or premade wordlists obtained from various sources.

Make sure each word is spelled correctly (if feasible) because these are the words the spellchecker will match to determine correct spelling.

One example is OpenMedSpel, a free medical word list containing medical words. This wordliast adds 50,000 medical terms to help ensure that many medical words are spelled correctly. Download the ASCII text version.

Other wordlists, dictionaries, and thesauruses are available. The more words, the more chances of detecting incorrect spellings.

## Create a Single Wordlist File

The idea is to create a single dictionary from all of the separate wordlist files. To start, let's build upon the existing dictionary file by copying it into a temporary file in home.

```
cp /usr/share/dict/words ~
```

A file containing the default dictionary is created. Open this file and copy and paste each word list into it. Or use something like **cat wordlist1 wordlist2 wordlist3 >> copied-wordlist** to
create a single word list file containing all words from all of the gathered wordlist files.

Ignore duplicates for now. Copy them anyway because they will be removed later. The result should be a single wordlist text file with one word per line.

## Removing Duplicates

Remove the duplicates lines. In a terminal, enter

```
sort -u single-wordlist > sorted-wordlist
```

or

```
sort single-wordlist | uniq > sorted-wordlist
```

The sorted wordlist should be arranged alphabetically with one unique word per line. This helps reduce the file size and eliminates redundancy.

## Remove Blank Lines

Chances are good that blank lines exist, so remove them too.

```
sed -i '/^$/d' sorted-wordlist
```

Sed is a a stream editor that processes text using regular expressions. The **-i** option edits the file in place, so any changes are made directly to the file itself.

## Remove Leading Spaces

Some words might contain whitespace before them. Each word should begin each line. Remove the leading whitespace.

```
sed -i 's/^[[:space:]]*//' sorted-wordlist
```

## Remove Oddities by Hand

Open the word list in a text editor and delete unnecessary words, garbage characters (they happen), and any other nonsense that might have snuck into the wordlist.

## Overwrite the Existing Wordlist

Once satisfied with the wordlist, copy it into the Linux dictionary.

```
sudo cp sorted-wordlist /usr/share/dict/words
```

The improved wordlist already contains all default words (it was copied to home earlier), so they will be preserved by overwriting the default dictionary file. Even though /usr/share/dict/words is a soft link, the correct locale file will be updated.

## Save the Wordlist

The newly-created wordlist is valuable and will probably be used again, so save the wordlist file so it can be used in future Linux installations. If any major updates or corrections are made to the wordlist, always backup the wordlist.

Now, gedit will see a larger set of words for spellchecking.

Source : https://delightlylinux.wordpress.com/2012/05/23/improving-the-built-in-dictionary/