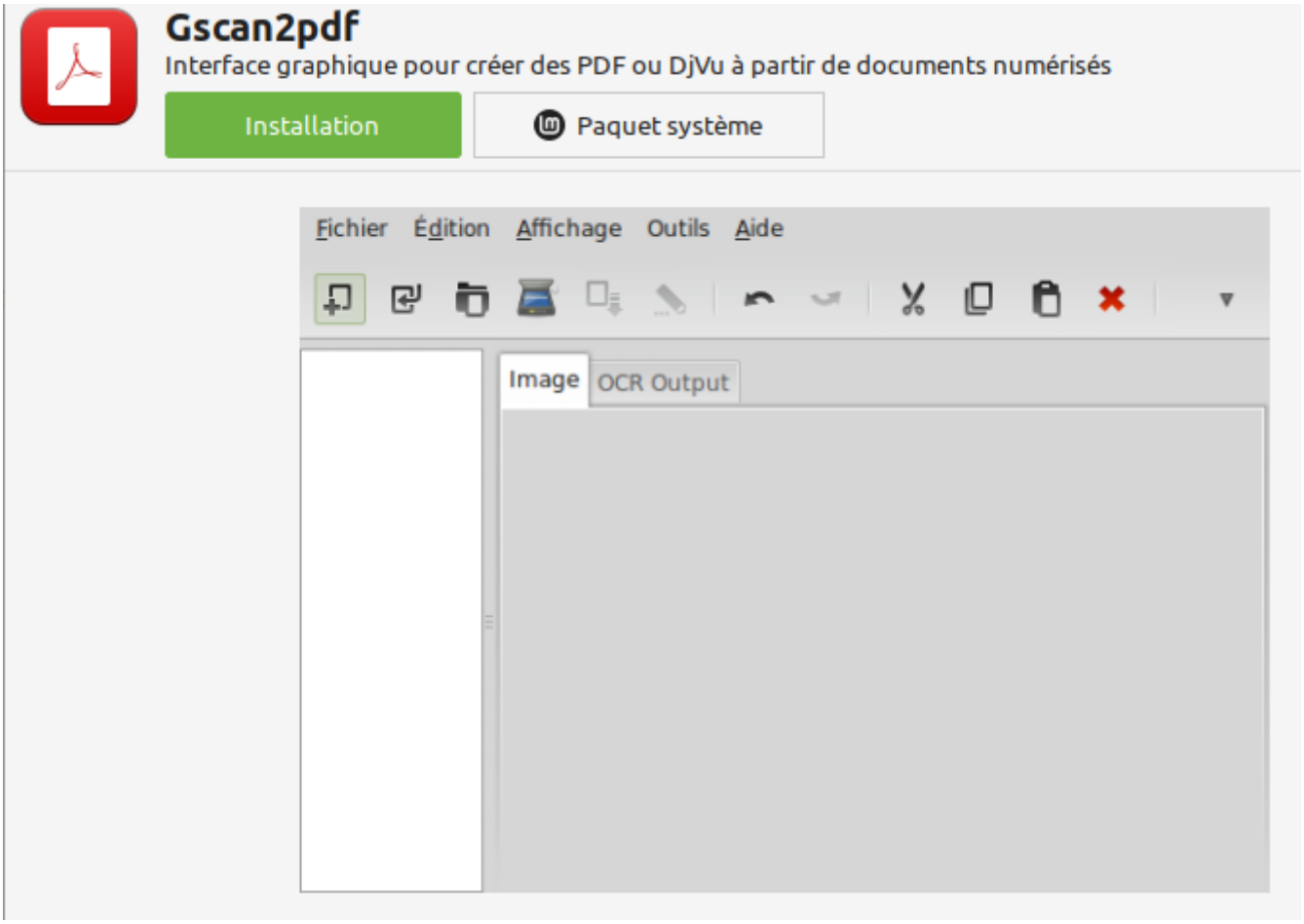
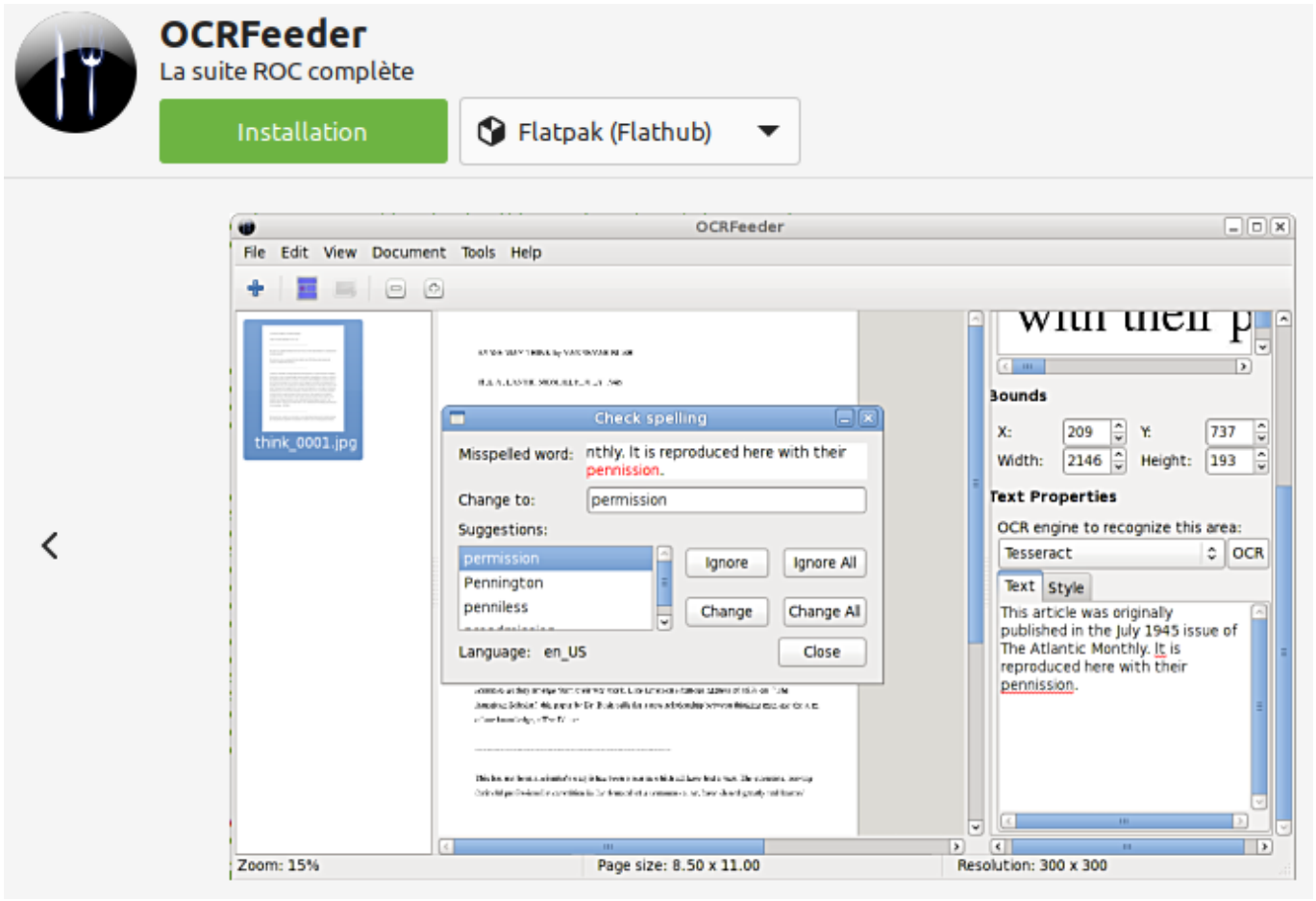
Tesseract OCR

Tesseract OCR est un moteur de reconnaissance de caractères. Il peut être utilisé soit en ligne de commande, soit par l'intermédiaire d'un programme graphique comme **gscan2pdf**



ou **OCRfeeder**



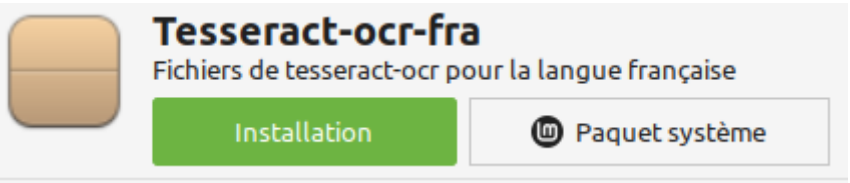
Le programme effectue la reconnaissance de caractères (OCR) à partir de fichiers images populaires : pif, png, jpeg, tiff, bmp, gif, pgm, ppm, ico, xbm et xwd.

Installer Tesseract

Tesseract OCR n'est pas installé par défaut. Il faut l'installer soit en ligne de commande :

```
sudo apt install tesseract-ocr-fra tesseract-ocr
```

soit à partir de la logithèque :



Utilisation

Nous allons étudier le fonctionnement de Tesseract OCR en ligne de commande.

Dans notre premier exemple nous effectuerons la reconnaissance de caractères (OCR) sur un fichier image JPG en utilisant la langue française :

```
tesseract -l fra mon-fichier.jpg mon-fichier-texte
```

La commande exécute **Tesseract OCR** sur le fichier image JPG “mon-fichier.jpg”, en utilisant le modèle de langue française, et extrait le texte reconnu, qu'il enregistre dans un nouveau fichier nommé “mon-fichier-texte.txt”. Vous obtiendrez ainsi une version textuelle du contenu du fichier image JPG.

Dans notre second exemple, nous effecturons la reconnaissance optique de caractères (OCR) sur plusieurs fichiers JPEG (avec l'extension .jpg) dans le répertoire courant.

Le texte généré est stocké dans des fichiers texte (.txt) portant le même nom que les fichiers “.jpg” d'origine. Par exemple, si vous avez un fichier “image1.jpg”, cette commande créera un fichier “image1.jpg.txt” contenant le texte extrait de “image1.jpg”.

```
for i in *.jpg ; do tesseract -l fra $i $i;  done;
```