PREPARATION TO WORKSHOP - INTRODUCTION TO AUTOMATED TEXT TO IMAGE CONVERSION

Welcome to this first session of the workshop series on Digital Humanities organized by Dagitab. My name is Rocío Ortuño and I am going to lead the first workshop. It is just an introduction to Transkribus, a programme to extract text from pictures of books, newspapers etc. Although Transkribus was designed to ‘transcribe’ handwritten texts, it is also a powerful tool to ‘transcribe’ printed texts. We are going to focus then on printed texts.

Transkribus is not difficult to use. However, installing it may be a bit tricky. This is why I recommend doing some activities before the virtual contact session.

A few tips even before the video and the activities:

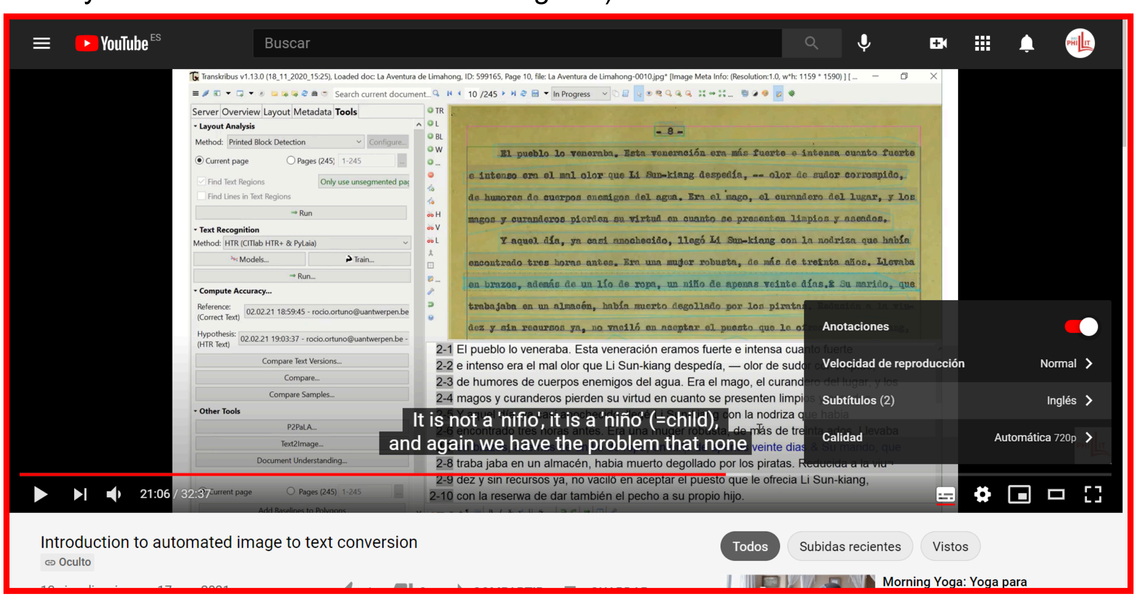
1) You need to FIRST register on the Transkribus website https://readcoop.eu/transkribus/ and SECOND download Transkribus.
2) From the downloaded package, if you have a Mac you need to install the Transkribus-1.13.0.jar file (the one with .jar at the end). If you have a Windows pc, you will need to install the .exe file Transkribus.exe
3) If you have any problems installing Transkribus or any questions, you can leave them on this form. I will reply to you as soon as possible (before the contact session).

Have fun!

ACTIVITIES

1. Watch this video https://youtu.be/OCI2D03dcz8 (it takes about 30 minutes). It is in Spanish with English subtitles, so you will also be refreshing your Spanish 10.

    *To turn the English subtitles on, please press on the subtitles option on the bottom of the image (white rectangle with small lines), and then on the tools icon. There you need to choose “subtitles > English”).



2. Do the following exercises:
    a. Sign up and download Transkribus
    b. Create a collection called "Dagitab Workshop"
    c. Share it with me (rocio.ortuno@uantwerpen.be) if you want me to share the model we trained in Antwerp on prose in Spanish.
    d. Upload a pdf document. It may well be one of the newspapers from the repository https://repository.mainlib.upd.edu.ph/
    e. Make a layout analysis of some pages of the documents and check that it is correct. If there are errors in the layout analysis (ie. the text areas marked, are marking stains on the paper or lines are not in the right order…), you can check here how to correct them https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/#Segmentation_-_Layout_Analysis , in the section "Correcting the results of automated segmentation ".
    f. Apply a transcription template to some pages (ideally 25) and correct them.
    g. Create your own model.
    h. Write to Transkribus for them to classify it as printed text, and apply it when ready to the rest of your pages. Export everything in txt.

LIVE SESSION:

Slides:
https://drive.google.com/file/d/1AGu0UmH4X2cfe5h4eELHnBLgxN9EifBf/view?usp=sharing

text for practising
https://drive.google.com/file/d/1IFirzLuU_xMRZe6qZ8HpY9DXEwrfrGW4/view?usp=sharing

> Regarding layout analysis:

I made a couple of mistakes: you only need to do the layout analysis (and correct it) for the pages that you are going to train. The others will be done automatically.

Please follow this video tutorial: https://www.youtube.com/watch?v=_r8woJQSyGE