Exchange

Ray Smith is the main developer of Tesseract since 1985.

Ray Smith
<theraysmith@gmail.com>
1 Oct 2013
to Colm, Sarah, Pierre, Ludi

The training process that we use involves rendering text in available fonts and building a language model.
The source text for both these processes is gathered from web pages that has been identified to be in the required language.
Some languages have received special treatment for specific problems, but Korean is not one of them.
Although people occasionally ask for the training data, it is just too big to host on the site even if we had all the necessary copyright clearances.
Instead we are working on opening up the training tools that we use so more people can enjoy automated training.

Hope that helps.
Regards
Ray.

Sep 28, 2013, Colm O'Neill:

Dear Ray Smith,

We're contacting you regarding Tesseract, some of the dev forums suggest you as a person to contact about language support in Tesseract.
We are OSP (Open Source Publishings), a working group based in Brussels, that works exclusively with Free / Libre Open Source Software in the context of Graphic Design, bringing together commissions, teaching and research.
http://osp.constantvzw.org/

We're currently in Seoul prepping for a workshop at the typography biennial Typojanchi and for this we're going to use Tesseract as common ground between our different alphabets, i.e. Latin and Hangul. Our focus is going to be on the training of Tesseract, going through the steps to build some new specific data for the Korean language.

With this, we're wondering about the origins of the current data available for Korean. We can't find out where the downloadable data originates from, on what kind of text images it was trained, etc.

The history of the tool obviously has something to do with this, so we're coming to you hoping you can provide us with information about the origins? (What are the /) Where could we find the original samples?

Any help would be appreciated.

(You're not in Seoul yourself next week by the wildest of chances?)

All the best,
Yours,

Sarah, Pierre, Ludi, Colm,
for OSP.