Description of the workshop as announced

Fancy reading machines and androids from science-fiction fantasies are embodied in our modern lower-profile world as OCR software packages. OCR means Optical Character Recognition and it is software that can extract text from image files. One of these softwares, the free and open source Tesseract[1] is composed of two parts that we can study, thanks to its license. There is the engine itself, and the training data for a language[2] partly based on what Tesseract called 'prototypes'. We could compare this 'before the type' (proto-type) to the culture a lecturer progressively gathers from his first lesson going from a novice to a fully grown expert. By following the limits between the blank surfaces and the dark pixels of the shapes of letters, Tesseract compares its journey with previous ones, on images already followed in the past. It starts by learning patterns and specificities of languages, rhythms and irregularities. It goes on to recognise the body of a glyph, then it works out, bit by bit, if this glyph is a letter, form is a word, and eventually it makes out phrases. Like all of us, Tesseract learns typography in this same process, in a completely intertwined way, as sentences, script and eventually, language.[3]

Tesseract follows rules by which it can make decisions. In a basic example from Latin script, if the software seems to be recognising something resembling to iii (three times the letter 'I'), specific rules kick in to suggest that it is most lightly the letter 'm' and not a triple consonant. Grammar and language coming in at a later stage, as it did for us, still following this unusual idea of teaching software to read.[4] The very specificities of typography and how each shape is drawn and could or couldn't be distinguised from another one arrives just after. As in the previous example the potential small parts that protrude from the I are more likely to be the arc of the m if the font is a serif one tha it is a sans serif one. This process becomes intertwined with the actual context: with time, the system becomes familiar, and extremely efficient with some specificities of a typeface. Its shape, its overall form and size now mean something. It would have to relearn an entirely new toolkit to be able to read a different typeface. With this, could the relations binding shapes to their meanings be noticed?

At young, naive and early stages of deciphering writing systems, slowly working out the building blocks to a

[1] - http://en.wikipedia.org/wiki/Tesseract_(software)
[2] - https://code.google.com/p/tesseract-ocr/
[3] - http://code.google.com/p/tesseract-ocr/wiki/TrainingTesseract3
[4] - http://code.google.com/p/tesseract-ocr/wiki/TrainingTesseract3#The_last_file_(unicharambigs)