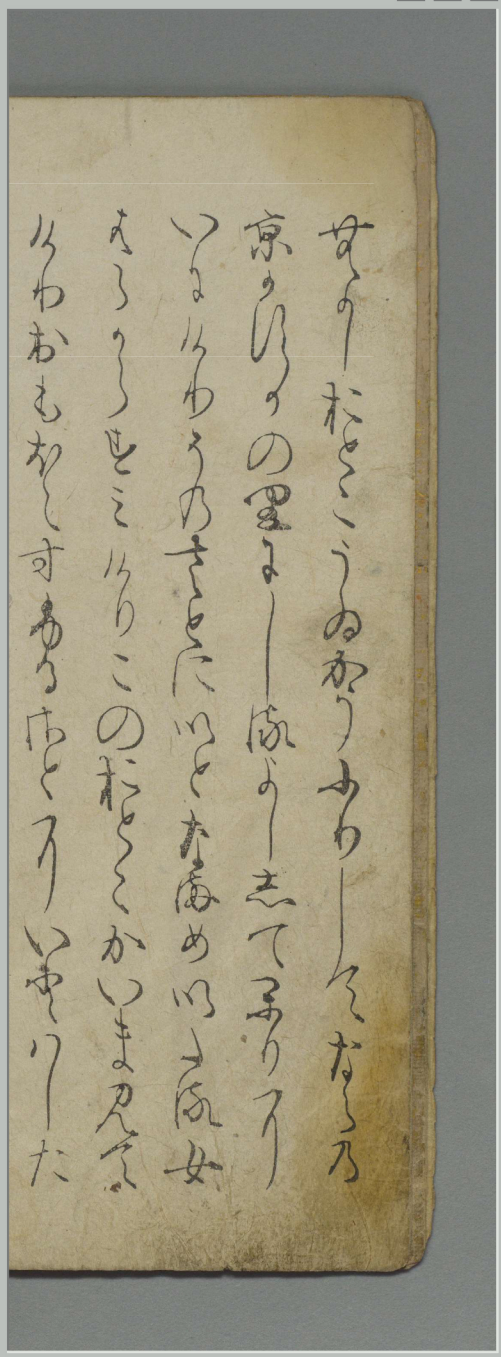

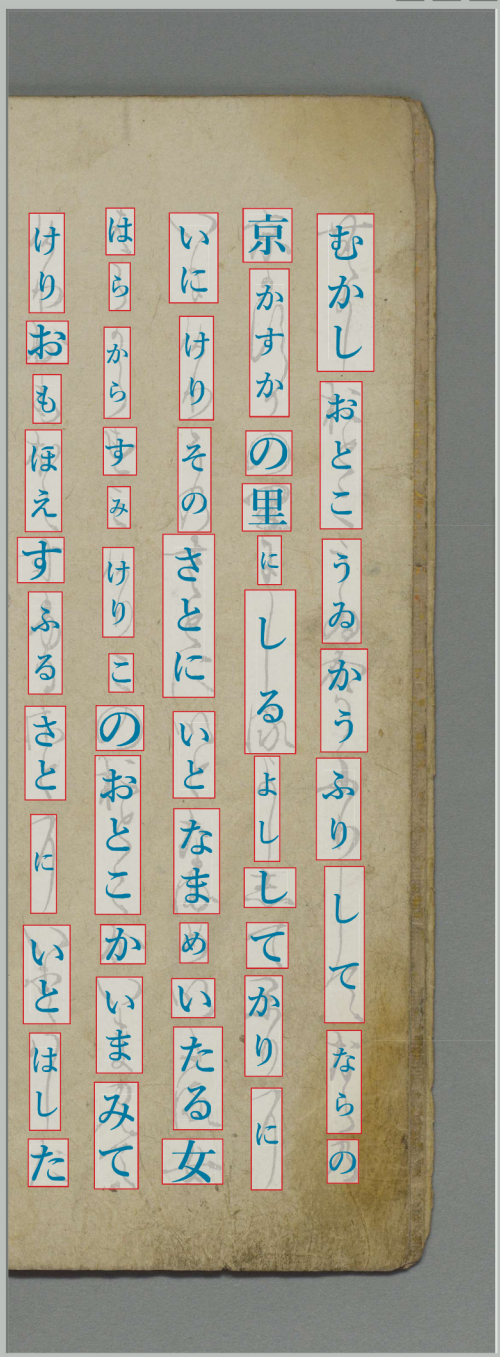

声に出して読めない日本語。

ご覧いただいているのは、江戸初期の「嵯峨本」と呼ばれるもの。手書きのようですが、実は木でできた「活字」なのです。つまり、手書き文字というアナログからの変換の試み…。いわば、400年前に日本で起きたデジタライゼイション。そう言えるかもしれません。トッパンが印刷博物館で所蔵していますが、私たちは嵯峨本を手がけた人々の、文化への思いも受け継ごうと考えています。たとえば今、日本人のほとんどは、この嵯峨本にもある「くずし字」を読むことが難しくなりました。一方、まだ読み解かれていない古典籍・古文書は数十億点とも…。解読できる専門家の数は限られ、日本人が紡いできた千年に及ぶ知恵や歴史は眠ったままなのです。そこで、トッパンが取り組んでいるのが、「くずし字OCR（光学文字認識）」。これは、文字の形状のパターンを認識する技術を応用し、スキャナーなどで入力した文書の画像の中から、文字を識別していくテクノロジー。ディープラーニングベースの認識エンジンを確立できたことにより、今、研究が一気に加速。さらなる解読精度の向上を目指して、文字の形、つまり字形のデータベースが、より正確かつ網羅的なものとなるよう磨きをかけています。

TOPPAN

https://www.toppan.co.jp/