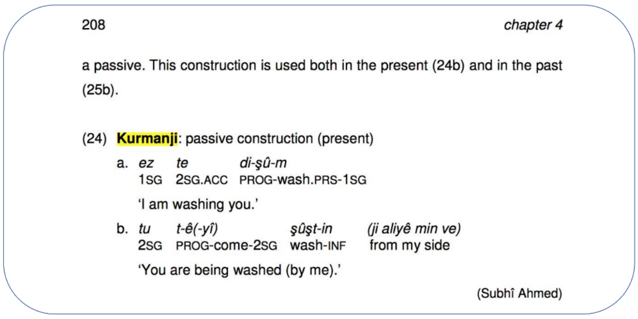
PDF-to-Text Reanalysis for Linguistic Data Mining

Michael Wayne Goodman **Ryan Georgi** **Fei Xia**
Linguistics Department
University of Washington
Seattle, WA, USA
{goodmami, rgeorgi, fxia}@uw.edu

Motivation

- **Many online academic publications are in the form of PDF files, and extracting information from them can be very useful:**
  - Ex: Millions of online linguistic documents that contain language data.
- **Extracting semi-structured text from PDF files is not trivial:**
  - Existing PDF-to-text converters have limitations
- **We develop a system (called Freki) for the reanalysis of PDF-extracted text:**
  - It performs block detection, respacing and tabular data analysis.

Interlinear Glossed text



208 *chapter 4*

a passive. This construction is used both in the present (24b) and in the past (25b).

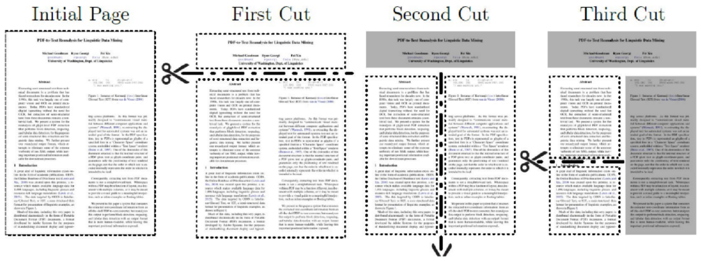
(24) **Kurmanji**: passive construction (present)

a. *ez* *te* *di-şû-m*
1SG 2SG.ACC PROG-wash.PRS-1SG
'I am washing you.'

b. *tu* *t-ê(-yî)* *şûşt-in* *(ji aliyê min ve)*
2SG PROG-come-2SG wash-INF from my side
'You are being washed (by me).'

(Subhî Ahmed)

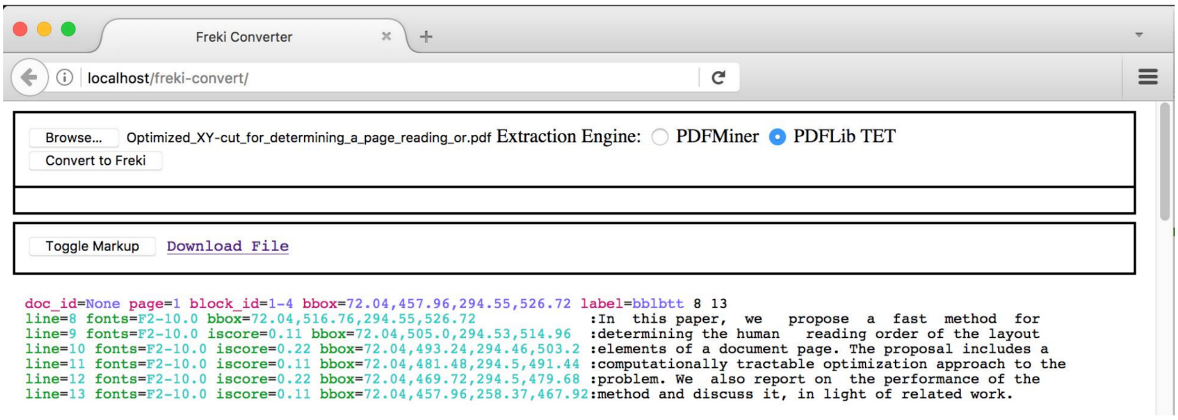
The XY-cut algorithm



Our approach

- **Built on top of existing PDF-to-text converters that produce XML format.**
- **Added functionality:**
  - Block detection
  - Respacing and tabular data analysis
  - Changes to the XY-cut algorithms
- **Used a new output format for less verbosity and more readability**

The interface



Extraction Engine: ○ PDFMiner ● PDFLib TET

```
doc_id=None page=1 block_id=1-4 bbox=72.04,457.96,294.55,526.72 label=bblbtt 8 13
line=8 fonts=F2-10.0 bbox=72.04,516.76,294.55,526.72                 :In  this paper,  we   propose  a  fast  method  for
line=9 fonts=F2-10.0 iscore=0.11 bbox=72.04,505.0,294.53,514.96    :determining the human   reading order of the layout
line=10 fonts=F2-10.0 iscore=0.22 bbox=72.04,493.24,294.46,503.2   :elements of a document page. The proposal includes a
line=11 fonts=F2-10.0 iscore=0.11 bbox=72.04,481.48,294.5,491.44   :computationally tractable optimization approach to the
line=12 fonts=F2-10.0 iscore=0.22 bbox=72.04,469.72,294.5,479.68   :problem. We  also report on  the performance of the
line=13 fonts=F2-10.0 iscore=0.11 bbox=72.04,457.96,258.37,467.92:method and discuss it, in light of related work.
```

Using Freki

- We ran Freki on millions of documents from the Web which may contain interlinear gloss text (IGT) data.
- From the extracted text, we performed IGT detection, language identification, etc.
- This leads to an extension to the Online Database of INterlinear text (ODIN), which contains IGT for thousands of languages.

Summary

- **PDF-to-text is not trivial, especially if we want to preserve the structure within the text.**
- **Our system, Freki, builds on top of existing converters and focuses on identifying blocks and respacing.**
- **The package is available at** github.com/xigt/freki

LREC 2018, May 7-12, 2018, Miyazaki, Japan