Linguistic Data Consortium

CAMIO: A Corpus for OCR in Multiple Languages

Michael Arrigo[1], Stephanie Strassel[1],
Nolan King[2], Thao Tran[2], Lisa Mason[2]
Linguistic Data Consortium (1)
U.S. Dept. of Defense (2)

Introduction

- CAMIO: Corpus of Annotated Multilingual Images for OCR
- New LDC resource to support the development and evaluation of optical character recognition-related technologies like script ID, language ID, text localization, OCR decoding, keyword search and end-to-end OCR
- Systematically constructed to address gaps in existing corpora*
- CAMIO includes machine-printed text, covering
  - 35 languages across 24 unique scripts, up to 2500 docs/language
  - Scanned docs and images in a broad set of document domains
  - Variety of genres, attributes and scanning/capture artifacts
  - Most documents subject to text localization and reading order
  - Subset of documents in 13 languages also transcribed
- Data to be published in LDC Catalog and used in future evaluations

*Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V. (2019). ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1516-20.

CAMIO Corpus Pipeline

	LDC Annotator	Crowd Worker	Native Speaker Required
1 Collection	✓	✓	
2 Auditing	✓		if collection by non-native
3 Text Localization	✓		
4 Reading Order	✓		✓
5 Transcription	✓		✓

- Each stage is a check on prior stage's quality, with feedback
- Additional quality review at end of pipeline

Linguistic Data Consortium

Data Collection and Auditing

CAMIO Languages

- 35 language-script pairs
- 24 unique scripts reflected
- 13 languages selected for transcription, covering 10 scripts

Amharic (Ge'ez)	Hebrew (Hebrew)	Swahili (Latin)
Arabic (Arabic)	**Hindi (Devanagari)**	Tagalog (Latin)
Armenian (Armenian)	Hungarian (Latin)	**Tamil (Tamil)**
Bengali (Eastern Nagari)	**Japanese (Japanese)**	Telugu (Telugu)
Burmese (Burmese)	**Kannada (Kannada)**	**Thai (Thai)**
Cambodian (Khmer)	**Korean (Hangul)**	Tibetan (Tibetan)
Chinese (Simplified)	Malayalam (Malayalam)	Tigrinya (Ge'ez)
Dari (Arabic)	Maldivian (Thanna)	Ukrainian (Cyrillic)
English (Latin)	Oriya (Oriya)	**Urdu (Arabic)**
Farsi (Arabic)	Pashto (Arabic)	Uyghur (Arabic)
Georgian (Georgian-Mkhedruli)	**Russian (Cyrillic)**	**Vietnamese (Latin)**
Greek (Greek)	Sinhalese (Sinhala)	

Collection Data Volume Targets

- Originally planned for 2500 images of machine-printed text collected per language
- Revised to target a variable number of images per language
  - Reflecting data and annotator availability
  - And addition of new annotation tasks (e.g., reading order)
  - As well as prevalence of text-heavy images for some languages

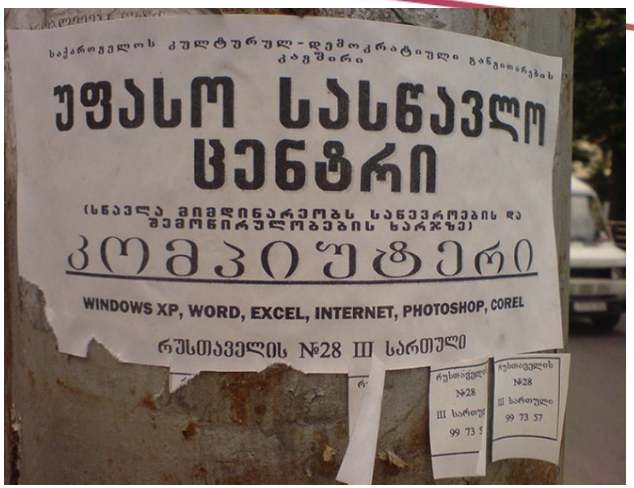
Images	Languages
2500	Arabic, Chinese, English, Farsi, Hindi, Japanese, Kannada, Korean, Russian, Tamil, Thai, Urdu, Vietnamese
2000	Amharic, Armenian, Burmese, Dari, Greek, Hungarian, Malayalam, Odia, Pashto, Swahili, Tagalog, Telugu, Ukrainian
1500	Bengali, Cambodian, Georgian
1000	Hebrew, Tibetan, Tigrinya
500	Maldivian, Sinhalese, Uyghur

Document Features

- Range of desired document features in collection, plus broad topical variety including both formal and informal content
- Attempted to achieve representation from every feature and plausible feature combination for all languages
  - Not intended to be a fully balanced corpus (cost prohibitive)
- Data scouting assignments specified desired feature combos

Document Genre	Document Domain	Document Attributes	Scanning/Capture Artifacts
Book	Text-heavy documents	Tables	Varied DPI/resolutions
Card/Slide	Unconstrained text	Multi-column	Color/grayscale/black & white
Periodical	Overlaid text	Fielded text	Warping
Record	Diagrams with text	Multi-script	Text runoff
Scene text	Varied content	Multilingual	Occluded text
Webpage		Handwriting	Distance from camera
		Text with images	Perspective
		Other complex layout	Lighting
			Skew, slant, rotation
			Noise

Document Examples



বি, বসু, এণ্ড কোম্পানীর

হাতীমার্কা সালসা।

ইহা ঠিক সালসা নহে, তবে সালসা নাম না দিলে. ইহার গুণাবলীর বিষয় কিছুই হৃদয়ঙ্গম করিতে সমর্থ হইবেন না, সেই জন্য সালসা নাম দিতে হইল। আমরা ইংরাজী-ভাবাপন্ন হইয়া পড়িতেছি, এই আয়ুর্ব্বেদীয় ঔষধের নাম তাই বিজাতীয় ভাষায় করিতে বাধ্য হইলাম—নচেৎ উপায় নাই। বলুন দেখি, সোমরস নাম দিলে সাধারণে কি বুঝিবেন?

বি, বসু এণ্ড কোম্পানীর

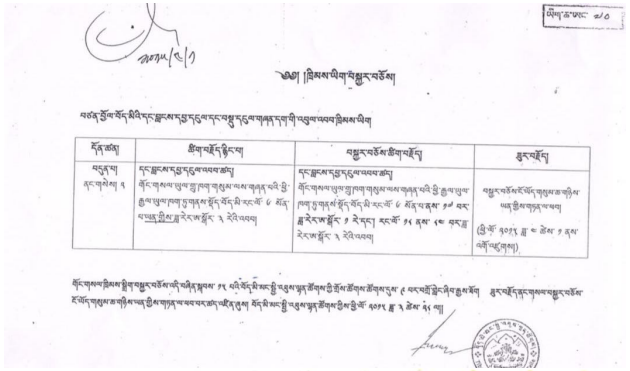
হাতীমার্কা সালসা

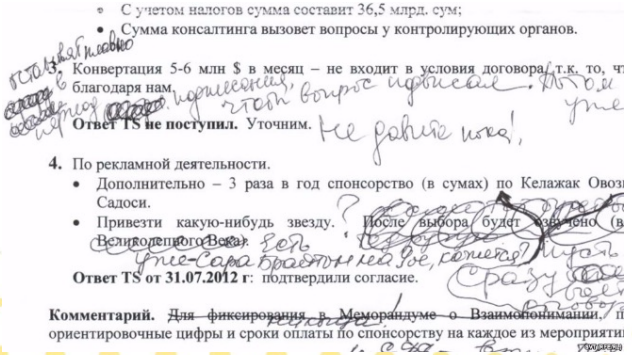
এক মহাভেজঃস্বরূপ। উত্তর চীনদেশ হইতে আনীত কোন লতা-বিশেষের এমন গুণ যে, এ সালসা সেবনের পাঁচ মিনিট পরেই দেহে এবং মনে মহাস্ফূর্ত্তি অনুভূত হইবে। মনে হইবে, শরীরে যেন কোন বৈদ্যুতিক ক্রিয়া নিষ্পন্ন হইল। এই মহাশক্তি-স্বরূপিণী সালসা-সুধাপানে মনঃপ্রাণ স্বর্গীয় সুখে বিভোর হইয়া উঠিবে। এ সালসা সহজ শরীরেও সেবনীয়! শীত, গ্রীষ্ম, বর্ষা, শরৎ, বসন্ত—সর্ব্বকালে সর্ব্বঋতুতে সেবনীয়

মূল্যাদি।

	মূল্য	ডাঃমাঃ	প্যাকিং
১নং আধপোয়া শিশি	৷৵০	৷০	৵০
২নং একপোয়া শিশি	১৵০	৸০	৵০
৩নং দেড়পোয়া শিশি	১৷৵০	১	৵০

ভ্যালুপেবলে লইলে মূল্য আরও দুই আনা বা চারি আনা অধিক পড়ে। তিন বা চারি শিশি অথবা এক ডজন একত্র লইলে ডাকমাশুল



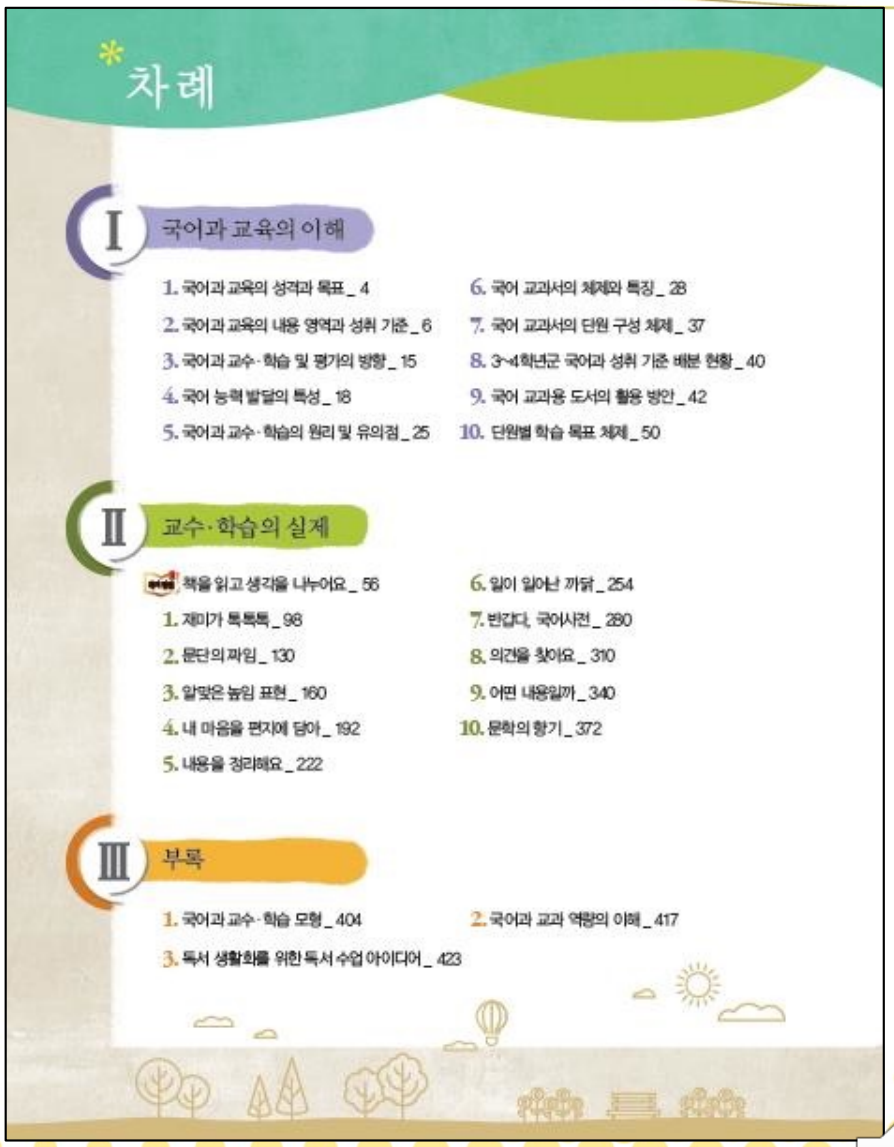


CAMIO Collection Methods

- Method 1: Trained data scouts search the web for existing images (scanned documents or photos) of text in their language on public websites
  - Primary method for most languages
- Method 2: Data from text repositories
  - Designed to address gaps in Method 1
  - Proved logistically difficult, not utilized
- Method 3: Crowdworkers upload existing images in their language
  - Designed to address gaps in Methods 1/2
  - Most fruitful for Bengali, Cambodian, Dari, Hebrew, Maldivian, Sinhalese and Uyghur

Auditing

- All collected images audited
- By a native speaker if collection by non-native speaker
- Verify that image is usable and in-language (if applicable)
- Verify or provide document feature judgments
- Simple web form

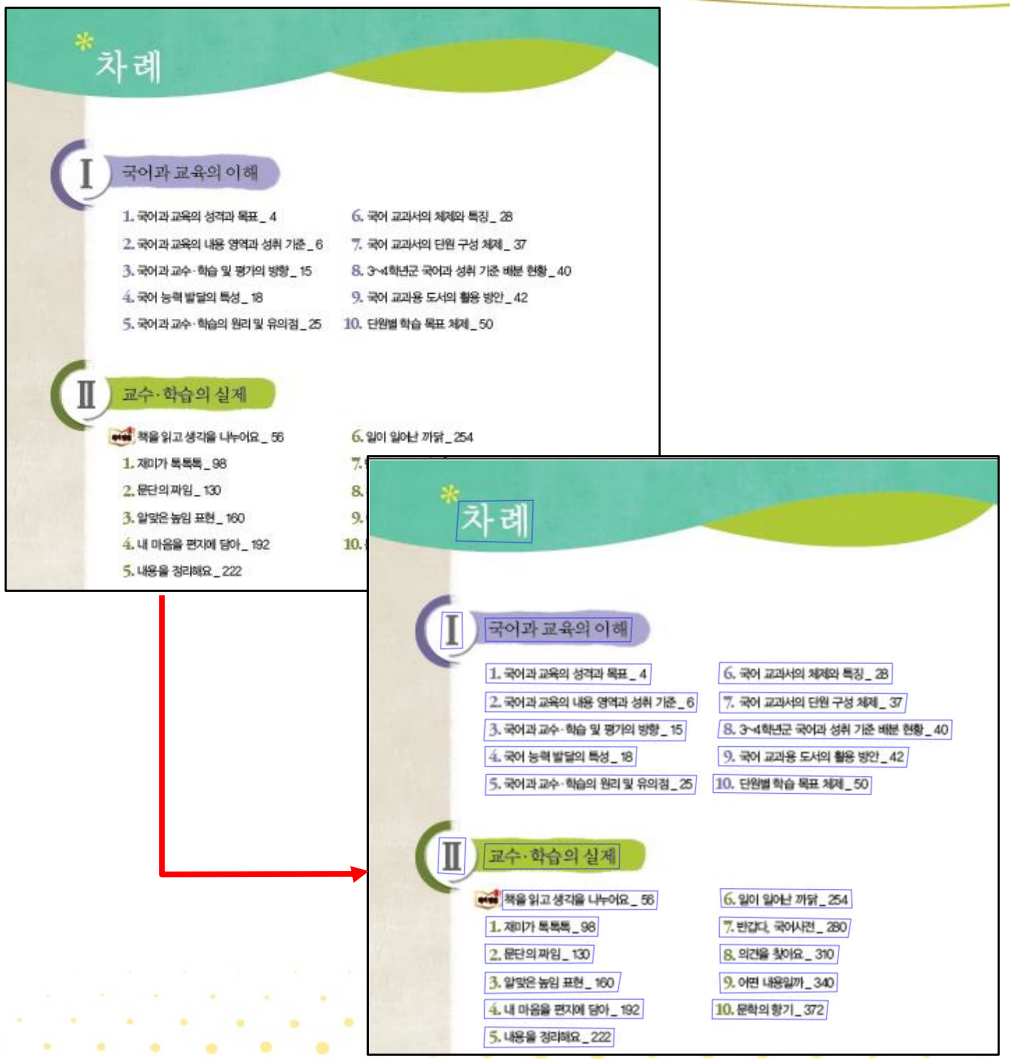


Linguistic Data Consortium

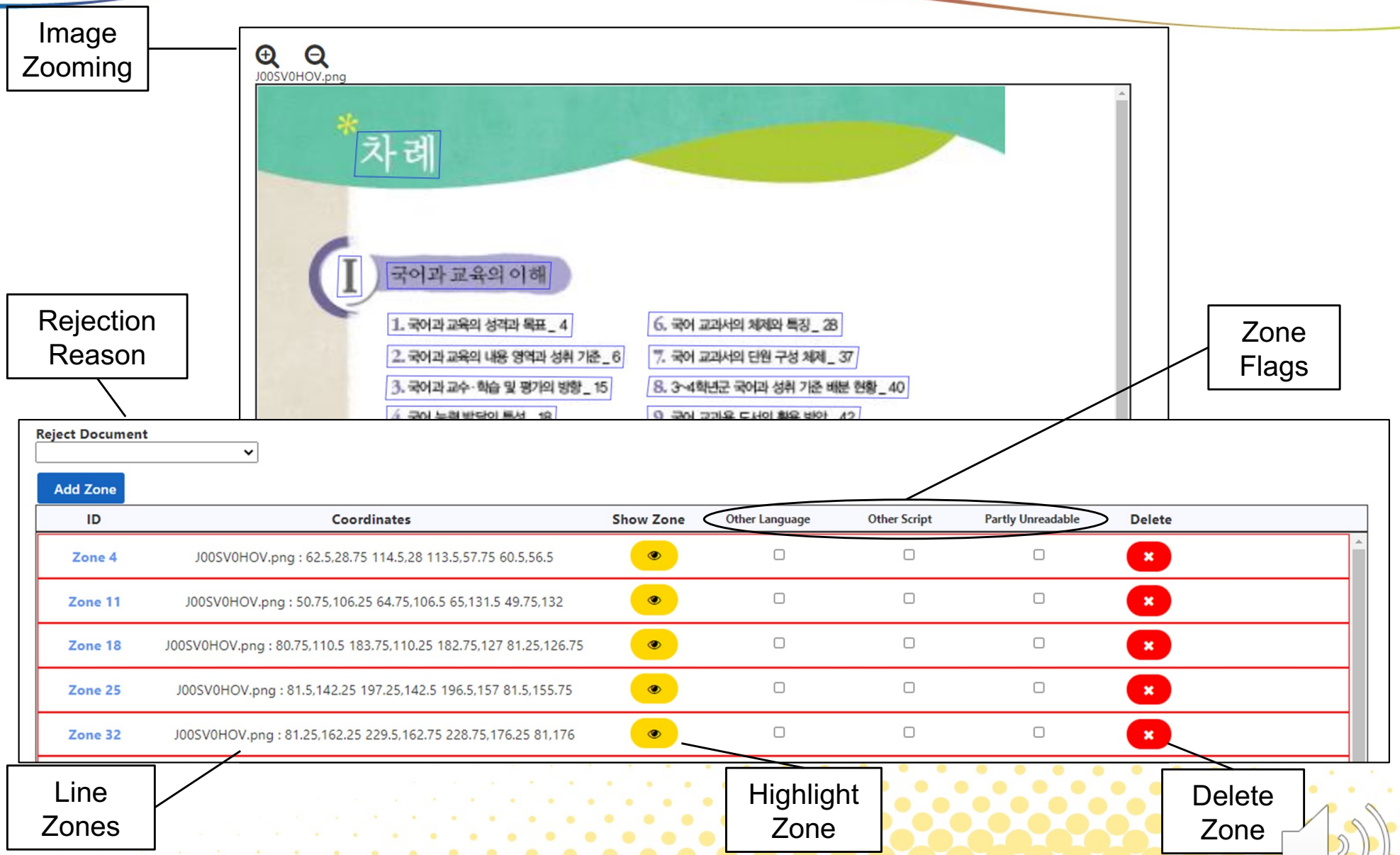
Annotation

Text Localization Task

- Draw 4-point bounding box around each line of machine printed text
- Specify attributes for each box, including the presence of non-target languages, other scripts, or illegible content
- Custom user interface, used for all stages of annotation

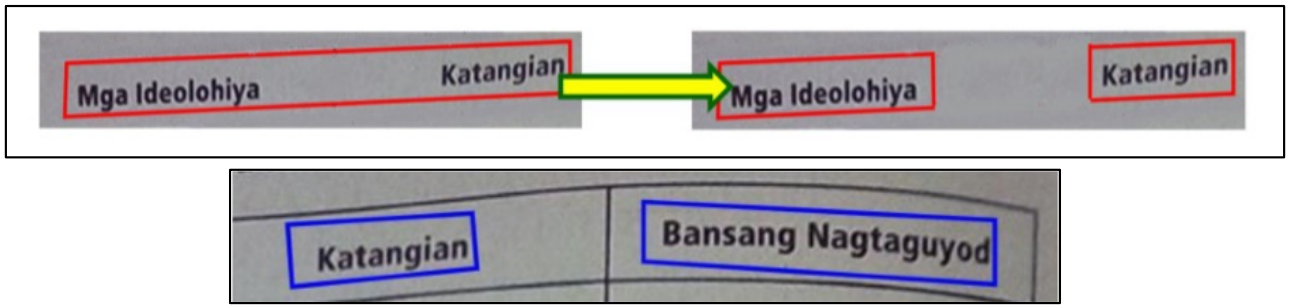


Text Localization GUI

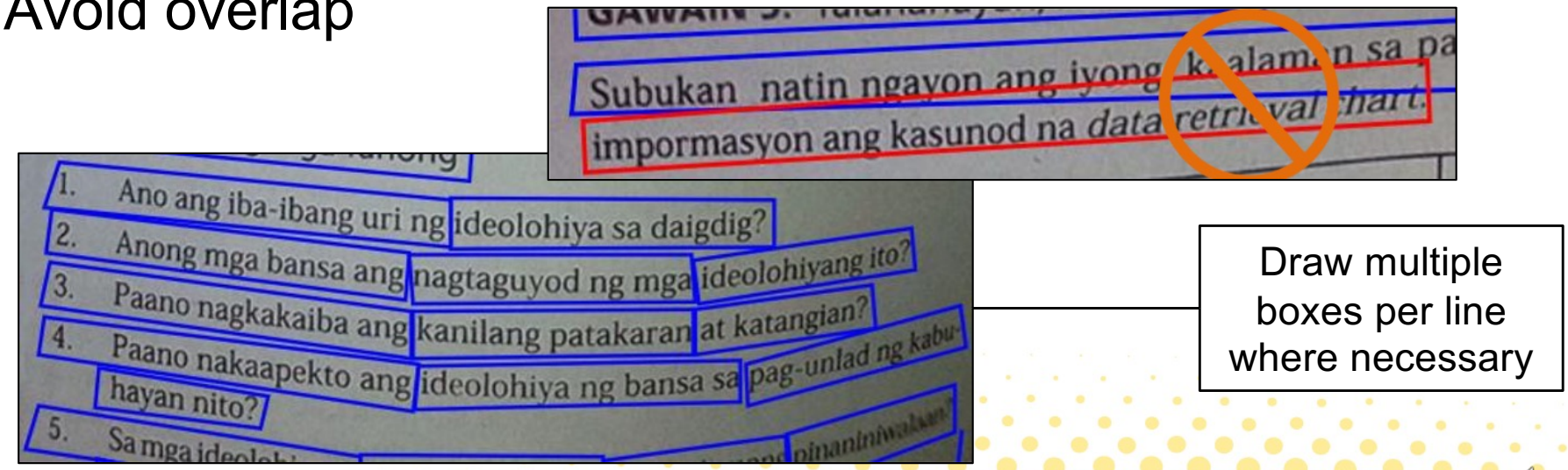


Text Localization Guidelines

- Avoid whitespace and non-textual boundaries



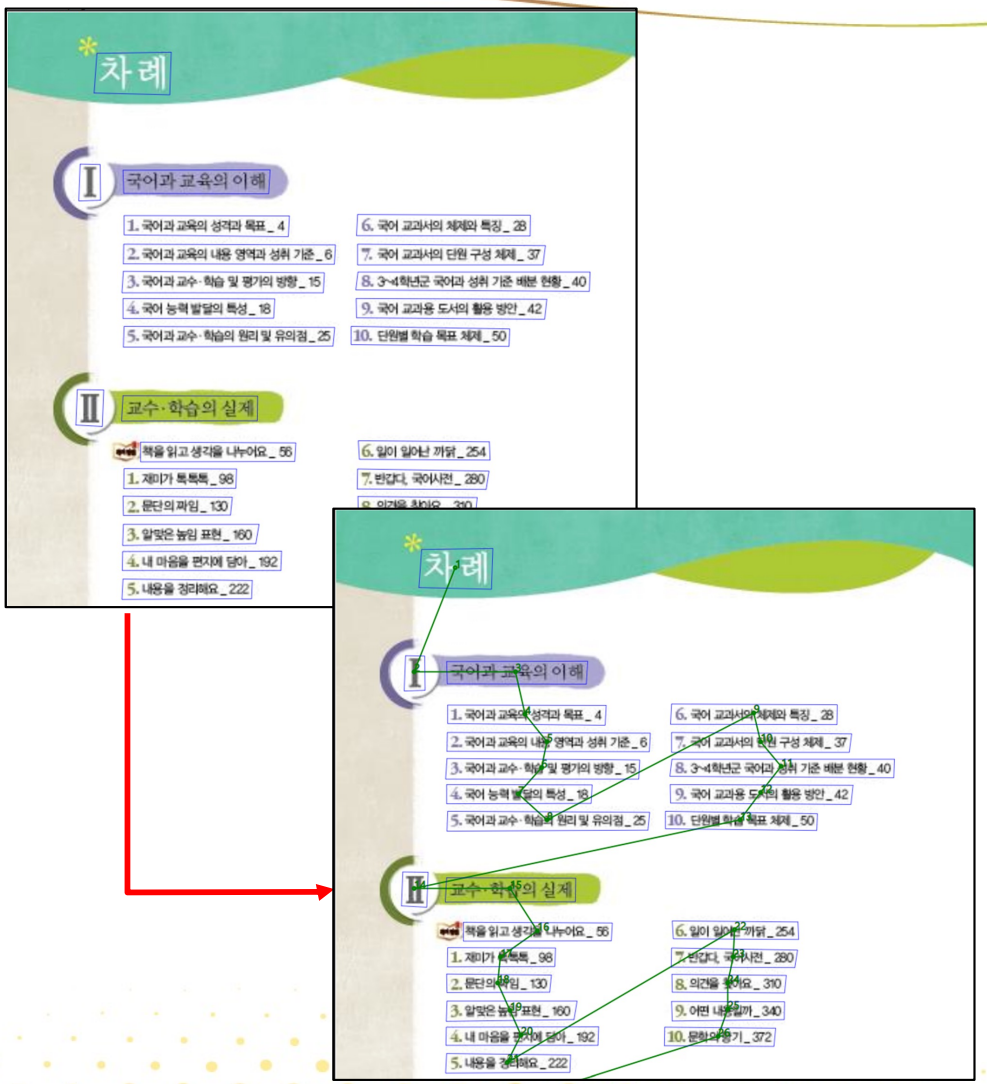
- Avoid overlap



Draw multiple boxes per line where necessary

Reading Order Task

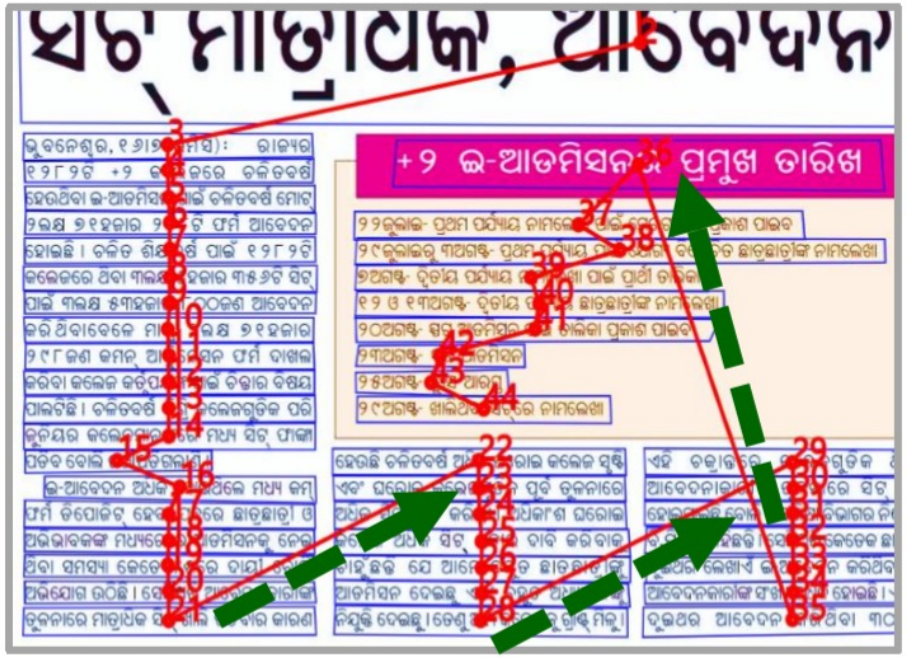
- Top text localization annotators do reading order annotation
- First verify and correct text localization annotation
- Then specify natural, logical reading order by applying *next_id* tag to each bounding box
  - Last line in doc has next_id = *NONE*



Reading Order Guidelines

- Do follow natural, logical flow of content within section
- Do not break flow of text by crossing between sections

Correct annotation



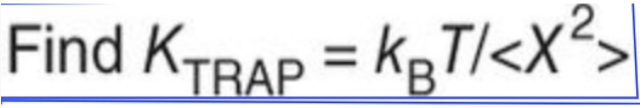
Incorrect annotation

Transcription Task

- For each of 13 transcription languages, 1250 images selected for transcription
- Native speaker annotators add orthographic transcript for each line of machine printed, bounding boxed text
  - Native orthography matching the image script
  - Retain capitalization and punctuation
  - Follow reading order
  - Transcribe only what's inside bounding box
  - Don't correct errors – transcribe as written
  - Add markup for special content and for orientation

Transcription Guidelines

Do not transcribe stylistic features


Find $K_{TRAP} = k_B T/<X^2>$

➡ Find KTRAP = kBT/<x2>

Transcribe only what's visible in each box

It's g

reater

Label the text orientation

Normal

Mirror

Upside Down

Vertical

Transcription Markup

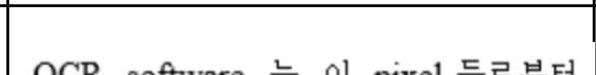
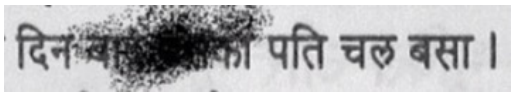
- Additional markup for special text features
- Detailed guidelines with examples in many languages for each category

Wrong Script

Unusual Characters

Need Context

Illegible

Lang	Image	Transcription
Korean		### ### 는 이 ### 들로부터
English		Paradiſe loſt. ***or*** [[Paradise]] [[lost]].
English		Took photos of the ((pagoda)).
Hindi		दिन +++ पति चल बसा ।

Annotator Management and Quality Control

- Over 200 staff and contract annotators on CAMIO, plus crowdworkers for data scouting
  - Dedicated team of project assistants to manage workers
- Pre-annotation QC
  - Crowdworker qualification test
  - Annotator aptitude screening, (in person or virtual) training and testing prior to work being assigned
  - Formal guidelines for each task
- Manual quality checks for all languages, tasks and workers (approx. 5%) plus spot checks to identify pervasive issues
- Pipeline design: each stage of annotation checks prior stage
  - Later stages required more skilled, experienced workers
- Additional automatic, manual sanity checks by external team

Linguistic Data Consortium

Challenges and Solutions

Team Management Challenges

- Very large annotation team
  - Increased need for quality control
  - Difficult to maintain consistent communication with contractors
  - Solution: Rely on more project assistants than initially planned to increase communication and boost productivity

- Scarceness of available native speakers and data for some languages
  - Recruitment took longer than expected
  - Solution: Hire additional Penn students who are native speakers

Data Challenges

- Text-heavy images
  - Decreased annotation rate for Text Localization and Reading Order
  - More annotations meant less data
  - Solutions
    - Improvements to annotation tool functionality and user-friendliness
    - Sequestering of very text-heavy documents
    - Additional collection to get appropriately sized images into pipeline

Thai

Linguistic Data Consortium

Results

Technology Baselines

- Produced baseline performance results for text localization and OCR decoding
- Using open-source Tesseract OCR engine for a subset of transcribed images
- 50/10/40 train/validation/test split
  - Nominally 625 train, 125 validation, and 500 test images per language

*Smith, R., Antonova, D., Lee, D. (Jul 25, 2009) Adapting the Tesseract open source OCR engine for multilingual OCR. Proceedings of the International Workshop on multilingual ocr, pp. 1-8

Text Localization Results

- Tesserocr python wrapper for Tesseract 4.0.0
- Comparing Tesserocr SINGLE_BLOCK output with ground truth
- Intersection over union (IoU) measures overlap between bounding boxes
- Overlap between Tesseract text boxes and ground truth annotation

CAMIO Test Set	F1 Score	Precision	Recall
Kannada	36.90%	30.40%	47.20%
Tamil	30.40%	26.20%	36.30%
Arabic	27.50%	22.60%	35.30%
Urdu	27.50%	22.90%	34.50%
Korean	27.00%	21.90%	35.40%
Hindi	25.20%	22.10%	29.40%
Farsi	23.80%	18.50%	33.40%
Russian	19.90%	18.20%	22.00%
Vietnamese	18.30%	14.00%	26.50%
English	17.30%	15.20%	20.10%
Thai	16.00%	15.00%	17.10%
Chinese (Simplified)	15.60%	13.00%	19.30%
Japanese	13.50%	11.30%	16.90%

Precision: ground truth boxes with IoU of at least 0.5 divided by total number of Tesseract boxes found

Recall: Tesseract boxes with IoU of at least 0.5 divided by number of ground truth boxes

F1: harmonic mean of precision and recall
2 Precision * Recall / (Precision + Recall)*

OCR Decoding Results

- Character Error Rate (CER) based on Levenshtein edit distance
- Minimum number of edits per line to modify system output to match the ground truth, divided by the number of characters in the ground truth
- Significant performance degradation possible with CER > 5%*

Normalization Procedure

1. Conversion of characters to lower-case

2. Removal of extra spaces (down to single spaces)

3. Removal of punctuation

4. Apply Unicode compatibility decomposition, followed by canonical composition (NFKC)

CAMIO Test Set	Character Error Rate (CER)
Urdu	68.20%
Japanese	39.80%
Chinese (Simplified)	34.90%
Tamil	24.50%
Arabic	24.00%
Farsi	23.00%
Korean	18.90%
Thai	18.10%
Vietnamese	17.00%
Hindi	16.40%
Russian	15.30%
Kannada	12.40%
English	12.10%

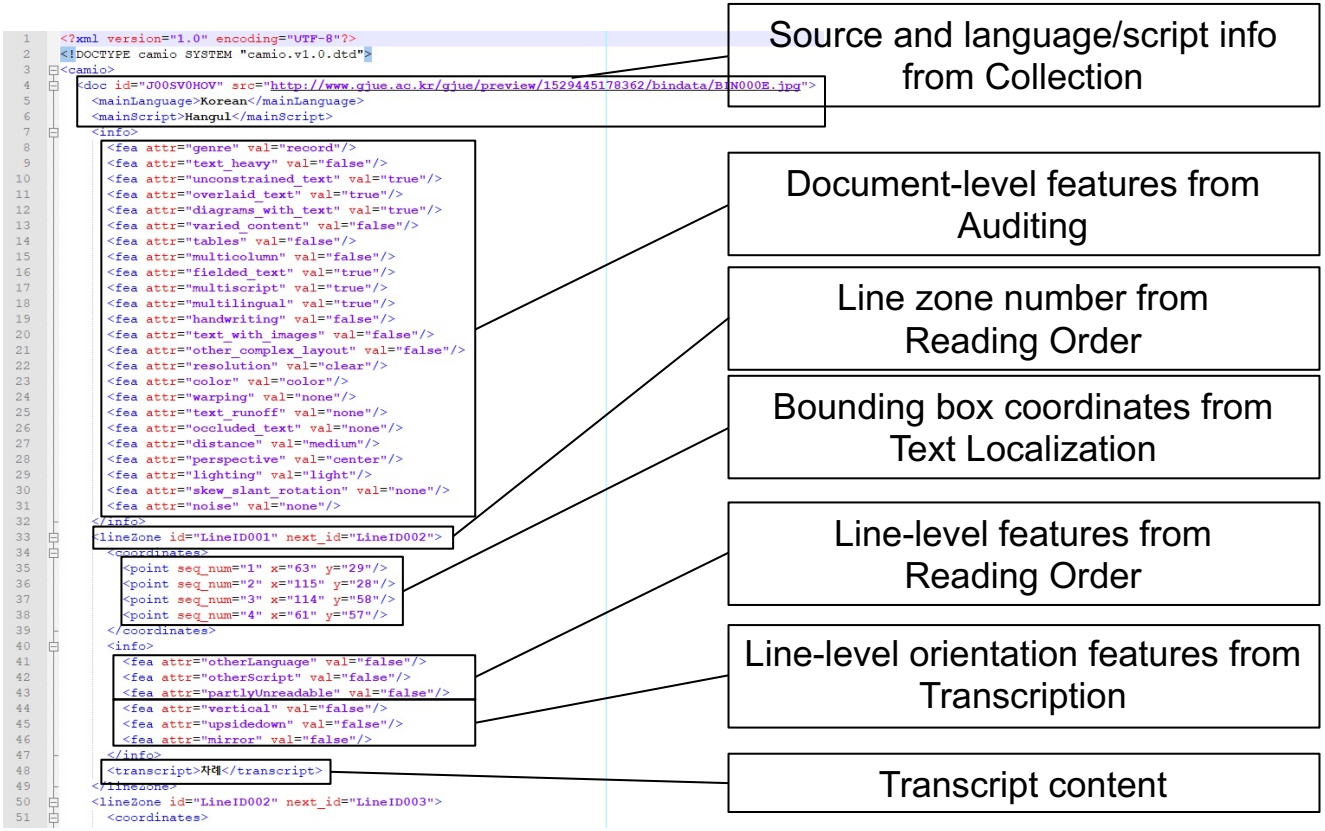
Likely due to stylized fonts like Nastaliq

Likely due to presence of vertical text

**Bazzo, G.T., Lorentz, G.A., Suarez Vargas, D., Moreira, V.P. (2020). Assessing the Impact of OCR Errors in Information Retrieval. In Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036.*

Corpus Release

- Image data in two versions: original (any format) and converted (PNG)
- Annotations in unified XML format
- Guidelines and documentation



Conclusion

- Corpus of Annotated Multilingual Images for OCR (CAMIO) is a new resource for computer vision and document analysis
- Poor baseline performance in text localization and OCR reflects challenging data that will stimulate future research
- CAMIO is expected to appear in LDC's catalog starting later this year
  - Part 1: Transcribed languages
  - Part 2: Untranscribed languages
  - Some data will remain unpublished for use in future evaluation campaigns
- More annotation planned including doc zoning, translation, transcription

CAMIO Corpus	Collection/ Annotation	Transcription
Languages	35	13
Total images	69,440	15,724
Boxed images	59,982	15,724
Boxes	2,339,358	311,619
Average boxes per image	~39	~20
Tokens	2,352,411	2,352,411