Linguistic Resources for OpenHaRT-13

Zhiyi Song[1], Stephanie Strassel[1], David Doermann[2], Amanda Morris[1]

1 – Linguistic Data Consortium

2 – Applied Media Analysis

Linguistic Data Consortium

Introduction to LDC

Model

- LDC is an open, international consortium hosted by the University of Pennsylvania
  - first organized at UPenn in 1992 via open, competitive DARPA solicitation
  - centralize distribution, archiving of language data; manage licenses & data distribution practices
- Business Model
  - developed by overseers from government, industry and academia
  - DARPA funding covered operations, corpus creation for 5 years
  - required to be self-sufficient via annual membership fees, data licenses
  - new grants fund LR creation, not maintenance; NSF, NIST early supporters
- Data Sources
  - donations, funded projects, community initiatives and LDC initiatives
- Membership
  - members provide annual support generally fees, sometimes data, services
  - receive ongoing rights to data published in years when they support LDC
  - reduced fees on older corpora, extra copies
  - access to LDC Online

Linguistic Data Consortium

Benefits

- Uniform licensing within & across research communities
  - 4 basic user license types, 1000s of instances
  - ~100 provider arrangements
  - no significant copyright issues in 21 years of operation
- Cost Sharing
  - relieves funding agencies of distribution costs, concerns
  - provides vast amounts of data to members
    - LDC annual membership benefit ~30 corpora
    - development cost for 1 corpus ≥ (LDC membership fee * 10 | 100 | 1000)
- Stable research infrastructure
  - LRs permanently accessible, across multiple platform changes
  - terms of use & distribution methods standardized & simple
  - members' access to data is ongoing
  - any patches available via same methods
  - tools, specifications, papers distributed without fee

Benchmarking

- Since inception in 1992, LDC has distributed
  - >84,000 copies
  - >1300 titles
  - >3168 organizations
  - >70 countries
- About half of the titles are e-corpora
  - developed for specific technology evaluation programs
  - published generally after use in the relevant communities
  - >4 years of publications “in queue”!!!
- 8309 academic papers relying on LDC Corpora
  - search for such papers is ~ 60% complete

LDC Roles

- Staff of 44 FT + 65 PT + several dozen ICs
  - supplemented by outsourcing, crowdsourcing and collaborations
- data distribution & archiving
- language resource production, including quality control
- intellectual property rights and license management
- human subject protocol management
- data collection and creation
- annotation and lexicon building
- creation of tools, specifications, best practices
- knowledge transfer: documentation, metadata, consulting, training
- corpus creation research (meta-research) and academic publication
- resource coordination in multisite research programs
- serving multiple research communities

Data Collection

- news text, journals, financial documents
- web text: newsgroups, blogs, discussion forums
- email, chat, SMS, tweets
- biomedical text & abstracts
- printed, handwritten & hybrid documents/images
- broadcast news & conversation, podcasts
- conversational telephone speech
- lectures, interviews, meetings, field interviews
- read & prompted speech
- task oriented speech, role play
- amateur video
- animal vocalizations
- …

Annotation

- data scouting, selection, triage
- audio-audio alignment; bandwidth, signal quality, language, dialect, program, speaker
- quick & careful transcription
- segmentation & alignment at story, turn, sentence, word level
- orthographic & phonetic script normalization
- phonetic, dialect, sociolinguistic feature & supralexical
- documenting zoning, handwriting transcription, OCR
- tokenization and tagging of morphology, part-of-speech, gloss
- syntactic, semantic, discourse function, disfluency, sense disambiguation
- fine and coarse-grained topic, relevance, novelty, entailment
- identification, classification of mentions in text of entities, relations, events, time, location & co-reference
- knowledge base population
- single & multi-document summarization of various lengths
- translation, multiple translation, edit distance, translation post-editing, quality control
- alignment of translated text at document, sentence, phrase & word levels
- physics of gesture
- identification, classification of entities and events in video

Program Services

- assess program needs: sponsors, developers, evaluators
- develop timelines for LR creation and system evaluation
- translate of “wish lists” into feasible action plan
- coordinate LR activities across & among programs
- maintain data matrices of LR features and availability
- maintain optimization, stabilization of data requirements
- incorporate technology into data production improving
- rapidly catalog, license, replicate, distribute program LRs
- broaden program impact through general distribution
- protection of restricted data

Programs

Program	Goal	NW	WB	SMS/Chat /Tweets	BN/BC	CTS	IV	Vid	OTHER
CALLHOME	STT					■			
CALLFRIEND	LR					■			
SWITCHBOARD	STT					■			
Mixer	SR					■	■	■	
LCTL	Translingual IR, MT	■	■			■			■
TDT	STT, MT, IR	■			■				
TIDES	STT, MT, IR, IE	■			■				
EARS	STT	■			■	■			
GALE	STT, MT, IR, IE, SUM	■	■		■				
BOLT	MT, IR	■	■	■		■			
MADCAT	OCR, MT	■	■						■
MR	NLU	■	■						
RATS	SAD, LID, SID, KWS				■	■			■
DEFT	Deep NLU	■	■	■	■	■			■
BEST	SR					■	■	■	
HAVIC	Video ED							■	
DOE Reading Enh.	Language Learning								■
DOE Dictionaries	Language Learning								■
LDC Online	Access	■				■			
Net-DC	Networking				■				
TalkBank	Networking					■		■	
Bio-IE	IE								■
SCOTUS	Access, Diarization								■
Digging into Data	Mining				■	■			■
PNG/BOLD	Fieldwork						■	■	

LDC Data in NIST Evaluations

	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13
LRE	✓							✓		✓		✓		✓		✓		
SRE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓		✓	
BN Re	✓	✓	✓	✓														
CTS Re		✓	✓		✓	✓												
RT							✓	✓	✓	✓	✓	✓		✓				
STD											✓							
SDR			✓	✓	✓													
TDT			✓	✓	✓	✓	✓	✓	✓									
ACE					✓	✓	✓	✓	✓	✓		✓	✓					
TAC KBP														✓	✓	✓	✓	✓
DUC						✓	✓	✓	✓	✓	✓	✓						
OpenMT						✓	✓	✓	✓	✓	✓		✓	✓			✓	
MetricsMaTr													✓		✓			
GALE Trans											✓	✓	✓	✓	✓	✓		
BOLT Trans, IR																	✓	✓
OpenHaRT															✓			✓
MADCAT													✓	✓	✓	✓	✓	✓
TRECVid SED													✓	✓	✓	✓	✓	
TRECVid MED															✓	✓	✓	✓
TRECVid MER																	✓	✓

Linguistic Data Consortium

Linguistic Resources for OpenHaRT 2013

LDC Data in Document Image Recognition Programs

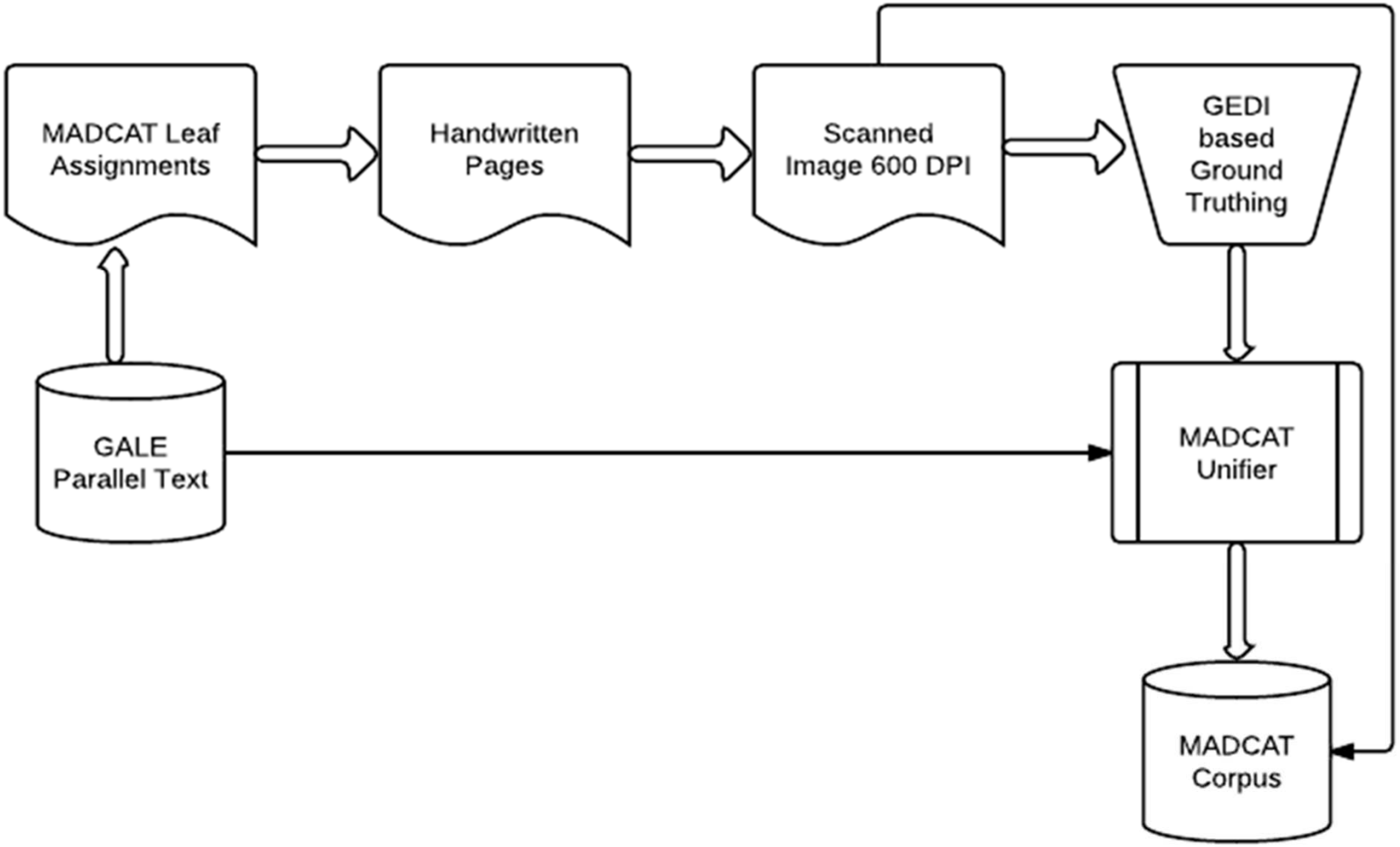
- DARPA MADCAT (Multilingual Automatic Document Classification, Analysis and Translation), 2008-2013
  - Goal: automatically convert foreign text images into English transcripts
  - LDC supports MADCAT by
    - collecting handwritten documents in Arabic and Chinese
    - scanning texts at a high resolution
    - annotating the physical coordinates of each line and token
    - transcribing and translating the content into English
    - post-editing machine translation system output during annual evaluations conducted by NIST to evaluate the MADCAT technologies
- OpenHaRT 2010, 2013
  - Goal: evaluation of transcription and translation technologies for document images containing primarily Arabic script.
  - LDC supports OpenHaRT by
    - providing training, devtest and eval data drawn from MADCAT Phase 1 to Phase 3
    - collecting additional handwritten data for possible future evaluations

Data for OpenHaRT 2013

- OpenHaRT 2013 training, eval data produced in MADCAT as follows:
  - LDC processed NW, WB parallel text from DARPA GALE to extract “leaf” pages for manual re-transcription by literate native speakers (scribes)
  - LDC conducted manual QC on scribe assignments prior to scanning & ground truthing
  - AMA scanned collected pages at 600-dpi, greyscale
  - AMA completed ground truthing annotation, aligning source tokens to polygons
  - LDC produced unified xml output containing text layer, image layer, scribe demographic layer, document metadata layer

Linguistic Data Consortium

Pipeline



Scribe Collection

- LDC recruits Arabic speakers in Philadelphia; supplemented by international partners to ensure regional variety
- Scribes register in person at LDC or remote collection site and take writing test to assess literacy and eligibility
- Demographic information collected
  - Name, address (for compensation purposes only)
  - Age, gender, level of education, occupation
  - Where born, where raised
  - Primary language of educational instruction
  - Handedness
- After registration scribes complete brief tutorial emphasizing need to copy text exactly and follow line/page breaks
  - No omissions or insertions, don't correct "wrong" source text

Scribe Demographics

	phase1_train	phase2_train	phase3_train	phase3_eval
Participant	72	230	53	24
Countries	15	10	10	8
Male	27	95	22	14
Female	45	133	30	10
Right hand	70	227	49	21
Left hand	1	2	3	3

Data Preparation

- Starting with sentence-segmented source documents, create scribe assignment or “kit”
  - **Step 1**: tokenize text, execute word and line wraps, paginate source text into kit pages
    - Maximum of 20 lines/page, 5 words/line to avoid page breaks or line wraps
  - **Step 2**: manually review generated pages for content and formatting
    - Political views in weblogs upsetting to some scribes
  - **Step 3**: generate alternate kits given a set of MADCAT pages and preselected kit parameters
    - Arabic: 50 pages/ kit, 2-7 versions of the same kit

Scribe Assignments

- Assignments are in the form of printed "kits"
  - 50 printed pages to be copied plus assignment table that specifies page order and writing conditions
    - Multiple scribes/kit, so conditions and order vary
  - Printed pages labeled with page and kit ID
  - Scribes affix label with scribe ID, page ID and kit ID to back of completed manuscript
    - To facilitate data tracking during scanning and post-processing
- Scribes supply their own paper and writing instrument to sample natural variation
- Quality control on submitted pages
  - Exhaustive check on first assignment
  - Spot check on remainder of assignments
  - Compensation only after kits are verified
- LDC Scribble tool supports scribe recruitment, testing, enrollment, data assignment and tracking, progress reporting, quality control and compensation both at LDC and remote collection sites

Writing Conditions

- Excepting Phase1_Train, targeted writing conditions were
  - 90% pen, 10% pencil
  - 75% unlined paper, 25% lined paper
  - 90% normal speed, 5% careful speed and 5% fast speed

Writing Condition		phase1_train	phase2_train	phase3_train	phase3_eval
utensil	pencil	3357	2773	455	66
	pen	6336	25041	4085	567
paper	line	2468	7163	1181	167
	unline	7225	20651	3359	466
speed	normal	8667	24578	3995	559
	fast	520	1611	272	39
	careful	507	1626	274	36

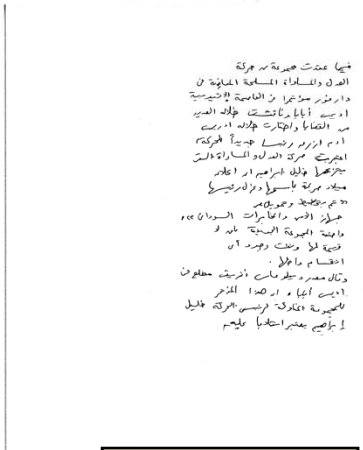
Scribe Writing Assignment Example

AAW_ARB_20070102.0045_1

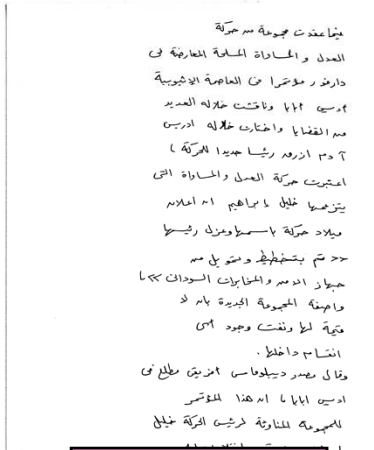
لا تنسخ أي شيء مكتوب فوق الخط

<1> فيما عقدت مجموعة من حركة
<2> العدل والمساواة المسلحة المعارضة في
<3> دارفور مؤتمرا في العاصمة الإثيوبية
<4> أديس أبابا وناقشت خلاله العديد
<5> من القضايا واختارت خلاله ادريس
<6> آدم ازرق رئيسا جديداً للحركة ،
<7> اعتبرت حركة العدل والمساواة التي
<8> يتزعمها خليل ابراهيم ان اعلان
<9> ميلاد حركة باسمها وعزل رئيسها
<10> « تم بتخطيط وتمويل من
<11> جهاز الأمن والمخابرات السوداني » ،
<12> واصفة المجموعة الجديدة بان لا
<13> قيمة لها ونفت وجود أي
<14> انقسام داخلها .
<15> وقال مصدر ديبلوماسي افريقي مطلع في
<16> اديس ابابا ، ان هذا المؤتمر
<17> للمجموعة المناوئة لرئيس الحركة خليل
<18> إبراهيم يعتبر انقلابا عليه .

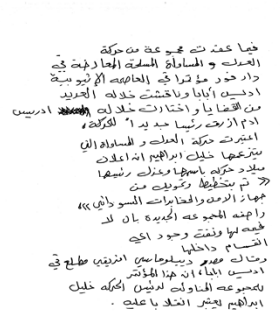
Assigned document



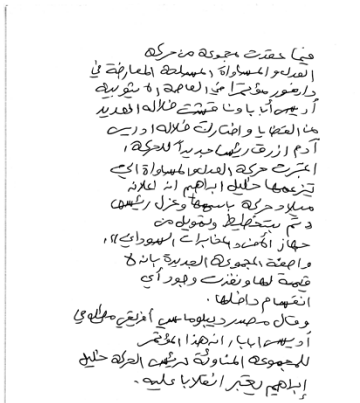
Scribe 1



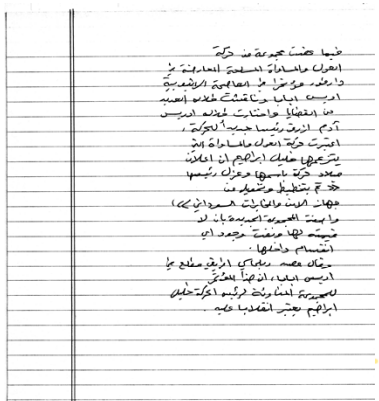
Scribe 2



Scribe 5

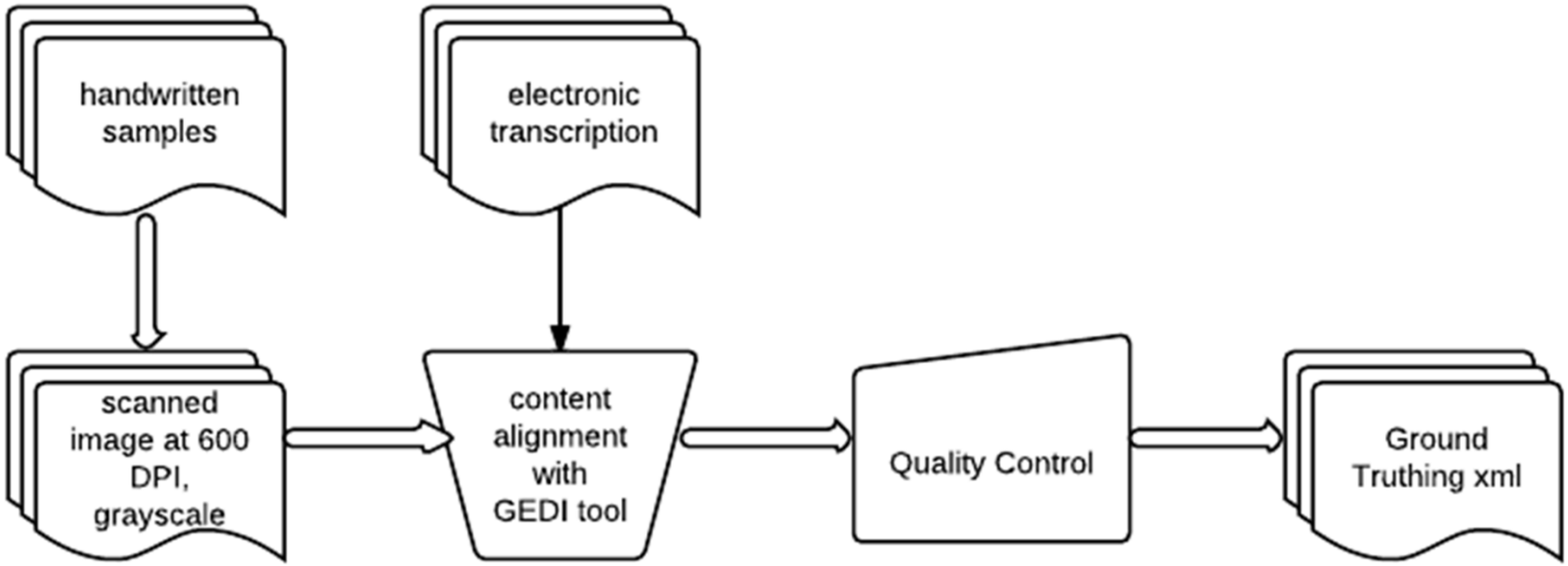


Scribe 3



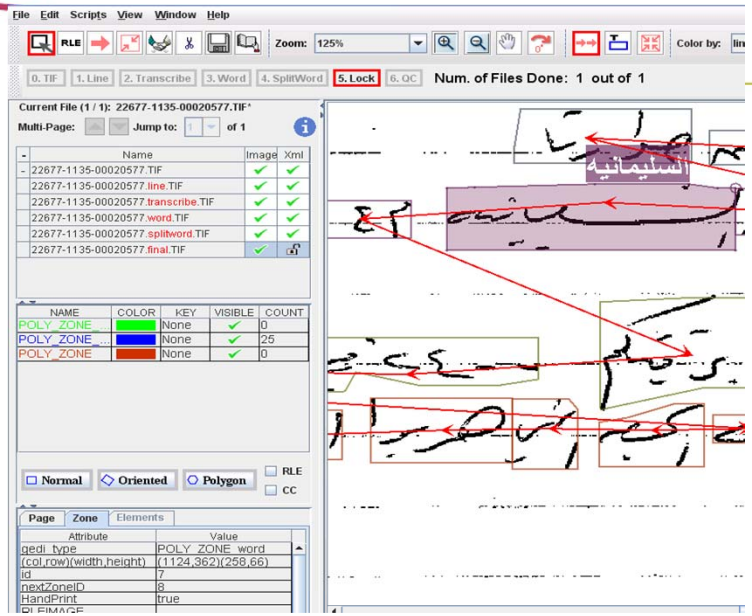
Scribe 4

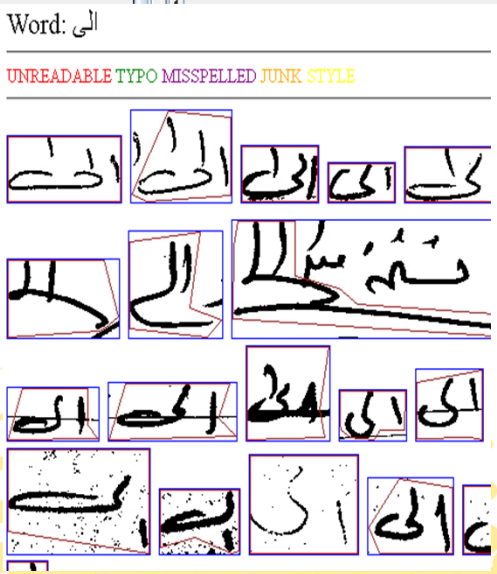
Annotation Pipeline



GEDI Tool

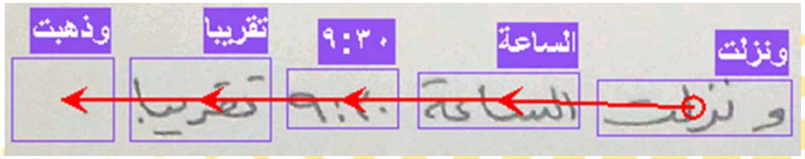
- Developed by Language and Media Processing Laboratory at the University of Maryland
  - A generic annotation tool that assists in ground truthing scanned text documents
  - Enforces constraints on reading order, text alignment and consistency of token attributes
  - Provides mechanisms for other QC procedures





Annotation Procedure

- Use GEDI to draw polygon bounding box around each line, word/character token with unique line and token ID assigned
- Aligning source text and image token
- Each token's page coordinates are recorded as the "ground truth"
- Reading order automatically added (Chinese L>R, Arabic R>L)
- Each token is reviewed, additional features are added to indicate status of extra token, typo, etc.
- Missing tokens in handwritten image are aligned with empty boxes in GEDI



Release Preparation

- Major challenge is logical storage of many layers of information across multiple versions of the same data
- Defined MADCAT-Unifier Process
  - Input: multiple disparate data streams
  - Output: single xml file integrating all information layers
    - Text layer containing source text, tokenization, sentence segmentation, translation
    - Image layer containing zone bounding boxes
    - Scribe demographics
    - Document metadata

MADCAT Data Summary

	Training			Eval
	Phase 1	Phase 2	Phase 3	Phase 3
Genre	Newswire & Web Text			
Origin of source docs	GALE Phase 1-4 Parallel Text Training			GALE Phase 4 Eval
Number of unique pages	2334	5773	635	211
Arabic tokens/page	unconstrained	unconstrained	<=125	<=125
Scribes/page	Up to 5	up to 5	up to 7	3
Total handwritten pages	9693	27,814	4540	633
Total tokens	758,936	2,964,266	507,689	74,553
Number of unique scribes	72	230	53	24
Scribe exposure	all exposed	all exposed	all exposed	half unexposed

Linguistic Data Consortium

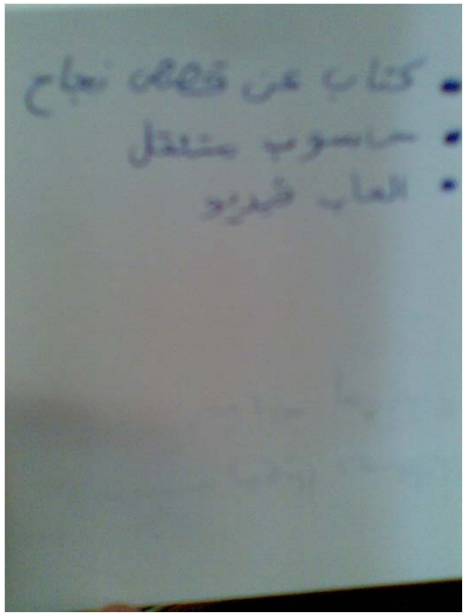
Current Activities and Conclusion

Chinese Collection

- Following MADCAT protocol
  - Data from GALE NW, WB parallel text with existing Treebank and/or word alignment annotation
- Pilot collection in 2011
  - 150 unique scribes
  - 15 scribes per page
  - 223,600 pages collected, annotated
- Additional collection in 2013
  - 40 scribes, half exposed and half non-exposed
  - >200 unique pages
  - 5 scribes per page
  - Collection complete
  - Ground truthing underway now

Arabic Collection via Crowdsourcing

- Crowd-sourced collection using Amazon Mechanical Turk
  - Pilot effort in 2011
  - Additional collection currently underway
- MTurk HITs posted in multiple Arabic-speaking countries
  - Write on specified topic; take photo; upload image
    - Topics include things like “recipe”, “shopping list”
    - No constraints on writing material or instrument
    - Image quality, resolution, orientation, etc. highly variable
- Each uploaded image is verified as handwritten, Arabic, on topic
  - Over half rejected
- Verified images transcribed, ground truthed, translated
- Collection, annotation and translation in progress



Conclusion

- To support OpenHaRT-13, LDC distributed 5 corpora originally created under MADCAT to 7 evaluation participants
- All MADCAT Arabic training data has been published in LDC's general catalog making it available to the larger research community
  - LDC2012T15 MADCAT Phase 1 Training Set
  - LDC2013T09 MADCAT Phase 2 Training Set
  - LDC2013T15 MADCAT Phase 3 Training Set
- We are completing production of additional Chinese MADCAT-type data and new collection/annotation of Arabic via crowdsourcing
- MADCAT evaluation data, Chinese data and Arabic crowdsourced collection will be added to LDC catalog after evaluation blackout period expires