New Approaches for Distributed Non-Expert Annotation and Collection at LDC

Stephanie Strassel (presenter), Ann Bies, Kira Griffitt, Xuansong Li, Justin Mott, Ann Sawyer, Zhiyi Song, Jennifer Tracey, Jonathan Wright

Linguistic Data Consortium

Introduction

Context for LDC Projects

- Linguistic resources for sponsored lang technology programs
- Multiple concurrent projects, tasks and languages
  - Amharic, Arabic, Bengali, English, Farsi, Hausa, Hindi, Hungarian, Indonesian, Mandarin, Russian, Somali, Spanish, Swahili, Tagalog, Tamil, Thai, Turkish, Twi, Uyghur, Uzbek, Vietnamese, Wolof, Yoruba, Zulu
  - Collection of conversational telephone speech, broadcast news and talk shows, amateur video, news text, blogs, Twitter, SMS/chat
  - Annotation ranging from simple
    - Orthographic Transcription, Language ID, Translation Quality Control
  - To complex
    - NP Chunking, Entities and Coref, Semantic Role Labeling, Belief/Sentiment
  - To expert
    - Syntactic Annotation (Treebanking), Lexicon Creation

The Workers

- Traditional model: hire staff and pay an hourly wage
  - Local (on site at least occasionally) with some lx expertise
  - Intensive, lengthy hands-on training
  - Long-term commitment in both directions
- Scope, number, complexity of projects has steadily increased
- Emphasis has shifted toward lower resource languages
  - Short lead time on new languages or tasks
- Need more people than ever, but it's harder to find them
  - Dozens or even hundreds of short-term workers at a time
- Remote, distributed work by non-experts is essential

Linguistic Data Consortium

Task Engineering

Decision Points

- What questions must be answered to complete the task?
- In what order should they be answered?
- What are the dependences among them?

- Decision Points inform all aspects of task design
  - Training approach including guidelines
  - Workflow
  - User interface design
  - Annotator testing and quality control

Relevance Assessment Example

- BOLT Information Retrieval Query Assessment
  - Given a natural language English query, BOLT system automatically processes multi-lingual informal web text corpora and returns answers in English

Query: What are the influences of the Euro financial crisis on China?

System response: Due to the spread of the European debt crisis has intensified, the us economic recovery is sluggish, further deterioration of the external environment in the development of the Asian economies.

- Poor annotator consistency given a single, simple question "Is the response relevant?"
  - Impact of (often poor) automatic translation
  - Varying degrees of leniency

Unpacking Decision Points

- Is the response relevant? → Up to five decision points

Q1: Can you answer relevance questions based on this English response alone, without consulting the original source text?

YES.

NO, because the translation is incomprehensible.

NO, because I need to see the source text to resolve pronouns, get more context and/or clarify the translation.

Q3A: Is the response informative?

YES, it adds information beyond restating the query.

NO, it does not add new information.

Q4: Does the response contain any relevant information?

YES, it addresses the query.

NO, it does not address the query.

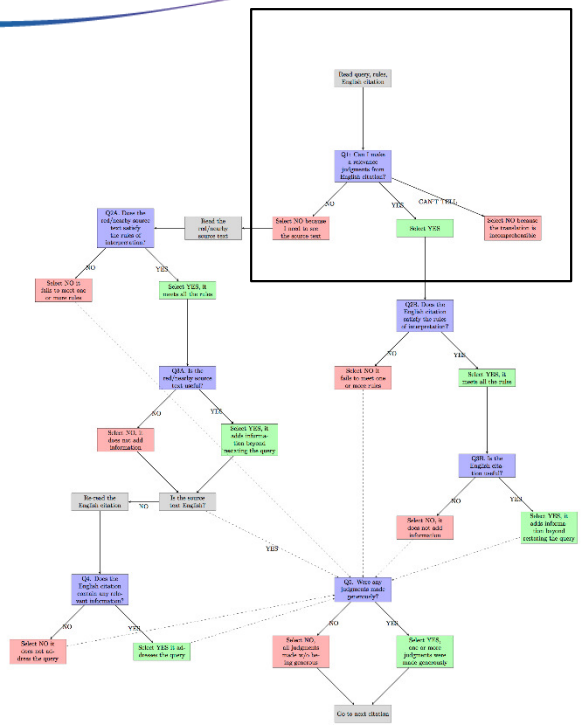
Q5: Were any of your judgments made generously?

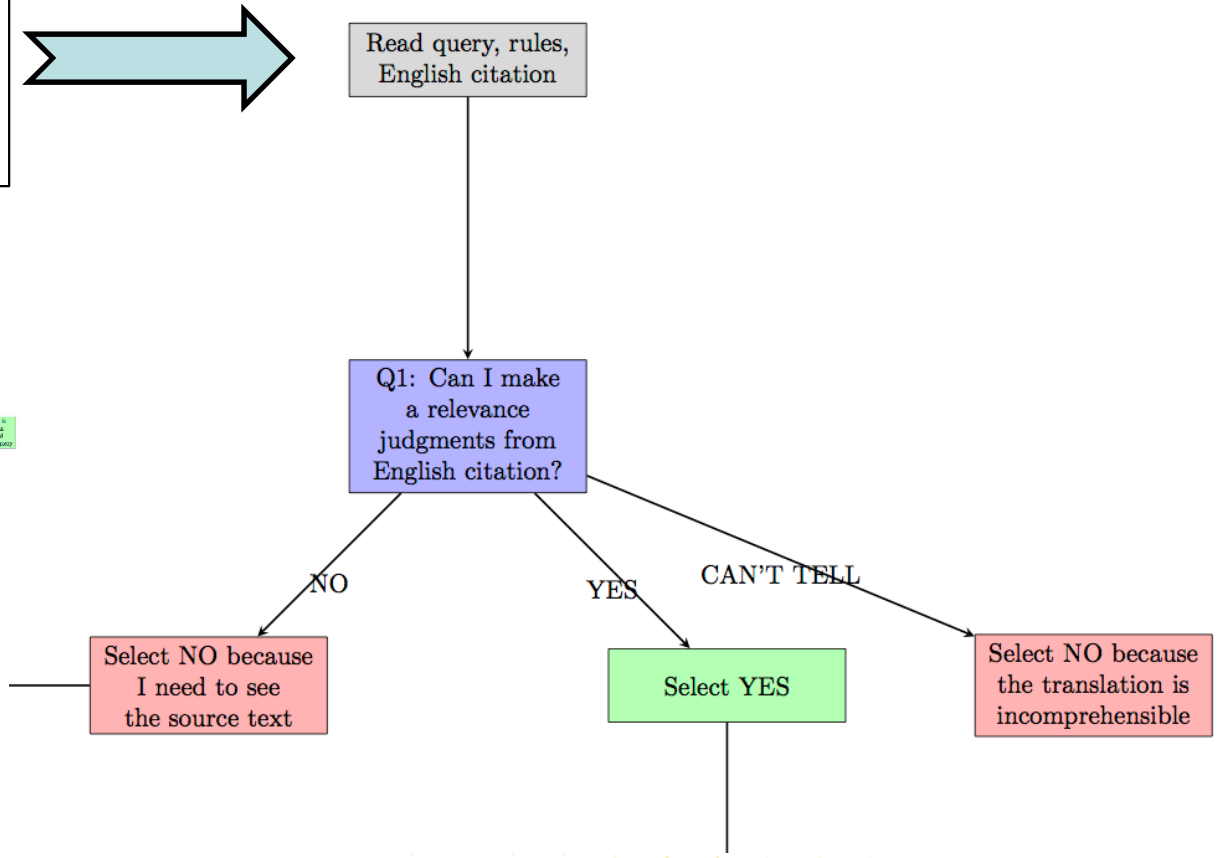
YES, one or more of the judgments was made generously.

NO, all of the judgments were made without being generous.

- Significant increase in annotator consistency using decision points model
- No significant decrease in annotator efficiency

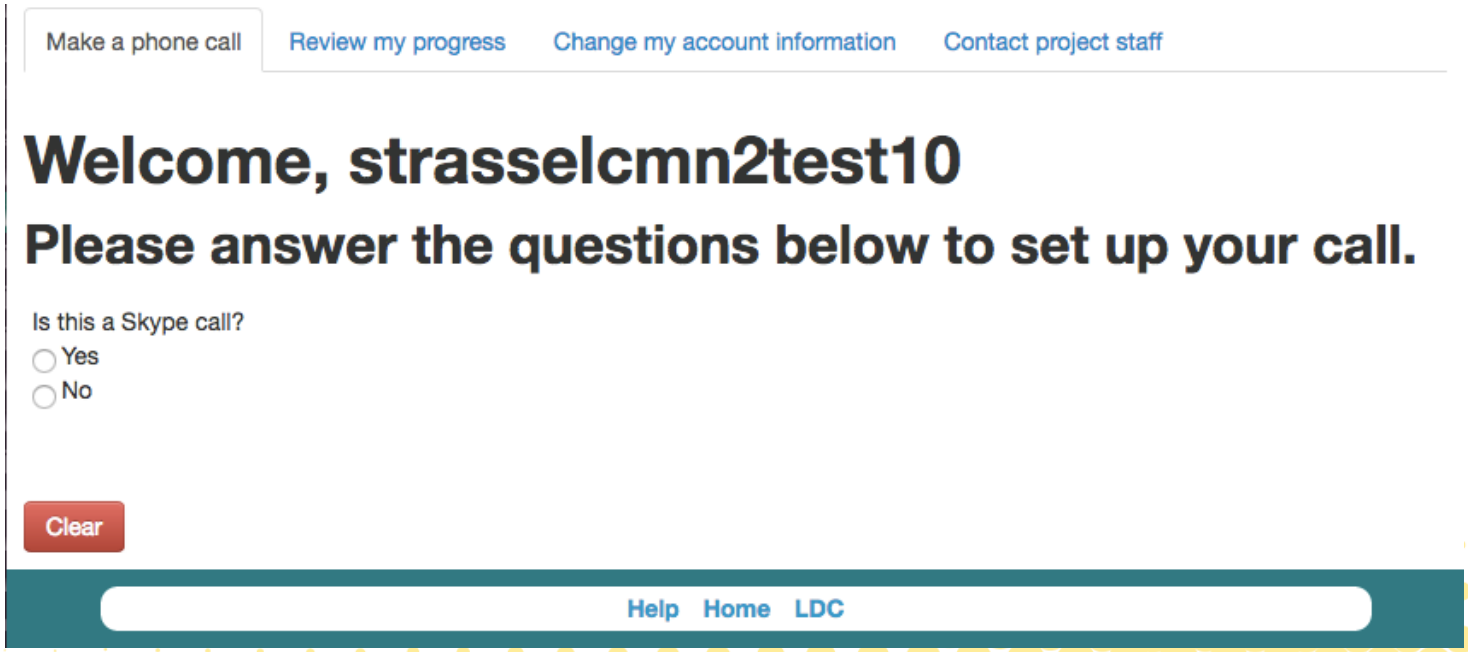
Decision Trees





GUI and Decision Points

- GUI design closely follows decision points
- Dynamic display and skip logic to walk worker through the decision making process
- Prevents illogical judgments, avoids wasting time on irrelevant decisions



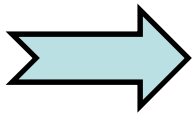
GUI Design Factors

- GUI design has a major impact on training time, efficiency, accuracy, even staff retention
- Consider user's computing background
- Are there assumptions about browser and OS requirements
- What are the network requirements
- Screen real estate
  - May be working on laptop or tablet
- Avoid wasted movements
  - Unnecessary scrolling
  - Mouse vs. keyboard
- Use standard browser shortcuts
- Constrain behavior (decision points)
- Constrain possible answers based on prior decisions
- Provide informative error reports/validation

Workflow Considerations

- Traditional: One annotator makes all task decisions
- Decision points present opportunity to distribute work across team of differently skilled workers

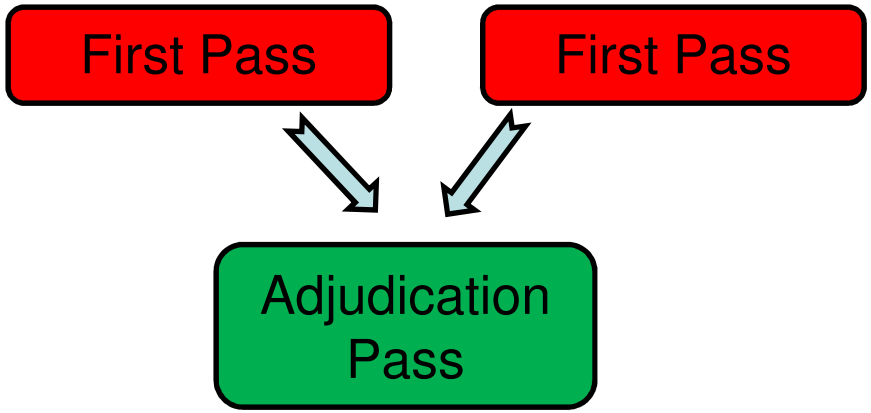
Full Entity Task	
✓	Is there an entity in this passage?
✓	Is it a specific person, place, organization, location or facility?
✓	Is it a name, nominal or pronoun?
✓✗	What is its extent (full NP)?
✓✗	If nom, what is its head?
✓	What is its type?
✓	Can it be linked to another entity?

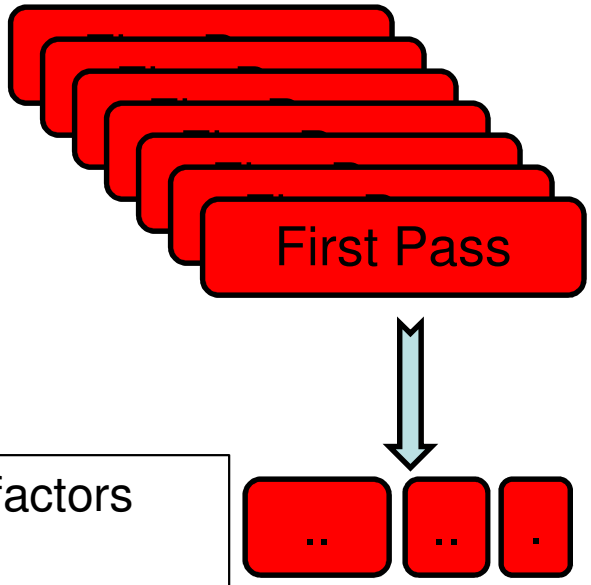


Entity Annotator
Is there an entity in this passage?
Is it a specific person, place, organization, location or facility?
Is it a name, nominal or pronoun?
Intuitive extent
What is its type?
Can it be linked to another entity?

NP Chunking Annotator
Label the full extent and head of this nominal mention

Many Possible Workflows





Workflow (and deliverable product) depends on many factors

- Cost, quality, timeline, data volume tradeoffs
- How will data be used?
- What is human consistency baseline?

Linguistic Data Consortium

Training

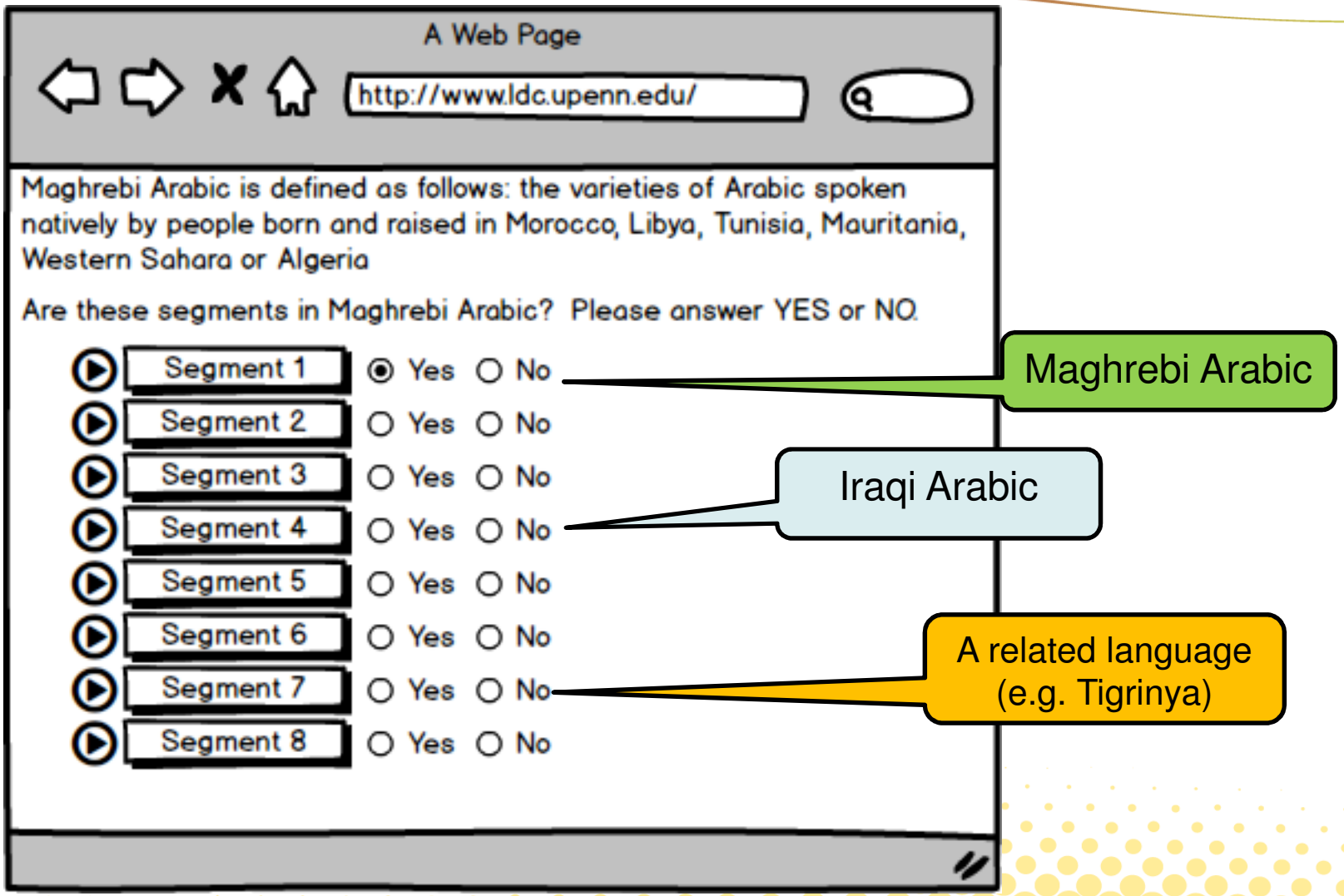
LORELEI Example

- Goal: dramatically advance the state of computational linguistics and human language technology to enable rapid, low-cost development of capabilities for low-resource languages
  - Improved situational awareness based on information from **any language**, in support of emergent missions such as humanitarian assistance/disaster relief, peacekeeping or infectious disease response
- LDC is building data packs for approximately 3 dozen LRLs
  - E.g. Amharic, Hausa, Somali, Uyghur, Zulu
- Remote native speakers collect relevant data (news, informal web text, Twitter etc) and perform linguistic annotation ranging from simple to very challenging
  - Named entities, full entity with coref, NP chunking, POS labeling, semantic annotation, etc.

Applicant Screening

- Applicants directed to online survey to assess …
- Eligibility for position
  - Paperwork, schedule, access to computer & internet
- Language use
  - How often do you …
    - Read, speak, Google, use social media etc
  - Did you grow up speaking ...
    - With family? In school? With friends?
- Survey is automatically "scored" and personalized (stock) reply sent
  - Standards vary depending on language, applicant pool
- Promising applicants automatically invited for aptitude test and/or Skype interview

Aptitude Testing



Self Directed Preliminary Training

- Primary goals
  - Introduce core concepts and terminology
  - Assess aptitude for specific linguistic annotation tasks
  - Give annotator a taste of the work they'll be asked to do
- Multiple units, multiple short modules per unit
  - No more than 1-2 hours per unit
- Logical progression based on task ordering, module complexity
- Narrated presentations for each module, followed by a quiz
  - Must complete the unit within 24 (48, 72) hours after starting
  - Must achieve acceptable score before moving on
  - Two strikes and you're out (for all tasks that build on this module)
- Can scale to effectively unlimited number of candidates
  - Labor-intensive advanced training limited to suitable workers

LORELEI : Ling Basics - Names

LDC Workshop - Training Demo Module III: Linguistics Branch 1

Chapter 18: Name

The next video explains what the term **name** means for us - what kinds of things are the types of names we care about for annotation.

Please watch the video below (as many times as you want to) then click *Next* to answer questions.

Prev Next

Identify Decision Points

- What questions must be answered to complete the task?
- In what order should they be answered?
- What are the dependences among them?

Data Scouting
Is the document suitable for collection?
Is the document relevant to the domain?
What is the genre?
What is the type of incident?
What is the name of the incident?
Is there a Wikipedia page about the disaster?
Does the document contain any eyewitness account(s) of the incident or its aftermath, impact, recovery efforts, etc.?
How specific/local is the information about the incident?
What topics are discussed?

Simple NE
Is there a name in this (sentence, paragraph, ...)?
Is it the name of a specific person, place, organization or facility?
What is its extent?
What is its type?

Simple Semantic Annotation
Does the passage describe an Act or State?
Does the passage describe a disaster-relevant Act?
Select the most important word or phrase that expresses the disaster-relevant Act
Select the word or phrase that corresponds to the Agent(s) of this Act
Select the word or phrase that corresponds to the Patient(s) of this Act
etc.

Identify Core Concepts

- What core concepts must be mastered in order to successfully navigate each decision point?

Data Scouting	Simple NE	Full Entity	NP Chunking	SSA
How much of the document needs to be in the target language?	What is a name	What is an entity	Key Concepts	What counts as an Act?
How to locate dates/times within text	What is an entity	What are specific entities	Tagging and taggability (general)	What counts as a physical Act?
How to locate names of disasters within text	What is name extent	What is an entity mention	Extents (general)	What counts as a State?
How to locate names of organizations within text	Entity typing	what is entity mention extent	Grammaticality	What counts as disaster-relevant State?
How to locate names of people within text	PER entity type	Name extents	Part of Speech (general)	What is a basic word?
How to locate names of places within text	ORG entity type	Nominal extents	Part of Speech: Noun	What is a phrase?
How to search for the web using relevant keywords	GPE entity type	Nominal heads	Part of Speech: Pronoun	What is an Agent?
Etc.				

Identify Shared Concepts

- Is the same core concept relevant to multiple tasks?

Data Scouting	Simple NE	Full Entity	NP Chunking	SSA
How much of the document needs to be in the target language?	What is a name	What is an entity	Key Concepts	What counts as an Act?
How to locate dates/times within text	What is an entity	What are specific entities	Tagging and taggability (general)	What counts as a physical Act?
How to locate names of disasters within text	What is name extent	What is an entity mention	Extents (general)	What counts as a State?
How to locate names of organizations within text	Entity typing	What is entity mention extent	Grammaticality	What counts as disaster-relevant State?
How to locate names of people within text	PER entity type	Name extents	Part of Speech (general)	What is a basic word?
How to locate names of places within text	ORG entity type	Nominal extents	Part of Speech: Noun	What is a phrase?
How to search for the web using relevant keywords	GPE entity type	Nominal heads	Part of Speech: Pronoun	What is an Agent?
Etc.				

Identify Shared Concepts

- Is the same core concept relevant to multiple tasks?

Data Scouting	Simple NE	Full Entity	NP Chunking	SSA
How much of the document needs to be in the target language?	What is a name	What is an entity	Key Concepts	What counts as an Act?
How to locate dates/times within text	What is an entity	What are specific entities	Tagging and taggability (general)	What counts as a physical Act?
How to locate names of disasters within text	What is name extent	What is an entity mention	Extents (general)	What counts as a State?
How to locate names of organizations within text	Entity typing	What is entity mention extent	Grammaticality	What counts as disaster-relevant State?
How to locate names of people within text	PER entity type	Name extents	Part of Speech (general)	What is a basic word?
How to locate names of places within text	ORG entity type	Nominal extents	Part of Speech: Noun	What is a phrase?
How to search for the web using relevant keywords	GPE entity type	Nominal heads	Part of Speech: Pronoun	What is an Agent?
Etc.				

Identify Shared Concepts

- Is the same core concept relevant to multiple tasks?

Data Scouting	Simple NE	Full Entity	NP Chunking	SSA
How much of the document needs to be in the target language?	What is a name	What is an entity	Key Concepts	What counts as an Act?
How to locate dates/times within text	What is an entity	What are specific entities	Tagging and taggability (general)	What counts as a physical Act?
How to locate names of disasters within text	What is name extent	What is an entity mention	Extents (general)	What counts as a State?
How to locate names of organizations within text	Entity typing	What is entity mention extent	Grammaticality	What counts as disaster-relevant State?
How to locate names of people within text	PER entity type	Name extents	Part of Speech (general)	What is a basic word?
How to locate names of places within text	ORG entity type	Nominal extents	Part of Speech: Noun	What is a phrase?
How to search for the web using relevant keywords	GPE entity type	Nominal heads	Part of Speech: Pronoun	What is an Agent?
Etc.				

One (Set of Related) Core Concept(s) Per Ordered Module

	Annotation Basics	Ling Basics	Ling Task Branch 1	Ling Task Branch 2
1	Annotation	What Counts as a Language?	Genre	Noun
2	Units of Annotation	Meaning	Domain	Pronoun
3	Annotation Task	Ambiguity	Topic	Constituency
4	Annotation Suitability	Token and Word	Name	Nesting
5	Tag/Label	Part-of-Speech Basics	Noun Phrases	Coreference
6	Tagging/Labeling	Structure	Entity: PER, ORG, GPE, LOC	
7	Extents	Grammaticality		
8	Annotation Rules			
9	Annotation Tool			
10	Annotation Pipeline			
11	Quality Control			
12	Each Language Is Different			

Task Specific Training

- Formal guidelines
  - Organized around decision points
  - As short as possible; minimize jargon; use diagrams
  - Intuitive basic principles plus a small number of rules of thumb
    - When in doubt, do it this way
  - Many naturally-occurring examples in language, genre, domain
  - Include glossary, FAQ
- Narrated presentations or videos explaining task and working through longer and/or more challenging examples
- Narrated presentations or videos showing how to use GUI
- One or more practice kits to get the hang of the task and GUI
- Self-scoring practice test(s)
- Final test resembles actual task
  - Scoring requirements depend on task, applicant pool

Ongoing Training and Testing

- Periodic insertion of “gold standard” work assignments
  - Can be helpful to use English translations of non-English data when language expertise not available
- Periodic insertion of “challenge” assignments to keep annotators attentive, especially for tedious, repetitive tasks
- Double-blind dual annotation of some data to monitor consistency and identify training needs
- “Shepherding” for high-value workers who are struggling
- Group interaction can be useful for consistency, morale
  - Email lists, video hangouts, chat
  - Interactive Q&A
  - Virtual team meetings

Linguistic Data Consortium

Incentives

Individual Incentives

- Hourly wage for long term, multi-skilled staff
- Pay by unit of work completed for distributed, short-term workers
- Graduated payments
  - Skills/experience: Training vs. junior vs. senior
  - Difficulty of tasking: topic scouting vs. transcription vs. lexicon development – but also noisy vs. non-noisy calls
- Lotteries
- Bonuses to address particular need
  - Completion of task
  - Data volume
  - Speed
  - Quality
- Combine as needed: e.g. Graduated + completion + speed
  - Get paid $2 per non-noisy call, $3 per noisy call
  - $100 after completing 10 calls
  - $50 if you complete all calls within 2 weeks of enrolling

Team Incentives

- Competition and Collaboration
- If team reaches goal, every individual on the team earns some bonus
  - *If your language team reaches the goal of completing 50 new 1P kits between midnight tonight and 11:59 PM Thursday, then anyone who works a minimum of 15 hours between midnight tonight and 11:59 PM Thursday will earn 5 extra "free" hours*
- If Team A beats team B, Team A gets a reward
- (Perception of) scarcity of work can also be a strong incentive
  - *Get 'em while they're hot! Supplies are limited!*
- Reporting progress, performance (quality, efficiency)
  - Against goal, against team, against self

Different Populations May Require Different Approaches

- BOLT: Large collection of SMS/Chat data in 3 languages
- Online recruiting methods worked well and quickly for English
- Chinese required greater reliance on personal contacts, social networks
- Egyptian challenges: fewer speakers, smaller web presence, more participant reluctance, more participant technical barriers

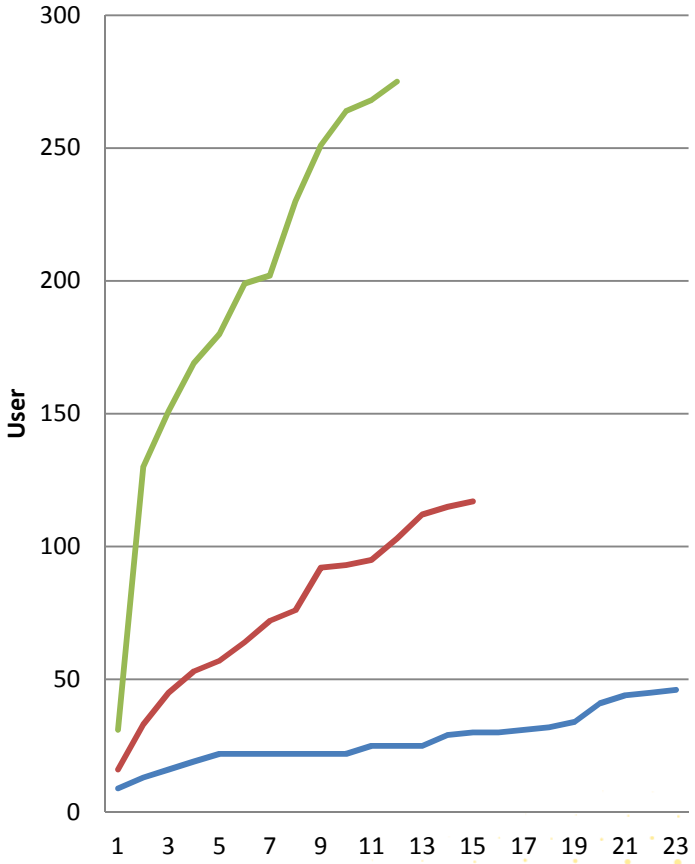
Approach	Egyptian	Chinese	English
In-country and US-based recruiters	9	3	3
Recruiting/participant care hours	>1000	300	250
Online advertising	92 cities	92 cities	92 cities
Social networking sites	60	500	200
Reach out to community orgs	2000	3000	4500
Fliers, postings	500	500	500
Word of mouth, friend-of-a-friend/recruiter	300	1000	300
Contact former collection participants	300	150	1500

Egyptian Collection Challenges

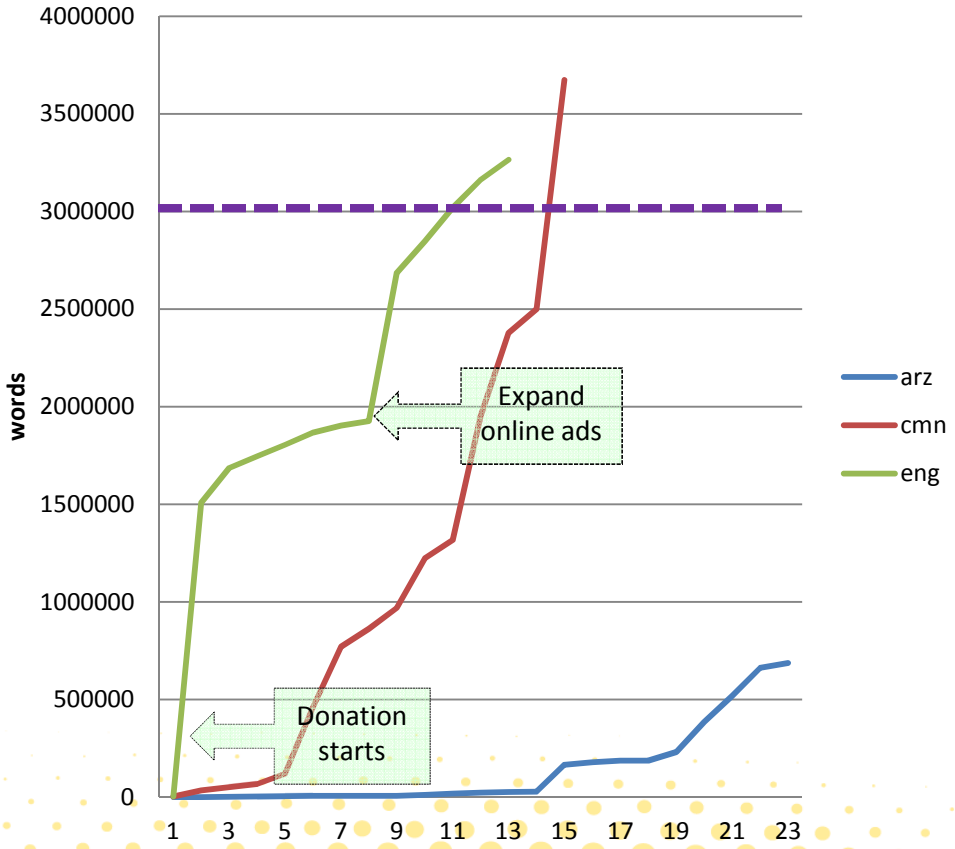
What we tried	What we learned
Meet with numerous cultural, social, industrial and academic organizations in Egypt and US to discuss possible partnership	Potential institutional partners were reluctant to collaborate given perceived risk in current political/social climate
Emphasize extensive privacy protections, UPenn reputation and oversight of collection, legitimacy of collection for language-related research	Despite careful explanations, most potential participants in Egypt remained suspicious of motives for collection
Employ young, tech-savvy recruiters with wide social networks in Egypt and among US Egyptians	Technical barriers were higher for Egyptian participants • Required more back-end development to simplify user experience • More direct hand-holding
Initially, compensate Egyptians (in Egypt) at same rate as English, Chinese participants (in US); incrementally increase compensation rates and offer bonuses	Even with significantly increased compensation, Egyptians reluctant to sign up due to greater perception of personal risk compared to other participants

Collection Progress Over Time

Recruiting Weekly Progress

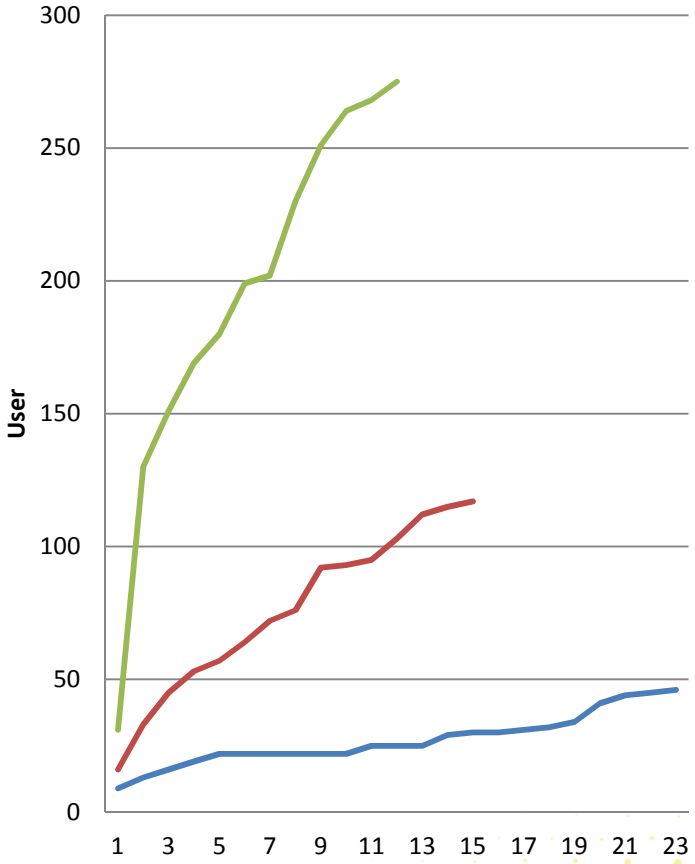


Collection Weekly Progress

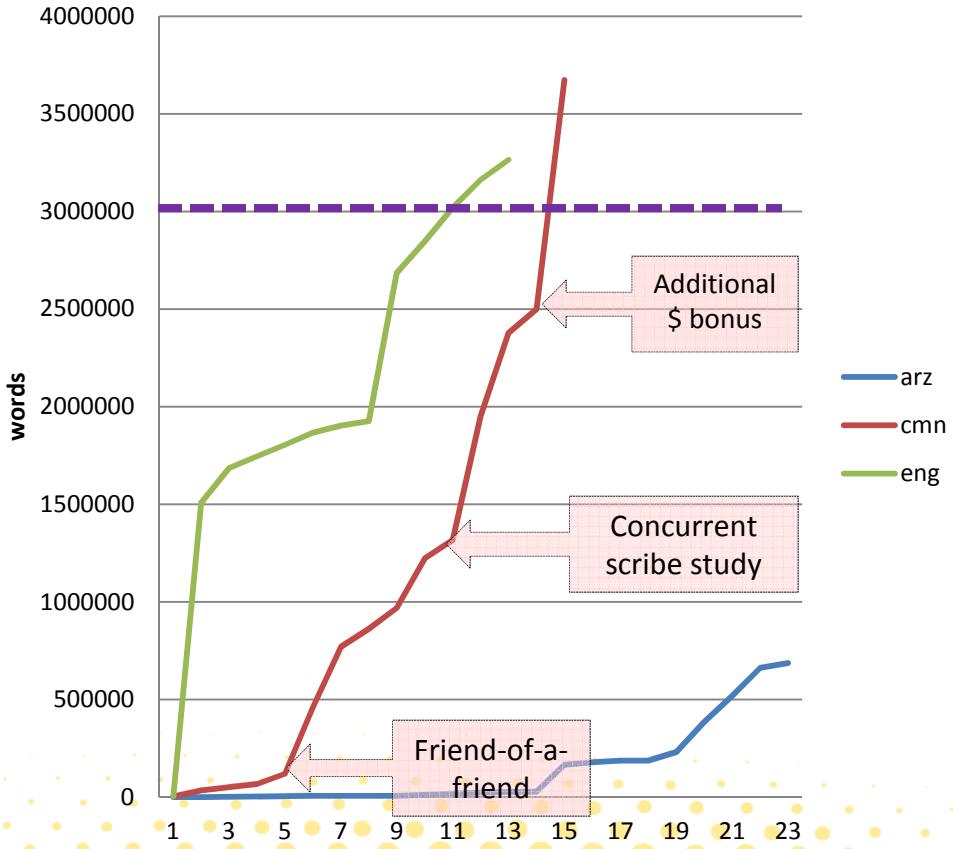


Collection Progress Over Time

Recruiting Weekly Progress

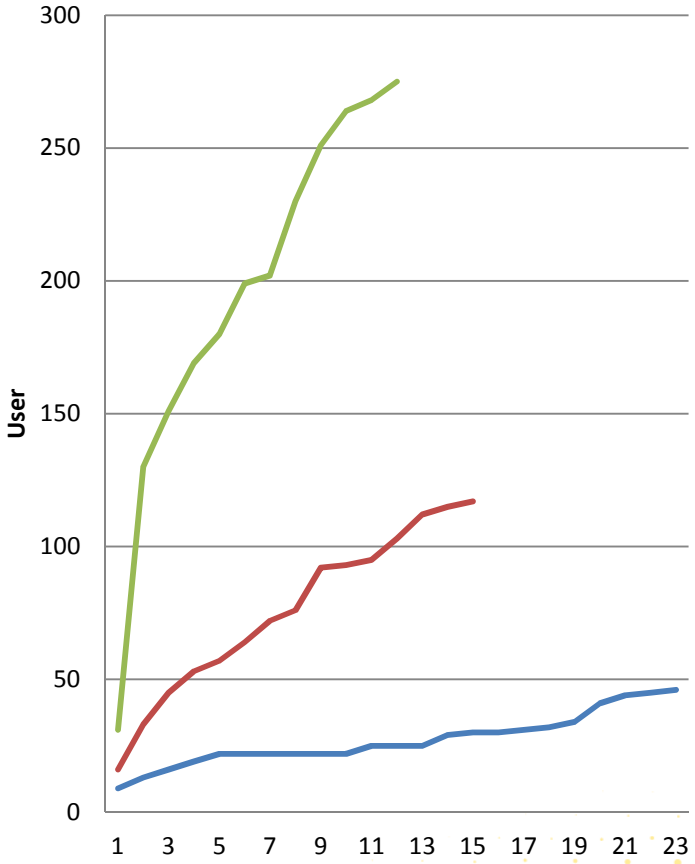


Collection Weekly Progress

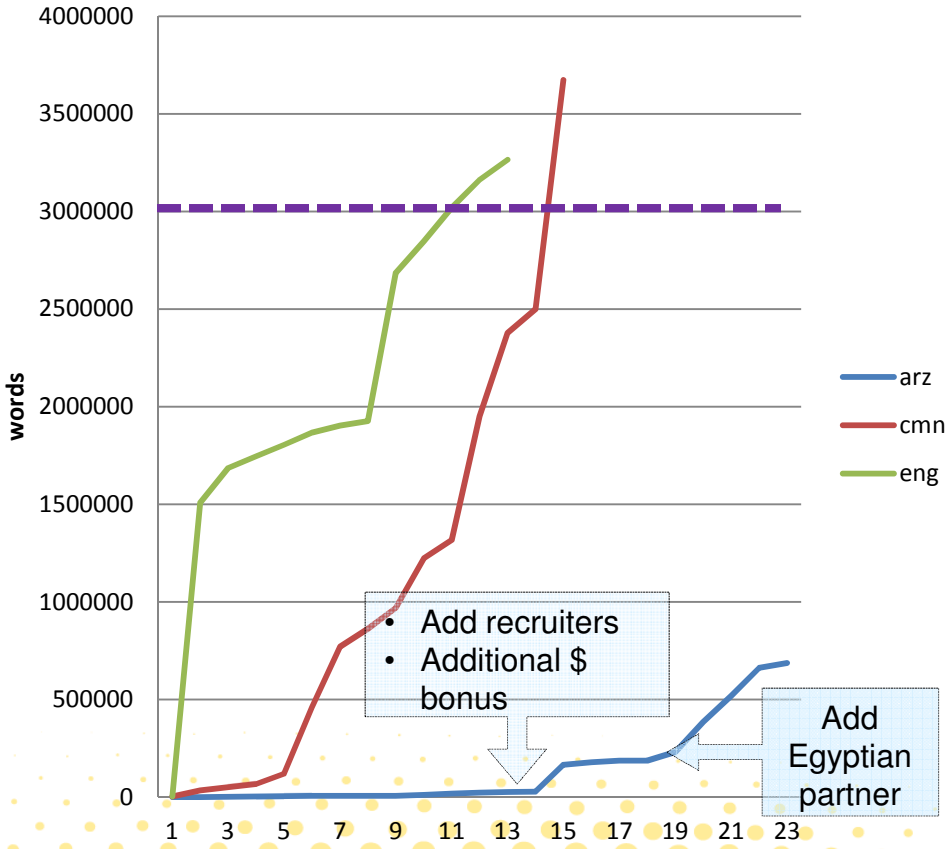


Collection Progress Over Time

Recruiting Weekly Progress



Collection Weekly Progress



Collection Results

- Large collection of SMS/Chat data in 3 languages
- Online recruiting methods worked well and quickly for English
- Chinese required greater reliance on personal contacts, social networks
- Egyptian challenges: fewer speakers, smaller web presence, more participant reluctance, more participant technical barriers

Lang	Enrolled Users	Collcted Messages	Collected Words	Accept Rate	Recruit Hours/ Partcpnt	Avg Words/ Msg	iMessage		WhatsApp		talk			
arz	46	161000	690000	69%	21.7	4	**		***	*			*	**
cmn	118	535000	3700000	70%	2.5	7	*	*			*	***		
eng	275	309000	3300000	78%	0.9	11	**	**	*		*		*	

*a few users; **several users; ***many users

Game-Based Incentives

- Especially effective when the work is monotonous or stressful
- TDT Project required millions of topic relevance: tens of thousands of stories against hundreds of topics
  - Tedious (repetitive news) and stressful (grim news)
  - Pop quizzes and Story of the Day
- HAVIC Project required searching for hundreds of thousands of topic-relevant video clips
  - Need for more variety and atypical videos
  - Scavenger Hunts: “POP”, “IF”, “GREEN”, “NEVER”, etc.
    - Big boost in productivity **and** variety
    - Videos about popcorn, videos about popping balloons, popping bubble wrap, popping bubble gum, videos with pop music, with mom & pop restaurants, with pop goes the weasel, with pop-up books, and more

Non-Financial Incentives

- Work is meaningful
  - A good cause
  - Supports my language or culture
- Work is fun/interesting
  - Meet people
  - Free therapy (through sociolx interviews!)
  - Language geek
- Build skills
  - Stepping stone for career
  - Prestigious employer on resume
- Recognition in lieu of financial incentive
  - Job title (Junior → Senior)
  - Certification
  - Public thank you

Researchers at Penn's Linguistic Data Consortium are starting a new project called SIREN to build linguistic resources for dozens of languages, as part of a larger effort to create language technology that can be utilized in disaster relief situations. You will work with project staff to identify appropriate texts in your language and label them for specific linguistic phenomena. The resources you help to develop will stimulate research in natural language processing technology for many new languages. This work could lead to significant advances in machine translation and related technologies that response teams rely on to provide better support during and after a natural disaster.

Linguistic Data Consortium

Conclusion

Conclusion

- Response to growing need in sponsored programs to rapidly hire, train and incentivize teams of non-expert native speaker consultants (and human subjects) in dozens of languages
- Decision points inform all aspects of task design and present opportunity to
  - Reconfigure traditional workflow for non-experts
  - Simplify training and better align it with worker skills
  - Design GUI for efficient, directed annotation
- Even with paid (traditional or contract) workers, novel incentives can focus effort on most important problems and make workers more engaged and productive