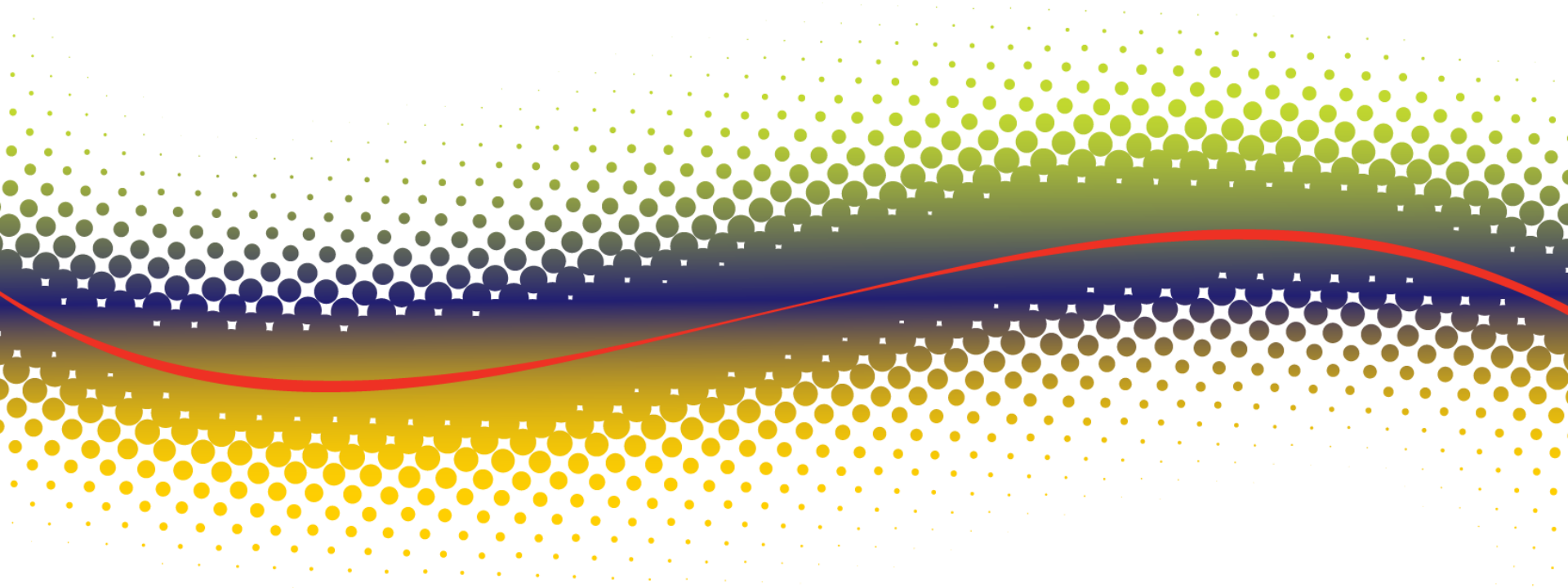
Linguistic Data Consortium

Creation of Language Packs for Low Resource Languages

Jennifer Tracey, Zhiyi Song (Presenter), Stephanie Strassel



Linguistic Data Consortium

Outline

- Introduction of the LORELEI program
- Data collection and translation
- Basic NLP tools and general resources
- Annotation resources
  - Entity
  - NP Annotation
  - Simple Semantic Annotation
  - Entity Linking
  - Situation Frame Annotation
- Discussion and Conclusion

Linguistic Data Consortium

Introduction

Linguistic Data Consortium

LORELEI Program

- LORELEI (Low Resource Languages for Emergent Incidents) is a multi-year DARPA Program, now in final year
- Goal: Improved Human Language Technology for low-resource languages → **language-independent** technology
- Use case: Rapid situational awareness in emerging situations like natural disasters or disease outbreaks
- 24 representative language packs to enable cross-language adaptation, transfer learning and projection research
- Additional Incident language packs (1-2/year) for annual Surprise Language evaluations
- Open Evaluation Campaign: NIST LoReHLT
  - Machine Translation, Entity Detection and Linking, Situation Frame

LORELEI Representative Languages

- Representative languages selected for typological and language family diversity and coverage
- **Representative data**, not training data
- 14 1-Belt-1-Road languages

Africa	Asia	Other
Akan (Twi)	Bengali	Arabic
Amharic	Farsi	Hungarian
Hausa	Hindi	Russian
Somali	Indonesian	Spanish
Swahili	Mandarin	Turkish
Wolof	Tagalog	English*
Yoruba	Tamil	
Zulu	Thai	
	Uzbek	
	Vietnamese	

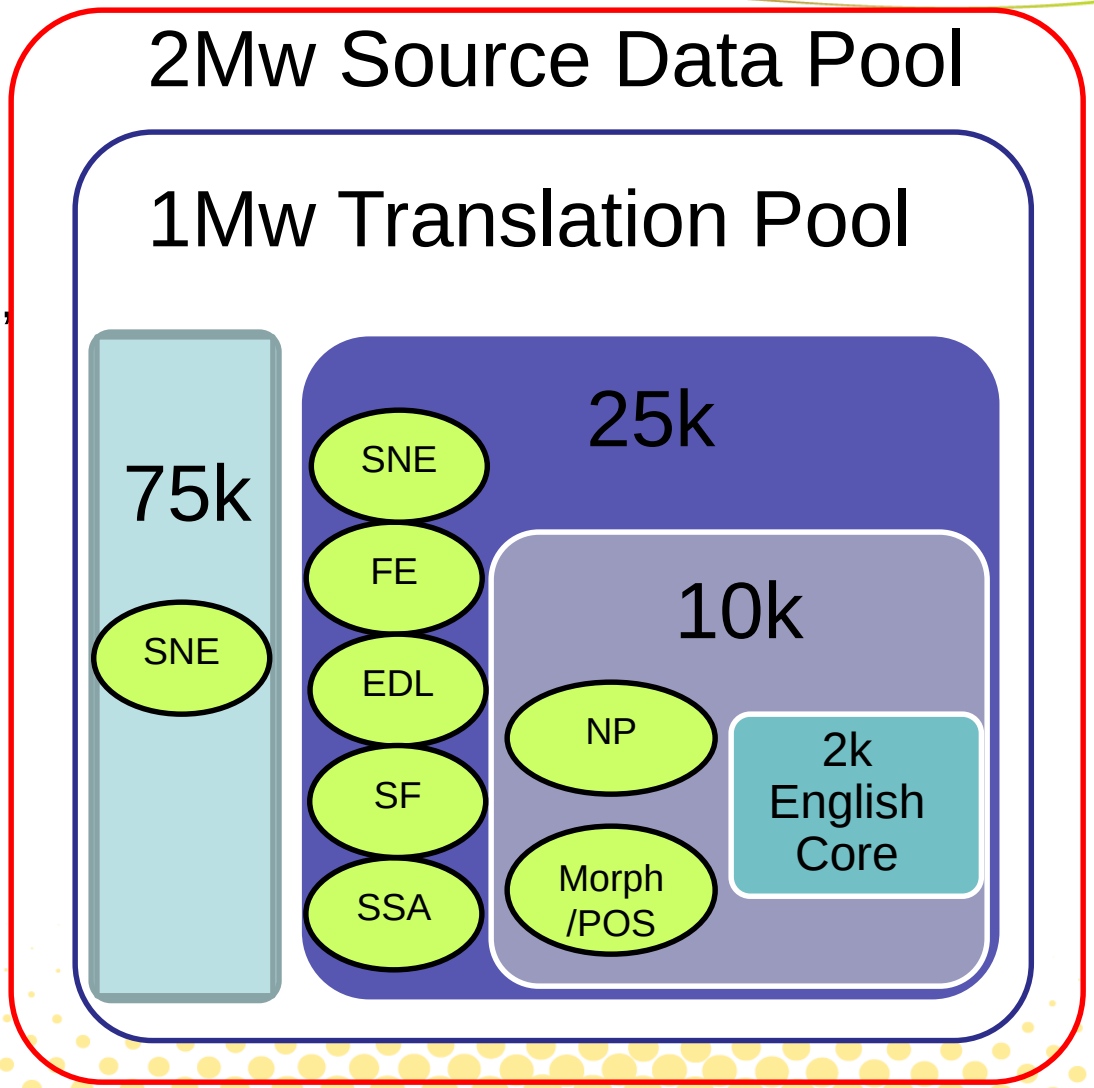
Linguistic Data Consortium

Representative Language Pack Components

Component	Volume
Mono Text	2Mw+ news, forums, blogs, tweets, etc.
Parallel Text	1Mw+ manual, found, crowdsourced
Constructed Lexicon	10K+ lemmas
Basic NLP tools	Tokenizer, segmenter, name tagger, name transliterator, encoding converter
Simple Named Entity	75Kw (from translation pool)
Full Entity	25Kw (from translation pool)
Entity KB Linking	25Kw (from entity pool)
NP Annotation	10Kw (from entity pool) – 10 languages
Morphological Segmentation	2Kw (from mono text) – 9 languages
Situation Frame Annotation	25Kw (from entity pool)
Simple Semantic Annotation	25Kw (from entity pool)
Other	Grammatical sketch, specialized wordlists, POS/Morph tagsets, annotation guidelines

Representative Language Pack Data Selection Principles

- Select data for maximal utility to LORELEI
  - Optimize for in-domain topic, required genre distribution
  - Annotation selection from translation pool
  - Maximize availability of multiple annotations on the same data



Surprise Language Evaluations

- Language, incident announced at start of evaluation
- No training data for evaluation languages
- Basic language pack distributed at start of eval, reflecting what might be available at the outbreak of a new incident

Evaluation Year	Language(s)	Incident(s)
2016	Uyghur	Xianjiang Earthquake
2017	Oromo, Tigrinya	Ethiopia flood/drought cycle, civil unrest, ethnic violence
2018	Kinyarwanda; Sinhala	Rwanda drought/flood cycle and civil unrest/refugee crisis; Sri Lanka floods and civil unrest
2019	TBD	TBD

Incident Language Pack Components

Released on Day 0	Volume
Mono Text	225Kw or more from pre-incident epoch
Found Parallel Text	300Kw from pre-incident epoch or additional comparable text
Found Parallel Dictionary	10K+ entries

Released on Day 0	To Evaluator
[illegible]	[illegible]

Released on Day 3 & 10	Volume
(5 of 8) Performers	50KW Simple NE grammar book
Mono Text	50KW post-incident epoch
Mono Text	50KW Entity Linking parallel IL <> English 50KW post-incident epoch 50KW Situation Frame

- monolingual IL gazetteer
- English gazetteer for incident region

Linguistic Data Consortium

Data Collection and Translation

Data Scouting and Harvesting

- Manual search for incident-relevant documents plus automatic collection of whole sites
  - Formal and informal genres: news websites, blogs, discussion forums, microblog feeds, etc.
  - Focus on sources whose terms of use permit redistribution
- Automatic LID with Google CLD2*
  - Poor performance on informal genres, esp. microblogs
  - Some manual auditing for language, content
- LDC's Webcol framework downloads, converts raw HTML into processed source documents for annotation and distribution
- Centralized tracking database maintains information about document properties and status in pipeline

*https://code.google.com/p/cld2/

Source Data Format Normalization

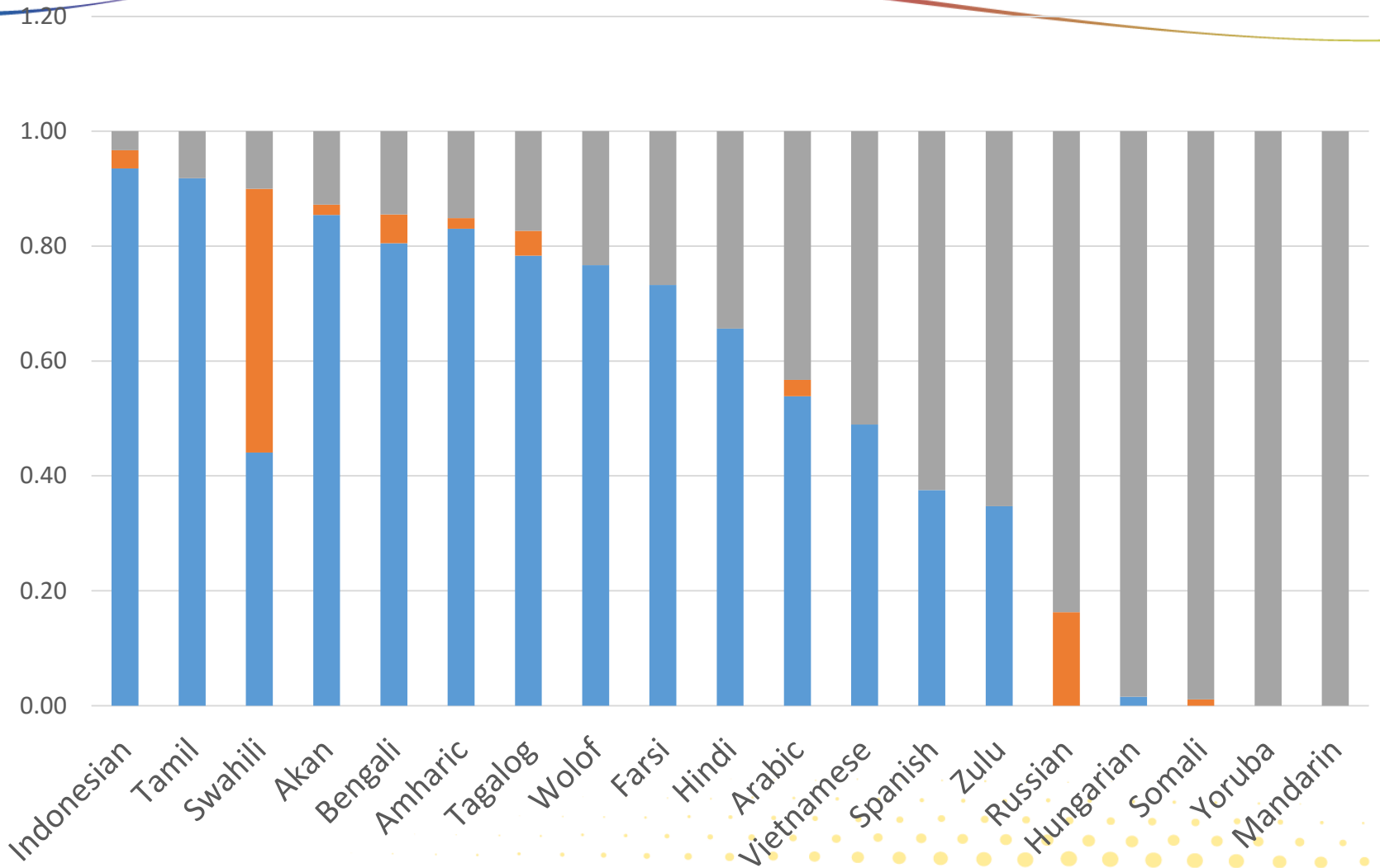
- Processing pipeline addresses numerous issues in source data, e.g.
  - Variable (lack of) compliance with established standards for markup, character encoding, orthography and punctuation
  - Absence or flexibility of orthographic standards in some languages
  - Unknown scope of variability in data input methods used by content authors
- Processing creates separate data streams for linguistic content and structural features for internal data pipeline
  - Raw linguistic content is simple, plain UTF-8 text only
  - Essential structure and metadata is kept in uniform stand-off XML
- Recombined data stream in language packs
  - Explicit markup for sentences, word/punctuation tokens

Representative Language Translation Approaches

- Variety of approaches; ratio varies by language
- Preferred: Found parallel text
  - BITS (Bilingual Internet Text Search) plus Champollion sentence-aligner create an end-to-end harvesting pipeline
    - Usually successful for NW, rarely WL/DF, never Twitter
- If necessary: Crowdsourcing
  - Not viable for all LORELEI languages
- Fallback: Translation vendors
  - Emphasize content-accurate over fluent
  - Re-use existing translations where available (e.g. Mandarin, Arabic from prior DARPA programs)

Linguistic Data Consortium

Translation Distribution



Linguistic Data Consortium

Incident Language Translation Approaches

- Gold standard translations for test set
  - 1-4 independent translations by vendors, plus QC
- Language pack data (released at start of eval)
  - Preferred: 300Kw found parallel text
    - Not always available for incident languages, particularly for in-domain data
  - Fallback: larger volumes of comparable text
    - Semi-automatic (date-informed) multilingual topic clusters
    - Clusters of topically related documents for IL and English

Linguistic Data Consortium

Basic NLP Tools and General Resources

Representative Language Grammatical Sketches

- Emphasize paradigms and basic grammatical description over deep theoretical discussion or nuanced explication of exceptional cases
- Grammatical issues impacting annotation are documented first
  - Are determiners attached to nouns? Is there white space around case markers and adpositions? Describe adjectival forms of names such as *American.*
- All grammars follow same template

1. About the language (basics: classification, ISO code, word order, etc.)
2. Orthography (characters, variation, word boundaries, etc.)
3. Encoding (Unicode chart, etc.)
4. Morphology (inflection and productive derivational morphology for major word classes, morphophonemics where relevant to orthography)
5. Syntax (constituent order, phrasal and clausal phenomena)
6. Specialized subgrammars (personal names and locations, numbers)
7. Variation (register/dialect where relevant to text, codeswitching/borrowing)
8. References

RL and IL Lexical Resources

- Goal: At least 10K lemmas, 90% coverage of mono text
- Representative Languages – build a lexicon
  - Search and harvest online resources
    - Evaluate structure and content
    - Standardize format
      - Citation-form orthography, part-of-speech, English gloss
      - Where necessary, language-specific features (gender, conjugation class, etc.) included via “generic” tags
  - Manually augment entries for highest-frequency terms in mono text corpus not yet covered in the lexicon
- Incident Languages – find a lexicon (digital or hardcopy)
  - Estimate entry count
  - Provide summary of structure, content
  - Rate/describe overall quality and coherence
  - Very minimal processing for digital lexicons

Representative Language NLP Tools

- Tokenizer
  - Custom tokenizer for whitespace-delimited languages
  - Existing tokenizers for non-whitespace languages
- Sentence segmenter
  - NLTK Punkt* algorithm with manual tuning for informal genres
- Custom Named Entity Tagger
  - CRF-based tagger trained on manual named entity annotation
- Custom Name Transliterator where required
  - Integration with lexicon to ensure coverage of most common variants
- Encoding converter where required

**Kiss, Tibor and Strunkt, Jan. (2006). "Unsupervised multilingual sentence boundary detection." Computational Linguistics 32: 455-525*

Linguistic Data Consortium

Annotated Resources

Guiding Principles for (Reasonable) Uniformity Across Languages

- Desire for consistency across all LORELEI languages
  - Uniform approach to all tasks across all language packs
  - While allowing for language-specific variation where needed
- Approach
  - Identify key questions about language features that influence annotation guidelines; address in grammatical sketch
    - Eg. tokenization, case markers, pronouns etc.
  - Develop guidelines template for each task that identifies the possible variance in guidelines based on key questions
  - Shared annotation rules across all tasks/languages wherever possible (e.g. marking text extents)
  - Uniform annotator training paradigm for all languages
    - Online tutorials with iterative training/testing

RL and IL: Entity Annotation

- Entity annotation is foundation for other annotation tasks

	Simple Named Entity	Full Entity
Status	RL and IL	RL Only
Types	Person, Organization, GeoPolitical Entity, Location (includes Facilities)	Same, plus Titles
Coverage	Names only	Names, Nominals, Pronouns
Coreference	No	Yes

- Extent boundaries always coincide with token boundaries
- Tag for usage

```
He works at the [University of Pennsylvania] ORG
He got off the bus at the [University of Pennsylvania] LOC
```

- Embedded names are not annotated

RL: NP Annotation

- Annotate maximal, non-overlapping Noun Phrases
- Also decompose, mark smaller NPs, e.g.

  `[The government] will send [aid workers] to [[the region] [that] was struck by [the earthquake] [last month]].`

  → Both [the region] and [the region that was struck... last month] are marked
- Follow surface syntactic form
  - Only label NPs that pass constituency tests

    `[North and South Korea]`
    not `*[[North] and [South Korea]]`
  - Decompose names when syntactic structure is present

    `[[University] of [Pennsylvania]]`

RL: Simple Semantic Annotation

- Capture basic understanding of what is happening and/or what is the case in a sentence
- Using broad predicate and argument categories
  - Not fine-grained semantic distinctions
- Label physical acts and domain-relevant states, their agents, their patients and their places
  - Identify taggable Act or State and select trigger word
    - Generally select head, but allow intuitive extents (e.g. for multiword expressions)
  - Identify Agent, Patient, Place
    - Select most informative mention (NAM > NOM > PRO)

RL and IL: Entity Linking Annotation

- Start with names labeled in Simple Named Entity task
  1. Link names to external knowledge base
  2. Perform cross-doc coreference for any unlinked names

- Single reference KB drawn from four distinct sources
  - **GPE, LOC** from GeoNames
  - **PER** from CIA World Leaders List
  - **ORG** from CIA World Factbook Appendix B
  - **Manual augmentation** for 100+ additional incident-, region- and/or domain-relevant PER and ORG entities that do not appear in the non-augmented KBs

Linguistic Data Consortium

RL and IL: Situation Frames

- Cornerstone evaluation task for LORELEI
- Goal: Enable information from different data streams to be aggregated into a comprehensive, actionable understanding of the basic facts needed to mount a response to an emerging situation
- Three primary information elements
  - What's happening
  - Where is it happening
  - What is the urgency (scope + severity) of the situation

Situation Frame Annotation

- **Need Frames** capture information about needs that may emerge in a disaster situation along with any response to those needs
  - The type of need
  - The place where the need exists, if known
  - The current status of the need and its resolution

Need Types		Issue Types
Evacuation	Infrastructure	Civil Unrest / Widespread Crime
Food Supply	Medical Assistance	Regime Change
Search/Rescue	Shelter	Terrorism or other Extreme Violence
Utilities, Energy or Sanitation	Water Supply	

  - The type of issue
  - The place where the issue exists, if known
  - The current status of the issue
- **Pilot:** Sentiment about the situation or those involved

IL Example Walkthrough

Landslide hit Guinsaugon in the south of the Philippine island of Leyte. *Reports say village totally flattened and housing destroyed.*

1. Label Named Entities and do within-doc coreference if needed

IL Example Walkthrough

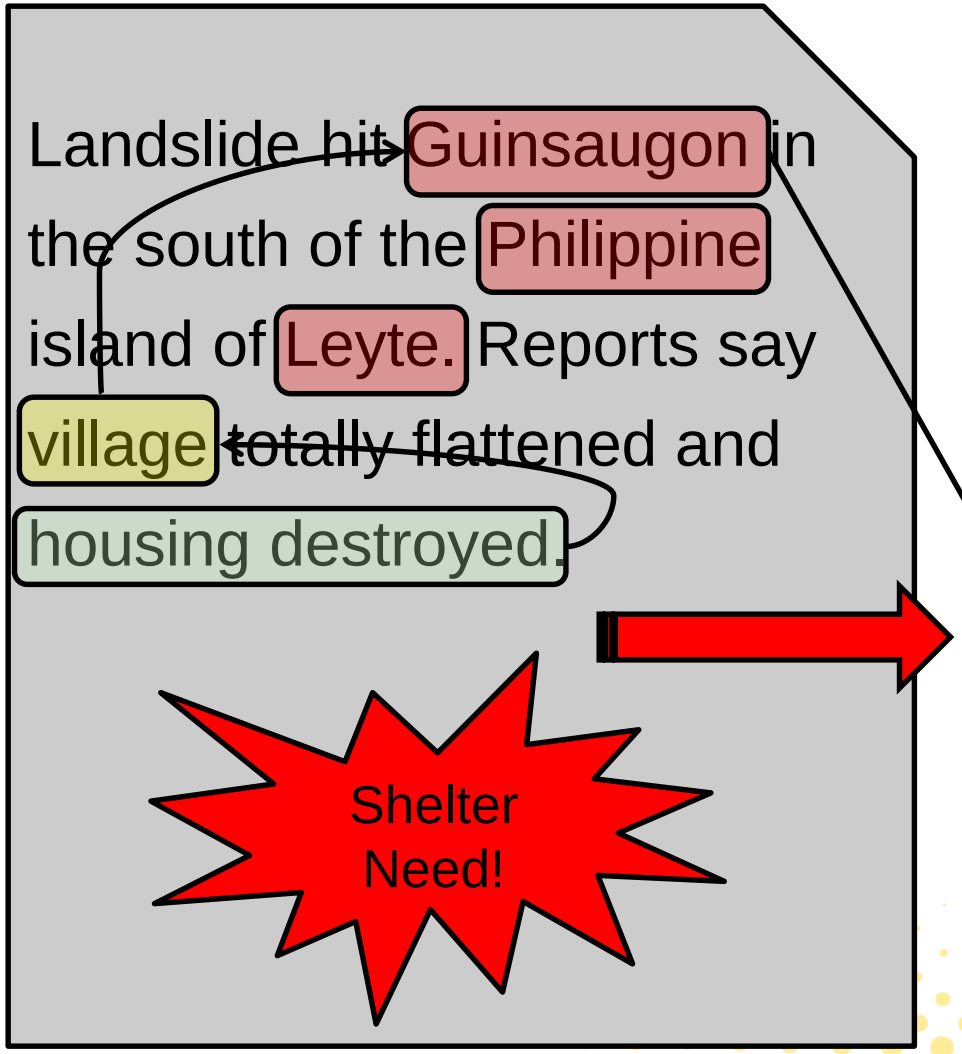
Landslide hit Guinsaugon in the south of the Philippine island of Leyte. *Reports say vi*
h

1. Label Named Entities and do within-doc coreference if needed

2. Link entities to Knowledge Base (based on GeoNames)

ID	Name	Feature class	Lat/Long
1694008	Republic of the Philippines	Independent political entity	N 13°00′00″ E 122°00′00″
9035710	Philippines	Populated place	N 7°23′00″ E 122°45′46″
1685725	Province of Southern Leyte	2nd order administrative division	N 10°20′00″ E 125°05′00″
1706802	Leyte Island	Island	N 10°49′58″ E 124°50′07″
1712304	Guinsaugon	Populated place	N 10°16′00″ E 125°11′00″
1712303	Guinsaugon	Populated place	N 10°21′06″ E 125°06′33″

IL Example Walkthrough



1. Label Named Entities and do within-doc coreference if needed
2. Link entities to Knowledge Base (based on GeoNames)
3. **Create Situation Frames**

Need Frame 1

- Type: Shelter
- Place: Guisaugon
- Need status: Current, Urgent
- Resolution status: Unknown
- Reported by: n/a
- Resolved by: n/a

RL Example Walkthrough

Landslide hit Guinsaugon in the south of the Philippine island of Leyte. Reports say village totally fl
housing destro

```
- Act: landslide
    - Patient: Guinsaugon
    - Place: Leyte
    - Place: Philippine
    - Place: south
- Act: hit
    - Agent: landslide
    - Patient: Guinsaugon
- Act: flattened
    - Agent: landslide
    - Patient: Guinsaugon
- Act: destroyed
    - Agent: landslide
    - Patient: housing
    - Place: Guinsaugon
```

[Land slide] hit [[Gui h] of [[the Philippine island] of [Leyte]]]]. [Report s] att en ed and [hous ing] destroy ed.

- **Named entities**
- **Nominal, pronominal entities**
- **Entity coreference**
- **NP Annotation and/or morph segmentation**
- **Simple Semantic Annotation**

Linguistic Data Consortium

Discussion and Conclusions

Linguistic Data Consortium

Challenges and Solutions

- Timeline: creating 30+ RL/IL language packs in 3 years
- Annotator management
  - Low resource languages → low availability of skilled annotators
  - Standardized, self directed training paradigm reduced management burden
  - Use of English “annotation shepherds” increased retention
- Annotation guidelines for 8+ tasks in each of 30+ languages
  - Standardized templates highlighting places where languages vary
  - Synchronization with grammatical sketch information
- Optimizing quality vs. quantity vs. complexity tradeoff
  - Where possible, use language-independent approaches
  - Where possible, simplify annotator decision making → better quality, faster annotation
  - Independent quality control review by external site prior to delivery

Linguistic Data Consortium

Conclusions

- All 24 Representative Language Packs will be completed by the end of 2018
  - Data already released to LORELEI and LoReHLT sites
  - Beginning to appear in LDC's catalog; two corpora per language
    - Part 1: Monolingual and Parallel Text
    - Part 2: Annotations and Other Resources
- Incident Language Packs will also be published in LDC catalog after end of the program
- Final LORELEI evaluation in July 2019
  https://www.nist.gov/itl/iad/mig/lorehlt-evaluations

Resources of LRLs in LDC's Catalog

- LDC has published audio, lexical and text resources for LRLs in the catalog

Amharic	Dschang	Kumanji Kurdish	Somali	Ukrainian
Assamese	Georgian	Lao	Swahili	Urdu
Bamanankan	Haitian	Malto	Tagalog	Vietnamese
Basque	Haitian	Maninkakan	Tamil	Vietnamese
Bengali	Hausa	Mawukakan	Telugu	Yoruba
Cantonese	Hindi	Ngomba	Tok Pisin	Zulu
Cebuano	Kazakh	Pashto	Turkish	

Acknowledgement and Reference

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0123. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

References

- Strassel, S., & Tracey, J. (2016). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. LREC 2016: 10th Edition of the Language Resources and Evaluation Conference Portoroz, May 23-28
- Strassel, S. M., Bies, A., & Tracey, J. (2017). Situational Awareness for Low Resource Languages: the LORELEI Situation Frame Annotation Task. SMERP 2017: First Workshop on Exploitation of Social Media for Emergency Relief and Preparedness, Aberdeen, April 9
- Christianson, C., Duncan, J. & Onyshkevych, B. Overview of the DARPA LORELEI Program. Machine Translation (2018) 32: 3. https://doi.org/10.1007/s10590-017-9212-4
- Griffitt, K., Tracey, J., Bies, A., & Strassel, S. M. (2018). Simple Semantic Annotation and Situation Frames: Two Approaches to Basic Text Understanding in LORELEI. LREC 2018: 11th Edition of the Language Resources and Evaluation Conference Miyazaki, May 7-12