LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages

Stephanie Strassel and Jennifer Tracey
Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104
E-mail: strassel@ldc.upenn.edu, garjen@ldc.upenn.edu

Abstract

In this paper, we describe the textual linguistic resources in nearly 3 dozen languages being produced by Linguistic Data Consortium for DARPA's LORELEI (Low Resource Languages for Emergent Incidents) Program. The goal of LORELEI is to improve the performance of human language technologies for low-resource languages and enable rapid re-training of such technologies for new languages, with a focus on the use case of deployment of resources in sudden emergencies such as natural disasters. Representative languages have been selected to provide broad typological coverage for training, and surprise incident languages for testing will be selected over the course of the program. Our approach treats the full set of language packs as a coherent whole, maintaining LORELEI-wide specifications, tag sets and guidelines, while allowing for adaptation to the specific needs created by each language. Each representative language corpus, therefore, both stands on its own as a resource for the specific language and forms part of a large multilingual resource for broader cross-language technology development.

Keywords: low resource languages, multilingual resources, situational awareness

1. Introduction

The goal of DARPA's LORELEI (Low Resource Languages for Emergent Incidents) Program is to improve the performance of human language technologies for low-resource languages, particularly in the context of a rapidly emerging and quickly evolving situation like a natural disaster or disease outbreak. LORELEI systems will be required to process information about topics, entities, events and sentiment found in the LORELEI data, with the goal of providing situational awareness within days or even hours of the outbreak of an incident.

Linguistic Data Consortium is building text language packs for LORELEI[1], comprising data, annotations, natural language processing tools, lexicons and grammatical resources for 23 Representative Languages as well as 12 Incident Languages, listed in Table 1. Representative Languages (RL) have been selected to provide broad typological coverage, while Incident Languages (IL) have been selected to enable development and testing of LORELEI system capabilities. The choice of evaluation Incident Languages remains unknown to performers until the start of the evaluation.

There is a growing interest in human language technology for low resource languages (LRLs). The IARPA Babel program targets improvements in Automatic Speech Recognition (ASR) and Keyword Spotting (KWS) system performance, and language packs for Babel primarily consist of transcribed and untranscribed speech in a variety of acoustic conditions (IARPA 2016). While Babel focuses on general improvements to speech processing technology for LRLs, LORELEI language packs (and the corresponding technology evaluations) are designed specifically with the goal of improved technology for situational awareness in emergent situations, and their composition reflects this.

Year 1 Representative Languages	Year 2 Representative Languages	Incident Languages
0: Uzbek	13: Akan (Twi)	1: Uzbek
1: Turkish	14: Bengali	2: Mandarin
2: Hausa	15: Hindi	3: Y1 eval (undisclosed)
3: Amharic	16: Indonesian	4-5: Y1 dev (undisclosed)
4: Arabic	17: Swahili	6: Y2 eval (undisclosed)
5: Farsi	18: Tagalog	7-8: Y2 dev (undisclosed)
6: Hungarian	19: Tamil	9: Y3 eval (undisclosed)
7: Mandarin	20: Thai	10-11: Y3 dev (undisclosed)
8: Russian	21: Wolof	12: Y4 eval (undisclosed)
9: Somali	22: Zulu	
10: Spanish		
11: Vietnamese		
12: Yoruba		

Table 1: LORELEI Program Languages

This focus also provides a contrast between LORELEI and LDC's earlier work building language packs for the REFLEX Less Commonly Taught Languages (LCTL) Project (Simpson et. al. 2008). While REFLEX and LORELEI language packs have much in common,

[1] A small amount of speech data will also be created for each LORELEI language, under a separate effort by Appen.

LORELEI adds several new kinds of annotation as well as a surprise language element, all with an eye toward the LORELEI use case. Finally, LDC's approach to LORELEI resource creation treats the full set of 35 language packs as a coherent whole, in order to enable research approaches focused on rapid adaptation through use of language universals and projection from related-language resources. LORELEI specifications, guidelines and tag sets have been informed by language universals while allowing for language-specific adaptation as needed. The sections that follow describe our approach to building the LORELEI text language packs in detail.

2. Representative Language Packs

Representative language packs for LORELEI contain monolingual text, parallel text, several types of annotation, tools for text processing, segmentation, and entity tagging, as well as lexicons and grammatical sketches. Annotations include two types of entity annotation, noun phrase chunking, simple semantic annotation of limited predicates and argument roles, and morphological and part-of-speech analysis. In most cases language pack components are newly produced for LORELEI, but the first three language packs (Uzbek, Turkish and Hausa) were produced under the DARPA BOLT Program and therefore may not be fully compliant with the final set of LORELEI requirements. Moreover, for a handful of languages some portion of the language pack is drawn from data first produced in REFLEX, with some updating of the content to meet current LORELEI standards. The sections below describe the RL packs in more detail.

2.1 Monolingual Text

LORELEI requires collection of at least 2 million words of monolingual text for each RL, including news (50%), blogs and discussion forums (40%), and microblogs like Twitter (10%). A special emphasis is placed on collection of data in the LORELEI domain (natural disasters and the like), to support the requirement that at least half of the material selected for downstream translation and annotation is in-domain. Data scouts in each language search the web for suitable sources, designating entire data sources or websites for collection as well as selecting individual documents that discuss specific incidents. Each website or document selected for inclusion in the corpus is then harvested and reviewed for Intellectual Property Rights (IPR) issues, using an extension of LDC's WebCol infrastructure first developed in the DARPA BOLT Program (Garland et. al. 2012). For sources like Twitter whose terms do not permit redistribution by LDC, we release a list of URLs or IDs plus utilities for performers to harvest, process and validate the data themselves.

Harvested text is automatically tokenized and sentence-segmented using a combination of open source tools and approaches developed by LDC for LORELEI. Data is converted to UTF-8 encoding, and original files are converted into a variety of derived formats to support subsequent translation, annotation and distribution. The conversion process is intended to address things like variable (lack of) compliance with established standards for markup, character encoding, orthography and punctuation; absence or flexibility of orthographic standards in some languages; and unknown scope of variability in data input methods used by content authors. Procedurally, we create separate data streams for linguistic content versus structural features for use in LDC's internal data pipeline. Raw linguistic content is preserved in a simple, plain UTF-8 text only representation. Essential document structure and metadata (e.g. paragraph boundaries, <quote> tags in threaded conversations) is kept in a uniform stand-off XML. We then create a recombined data stream for inclusion in language packs, with tokenization, sentence segmentation and other post-processing applied as required.

Processed linguistic content files are run through the Google CLD2 language detector with some subsequent manually verification of language content. At least for some languages, code switching and orthographic variation are expected; for instance Uzbek text may be written in Cyrillic, Latin or Perso-Arabic script depending on the source. This presents a challenge for LID, as does the prevalence of highly informal short message data like Twitter. Therefore, automatic LID is intended to identify pervasive problems with a given data source, and some amount of language mixture in the monolingual text for representative language packs is expected.

Finally, we standardize file naming and register each document in a LORELEI-wide tracking database. A portion of the processed monolingual data is added to the translation queue, and a portion of the files in the translation queue are further added to the annotation queue, with some manual review to confirm that the data is suitable for further treatment.

2.2 Parallel Text

Each representative language requires 900,000 words from the monolingual text corpus to be translated into English. In addition, a fixed set of approximately 100,000 words of text is translated from English into each representative language. This "English Core" includes a set of domain-focused news text as well as some general news, a phrasebook containing everyday colloquial phrases, and an elicitation corpus of sentences designed by a team at Carnegie Mellon University to elicit various linguistic structures (Alvarez et. al. 2006).

Translations of the monolingual data are acquired through a combination of methods. To the greatest extent possible, we locate and harvest parallel text from the web using LDC's Bilingual Internet Text Search (BITS) tool; the harvested parallel text is then sentence-aligned using Champollion (Ma 1999; Ma & Liberman 2006). We also use crowdsourcing to obtain a portion of the translation

for some languages, although not all languages have sufficient numbers of crowd workers with the required skills to perform translation. However, based on previous research in crowdsourcing translation, we expect coverage of at least some of the representative languages (Pavlick, et al. 2014), although recent changes to Amazon's policies for workers on Mechanical Turk appear to have reduced the number of workers available through that site. For languages where crowdsourcing is not viable, and for portions of the data that require more translation expertise (like the elicitation corpus), we rely on professional translators. Prior to manual translation (whether crowdsourced or professional), data is automatically segmented, and translators preserve the alignment between the source language segment and the resulting target language segment.

The heavy reliance on found data and crowdsourcing in LORELEI compared to recent DARPA machine translation programs like GALE and BOLT yields highly variable translation quality, reflecting the LORELEI use case in which few high-quality, high-volume resources are likely to be available at the start of an incident. LDC conducts limited quality control on the translations, using methods that vary based on the translation type, but which are primarily automated rather than manual and which serve primarily to exclude egregiously bad data rather than improve the quality to a true "gold standard".

2.3 Annotation

LORELEI RLs also require several types of linguistic annotation, which vary in complexity and level of required linguistic knowledge. A portion of data in the translation pool for each representative language is selected for manual annotation. Most of the data comes from the source data translated *into* English (following the same genre and domain distribution as the general translation pool), but approximately 2000 words per language is drawn from the English Core set translated *from* English, thereby creating a small set of parallel annotated data across all RLs. This English Core set is not included in the annotation for Turkish, Uzbek or Hausa since those language packs were created under BOLT. To the greatest extent possible, the same set of data is annotated for all LORELEI annotation tasks, as shown in Figure 1 below.

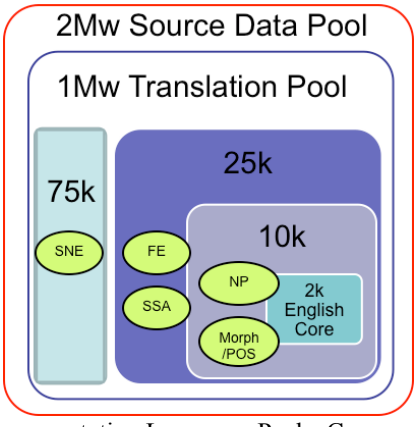

Figure 1: Representative Language Packs Composition

One of the key challenges in building LORELEI RL packs is the need to rapidly train native speakers in nearly two dozen languages to perform annotation tasks that require fairly sophisticated understanding of linguistic issues. The project timelines are quite compressed, with only a few months turnaround between starting annotation and delivering completed, quality-checked language packs, which means that annotator training also has to be very efficient. Further, for at least some LORELEI languages, the pool of available annotators is already small, and finding annotators with a strong linguistics background isn't feasible. Accordingly, annotation tasks have been designed with a novice, non-linguist annotator in mind. Each annotation task is broken down into a series of simple decision points (e.g. Is there a name to tag in this sentence? Where does the name start and end? Is it the name of a person, organization, geopolitical entity or location? etc.). Annotator training, guidelines and user interfaces have been designed to directly reflect these decision points.

2.3.1. Entity Annotation

We perform two types of manual entity annotation for the Representative Languages. In Simple Named Entity (SNE) annotation, text is labeled for person names, organization names, location/facility names, and geopolitical entity names. Full Entity (FE) annotation includes nominal and pronominal entities in addition to names and includes annotation of titles as well as the other entity types used in Simple Named Entity annotation. Full entity annotation also includes within-document entity co-reference. For both Simple Named Entity and Full Entity annotation, entities that can function as either Organizations or Locations/Facilities are tagged depending on their usage in a given sentence. Embedded names are not annotated, so that the name "Africa" would not be separately labeled in the organization name *Africa Rice Center*.

A total of 75,000 words per language is labeled for SNE, while an additional 25,000 words is labeled for FE; this 25Kw FE set includes the 2000-word English Core. This labeled data is in turn used to train the named entity taggers described in Section 2.4 below.

2.3.2. Simple Semantic Annotation

All of the data labeled for FE is also subject to Simple Semantic Annotation (SSA). The goal of SSA is to capture a basic understanding of what is happening and/or what is the case in a sentence. Using broad predicate and argument categories, annotators label specified types of Acts (events) and States (situations) along with their associated arguments. Procedurally, annotators first identify a taggable Act or State in the sentence and select the "trigger word" that most directly evokes the Act or State. This is typically the head of a Verb Phrase or Noun Phrase, but annotators may select multiple words when this seems intuitively preferable (for instance in the case of multi-word expressions).

Annotators then label three types of arguments for each Act or State, selecting the most salient and informative minimal text string for each argument type. An Agent argument is defined as the entity, event or situation that does or causes an Act/State to occur. Patient is the undergoer, receiver or experiencer of an Act/State, or the goal of an Act/State. Place is the place where the Act/State occurred, place headed to, or place leaving from.

The current version of SSA limits annotation to Physical Acts and Domain-Relevant States. Physical Acts are concrete events, actions or activities that take place in the observable, material world. This includes human-caused and non-human caused events but excludes abstract, cognitive and verbal/attribution events. Domain-Relevant States are situations that describe, are caused by or provide information about LORELEI domain events, like natural disasters. This restriction to Physical Acts and Domain-Relevant States was adopted after the creation of the Turkish and Uzbek language packs, where we found that the inclusion of abstract Acts and non-domain States slowed down annotation to an unacceptable pace.

An example of a Turkish sentence annotated for SSA appears in Figure 2 below.

```
Yaklaşık 1,2 milyon insan
depremden dolayı evsiz kaldı;
yardım çalışanları, Mayıs ayında
başlayacak olan şiddetli yağmur
sezonundan önce, bu insanlara
barınak sağlamak için süratle
çalışıyorlar.

- Act: depremden [earthquake]
- State: evsiz [homeless]
    Agent: depremden [earthquake]
    Patient: insan [people]
- Act: başlayacak olan [start]
    Patient: sezonundan [season]
- Act: sağlamak [provide]
    Agent: çalışanları [workers]
    Patient: barınak [shelter]
    Patient: insanlara [people]
- Act: çalışıyorlar [working on]
    Agent: çalışanları [workers]
    Patient: sağlamak [provide]
```

Figure 2: Turkish Sentence Labeled for SSA

2.3.3. NP Chunking

A 10,000-word subset of the data labeled for both FE and SSA, including the 2000-word English Core, is further labeled to identify maximal, non-overlapping Noun Phrases. Some NPs are also decomposed to mark smaller NPs within them, resulting in annotation like the following sentence:

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

Annotators follow surface syntactic structure, applying specific tests such as constituency, to determine which NPs to mark. Unlike Entity annotation tasks, names within larger NPs are extracted/labeled as their own NPs when syntactic structure dictates, as in: [[University] of [Pennsylvania]].

2.3.4. Morph and POS Annotation

The final and most challenging annotation task for the RL packs is part-of-speech and morphological annotation. The same 10Kw labeled for NP Chunking is also subject to morph/POS annotation.

In the three LORELEI languages produced under BOLT (Uzbek, Hausa and Turkish), our approach to morphological annotation was tightly integrated with creation of language-specific analyzers at LDC (Kulick and Bies, 2016). For subsequent LORELEI RL packs we will utilize a universal analyzer being developed by LORELEI performers, rather than creating custom analyzers for each language. The universal analyzer is expected to produce multiple (possibly ranked) analyses for each token, consisting of a lemma plus features (possibly with individually segmented, labeled morphemes). Annotators select the best solution from the list, or choose "unanalyzable" if the analyzer has no correct solution for the token. Part-of-speech labels are not directly annotated, but are instead derived from the morph annotation. In cases of unanalyzable tokens, annotators choose the appropriate POS label from the "Google 12" Universal POS set (Petrov et. al. 2011). Prior to manual annotation, mappings are created between the universal analyzer feature set and a simplified set of labels that are more suitable for non-linguist human annotators.

One additional type of morphological annotation appears in the Turkish and Uzbek representative language packs: morpheme alignment. This task was designed to identify translational correspondence at the morpheme level in parallel text. Although the task was completed for these two preliminary languages, it proved extremely challenging and costly for non-linguist annotators. Moreover, feedback from LORELEI performers suggested that this resource was lower priority than some other options under consideration; therefore, a decision was made to exclude morphological annotation from remaining RL packs in favor of additional evaluation resources.

2.3.5. Language Universal Annotation Principles

To encourage consistency across different RL packs, and across different annotation tasks within each RL pack, we have enumerated the types of language features that directly affect LORELEI annotation decisions; this includes things like the presence or absence of whitespace around words, the use of clitics and/or contractions, and the syntax of possessive noun phrases

(genitive case, possessive case, compounding, idafa-like constructions, adpositions, etc.). These issues are initially documented in the grammatical sketches (Section 2.5) for each language, and then folded into annotation guidelines as required.

All annotation tasks begin with a language-independent guidelines template, laying out the expected sections for each task, including generic verbiage that can be copied into language-specific versions of the guidelines and placeholders for language-specific examples. But the template guidelines also contain indications of areas where language-specific decisions must be made, outlining the various "annotation rules" that can be invoked in the language-specific version of the guidelines, depending on the particular features of a given language. For instance, one such rule states that:

Pronominal mentions of entities are only annotated if they are separate words. If they appear as verb morphology, they are not annotated.

Implementation of these rules is designed to produce more consistent treatment of similar phenomena across languages.

Care was also taken to ensure that similar concepts across different annotation tasks are treated in the same way, unless there is a need for variation in order to support the goals of a specific task. For example, name is a relevant concept in multiple annotation tasks, and decisions about such details as whether to include a definite article in the name of an organization such as "the Red Cross" are synchronized across tasks, so that annotators in Simple Semantic Annotation (SSA) and both Simple Named Entity (SNE) and Full Entity (FE) annotation exclude the definite article from the extent; Noun Phrase (NP) annotators, on the other hand, include the definite article because it is clearly part of the noun phrase. Figure 3 illustrates the treatment of named and nominal entity extents across different annotation tasks for an English example.

The Red Cross sent aid workers to the region.

	Red Cross	The Red Cross	workers	aid workers	region	the region
SNE	YES	NO	n/a	n/a	n/a	n/a
FE	YES	NO	NO	YES, mark head	NO	YES, mark head
SSA	YES	NO	select most salient, typically the head			
NP	NO	YES	NO	YES	NO	YES

Figure 3: Cross-task Treatment of Entity Extents

2.4 Tools

In addition to monolingual, parallel and annotated text, LORELEI RL packs include some simple NLP tools. These tools are not intended to provide state of the art results, but rather to simulate the kinds of baseline tools that may be available at the outbreak of a disaster situation involving some new language.

To the extent that original sources of data use encodings other than UTF-8, we provide a simple encoding converter. Where needed, we also create name transliterators, which are integrated with the lexicon to ensure coverage of most common name variants. Language packs also include baseline tokenizers and sentence segmenters. For whitespace-delimited languages we create a custom tokenizer that operates on a series of regular expressions that dictate how to tokenize; special attention is given to handling web text artifacts like hash tags and URLs. For non-whitespace languages we rely on existing widely-used tokenizers.

Sentence segmentation utilizes an implementation of the Punkt algorithm based on the version found in NLTK (Kiss et. al. 2006). This is an unsupervised, re-trainable algorithm and considerable tuning is required to handle the types of informal data prevalent in LORELEI. Finally, for each representative language we create a custom conditional random field-based named entity tagger.

In addition to the basic NLP tools, we also provide utilities for downloading and processing text from the web so that LORELEI performers can replicate LDC's data processing pipeline internally. This is particularly important given that during a real-world (or simulated) disaster situation, systems will be expected to process data "in the wild" instead of relying on pre-processed data.

2.5 Lexicon and Grammatical Sketches

Each RL pack also includes a lexicon encompassing an inventory of at least 10,000 headwords/lemmas with part-of-speech, English gloss, and optionally (where appropriate and available) morphology. The lexicon is comprised of found resources (existing online dictionaries, etc.) and existing LDC resources, with limited manual effort to create new entries as needed. The lexicon database is tightly linked with morphological annotation of selected text, so that we can measure the coverage in terms of both type count and token count. Manual annotation to supplement lexical entries is driven primarily by token frequency, to ensure that the coverage target can be reached as quickly as possible.

Language packs also include grammatical sketches intended to convey practical information about how to work with the language, focusing on paradigms and basic grammatical descriptions over deep theoretical discussions or nuanced explication of exceptional cases. Sketches for all languages follow a single template, so that the same topics are covered across languages and can be found in predictable sections within the sketch. Issues impacting annotation are documented first, addressing questions like: *Are determiners attached to nouns? Is there white space around case markers and adpositions? Describe adjectival forms of LOC, ORG, and GPE names such as "American".* The answers to annotation-relevant questions are passed on to annotation teams for use in guidelines development, and are added to one of the eight chapters of the grammatical sketch:

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

The grammatical sketch template also includes suggestions to the sketch author for the relative level of effort and approximate number of pages for each chapter, with a targeted page length of around 50 pages per language. Sketches are typically authored by theoretical linguists in consultation with native speakers, and are independently reviewed for structural and content completeness and cohesion prior to distribution in the RL packs.

3. Incident Language Packs

In addition to the representative language packs, we will produce language packs for 12 incident languages (IL) over the course of the program, with one IL per year designated for evaluation and the remainder to be used for system development. IL packs are intended to reflect the kind of data that might be available at the outbreak of an incident involving a low-resource language. Compared to representative language packs, IL packs contain smaller volumes of monolingual text and found parallel text, plus an assortment of grammatical resources. Additional evaluation data is created for one language per year. LORELEI performers are evaluated annually on a variety of component tasks; in 2016, evaluations include Machine Translation, Named Entity Recognition and Topic Labeling. LORELEI task evaluations are conducted by the National Institute of Standards and Technology (NIST). A corresponding open evaluation campaign, LoReHLT, allows non-LORELEI performers to participate in the same evaluations using the same data conditions (NIST 2016).

The LORELEI evaluations include three different checkpoints at which system output is delivered to NIST. The amount of data available for system training and development prior to each checkpoint varies, as does the amount of time between checkpoints. Prior to the start of the evaluation, performers receive an encrypted evaluation Incident Language pack. The IL pack components described in Section 3.1 are de-encrypted at the start of the evaluation, as is the monolingual IL test set described in Section 3.2. Additional monolingual IL text that post-dates the specific evaluation incident is de-encrypted just after the first checkpoint, with more post-incident IL monolingual text as well as post-incident English monolingual text de-encrypted after the second checkpoint. Gold standard annotations on the test set are used by NIST for scoring system submissions, and remain blind to performers until after the evaluation.

3.1 Common IL Components

All Incident Language packs, whether evaluation or development languages, share four components designated for system development and training.

First, IL packs include a minimum of 225,000 of monolingual text, nominally comprising 100Kw of news, 75Kw of blogs, discussion forums or other informal text, and 50Kw of microblogs. Some ILs may have no available microblog text, in which case the other genres will be increased. In most IL packs the amount of monolingual text is exceeded by 500% or more.

Next, IL packs include 300,000 of found parallel text, equally divided between news, informal text and microblogs. It is important to note that this is *found* parallel data harvested from the web; no manual or crowdsourced translations are included in the IL packs. When sufficient volumes of found parallel text are not available, larger volumes of comparable text are provided.

Third, all IL packs include a found IL-English dictionary containing at least 10,000 lemmas. This parallel dictionary is not a full-fledged lexicon of the type created for the Representative Language packs (Section 2.5), and the quality and structure of this component is expected to be highly variable across languages.

Finally, all IL packs include a set of found grammatical resources. There are eight types of allowable resources for this category, of which at least five must appear in all incident language packs; these include monolingual IL, regional English or bilingual IL-English gazetteers, bilingual IL-English or monolingual IL grammars, monolingual IL dictionaries or dictionaries that are parallel with a language other than English, and monolingual IL primers. Where possible we harvest the resource for direct inclusion in the IL packs, or provide a URL pointer where the resource can't be harvested and redistributed. For ILs that lack available digital or online resources of this type, we acquire hardcopies for distribution to performers concurrent with distribution of the IL packs.

3.2 Evaluation-Only IL Components

Creation of the evaluation data begins with identification of a specific real-world incident for the evaluation IL, e.g. a recent disaster that took place in the region where the language is primarily spoken. Using the selected evaluation incident as a guide, we produce a 200,000-word monolingual test set in the evaluation IL, with half of the data drawn from news and half drawn from informal text and/or microblogs. Approximately half of this monolingual test set is LORELEI domain-relevant, with a portion of that domain-relevant text discussing the specific evaluation incident. (Note

that all of the monolingual and parallel text described in Section 3.1 must pre-date the evaluation incident, since that data is to be used for system development and training.)

A portion of the monolingual test set (75Kw total, divided across the genres and domains) is manually translated, with 4 independent translations per document. This provides reference data for evaluation of machine translation technology.

The entire 200Kw test set is also labeled for the Simple Named Entity annotation task, using the same guidelines that are used in the Representative Language packs. This provides reference data for evaluation of named entity recognition. A portion of this test set is independently labeled for SNE by two separate annotators, to provide baseline information about human agreement.

3.2.1. Situation Frame Annotation for Topic Labeling Evaluation

Finally, a portion of the 200Kw test set is labeled for Situation Frames, providing reference data for evaluation of topic labeling. For each document, annotators create one or more situation frames describing "actionable" situations discussed in the document. Annotators label three information elements per Situation Frame: a situation type drawn from a fixed inventory, with one type per frame; a localization of the situation (limited to named entities in Year 1); and any sentiment, emotion or cognitive states relevant to that situation.

There are roughly a dozen Situation Frame types defined in Year 1, covering a range of possible incident types. For instance, the type "Infrastructure" is defined as *Any issue involving buildings, roads, bridges, facilities, or other permanent or semi-permanent physical infrastructure that has been damaged or made non-operational*; when this kind of incident is detected the expected action from a mission planner might be to send building materials or equipment to the scene. Annotators begin Situation Frame annotation by identifying all the taggable situations in the document and assigning each one a type. Multiple mentions of the same event/situation are part of same frame, but more than one frame with same situation type is possible. For instance, if the document discusses a landslide that destroyed critical infrastructure in 2015, and compares it with an earthquake that also destroyed infrastructure in 2012, the annotator would produce two separate situation frames of the type "Infrastructure". Completed, planned and near-future events can generate Situation Frames, but imagined hypothetical events cannot. This means that speculation about the chances of a meteor impact destroying all the bridges in New York City does not yield an Infrastructure Situation Frame, but expressions of concern over likely destruction of bridges in a town just hit by an earthquake does.

After identifying all the taggable frames in the document, the annotator then adds Sentiment/ Emotion/Cognitive State (SEC) attributes to the frame as needed to reflect SEC that is conveyed or expressed in the document about the situation. There are three possible attributes for SEC, outlined in Table 2 below.

SEC	Description
Positive	Approval, support, happiness, etc.
Negative	Disapproval, antagonism, displeasure, distress, etc.
Activation	Excitement, intensity, urgency, etc.

Table 2: SEC Attributes for Situation Frames

For instance, given the following sentence:

Local officials told CNN that they feared the town would experience significant damage to roads, bridges and other transportation infrastructure in the flood-affected areas.

annotators would add a Negative SEC attribute to the Infrastructure situation frame, to reflect the fear expressed by the officials about this situation.

Finally, annotators "Localize" the situation frame by linking any entity associated with the situation to the frame. In Year 1, localization is limited to entities that are named somewhere in the document. So in the example above, even though "the town" isn't named within the sentence, as long as it is named somewhere in the document then that entity (which encompasses both "the town" and the named mentions of the town) is linked to the situation frame, providing information about the location of the Infrastructure situation.

Language-specific guidelines for Situation Frame annotation include detailed examples and rules of thumb for dealing with common challenges, for instance how to handle cases where multiple nested Geopolitical and Location Entities are mentioned in connection with the same situation, as in the following sentence:

Eyewitnesses said a landslide hit the village of Guinsaugon in the south of the Philippine island of Leyte. Governor Rosette Lerias described the village as totally flattened with virtually all of housing destroyed.

In this example, the village of Guinsaugon (GPE) is located on the island Leyte (LOC) in the country Philippines (GPE). The guiding rule for annotation is that annotators should always tag the most specific entity that is associated with the situation, so in this case the Guinsaugon entity is associated with the situation frame, while the Leyte and Philippines entities are not.

As with the SNE evaluation data, a portion of the Situation Frame evaluation data is independently labeled by two separate annotators to provide baseline human agreement numbers.

4. Distribution of Language Packs

All completed language packs (both RL and IL) are subject to sanity checks and validation at LDC, followed by independent quality control by the University of Maryland Center for Advanced Study of Language (CASL), prior to their release to LORELEI performers.

Checks on monolingual text include language ID verification using standard character n-gram based methods, automated dictionary lookup and spot checking by native speakers. Character sets are validated against a list of known valid code points for language. Manual spot-checking by linguists and/or native speakers will also be used to identify any systematic issues with language packs concerning content of domain text; tokenization or segmentation; parallel text accuracy and fluency; consistency of POS and morph tagsets; and annotation quality. Grammatical sketches are also validated for stylistic and content issues.

To date we have completed and distributed Representative Language packs for Turkish, Hausa and Uzbek, and Incident Language packs for Uzbek and Mandarin. Because (parts of) of these language packs were created under BOLT they do not always reflect the current LORELEI requirements. Some components (e.g. Simple Named Entity annotations) will be updated to reflect the current guidelines, while other components (e.g. SSA and NP annotation) will remain slightly out of alignment with current standards.

Representative and development language packs are delivered to LORELEI at the end of each program year. Representative language packs are also deposited in the LDC Catalog as they are completed, while incident language packs are published after they are no longer sequestered for use in LORELEI or LoReHLT evaluations. LORELEI Performers and LDC members will receive language packs at no cost. Members of the general research community will pay a minimal fee to defray the costs of data curation, storage and distribution. All deliverables are provided to the government under LDC's existing government-wide license. The first set of LORELEI language packs is expected to appear in LDC's catalog in late 2016.

5. Acknowledgements

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.

6. References

Alison Alvarez, Lori Levin, Robert Frederking, Erik Peterson, Simon Fung. 2006. Tools for Elicitation Corpus Creation. Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art. Lansing, MI. June 20-22.

CDL2 website, retrieved October 25, 2015. https://github.com/CLD2Owners/cld2

DARPA LORELEI website, retrieved October 25, 2015. http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee. 2012. Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT. Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 21-27.

IARPA Babel website, retrieved March 8, 2016. http://www.iarpa.gov/index.php/research-programs/babel

Tibor Kiss, Jan Strunkt. 2006. Unsupervised multilingual sentence boundary detection. Computational Linguistics 32: 455-525

Seth Kulick, Ann Bies. 2016. Rapid Development of Morphological Analyzers for Typologically Diverse Languages. Proceedings of LREC 2016: 10th International Conference on Language Resources and Evaluation, Portorož, May 23-28.

Xiaoyi Ma. 2006. Champollion: A Robust Parallel Text Sentence Aligner. Proceedings of LREC 2006: 5th International Conference on Language Resources and Evaluation, Genoa, May 22-28.

Xiaoyi Ma, Mark Liberman. 1999. BITS: A Method for Bilingual Text Search over the Web Machine Translation Summit VII: Singapore, September 13-17.

NIST LoReHLT 2016 Evaluations website, retrieved March 8, 2016. http://www.nist.gov/itl/iad/mig/lorehlt16.cfm

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. Transactions of the Association for Computational Linguistics 2: 79-92.

Slav Petrov, Dipanjan Das, Ryan McDonald. 2011. A Universal Part-of-Speech Tagset. Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 21-27.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, Boyan Onyshkevych. 2008. Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned toward Creation of Basic Language Resources. Proceedings of LREC 2008, Workshop on Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages.