Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 277–284
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Basic Language Resources for 31 Languages (Plus English): The LORELEI Representative and Incident Language Packs

Jennifer Tracey, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
garjen@ldc.upenn.edu, strassel@ldc.upenn.edu

Abstract

This paper documents and describes the thirty-one basic language resource packs created for the DARPA LORELEI program for use in development and testing of systems capable of providing language-independent situational awareness in emerging scenarios in a low resource language context. Twenty-four Representative Language Packs cover a broad range of language families and typologies, providing large volumes of monolingual and parallel text, smaller volumes of entity and semantic annotations, and a variety of grammatical resources and tools designed to support research into language universals and cross-language transfer. Seven Incident Language Packs provide test data to evaluate system capabilities on a previously unseen low resource language. We discuss the makeup of Representative and Incident Language Packs, the methods used to produce them, and the evolution of their design and implementation over the course of the multi-year LORELEI program. We conclude with a summary of the final language packs including their low-cost publication in the LDC catalog.

Keywords: low resource languages, machine translation, information extraction, situational awareness

1. Introduction

The DARPA Low Resource Languages for Emergent Incidents (LORELEI) Program was a multi-year research program aimed at improving the utility of language technology in the context of rapidly emerging incidents like natural disasters. Language technology has the potential to provide crucial situational awareness to those responding to a crisis. For instance, during a disaster the affected population may turn to social media or use text messaging to report urgent needs and issues. When entity taggers and machine translation systems exist for the language(s) spoken in the disaster zone, this data can be processed to generate "heat maps" showing English-speaking mission planners what kind of help is needed where. Most of the world's languages are considered "low resource" when it comes to language technology (META-NET, 2012), and the unpredictable nature of disasters and other emergent situations means that it is not feasible to make language-specific investments in the development of technologies for every language that might become suddenly important in an emergency. LORELEI was designed to address this challenge by building technologies that can be rapidly transitioned to new languages with little to no new training data, for instance by exploiting language universals and using cross-language projection techniques. Evaluation of LORELEI systems was carried out under the NIST Low Resource HLT (LoReHLT) evaluation program (https://www.nist.gov/itl/iad/mig/lorehlt-evaluations); results of the evaluation are not discussed in this paper.

In order to support the research and evaluation requirements of the LORELEI Program, the Linguistic Data Consortium (LDC) created linguistic resources for text data in 31 languages (plus English)[1]. Corpora created for LORELEI consisted of two types of language packs: representative language packs for system training and development, and incident language packs for system evaluation. This paper describes the completed set of LORELEI language packs across all 31 languages. After presenting the design plan for representative and incident language data, we describe each component of the language packs in detail, including methods used to collect and annotate the data, the evolution of requirements and approaches and over the life of the program, and final results of the data creation effort.

2. Representative and Incident Languages

The 24 Representative Language (RL) Packs created for LORELEI provide basic text resources for system training and development, including large volumes of monolingual and parallel text, several types of annotation designed to support situational awareness, a lexicon, a grammatical sketch, and basic text processing and NLP tools. The seven Incident Language (IL) Packs were created specifically for system evaluation and include very limited amounts of monolingual and parallel text for system adaptation, pointers to known lexical and grammatical resources for the language, and a small set of reference translations and annotations used for scoring.

2.1 RL Pack Content Overview

By design, RLs did not provide training data for specific LORELEI evaluation tasks and languages, Instead, they are intended to enable cross-language and transfer learning methods as well as research into using language universals. The 24 RLs were carefully selected prior to the start of the program, with an eye to covering a broad range of language families, typological characteristics and geographic regions. They include relatively high resource languages (e.g., Spanish, Russian, Mandarin) as well as some very low resource languages (e.g., Wolof, Akan). Due to the need to collect significant amounts of digital text data, languages in which most speakers do not regularly read and write were not selected as either RLs or ILs.

[1] A small speech corpus was also created for each LORELEI language by another organization; those resources are not

RL Packs were designed to include the following components[2]:

- At least 2M words of monolingual text
- At least 900K words of parallel text
- 100Kw of English translated into the RL
- 10Kw of noun phrase chunking
- 75K words of named entity annotation
- 25K words of full entity annotation (names, nominals, pronouns and coreference)
- 25K words of simple semantic annotation (light semantic role labeling)
- 25K words of situation frame annotation (an information extraction task aimed at "situational awareness" in disaster settings)
- A lexicon containing at least 10K lemmas
- A grammatical sketch
- Sentence segmenter, tokenizer, named entity tagger, transliterator and other basic tools

Some of the RL Packs also included manual morphological analysis, morphological alignment and/or part of speech tagging, but these were not required elements. Later iterations of the RL language packs dropped the NP chunking task and added entity linking, where labeled entity mentions were linked to an external knowledge base. Three of the RL language packs were produced prior to program kickoff and were made available to performers at the start of LORELEI, while the remaining RLs were distributed incrementally over the course of the program. The types of RL data and annotation are described in more detail below.

2.2 IL Pack Content Overview

LORELEI ILs were selected by DARPA from a set of candidates developed by LDC and were used to test system capabilities. Because the LORELEI use case requires systems to quickly transition to a previously unseen language during an emergent situation, the identity of the incident languages for each year's evaluation were concealed until the start of the evaluation, and results were due at a series of "checkpoints", the first of which was as soon as 24 hours after the release of the evaluation data. In general, ILs were lower resourced than (most of) the RLs, but they still needed to have sufficient volumes of digital text available to support the evaluation requirements, and have a large enough pool of accessible native speakers who could be trained to produce the gold standard annotations used in scoring. ILs also had to have a related RL language pack (e.g. Ilocano IL ~ Tagalog RL). The first LORELEI evaluation included one IL, while all subsequent evaluations included two ILs. IL Packs were designed to contain the following components.

1) Data released to performers at the start of each evaluation for use in for system adaptation and training, consisting of the kinds of found resources that might be discoverable at the outset of an incident, but contains no annotated training data:

- At least 1.3 million words of monolingual text
- At least 300Kw of found parallel text
- Pointer to a 10Kw found IL-English dictionary
- Pointers to at least 5 of the following types of grammatical resources:
  - o English Gazetteer for Region
  - o Parallel IL-English Gazetteer
  - o Monolingual IL Gazetteer
  - o 10Kw IL-Non-English Dictionary
  - o Monolingual IL Primer
  - o Parallel IL-English Grammar
  - o Monolingual IL Grammar
  - o 10K Monolingual IL Dictionary

2) Test data processed by the LORELEI systems

- 200Kw per IL[3], comprising data from multiple genres focused on one or more real-world incidents in the region where the language is used; incidents were selected based on having sufficient coverage in digital text and giving rise to the types of needs and issues covered by the Situation Frame annotation task (see section 4.3).

3) Reference annotations and translations used in scoring (not released to system developers)

- 75Kw professional quality translation
- 50Kw Simple Named Entity annotation
- 50Kw Situation Frame annotation

IL language packs for year 2 and beyond also contained Entity Linking reference annotation.

Table 1 lists all of the RL and IL languages.

Pre-LORELEI RLs	Hausa, Turkish, Uzbek
RLs	Akan, Amharic, Arabic, Bengali, Farsi, Hindi, Hungarian, Indonesian, Mandarin, Russian, Somali, Spanish, Swahili, Tagalog, Tamil, Thai, Ukrainian, Vietnamese, Wolof, Yoruba, Zulu; English (partial, annotation only)
ILs	Ilocano, Kinyarwanda, Odia, Oromo, Sinhala, Tigrinya, Uyghur

Table 1 : All LORELEI Languages

3. Monolingual and Parallel Text

The collection approach for monolingual text was the same across both RLs and ILs, while the approaches taken to translation varied slightly between RL and IL packs, due to evaluation requirements for the ILs.

3.1 Monolingual Text

Monolingual text was collected using a combination of manual and automatic methods. Native speakers for each

[2] While there are resources available for some of the RLs, for example, through the Universal Dependencies Project at https://universaldependencies.org, LORELEI language packs were designed to have a uniform set of contents across all languages; researchers in the LORELEI program were free to make use of such existing resources, but they were out of scope for inclusion in the language packs produced by LDC for LORELEI.

[3] In the last two LORELEI evaluations, systems also processed a corresponding 50Kw English set for each IL, and were scored against English references in addition to IL references.

language were tasked with finding documents in their language on the web, focusing largely on events in the LORELEI domain (natural disasters and other kinds of emergent situations) in both formal (e.g., news reports) and informal (e.g., blogs, discussion forums, Twitter) genres. For RLs, they searched for a variety of event types and there were no constraints on the date of the material collected. For ILs, target incidents were selected in advance for each language, and searches focused on documents about those specific incidents. During data scouting we also recorded information about individual documents that would help inform data selection for downstream tasks: genre, incident type, incident name, incident Wikipedia page link if available, country/region/location, date, presence of eyewitness accounts, how specific the document was about people/places/organizations/dates and times, and topics (e.g. evacuation, food, shelter).

In addition to collecting individual documents identified by native speakers, wherever possible other documents from the same website were also harvested to provide a pool of background data that was not specifically about LORELEI domain topics. All collected documents were then processed into a uniform tokenized and sentence-segmented xml format to support downstream annotation and evaluation pipelines. LDC acquired rights to use and distribute third party data by a combination of explicit permission from the data provider, a compatible license attached to the data by a data provider, or under the doctrine of fair use which is supported by a fair use analysis commissioned by LDC and conducted by independent counsel. For a handful sources including Twitter, pointers to the source document were provided along with utilities for interacting with the data provider's API to download the documents and convert them into the format used by LDC for annotation.

Because we were collecting tens or even hundreds of millions of words of data in some languages, manual review of every document to ensure that it was in the target language was not possible. Instead, all documents selected for translation and annotation were manually audited for language and other features, while the rest of the monolingual text was subject to automatic language identification using Google CLD2 to filter out documents which were clearly out-of-language.

Monolingual text data volumes included in the corpora vary widely. For ILs, all language packs exceeded the 1.3 million word target. Kinyarwanda included 10 million words of monolingual text, and Uyghur included 27 million words. All of the remaining ILs had between 3 and 7 million words of monolingual text. For RLs, only Wolof fell short of the 2 million word goal. Tagalog, Swahili, Zulu, Akan, Arabic, Yoruba, and Hausa all exceeded the goal but had less than 10 million words of monolingual text. Thai, Tamil, Hindi, Indonesian, Spanish, Somali, Amharic, Mandarin, and Uzbek all exceeded 10 million words, and Bengali, Russian, Vietnamese, Hungarian, Farsi, Ukrainian and Turkish all exceeded 100 million words of monolingual text.

3.2 Parallel Text

Parallel text resources in the LORELEI language packs were produced using three methods: found parallel text, crowd translation, and professional translation. For ILs, found parallel text was provided in the training partition of the language pack, while professional translation was provided in the evaluation references. For RLs each pack contained a mixture of found, crowd, and professional translation, depending on the availability of found parallel text and crowd workers for the language. The goal for RLs was to first maximize the amount of found translations, then use crowdsourcing if viable, and only use professional translation as much as required to make up the remaining amount. Therefore, the RL packs differ considerably in the proportion of professional, crowd and found translation.

3.2.1 Found Parallel Text

Parallel text was harvested from the web using a combination of approaches. When possible, native speaker annotators provided information about sites containing parallel text in their language. We also used Bilingual Internet Text Search (BITS) (Ma and Liberman, 1999) to locate additional data by scanning potential parallel sites and using a translation lexicon to identify pairs of documents that are translations[4]. Champollion (Ma, 2006) was then used to align the documents at the sentence level. In some cases, the amount of parallel text harvested far exceeded the total parallel text target.

3.2.2 Crowdsourced Translation

LDC's goal was to use crowdsourcing wherever possible for RLs. In the first year of the program, 10 of the RLs were scheduled for active collection. We posted hits using Amazon Mechanical Turk (www.mturk.com), but only had good yield for Spanish, Arabic, and Russian. For the second year, we focused our crowdsourcing efforts on Hindi and Bengali, for which we expected to find a reasonable number of crowd workers, and we switched to using CrowdTrans (https://crowdtrans.com/), a platform first developed under LORELEI and designed to allow the workflow to be customized to LORELEI requirements. Sentences that received at least three crowd translations were subjected to a ranking task in which qualified workers gave a best-to-worst ranking of the translations. Sentences with fewer than three translations were included, but did not have the ranking information.

3.2.3 Professional Translation

Professional translation was used to some degree for every language pack. For ILs, translation references used in scoring were all professional quality translations, with 4 independent translations produced for the first year's evaluation, 2 in the second year, and one in the third and fourth years. For the RLs, some data was professionally translated even for languages where the found parallel text was sufficient to meet or exceed the target volume. This was to ensure that an adequate amount of LORELEI domain material was translated for each package and to ensure appropriate genre distribution in the translation data, since the topic content and genre of found parallel text was not necessarily aligned with the targets for the LORELEI language packs. In addition to the data

[4] BITS was used rather than more recently developed tools since it was already fully integrated into LDC's collection infrastructure and its results were sufficiently good to satisfy requirements.

translated into English for each language, RL packs also contain a "core" set of English documents that were translated into each RL, so that when all RL packs are combined, these documents form a 25-way parallel corpus (24 RLs plus English). The core English set consists of approximately 80,000 words of news text plus a short phrasebook of conversational sentences and an elicitation corpus designed to highlight various grammatical and morphological features of languages for a total of approximately 100,000 words. A small set of the translated documents from this core set (2000 words) was annotated for each annotation task in every language, so that there is also a small multi-way parallel annotation dataset across all the RLs.

For ILs, two languages (Tigrinya and Oromo) fell short of the goal of 300,000 words of parallel text, and while all other ILs had between 2 and 6 million words of parallel text, much of that was religious text rather than news, informal web text, social media, or other target genres. For RLs, Amharic and Wolof fell below the target volume of 900,000 words of parallel text, while Thai, Bengali, Zulu, Akan, Spanish, Russian, Vietnamese, Hungarian, Uzbek, and Farsi had between 1.3 million and 8.6 million words. All other RLs hit the target volume or exceeded it by 100,000 words or less. These volumes include the combined total of found, crowdsourced, and professional translations; where languages greatly exceeded the target volume, it was due to large amounts of found parallel text.

4. Annotation

In addition to monolingual and parallel text, each language pack contains several types of annotation, described in the following sections.

4.1 Entity Annotation

Three types of entity annotation were performed: simple named entity, full entity and entity linking.

4.1.1 Simple Named Entity and Full Entity

The most basic of the entity annotation tasks for LORELEI is Simple Named Entity (SNE), in which annotators label all the named person, organization, location/facility, and geopolitical entities in a document. All annotations were grounded in text extents. The output of Simple Named Entity annotation serves as input to the subsequent Entity Linking and Situation Frame annotation tasks.

In the Full Entity (FE) task, annotators label entities of the same types labeled for SNE, but in addition to named mentions, all nominal and pronominal mentions are tagged as well. There is also an additional category of title, which is used to capture professional or honorific titles of persons. All annotations were grounded in text extents. FE also includes within document coreference of entity mentions.

All RLs hit the target volume for SNE and FE annotation except for Hausa, which had the full amount of SNE but only 13,000 words of FE due to limited availability of annotators. All ILs met the target volume for SNE annotation.

4.1.2 Entity Linking

In the final entity task, Entity Discovery and Linking (EDL), entity mentions labeled in either SNE or FE[5] are linked to a Knowledge Base (KB) developed especially for use in the LORELEI Program by merging three existing knowledge resources and then performing manual augmentation. Most LORELEI KB entities come from the Geonames[6] database, which contains millions of entries for geographical names, along with information about the places, such as population, latitude and longitude, etc. The CIA World Factbook World Leaders List[7] is the source for person entities, and includes cabinet-level leaders from countries around the world. The CIA World Factbook Appendix B[8] consists of a list of international organizations and groups such as trade organizations, economic development groups, etc. While the Geonames entries cover a huge number of locations and geopolitical entities, the number of persons and organizations covered by the CIA World Factbook sources are quite small, and in order to have better coverage of entities likely to appear in the annotated data, manual augmentation of the KB was performed for each language. New person and organization entities were added to the KB for ILs based on their relevance to the specific incidents selected as the focus of the evaluation. For RLs, since there was no focus on a specific incident, new entities were added to the KB based on frequency of appearance in the annotated SNE data.

During Entity Linking annotation, annotators checked the KB against each entity labeled during the prior SNE (or FE) task. When that entity was found in the KB they created a link by associating the KB ID with the entity ID. In some cases, ambiguity in the labeled data or in the KB itself made it impossible for annotators to determine a single correct KB link; in these cases they created links to all possibly correct KB entries. If no match for the entity was found in the KB, the annotator marked it as a "NIL" entity. After annotation on all documents was complete, the NIL entities were examined to see if any of them were coreferent, and all coreferent entities were clustered together under a new unique KB ID.

Entity linking was added in the second year of the program, and so language packs completed prior to that point (Uzbek, Turkish, Hausa, Somali, Yoruba, Uyghur) do not contain any Entity Linking annotation. All other RLs and ILs have the target volume of annotation.

4.2 Simple Semantic Annotation

In Simple Semantic Annotation (SSA), documents were labeled for predicates (Acts and States) and their basic arguments (Agents, Patients, and Locations). Existing predicate-argument based semantic annotation protocols such as Abstract Meaning Representation (Banarescu et al., 2013) or PropBank (Palmer et al., 2005) are fairly

[5] In the year 3 evaluation, EDL was performed on a set of English documents labeled for Full Entity, such that both named and nominal entity mentions were linked to the KB. In all other data sets, EDL used only name mentions from SNE.

[6] www.geonames.org

[7] www.cia.gov/library/publications/world-leaders-1/

[8] www.cia.gov/library/publications/the-world-factbook/appendix/appendix-b.html

complex and require annotators to have a background in linguistics and/or require a substantial amount of annotator training time.

SSA Annotation : Physical Acts

Breaking News: Landslide hits Guinsaugon in the south of the Philippine island of Leyte. Reports say village totally flattened and housing destroyed.

Act : landslide
- Patient : Guinsaugon
- Place : Leyte
- Place : Phillipine
- Place : south

Act : hit
- Agent : landslide
- Patient : Guinsaugon

Act : flattened
- Agent : landslide
- Patient : Guinsaugon

Act : destroyed
- Agent : landslide
- Patient : housing
- Place : Guinsaugon

Figure 1: SSA Annotation Example

In contrast, SSA was designed specifically for LORELEI as a way to get some light semantic role labeling from non-expert annotators with limited time available for training. In order to scope the annotation task (both to constrain the amount of effort and to focus that effort on the most LORELEI-relevant parts of the annotated documents), only physical acts and disaster-relevant states were annotated. Each predicate that falls into one of these categories was first labeled as either an act or state, and then the agent, patient, and location arguments were added.

All annotations were grounded in text extents, and the scope of annotation was the sentence, with the full document available to annotators for context. Figure 1 shows an example of physical acts annotated for SSA.

Ukrainian does not have any SSA, and Hausa has less than the target volume; all other RLs met the target volume for this task. No SSA was performed on any IL.

4.3 Situation Frame

The Situation Frame (SF) annotation was a new task designed specifically for LORELEI; its purpose was to test system capabilities for providing situational awareness about the types and locations of needs (e.g. water, infrastructure) and issues (e.g. civil unrest) present in the source data, as well as the urgency and scope of the situation and any entities involved in reporting or resolving the situation. During Situation Frame annotation native speakers read each document and indicated whether any of needs or issues from a fixed list were present in the document, and when present, they filled in the rest of the situation "frame" with information about location, urgency, status, and so on. Locations and other entities involved in the situation were selected by the annotator from the list of names annotated during SNE. Annotations for this task are produced at the document level and do not involve anchoring the elements of a situation frame in a text span. Figure 2 shows an example of a labeled need frame.

SF Annotation : Need Frame

Breaking News: Landslide hits Guinsaugon in the south of the Philippine island of Leyte. Reports say village totally flattened and housing destroyed.

Need
- Type(s): Shelter, Infrastructure
- Place : Guinsaugon
- Status : Current
- Scope: Large
- Severity : High
- Reported by : N/A
- Resolved by : N/A

Figure 2: SF Annotation Example

SF Annotation	Y1: Uyghur	Y2: RLs, Tigrinya, Oromo
Reference Annotations	1	RLs: 1 ILs: 3
Types	8 need types 3 issue types	Tweak definitions
Place Values	NAM	NAM
Resolution/ Status Values	3	2
Default Issue Status	None	Currently Relevant
Urgency	Binary	Binary

SF Annotation	Y3: Kinyarwanda, Sinhala	Y4: Odia, Ilocano
Reference Annotations	ILs: 2 Eng: 3	ILs: 2+ Eng: 3
Types	No Change	No Change
Place Values	NAM + NOM (Eng)	NAM
Resolution/ Status Values	No Change	No Change
Default Issue Status	No Change	No Change
Urgency	Scalar: Scope & Severity	No Change

Table 2: SF Task Evolution

The types used for labeling situation frames were developed for LORELEI in consultation with program stakeholders, and were inspired by types used in real-world disaster response projects such as MicroMappers (Imran et al., 2014). Need types include Evacuation, Infrastructure, Food Supply, Medical Assistance, Search/Rescue, Shelter, Utilities/Energy/Sanitation and Water Supply. Issue types include Civil Unrest/ Widespread Crime, Regime Change and Terrorism or other Extreme Violence.

The SF task evolved significantly over the life of the LORELEI program to address new requirements and to improve the utility of the labeled data for evaluation. For example, initial versions of the task included a binary judgment of whether or not a situation was urgent, but poor inter-annotator agreement on this decision led to a change in which urgency was decomposed into scalar judgments about the scope (size of affected region/population) and severity (low = inconvenience; high = death) of the situation. Table 2 shows the elements of the SF annotation task over the four years of the LORELEI program.

Situation Frame was defined as a task after the LORELEI program was already underway, and so language packs completed prior to that point contain no SF annotation (Turkish, Uzbek, Hausa, Somali, or Yoruba). All other RLs and ILs met the target volumes for Situation Frame annotation.

4.4 Morphosyntactic Annotations

Several types of morphological and syntactic annotations were performed on subsets of the RLs. These annotations were intended to serve as resources for system development and were not directly evaluated, so they only appear in RLs. In response to input from DARPA and LORELEI performers, most morphosyntactic annotation was dropped after the first set of language packs in favor of greater effort elsewhere (e.g. Situation Frame), though one new morphological annotation task was added in subsequent phases of the program.

The first type of morphosyntactic annotation was Noun Phrase Chunking, which was originally planned for all RLs, but was later dropped in favor of additional EDL and SF annotation on the RLs. In this task, annotators identified maximal, non-overlapping noun phrases in the text, following surface syntactic structure and applying constituency tests to determine where to mark the boundaries of the noun phrases. Noun Phrase chunking was performed on Uzbek, Turkish, Hausa, Mandarin, Amharic, Somali, Farsi, Hungarian, Vietnamese, Yoruba, Russian, Spanish, and Arabic (10,000 words per language).

In the three languages for which resources were developed prior to the start of LORELEI (Turkish, Uzbek, and Hausa), morphological analysis and alignment were also performed, tightly coupled with the creation morphological analyzers for each of these languages (Kulick and Bies, 2016). In this task, annotators were presented with a list of possible solutions from the analyzer and were required to select the best one from the list, or choose "unanalyzable" if no correct solution for the token was present. Part-of-speech labels were not directly annotated, but were instead derived from the morphological annotation. For Turkish and Uzbek, morpheme alignment between the Turkish/Uzbek text and an English translation was also performed to identify translational correspondence using the same general approach applied in previous word alignment annotation tasks (e.g., Li et al., 2010) but adapted to align the translations at the morpheme level rather than the word level. These task were dropped from the plan for RLs at the very start of LORELEI and so are not present in any of the subsequent language packs.

However, in collaboration with other LORELEI researchers, some additional morphological segmentation annotation was performed on 9 languages toward the end of the program (Mott et al., 2020). The languages selected for this task include a variety of morphological features of interest such as case marking and noun class systems, infixes, circumfixes, etc. For Akan, Hindi, Hungarian, Indonesian, Russian, Spanish, Swahili, Tagalog, and Tamil, 2000 tokens per language were segmented at morpheme boundaries, and markup was added to indicate substitution (as in *come/came* in English). Due to the difficulty of this type of task for non-expert annotators, the segmentation was performed by a trained linguist working in tandem with a native speaker annotator.

5. Lexical and Grammatical Resources

For each of the RLs, LDC developed a lexicon containing at least 10,000 entries, with part of speech and English glosses. The lexicon was created using a combination of found resources such as existing online dictionaries, and manual effort by native speakers to create new entries. Manual effort focused on adding high frequency tokens that were missing from the found resources, and the amount of manual annotation required varied significantly by language. Lexicons were augmented for several languages by integrating the detailed morphological analysis information in a separate word forms table that is indexed to the entries in the main lexicon. For Arabic, this information was extracted from the Penn Arabic TreeBank (Kulick, et al., 2010); for Amharic, Farsi, Hungarian, Russian, Somali, Spanish and Yoruba, morphological information comes from the Unimorph Project (Kirov, et al., 2018).

Each RL pack also includes a grammatical sketch created specifically for LORELEI. The grammatical sketches provide basic linguistic information including paradigms and brief grammatical descriptions. The sketches were intended to contain the kind of practical information that would be useful in working with the language, rather than the kind of deep theoretical analysis and description of exceptional cases that might be found in a full length academic grammar. A single template was used for all languages so that each sketch includes basic information about the language (e.g., classification, ISO code, word order), orthography, encoding (Unicode chart, etc.), morphology, syntax, and specialized sections for personal names, locations, numbers, and variation, as well as references to in-depth grammars.

For ILs, no LORELEI-specific lexicon or grammatical sketch was produced, but pointers to several types of existing lexical and grammatical resources (monolingual and parallel dictionaries, grammars, and gazetteers) were provided.

6. Tools

Each LORELEI RL pack also includes a set of basic NLP tools and text processing utilities for working with the data. For all RL languages, tokenizers, sentence segmenters and named entity taggers were provided, along with utilities to download and process any data that could not be directly redistributed by LDC. For languages that use whitespace to delimit word boundaries, LDC provided a regular expression-based tokenizer designed to separate words and punctuation while preserving certain kinds of web-text artifacts, such as urls and hashtags, as single tokens. For languages that do not use whitespace at word boundaries, we turned to widely-used existing tokenizers. Sentence segmenters were created using an implementation of the Punkt algorithm based on the version found in NLTK (Kiss and Strunkt., 2006). Custom conditional random field-based named entity taggers were produced for each of the RLs and trained using the SNE annotations described above. For all RLs written in non-Roman scripts, a transliterator was also provided.

7. Quality Control and Validation

Quality control was performed by senior and lead annotators for each stage of collection, translation and annotation. Collection and translation QC involved manual spot checking on individual files, searching for outliers across the corpus, and running a suite of automated sanity checks to identify and correct known issues. Annotation quality control involved formal training and testing of annotators, detailed procedural guidelines for each task, and custom user interfaces for each task that included many types of validation to prevent certain types of error. All annotations were subject to review passes by senior annotators, and corpus-wide checks were conducted at the conclusion of each task. Inter-annotator agreement was measured for tasks/languages with a large enough annotator team, timeline and budget to permit dual annotation. For SNE, agreement for ILs fell within expected ranges given inexperienced annotators working in low resource languages (67-90%). For SF, agreement rates in the first year were poor (under 60%). Analysis of the data revealed that multiple correct answers existed given differences in use of inference and world knowledge, and a decision was made to utilize multiple human references in scoring rather than attempting to redesign the task for better agreement (Strassel et. al., 2017). Other tasks were not subject to IAA analysis given resource constraints. Finally, all RL and IL language packs were subject to independent review by the University of Maryland Center for Advanced Study of Language (CASL) prior to release.

8. Conclusion

We have presented two types of new resources developed by LDC for the DARPA LORELEI Program.

Representative Language Packs and Incident Language Packs are designed to enable creation and evaluation of systems capable of providing language-independent situational awareness in emergent incidents. All language packs have now been completed, covering 31 typologically and geographically diverse languages, plus English. Table 3 summarizes the contents of every final language pack.

KEY ✓ met target ▬ fell short ◎ not done	Incident Languages							Pre LORELE			X
	Uyghur	Tigrinya	Oromo	Kinyarwanda	Sinhala	Odia	Ilocano	Uzbek	Turkish	Hausa	English
Mono Text	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Parallel Text	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	■
POS Tagged	■	■	■	■	■	■	■	✓	✓	▬	■
NP Chunking	■	■	■	■	■	■	■	✓	✓	▬	■
Morph Analysis	■	■	■	■	■	■	■	✓	✓	▬	■
Morph Align	■	■	■	■	■	■	■	✓	✓	◎	■
Morph Seg	■	■	■	■	■	■	■	■	■	■	■
SNE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Full Entity	■	■	■	■	■	■	■	✓	✓	▬	■
Entity Linking	■	✓	✓	✓	✓	✓	✓	■	■	■	✓
SSA	■	■	■	■	■	■	■	✓	✓	▬	✓
Situation Frame	✓	✓	✓	✓	✓	✓	✓	■	■	■	✓

KEY ✓ met target ▬ fell short ◎ not done	Year 1 RLs										
	Amharic	Arabic	Farsi	Hungarian	Mandarin	Russian	Somali	Spanish	Ukrainian	Vietnamese	Yoruba
Mono Text	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Parallel Text	▬	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
POS Tagged	■	■	■	■	■	■	■	■	■	■	■
NP Chunking	✓	✓	✓	✓	✓	✓	✓	✓	■	✓	✓
Morph Analysis	■	■	■	■	■	■	■	■	■	■	■
Morph Align	■	■	■	■	■	■	■	■	■	■	■
Morph Seg	■	■	■	✓	■	✓	■	✓	■	■	■
SNE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Full Entity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Entity Linking	✓	✓	✓	✓	✓	✓	■	✓	✓	✓	■
SSA	✓	✓	✓	✓	✓	✓	✓	✓	■	✓	✓
Situation Frame	✓	✓	✓	✓	✓	✓	■	✓	✓	✓	■

KEY ✓ met target ▬ fell short ◎ not done	Year 2 RLs									
	Akan	Bengali	Hindi	Indonesian	Swahili	Tagalog	Tamil	Thai	Wolof	Zulu
Mono Text	✓	✓	✓	✓	✓	✓	✓	✓	▬	✓
Parallel Text	✓	✓	✓	✓	✓	✓	✓	✓	▬	✓
POS Tagged	■	■	■	■	■	■	■	■	■	■
NP Chunking	■	■	■	■	■	■	■	■	■	■
Morph Analysis	■	■	■	■	■	■	■	■	■	■
Morph Align	■	■	■	■	■	■	■	■	■	■
Morph Seg	✓	■	✓	✓	✓	✓	✓	■	■	■
SNE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Full Entity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Entity Linking	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SSA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Situation Frame	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 3: Final LORELEI Language Packs

The linguistic resources described here have been distributed to LORELEI performers and to participants in the NIST Open Low Resource Human Language Technologies (LoReHLT) evaluation (NIST, 2019). While some initial LORELEI resources have already been published by LDC, starting in early 2020 the final LORELEI language packs will begin appearing in the LDC catalog at the rate of 1-2 per month. Language packs are available to LDC members at no cost, while non-members pay a minimal fee to defray the costs of data curation, storage and distribution.

Taken as a whole, the LORELEI Language Packs comprise over 2.5 billion words of monolingual text, 60 million words of parallel text, 100 grammatical/lexical resources, and 3 million annotation decisions by over 250 native speakers.

9. 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. The authors gratefully acknowledge Lori Levin, Chris Callison-Burch, Song Chen, Ann Bies, Kira Griffitt, Justin Mott, Dana Delgado, Michael Arrigo, Neil Kuster, Dave Graff, Seth Kulick, Neville Ryant, Jonathan Wright, Brian Gainor, Christopher Caruso, Alex Shelmire, University of Maryland Applied Research Laboratory for Intelligence and Security (ARLIS), formerly UMD Center for Advanced Study of Language (CASL), and the hundreds of LORELEI native speaker annotators for their contributions to this research.

10. Bibliographical References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse.

Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S. (2014). AIDR: Artificial Intelligence for Disaster Response. In: *WWW'14 Companion*. International World Wide Web Conference Committee (IW3C2)

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., Hulden, M. (2018). UniMorph 2.0: Universal Morphology. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18), Miyazaki, Japan, May 7-12. European Language Resource Association (ELRA).

Kiss, T., Strunkt, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics* 32: 455-525.

Kulick, S., Bies, A., Maamouri, M. (2010). Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valetta, Matlta, May 17-23. European Language Resource Association (ELRA).

Kulick, S., Bies, A. (2016). Rapid Development of Morphological Analyzers for Typologically Diverse Languages. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroz, Slovenia, May 23-28. European Language Resource Association (ELRA).

Li, X., Ge, N., Grimes, S., Strassel, S., Maeda, K. (2010). Enriching Word Alignment with Linguistic Tags. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valetta, Matlta, May 17-23. European Language Resource Association (ELRA).

Ma, X. (2006). Champollion: A Robust Parallel Text Sentence Aligner. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, May 22-28. European Language Resource Association (ELRA).

Ma, X., Liberman, M. (1999). BITS: A Method for Bilingual Text Search over the Web. In Proceedings of Machine Translation Summit VII, Singapore, September 13-17.

META-NET. 2012. META-NET White Paper Series, https://link.springer.com/bookseries/10412, accessed March 20, 2020.

Mott, J., Bies, A., Strassel, S., Kodner, J., Richter, C., Xu, H., Marcus, M. (2020). Morphological Segmentation for Low Resource Languages. To Appear, LREC 2020.

NIST LoReHLT Evaluations Website. https://www.nist.gov/itl/iad/mig/lorehlt-evaluations. Retrieved February 14, 2020.

Palmer, M., Gildea, D., Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, 31:1.

Strassel, S., Bies, A., Tracey, J. (2017). Situational Awareness for Low Resource Languages: the LORELEI Situation Frame Annotation Task. First Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP), Aberdeen, Scotland, April 8-13.

Strassel, S., Tracey, J. (2016). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroz, Slovenia, May 23-28. European Language Resource Association (ELRA).