Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources

Heather Simpson*, Christopher Cieri*, Kazuaki Maeda*, Kathryn Baker†, Boyan Onyshkevych†

*Linguistic Data Consortium
University of Pennsylvania
3600 Market St., Suite 810, Philadelphia PA, 19104, U.S.A.
{hsimpson, ccieri, maeda}@ldc.upenn.edu
†U.S. Department of Defense

Abstract

The REFLEX-LCTL (Research on English and Foreign Language Exploitation) program, sponsored by the United States government, was a medium-scale effort in simultaneous creation of basic language resources for several less commonly taught languages (LCTLs). To address some of the gaps in language technologies and resources, and to spur new research in this area, two REFLEX-LCTL sites constructed language packs for 19 LCTLs, and distributed them to research and development also funded by the program. This paper will focus on the work done at the Linguistic Data Consortium (LDC). LDC created language packs for 13 out of the 19 languages: Amazigh (Berber), Bengali, Hungarian, Kurdish, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Urdu, Uzbek, and Yoruba. Discussed are the goals and reasoning behind the language choice and language pack construction, and more in depth on the human resource and technology challenges in creating these language packs.

1. Introduction

The past decade has seen increased interest across multiple disciplines in resource creation for a growing number of languages. The new languages of focus have been grouped under several terms, including minority languages, less commonly taught languages, less resourced languages and endangered languages. Each term encodes differences in traditions, goals and approaches. A researcher working on an endangered language may seek to document that language and reinvigorate its use while a researcher working in less commonly taught languages (LCTLs) may seek to enable basic linguistic technologies or build language-aware applications.

The REFLEX-LCTL (Research on English and Foreign Language Exploitation) program, sponsored by the United States government, was a medium-scale effort in simultaneous creation of basic language resources for several LCTLs. To address some of the gaps in language technologies and resources, and to spur new research in this area, two REFLEX-LCTL sites constructed language packs for 19 LCTLs, and distributed them to research and development also funded by the program. The data sites are: the Linguistic Data Consortium (LDC), and the Computing Resource Laboratory (CRL) of the New Mexico State University (NMSU). This paper will focus on the work done at LDC.

The LCTL language packs address three goals. The first is to enable porting of existing technologies to new languages by providing training data and component technologies such as part-of-speech tagging and named entity extraction.. The second goal is to seed new research specifically on achieving better performance with fewer resources and on simplifying the process of porting of technologies to LCTLs when needed. Finally, the third goal is for the community to test and refine the choice, size and nature of the resources, contained in the language packs.

This third goal is directly related to the work of institutions ELSNET and ELRA (Evaluations and Language Resources Agency) in their definition of the BLARK (Basic Language Resource Kit) matrices. LCTL language packs contain 15 deliverable components including 6 of the 9 text resources and tools in 4 of the 15 text-based modules listed in the current BLARK matrix (ELDA, 2008).

2. Overview of Created Resources

2.1. Languages

LDC (http://projects.ldc.upenn.edu/LCTL) created resources for 13 of the 19 REFLEX-LCTL languages. These are: Amazigh (Berber), Bengali, Hungarian, Kurdish, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Urdu, Uzbek, and Yoruba.

CRL (http://crl.nmsu.edu/say) created resources for: Amharic, Burmese, Chechen, Guarani (spoken in Paraguay and Argentina), Maguindanao (Phillipines) and Uighur (Xinjiang, China).

The choice of REFLEX-LCTL targets addresses a number of criteria while still fitting within a fixed budget. All meet the basic criteria of being significant in terms of the number of native speakers but poorly represented in terms of available language resources.

Some of the languages (Thai, Urdu) were chosen to exercise a resource collection paradigm in which raw text is available digitally in sufficient quantity; others (Amazigh, Guarani, Maguindanao) were chosen to force the program to deal with cases in which it certainly is not. The cluster of Indic languages (Bengali, Punjabi, Urdu) was chosen to give researchers the opportunity to experiment with bootstrapping systems from material in related languages. Amazigh, Hungarian, Pashto, Tamil, and Yoruba were chosen to take advantage of existing collaborations in order to reduce costs.

Finally there was a general desire to select languages that are quite different from each other and from well-resourced

languages in order to maximize the generality of our methods. As a group, the LCTL languages are linguistically and geographically diverse; they include the national languages of fourteen different countries, representing eleven major language families, in Central, South and Southeast Asia, Austronesia, North, East and West Africa, the Middle East, Eastern Europe and South America.

2.2. Contents of Language Packs

The evolution of the planning of the LCTL language packs followed a path that has become somewhat familiar. The early phase was characterized by an appreciation of the difficulty of the endeavor and a strict balance in the distribution of resources across languages. As the work progressed, optimism inspired by some early successes and recognition of the differences in supply and demand of resources in the LCTLs led to modifications in the resource plan. The volume goals for some languages increased and specifications were refined to make the end result more useful across a broad range of HLTs, by converting found data from the original form into XML formats that were more easily integrated.

To control costs, we planned to take advantage of as much online data as possible. To this end we implemented a series of "Harvest Festivals"; intensive half day sessions where the entire LDC LCTL team, along with native speaker informants, convened to search the web for useful resources for each deliverable. By combining native speakers, linguists, programmers, information managers and projects managers in the same room, we were able to reduce communications latency nearly to zero, brainstorm jointly, and rapidly build upon each other's efforts.

This approach was generally quite successful, especially for the text corpora and lexica, and led us to some of our most useful data. Ideally the Harvest Festival would be the first step in language pack creation when the hope is to use raw online resources. Although it was not always possible to make it the preliminary step, we conducted a Harvest Festival at some point in the project for all but two of the 13 languages.

2.3. Text Corpora

Monolingual text serves as a basis for all of the other resources in the language pack and allows for small scale language modeling. For most of the LCTLs, this corpus was created by identifying and harvesting available resources from the internet, such as news and weblogs in the target language. Any source specific tags were removed from the harvested text, and it was converted into a standard digital representation for the LCTL, typically UTF8 encoded Unicode, and then tokenized.

Parallel Text supports the induction of translation lexicons and serves as both training and test material for machine translation technologies. Parallel text may be found and sentence aligned, or created from monolingual text by sentence segmenting and then having humans translate each sentence of source into one or more sentences in the target language. Our original concentration was on utilizing found Parallel Text, but we were not able to find a substantial amount for many of the LCTLs.

Additionally, although there are fewer steps involved in the found text processing, the alignment step can prove exceedingly difficult if there are deficiencies in either the segmentation in the original data, or in the sentence segmentation tool used to process the data.

In the end, most of our Parallel Text was created through outsourcing translation of our harvested Monolingual text to translation agencies. About 85,000 tokens of the Parallel Text for each language is English-to-LCTL translation. The English source text is shared across all 13 Language Packs, which will allow for comparison between these languages.

2.4. Lexica

Bilingual Lexicons support a variety of technologies including translation, tagging, information extraction and translingual information retrieval. The initial goal for this project was a lexicon, found or created, of at least 10,000 lemmas that included glosses and parts of speech. For most of the LCTLs, we were able to consult existing lexica, either digital or printed, to provide basic data for a subset of the lexical entries; however, in all cases we needed to process them substantially before they could be used efficiently. Processing steps included checking, normalizing and adding parts of speech and glosses, adding entire entries and removing irrelevant entries.

2.5. Tools for Conversion/Segmentation

The goal for this project was to include whatever encoding converters were needed to convert all of the raw text and lexical resources collected or created into the standard encoding selected for that LCTL.

Dividing text into individual sentences is a necessary first step for many processes including the human translation that dominated much of our effort. Simple in principle, LCTL sentence segmentation can prove tantalizingly complex. Our goal was to produce a sentence segmenter that accepts text in our standard encoding as input and outputs segmented sentences in the same encoding.

Word segmentation, or tokenization, is also relatively challenging for many LCTLs. Our goal for this project was to find or develop tokenizers that would produce word lists from texts in our standard format.

2.6. Annotated Corpora and Taggers

In order to support downstream processing, we also set out to produce three sets of internally coordinated resources: a part-of-speech tagger and tagged text, a morphological analyzer and tagged text and a named entity tagger and tagged text.

The project included the specific requirement that the morphological analyzer use the same tagset as the bilingual lexicon. Over time it became obvious that coordination among all of these resources was desirable and the work could be done most efficiently at the data sites. Unfortunately, we never found resources with this level of coordination. As a result we invested considerable time in creating or revising whatever resources we found for entity, part-of-speech, or morphology tagging. We found that at least 60,000 tokens of part-of-speech tagged text was the optimal amount for training our tagger, and we had to create this in-house

for almost every language. The named entity tagged text was also created in-house for all but the three outsourced languages.

2.7. Name Transliterators

The spelling of person names, particularly those foreign to the language under study, exhibit wide ranging variation in digital text and constitute a large percentage of the out-of-vocabulary terms in any HLT. To partially address this problem, we set out to create a personal name transliterator for each LCTL.

2.8. Grammatical Sketches

Finally, in order to identify for technology developers the challenges specific to the LCTLs, we undertook to create Grammatical Sketches for each. These are short outlines, approximately 50 pages, of the features of the written language and were based on existing grammars and experiences garnered in the work described above. The target audience included the other research groups participating in the REFLEX program, HLT developers who could be expected to have an understanding of basic concepts in linguistics.

2.9. Summary of LCTL Language Packs

We have completed a Language Pack for each of the 13 LCTL languages. 10 of them met our original requirements for project deliverables. Three of the Language Packs, Yoruba, Tigrinya, and Berber, fall short of our original requirements for some deliverables though they meet secondary requirements for others. Where these Language Packs do not meet original requirements, it was typically because the extreme dearth of resources existing for those languages made it impossible to do so given timeline and cost restraints. Table 1 and Table 2 summarize the contents of the Language Packs.[1]

Some of the Language Packs have already been distributed to REFLEX program members. Others are being held in reserve for possible use in technology evaluations. For example the Urdu Language Pack will be used in the NIST Open-MT evaluation campaign in 2008. Once a Language Pack has been exposed, it will be placed in the LDC publication queue for future release through the usual mechanisms.

3. Challenges and Solutions Toward Efficient Collaboration

3.1. Collaboration with Trained Researchers

As mentioned above, the extreme lack of available resources for Yoruba, Tigrinya, and Berber made it impossible for us to complete our requirements for some deliverables within the project's original time and budget.

For Yoruba and Berber, we found there simply was not enough harvestable digital text written in those languages to meet our Monolingual text requirement. We compensated for the lack of available Monolingual text by creating much of the data ourselves or under contract.

In the case of Yoruba, printed newspapers were physically collected and sent to us from Nigeria, which we then sent out to an outside agency to manually keyboard into digital text. The resulting corpus comprises 45% of our total Monolingual text for Yoruba.

In the case of Berber, we relied heavily upon our collaboration with the Institut Royal de la Culture Amazighe (IRCAM), in Morocco. IRCAM is working to develop and promote literacy and use of the Amazighe language. Two IRCAM researchers were able to come to LDC for a month, and shared their expertise and their resources with us. We were able to create tools to provide encoding conversion between IRCAM's standardized Latin-based transliteration of Berber, several other Latin-based transliterations, and Tifinagh, which we shared with IRCAM.

We also worked with Lori Levin at Carnegie Mellon University to help create our English-to-LCTL source text. She provided us with Elicitation Corpus, which she and her team specifically designed to elicit lexical distinctions in translations that do not occur in English (Alvarez et al., 2006).

Three of our LCTL Language Packs, Hungarian, Uzbek, and Kurdish, were entirely outsourced to the Media Research Centre at Budapest University of Technology and Economics (BUTE). This had the advantage that the team at BUTE was already working on or had access to many of the resources required for the language packs.

3.2. Working with Non-Specialist Native Speakers

We were dependent on finding native speaker assistance to create our annotated corpora and help identify harvestable online resources for most of the LCTL languages. Intensive recruiting efforts were conducted for native speakers of each non-outsourced LCTL language. Our recruiting strategy utilized such resources as online discussion boards and student associations for those language communities, and we were also able to capitalize on the diversity of the student/staff body of our host organization, the University of Pennsylvania, to recruit some native speakers internally.

We received a relatively high level of interest from most of our online advertising, from native speakers who seemed very excited that research attention was being paid to their languages. However, as might be expected, most of our respondents were not local to the Philadelphia area, and many were international. Though we did have support for remote work on some of our project tasks (as described in the Software Tools section below), we did not have the infrastructure to support complete outsourcing of annotation tasks to independent contractors. The creation of more comprehensive guidelines for non-specialist native speakers, and porting of more tasks into annotation tools such as the Annotation Collection Kit Interface (ACK), would perhaps make this a feasible option for a future effort of this kind.

We did find help from in-house native speakers all 10 non-outsourced languages. However, Berber and Yoruba were assisted by trained researchers who had limited time to spend on our particular needs, and our single Tigrinya native speaker informant also had time constraints. This resulted in a negative effect on completion of the Parallel Text, Part-of-Speech Tagger, and Named Entity Annotation

[1] The numbers represent the number of tokens.

	Large Languages		Small Languages				
	Urdu	Thai	Bengali	Tamil	Punjabi	Hungarian	Yoruba
Mono Text	14,804,000	39,700,000	2,640,000	1,112,000	13,739,000	1,414,000	363,000
Parallel Text (L ⇒ E)	1,300,000	694,000	237,000	308,000	221,000	70,000	
Parallel Text (Found)	947,000	1,496,000	243,000		230,000	2,338,000	78,600
Parallel Text (E ⇒ L)	65,000	65,000	65,000	65,000	65,000	65,000	65,000
Lexicon	26,000	232,000	482,000	10,000	108,000	182,400	128,200
Encoding Converter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sentence Segmenter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Word Segmenter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagger	Yes	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagged Text	5,000	5,000		59,000			
Morphological Analyzer	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Morph-Tagged Text	11,000			144,000			
NE Annotated Text	233,000	218,000	138,000	132,000	157,000	269,000	189,000
Named Entity Tagger	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name Transliterator	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Descriptive Grammar	Yes	Yes	Yes	Yes	Yes	Yes	

Table 1: LCTL Language Packs (Phase 1)

	Small Languages					
	Tagalog	Tigrinya	Pashto	Uzbek	Kurdish	Berber
Mono Text	774,000	617,000	5,958,000	790,000	2,463,000	181,000
Parallel Text (L ⇒ L)	203,000	139,000	180,000	206,000	163,000	26,000
Parallel Text (E ⇒ L)	65,000	65,000	65,000	65,000	65,000	65,000
Lexicon	18,000	0	10,000	25,400	6,500	Active
Encoding Converter	Yes	Yes	Yes	Yes	Yes	Yes
Sentence Segmenter	Yes	Yes	Yes	Yes	Yes	Yes
Word Segmenter	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagger	Yes	Yes	Yes	Yes	Yes	
POS Tagged Text						
Morphological Analyzer	Yes	Active	Yes	Yes	Yes	Active
Morph-Tagged Text						
NE Annotated Text	136,000	123,000	165,000	93,000	62,000	60,000
Named Entity Tagger	Yes	Yes	Yes	Yes	Yes	Yes
Name Transliterator	Yes	Yes	Yes	Yes	Yes	Active
Descriptive Grammar	Yes	Yes	Yes	Yes	Yes	No

Table 2: LCTL Language Packs (Phase 2)

deliverable requirements for those three languages. Though we were able to find translation agencies who could deliver Parallel Text for Yoruba and Berber, turn-around and cost precluded us from meeting our goal quantities of text corpora.

3.3. Software Tools

3.3.1. Overview

In creating the language resources included in the LCTL language packs, we developed a variety of software tools for helping humans provide data needed for the resource creation efforts. The following are some of the examples.

3.3.2. Annotation Collection Kit Interface (ACK)

Probably the most important of the annotation tools for the LCTL project was the Annotation Collection Kit Interface (ACK), developed by LDC (Maeda et al., 2008). ACK facilitates remote creation of multiple types of text-based annotation, by allowing individual "kits" to be uploaded onto a specific server URL which any remote user can access. Using this tool we were able to support native speaker annotators working on part-of-speech (POS) annotation from Thailand.

When annotators make judgments in ACK, they are stored in a relational database. The results can be downloaded in CSV (comma-separated value) or XML format, so anyone with secure access to the server can easily access the results.

Anyone with a relatively basic knowledge of a scripting language such as Perl or Python would be able to create the ACK annotation kits. They are essentially a set of data corresponding to a set of annotation decisions in the form of radio buttons, check boxes, pull-down menus, or comment

fields, so they are currently limited in scope, but creative use of this format can yield a great deal of helpful types of annotation.

For POS annotation, the annotators were given monolingual text from our corpus, word by word, in order, and asked to select the correct part of speech for that word in context. We also used ACK to add/QC glosses and parts of speech for lexicon entries and do morphological tagging, and many other tasks that require judgment from native speaker.

3.3.3. Named Entity Annotation Tool

LDC also developed an named entity (NE) annotation tool, called SimpleNET (Maeda et al., 2006). SimpleNET requires almost no training in tool usage, and annotations can be made with the keyboard or the mouse. The NE annotated text in the LCTL language packs was created with this tool.

3.3.4. POS and NE Taggers

The annotated text created with ACK and SimpleNET was used in the development of the part-of-speech (POS) taggers and named entity (NE) taggers included in the language packs. Most of these POS and NE taggers were created using a common development infrastructure, which was centered around the MALLET toolkit (McCallum, 2002). By using the common infrastructure, we minimized the duplicated effort in creating these tools.

3.3.5. Encoding Conversion Tools

We encountered difficulties relating to the lack of usage of standardized orthography for some of the LCTL languages, as mentioned earlier. Our Berber Encoding Converter supports conversion between 6 different romanizations/encodings, and there are still more out there that we did not have time or resources to include. There would have been more Berber Monolingual Text in our corpus if we had had the ability to decipher every idiosyncratic encoding and add to the converter.

4. Conclusion

Despite numerous challenges, we have successfully created large, and in some cases unique resources for each of the 13 LCTL languages that we hope will provide valuable support for research and technology development for these previously under-supported languages. At least some of the challenges we have undergone would surely be encountered during a similar effort with different LCTLs. We hope that others may be able to learn from our mistakes and from our solutions to make their project a more successful endeavor in HLT development for under-resourced languages.

5. References

Alison Alvarez, Lori S. Levin, Robert E. Frederking, Simon Fung, and Donna Gates. 2006. The MILE corpus for less commonly taught languages. In *Proceedings of HLT-NAACL 2006*.

ELDA. 2008. BLARK Resource/Modules Matrix. From Evaluations and Language Resources Distribution Agency (ELDA) web site http://www.elda.org/blark/matrice_res_mod.php , accessed on 2/23/2008.

Kazuaki Maeda, Haejoong Lee, Julie Medero, and Stephanie Strassel. 2006. A new phase in annotation tool development at the Linguistic Data Consortium: The evolution of the Annotation Graph Toolkit. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker, and Stephanie Strassel. 2008. Annotation tool development for large-scale corpus creation projects at the Linguistic Data Consortium. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.