Collection of Bilingual Data for Lexicon Transfer Learning

Leanne Rolston, Katrin Kirchhoff

UWEE Technical Report
Number UWEETR-2016-0000
March 2016

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Collection of Bilingual Data for Lexicon Transfer Learning

Leanne Rolston, Katrin Kirchhoff

University of Washington, Dept. of EE, UWEETR-2016-0000

March 2016

Abstract

This technical report describes the collection and format of a dataset of bilingual lexicons for 50 languages, undertaken as part of the DARPA LORELEI project on developing language technology for low-resource languages. We describe the data sources, collection method, and types of linguistic information included in the lexicons.

1 Introduction

The following is a brief description of a data collection effort for Phase I of the DARPA Low Resource Languages for Emergent Incidents (LORELEI) project, carried out at the University of Washington. The DARPA LORELEI project aims at providing rapid, low-cost natural language processing (NLP) capabilities for low-resource languages that lack the large annotated speech and text corpora typically needed to develop NLP systems. One of the desired capabilities is machine translation for the purpose of information extraction. Our intended contribution towards that goal is the provision of translation lexicons for incident languages from a set of related languages. That is, given a word list in an incident language, we aim to produce valid word translations into English for each word in the list. Translations will be inferred from a network expressing language similarities at the lexical level. To support this goal, a set of bilingual lexicons for 50 languages was collected.

We expect this data collection to be useful beyond the immediate goals of the LORELEI project, enabling e.g., the development of other types of NLP components or research in synchronic and diachronic linguistics.

2 Languages

Languages were chosen to maximize coverage of all major language families in the world, with special consideration given to languages and language families that may be relevant to the intended incident languages in the LORELEI project. These include languages that are related to incident languages through genetic relationship (same language family), socio-cultural ties, or geographical proximity. Additional consideration was given to languages that have been shown to be highly influential and would serve as good overall bridge languages for a variety of transfer learning experiments. For example, where possible a more general (historically older) rather than a more specific representative of a given language family was chosen.

Table 1 shows the languages covered by our our data collection. The ISO 639-3 language codes in column 2 represent the macrolanguage, or the more standardized form of the language. When searching for relevant data online, initially the more common language name or macrolanguage was used. Where necessary, due to lack of data or more detailed language categorization in the source, subsequent searches targeted sub-dialects or alternate language names.

Language	ISO639-3	Alternate Name	Language	ISO639-3	Alternate Name
Akan	aka		Kurdish	kur	
Albanian	sqi		Kyrgyz	kir	Kirghiz
Amharic	amh		Mandarin	cmn	
Arabic	ara		Nepali	nep	
Bengali	ben	Bangla	Pashto	pus	
Berber	ber		Portuguese	por	
Bulgarian	bul		Romanian	ron	
Burmese	mya		Russian	rus	
Cantonese	yue		Sinhalese	sin	Sinhala
Croatian	hrv		Somali	som	
Dinka	din		Spanish	spa	
Dutch	nld		Swahili	swa	Kiswahili
English	eng		Tagalog	tgl	
Persian	pes	Farsi	Tamil	tam	
French	fra		Thai	tha	
Fulfulde	ful	Fula, Fulani, Pulaar, Peul	Tigrinya	tir	Tigrigna
Greek	ell		Turkish	tur	
Guarani	grn		Turkmen	tuk	Torkoman
Hausa	hau		Urdu	urd	
Hebrew	heb		Uyghur	uig	Uighur
Hindi	hin		Uzbek	uzb	
Hungarian	hun		Vietnamese	vie	
Indonesian	ind		Wolof	wol	
Japanese	jpn		Yoruba	yor	
Kannada	kan		Zulu	zul	

Table 1: Languages included in the lexicons

Figure 1 shows the languages by language family, as listed by Ethnologue [5]. Phylogenetic relationship is given only in enough detail to demonstrate basic language family relatedness. Figures 2, 3, and 4 offer a zoomed in view of the Indo-European, Afro-Asiatic, and Niger-Congo branches, respectively.

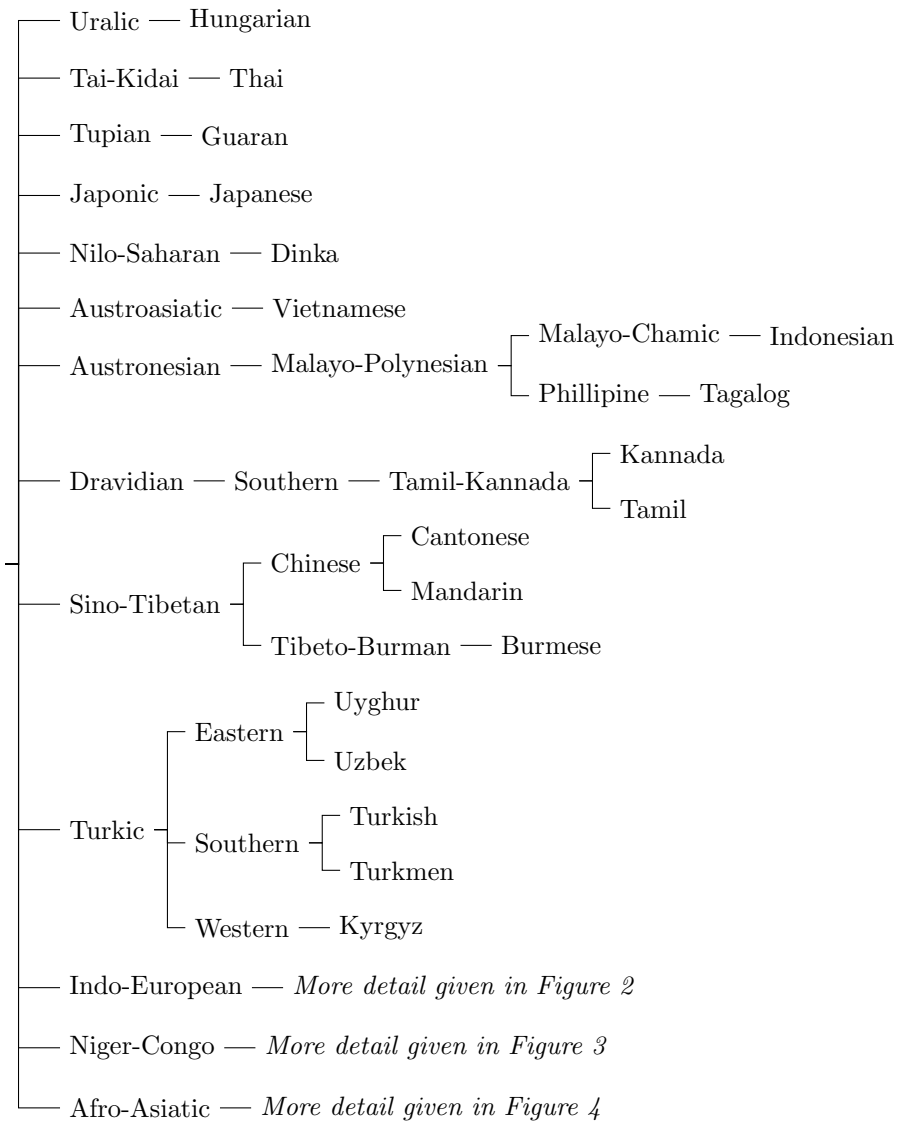

Figure 1: Languages Represented in the Master Lexicon Files by Language Family

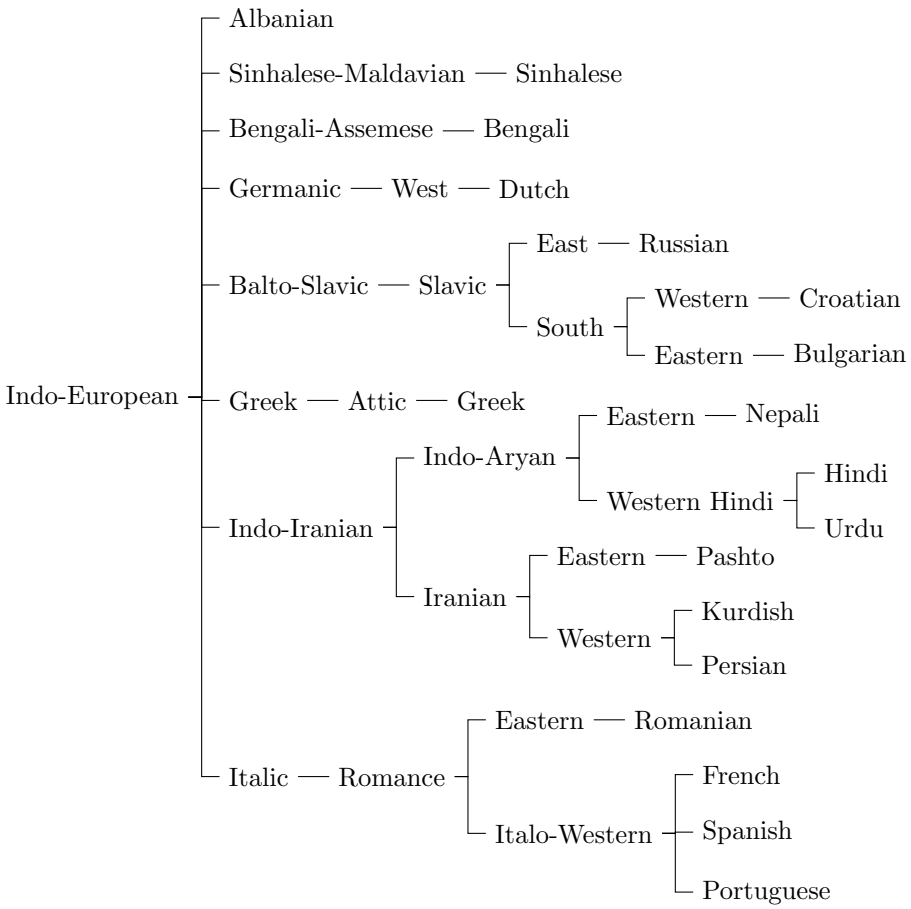

Figure 2: Indo-European Languages in the Master Lexicon Files

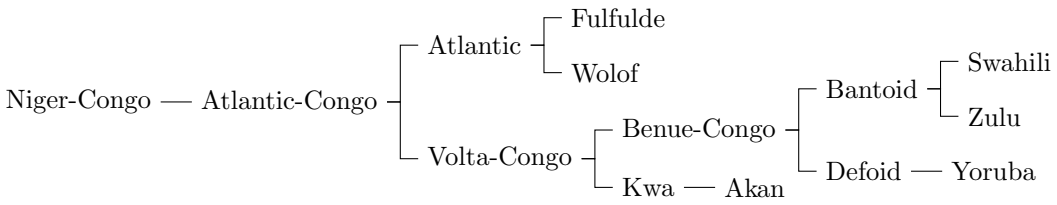

Figure 3: Niger-Congo Languages in the Master Lexicon Files

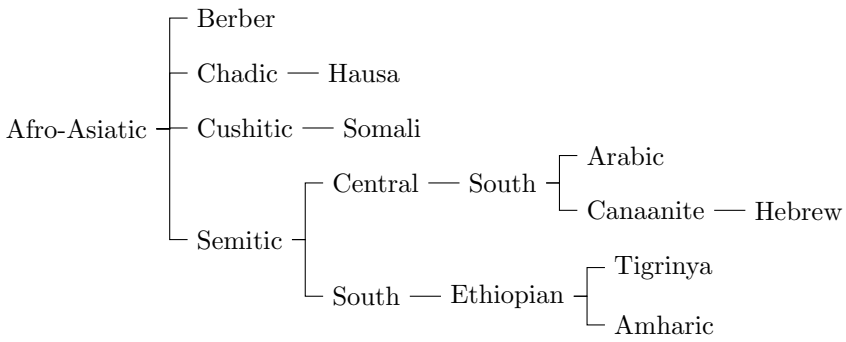

Figure 4: Afro-Asiatic Language in the Master Lexicon Files

3 Data Sources

In addition to the data provided specifically to the LORELEI Project, additional data was found through in publicly available language corpora and Internet resources such as Wikipedia, Wiktionary, language learning sites, travel sites, indigenous folklore resources, and collections gathered by educational institutions. Source data includes word lists, glossaries, and parallel texts. The following sources are included in our distribution as of the date of publication:

- Blench [2]
  - **URL**: `http://www.rogerblench.info/Language/Nilo-Saharan/Nilotic/Comparative\%20Dinka\%20lexicon\%20converted.pdf`
  - **Licensing**: None listed
  - Dinka-English Dictionary gathered from various sources by Roger Blench
- Chaihana [4]
  - **URL**: `www.chaihana.com/dict.pdf`
  - **Licensing**: may be freely distributed for educational purposes
  - Dictionary created by Peace Corps Turkmenistan
- DictionaryforMIDs
  - **URL**: `http://dictionarymid.sourceforge.net/dict.html`
  - **Licensing**: GNU General Public License. All dictionaries that can be downloaded from the site are free to use.
  - Multilingual dictionaries for cell phones and PDAs
- DY
  - **URL**: `http://www.denizyuret.com/2006/11/turkish-resources.html`
  - **Licensing**: None listed
  - English-Turkish dictionary, originally hosted at `www.fen.bilkent.edu.tr/~aykutlu`
- en_wiktionary
  - **URL**: `https://en.wiktionary.org/`
  - **Licensing**: Creative Commons Attribution-ShareAlike License

  - Free multilingual dictionary produced by online collaborative user community. The data attributed to Wiktionary was gathered through searching the English Wiktionary (`https://en.wikipedia.org`) for entries that included a translation section, as determined through automatic parsing of the section headings. Translation pairs, part-of-speech and pronunciation information were extracted.

- IATE (Inter-Active Terminology for Europe)
  - **URL**: `http://iate.europa.eu`
  - **Licensing**: data can be freely downloaded and reproduced
  - The EU's inter-institutional terminology database. Includes domain labels.
- ICD (Intercontinental Dictionary Series)
  - **URL**: `http://lingweb.eva.mpg.de/ids/`[1]
  - **Licensing**: Creative Commons
  - Large database of vocabulary lists in various languages gathered by the Max Planck Institute for Evolutionary Anthropology
- LDC2008L03
  - **Licensing**: LDC
  - Global Yoruba Database
- LDC2015E14
  - **Licensing**: LDC
  - Translations provided in V1.1 of the REFLEX Less Commonly Taught Languages (LCTL) Urdu Language Pack
- LDC2014E115
  - **Licensing**: LDC
  - Translations provided in V2.1 of the BOLT Low Resource Language (LRL) Turkish Representative Language Pack
- LDC2015E70
  - **Licensing**: LDC
  - Translations provided in V1.1 of the BOLT Low Resource Language (LRL) Hausa Representative Language Pack
- LDC2015E13
  - **Licensing**: LDC
  - Translations provided in V2.1 of the REFLEX Less Commonly Taught Languages (LCTL) Bengali Language Pack
- LDC2015E82
  - **Licensing**: LDC
  - Translations provided in V1.1 of the REFLEX Less Commonly Taught Languages (LCTL) Hungarian Language Pack
- LDC2015E83
  - **Licensing**: LDC
  - Translations provided in V1.1 of the REFLEX Less Commonly Taught Languages (LCTL) Tamil Language Pack

[1] This content was retrieved in late 2015. This content has been moved as of January 4th, 2016 and is now password protected.

- LDC2015E84
  - **Licensing**: LDC
  - Translations provided in V1.1 of the REFLEX Less Commonly Taught Languages (LCTL) Thai Language Pack
- LDC2015E89
  - **Licensing**: LDC
  - Translations provided in V1.1 of the BOLT Low Resource Language (LRL) Uzbek Incident Language Pack.
- LDC2015E90
  - **Licensing**: LDC
  - Translations provided in V1.1 of the REFLEX Less Commonly Taught Languages (LCTL) Tagalog Language Pack
- LDC2015E91
  - **Licensing**: LDC
  - Translations provided in V1.1 of the REFLEX Less Commonly Taught Languages (LCTL) Yoruba Language Pack
- LDC2015G02
  - **Licensing**: LDC
  - Translations provided in V1.1 of the REFLEX Less Commonly Taught Languages (LCTL) Amharic Language Pack
- OMWN (Open Multilingual Wordnet)[3]
  - **URL**: `http://compling.hss.ntu.edu.sg/omw/`
  - **Licensing**: can be freely used, modified, and shared by anyone for any purpose.
  - Wordnets in a variety of languages.
- Panlex [6]
  - **URL**: `http://panlex.org`
  - **Licensing**: Creative Commons CC0 1.0 Universal
  - Panlingual lexical translation database
- parallel_texts
  - Lexicons inferred from parallel texts: Universal Declaration of Human Rights (http://www.unicode.org/udhr/), Europarl (http://www.statmt.org/europarl/), and Parallel Bible Corpus (http://paralleltext.info/data/all/)
  - **License**: None listed
- PeaceCorps
  - **Licensing**: None listed
  - 2000 word glossary; early draft of what eventually became the Chaihana Turkmen English Dictionary Project

- TaaS (Terminology as a Service) [1]
  - **URL**: `http://www.taas-project.eu`
  - **License**: None listed
  - European multilingual terminological data
- UCLA
  - **URL**: `http://archive.phonetics.ucla.edu`
  - **License**: None listed
  - UCLA Phonetics Lab archive
- Wiktionary Swadesh
  - **URL**: `https://en.wiktionary.org/wiki/Appendix:Swadesh_lists`
  - **License**: Creative Commons Attribution-ShareAlike License
  - Swadesh lists archived on Wiktionary

4 Data Extraction

Translation pairs were extracted directly from the source using meta data, or they were inferred through word alignment of parallel texts. Any translation pairs resulting from direct data extraction that were longer than 4 tokens were added to the parallel text. Supplementary information, such as part of speech or lemma, was identified by analyzing the HTML markup of the source, or through meta information in the source data, such as column headings.

The parallel text processing pipeline was based on from [1]. The first step was word alignment using GIZA++. The raw word translation lexicons generated in this step were subsequently filtered down to include only probable translation pairs. Filtering included e.g., removing words aligned with with punctuation as well as source and target words whose relative inverse document frequency values differed by more than a threshold value. A subsequent check on the orthography of the entries was completed, however some non-standard orthographies, for instance Russian text presented in the Roman alphabet, were allowed due to the high incidence in the data. The lexicon files from individual sources were then merged into a single "master" lexicon file for each language, with duplicate entries consolidated if source word and translation matched exactly.

5 Master Lexicon Format

Master lexicon files are named 'CODE-eng.masterlex.txt' where CODE is a three-letter ISO 639-3 language code. Where the source language represents a macrolanguage, data for all individual sub-language varieties as defined by Ethnologue [5], was sought. The lexicons are flat text files (no database model), with the tab-separated columns: listed in Table 2.

Field	Description
1	Orthographic form
2	Lemma
3	Part-of-Speech
4	Transliteration
5	Pronunciation
6	Translation
7	Score
8	Dialect/linguistic variety
9	Domain label
10	Data source
11	Morphological variants (if available)

Table 2: Fields in the master lexicon files.

Fields 2,3,4,5,7 and 8 have meaningful values for only a subset of the entries since most data sources did not include this information. Where data was unavailable, the value of the respective field is "N/A".

The orthographic form (Field 1) may consist of one or more whitespace-delimited words. Some of the original data sources only provided a phonetic transcription of the word but no orthographic form; in these cases Field 1 is "N/A".

Lemma (Field 2) is the base form of Field 1, if available from the original source. No lemmatization was performed on the original data.

Part-of-Speech (Field 3) is the part-of-speech extracted from the original source and mapped one of Petrov's universal part-of-speech tags as defined in [8].

Transliterations (Field 4) have been provided for some entries. These stem either from the original source, or they were created using a variety of open-source and in-house transliteration tools. They do not necessarily follow any standard transliteration model and might be replaced by a more consistent format in the future.

Pronunciations (Field 5) were either collected from the original source, extracted from an on-line monolingual dictionary, or they were produced using a simple grapheme-to-phoneme mapping procedure. The latter was based on grapheme-to-phoneme correspondence rules given by the Wikipedia page for the language in question. Pronunciations are specified in International Phonetic Alphabet (IPA) unless the original source used a different "in-house" phonetic representation.

Translations (Field 6) can be single words or multiword glosses in English. These are taken directly from the data source. Where scores were given along with these translations, they are listed in Field 7. Scores do not necessarily represent proper probability distributions.

The dialect or sub-language, if available, is indicated in Field 8.

Field 9 represents a domain label, as indicated in the original data source. Most domain labels are derived from the IATE terminology database.

Field 10 indicates the data source.

Field 11 lists other morphological variants of the word, where available.

6 Statistics

Table 3 shows the count of entries in each master lexicon file as of the date of publication. The second column shows the number of lexical entries; the third column shows the number of unique source words covered (multiple translations of the same source word are listed as separate entries in the lexicon).

Language	# of Entries	# Unique Words	Language	# of Entries	# Unique Words
Akan (aka)	777	739	Kyrgyz (kir)	10,291	6941
Albanian (sqi)	97,196	33,158	Mandarin (cmn)	2,041,076	1,412,647
Amharic (amh)	864	864	Nepali (nep)	4994	4185
Arabic (ara)	224,216	224,216	Pashto (pus)	1497	548
Bengali (ben)	67,470	67,470	Portuguese (por)	853,207	637,620
Berber (ber)	196	150	Romanian (ron)	293,028	202,050
Bulgarian (bul)	283,007	185,564	Russian (rus)	1,424,770	866,797
Burmese (mya)	6891	4362	Sinhalese (sin)	4330	2965
Cantonese (yue)	59,412	19,976	Somali (som)	3042	2467
Croatian (hrv)	311,380	178,550	Spanish (spa)	1,017,963	712,466
Dinka (din)	6337	3400	Swahili (swa)	82,284	33,154
Dutch (nld)	771,585	585,959	Tagalog (tgl)	51,816	21,629
Persian (pes)	161,779	97,902	Tamil (tam)	174,782	83,897
French (fra)	1,455,653	1,086,312	Thai (tha)	507,742	20,102
Fulfulde (ful)	5203	3535	Tigrinya (tir)	1027	200
Greek (ell)	667,857	508,729	Turkish (tur)	501,138	198,782
Guarani (grn)	2090	607	Turkmen (tuk)	34,657	18,393
Hausa (hau)	61,398	29,356	Urdu (urd)	167,921	48,964
Hebrew (heb)	79,009	58,879	Uyghur (uig)	5213	3927
Hindi (hin)	204,137	123,881	Uzbek (uzb)	53,316	28,136
Hungarian (hun)	1,140,444	673,885	Vietnamese (vie)	256,337	134,464
Indonesian (ind)	126,171	55,316	Wolof (wol)	3661	2946
Japanese (jpn)	750,257	256,943	Yoruba (yor)	369,059	165,316
Kannada (kan)	6044	4463	Zulu (zul)	7603	4852
Kurdish (kur)	8107	3891			

Table 3: Counts of entries and unique source words covered, by language.

References

[1] Ahmet Aker, Monica Lestari Paramita, Marcis Pinnis, and Robert J Gaizauskas. Bilingual dictionaries for all eu languages. In *LREC*, pages 2839–2845, 2014.

[2] Roger Blench and Arthur Nebel. Dinka-english and english-dinka dictionary. 2005.

[3] Francis Bond and Kyonghee Paik. A survey of wordnets and their licenses. *Small*, 8(4):5, 2012.

[4] Jonathan Garrett, Greg Lastowka, et al. Turkmen-english dictionary: a spa project of peace corps turkmenistan. 1996.

[5] Raymond G Gordon Jr. Ethnologue: languages of the world, dallas: Sil international. *Online version: http://www.ethnologue.com*, 2005.

[6] David Kamholz, Jonathan Pool, and Susan M Colowick. Panlex: Building a resource for panlingual lexical translation. In *LREC*, pages 3145–3150, 2014.

[7] Dr. Boyan Onyshkevych. Low resource languages for emergent incidents (lorelei). `http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents`, 2014.

[8] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.