UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Ainu Language

Language Resource Assessment

Mayumi Ohta

May 11, 2016

Institute for Computational Linguistics, Heidelberg University

Table of contents

1. Ainu Language

2. Resources

3. Projects

Ainu Language

Overview

ethnicity	Ainu
location	Hokkaido, Sakhalin, Kuril
ethnic population	15,000 [1]
♯ native speakers	less than 10 (80+ years old) [2]
♯ fluent speakers	less than 100 (60+ years old) [3]
endangerment status	8b (critically endangered)
writing system	-
transcription	Katakana script, Latin alphabet
ISO 639-3	ain [4]

[1] ethnologue 8th ed. (1974)

[2] Bradley (2007)

[3] Hohmann (2008)

[4] ≠aib (Ainu language spoken in western China)

Overview

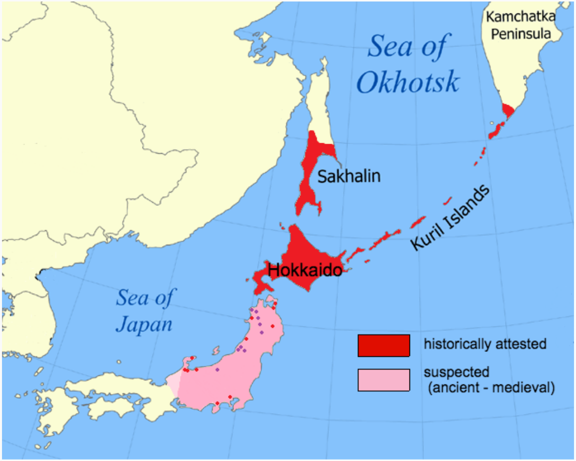


Figure 1: Distribution of the ethnic group Ainu[1]

[1] `http://commons.wikimedia.org/wiki/File:Historical_expanse_of_Ainu.png` (legend added)

Overview

ethnicity	Ainu
location	Hokkaido, Sakhalin, Kuril
ethnic population	15,000 [1]
♯ native speakers	less than 10 (80+ years old) [2]
♯ fluent speakers	less than 100 (60+ years old) [3]
endangerment status	8b (critically endangered)
writing system	-
transcription	Katakana script, Latin alphabet
ISO 639-3	ain [4]

[1] ethnologue 8th ed. (1974)

[2] Bradley (2007)

[3] Hohmann (2008)

[4] ≠aib (Ainu language spoken in western China)

Overview

ethnicity	Ainu
location	Hokkaido, Sakhalin, Kuril
ethnic population	15,000 [1]
♯ native speakers	less than 10 (80+ years old) [2]
♯ fluent speakers	less than 100 (60+ years old) [3]
endangerment status	8b (critically endangered)
writing system	-
transcription	Katakana script, Latin alphabet
ISO 639-3	ain [4]

[1] ethnologue 8th ed. (1974)

[2] Bradley (2007)

[3] Hohmann (2008)

[4] ≠aib (Ainu language spoken in western China)

Overview

ethnicity	Ainu
location	Hokkaido, Sakhalin, Kuril
ethnic population	15,000 [1]
♯ native speakers	less than 10 (80+ years old) [2]
♯ fluent speakers	less than 100 (60+ years old) [3]
endangerment status	8b (critically endangered)
writing system	- oral language !
transcription	Katakana script, Latin alphabet
ISO 639-3	ain [4]

[1] ethnologue 8th ed. (1974)

[2] Bradley (2007)

[3] Hohmann (2008)

[4] ≠aib (Ainu language spoken in western China)

Brief History

∼ Middle Ages fishing-hunting-gathering culture

∼ Pre-modern trade with Japanese people

19c: assimilation by the Japanese government

- persecution and discrimination

→ Ainu language sifted to Japanese

the oldest surviving record: by western missionary

- 1897 Bible translated into Ainu language

20c: ”discovery” by anthropologists, linguists

- 1906 Ainu-En-Ja Dictionary (20k entries)

after WWII revival

- 1997 Ainu Culture Law
- 2007 UN Declaration on the Rights of Indigenous Peoples
- 2009 UNESCO Red Book of Endangered Language

Classification, Topology

Language family: language isolate (origin unknown)

Topology:

- SOV word order
- post-positions
- head-final
- case-marking

- verb affixes mark person, number
- polysynthetic
- CVC syllables
- non-tonal

Resources

Primary resources

Type: spoken resources (audio media)

Total amount: unknown (ca. 50h?? in index)

Time span: 60’s - 90’s

Format: open-reel, cassette tape …

- out-of-date technology, but well-preserved

Genre: folklore, epic poetry, monologue (i.e. play on words)

Meta data: field work report

- speaker’s biography
- when, where, by whom recorded
- background info about the content
- info about the irregular situations

(not standardized, digitalized, indexed, …)

Access: closed

Secondary resources

Type: written resources (print media)

Time span: end of 19c $\sim$

Products:

- dictionaries (Ainu-Ja, Ainu-Ja-En)
- word lists (i.e. Place name lexicon)
- collections of stories
- grammar books

Remarks:

- numerous errors especially in old resources
  - typo
  - grammatical errors
  - misunderstanding
  - errors in meta data
  - propagated to newer resources
- developed by researchers
- various transcribing rules, letters, diacritics, …

Learning environment

- (hand down by word-of-mouth)
- self-study books with CDs
- radio lesson program (free!)
    - new episode every week (not rerun)
    - podcast archive
    - PDF textbook
- language courses (face-to-face classes)
    - a few universities, culture schools in Hokkaido (and Tokyo)
    - designed for adult Japanese native speakers
    - lack of instructors, instructor training

Development environment for digital transcription of Ainu

- standard transcribing rule: **”AKOR ITAK” rule**
- standard character set: **expanded Katakana**
  - Unicode 2.3 (substitute: Japanese half-width Katakana)
  - Input Method
- gloss tag set for Ainu language
- web-based Katakana-Alphabet converter
- Aozora Bunko:
  open digital library for out-of-copyright literature in Japanese
  - digitalization workflow
  - platform to find collaborator (i.e. proofreading volunteer)

Recent movements

Revise existing data:

- animation
- comic books
- pop songs

Create new data:

- new words, new usage
  - imeru [God’s light, God’s shine ⇒ **electricity**]
  - imeru kampi [**email**]
  - imeru pasuy [**cell phone**]
- new text
  - magazine **”AINU TIMES”** (everyday life topics)
- new recordings
  - storytelling contest

Projects

Ainu Language Archiving Project (2014-15) Agency for Cultural Affairs

focus preservation of language

objectives

1. survey of primary (audio) resources in existence
2. digitalization of the NIBUTANI collection
    - meta data
    - sanitized digital audio
    - transcription (Katakana/Alphabet)
    - translation (Ja/En)
    - gloss
    - footnotes
3. establish a working procedure
4. costs estimation, evaluation
5. open source access
6. develop learning materials

final product not yet available...

Ainu Language Archiving Project (2014-15) Agency for Cultural Affairs

focus preservation of language

objectives

1. survey of primary (audio) resources in existence
2. digitalization of the NIBUTANI collection
    - meta data
    - sanitized digital audio
    - transcription (Katakana/Alphabet)
    - translation (Ja/En)
    - gloss
    - footnotes
3. establish a working procedure
4. costs estimation, evaluation
5. open source access
6. develop learning materials

final product not yet available...

Ainu Language Archiving Project (2014-15) Agency for Cultural Affairs

focus preservation of language

objectives

1. survey of primary (audio) resources in existence
2. digitalization of the NIBUTANI collection
    - meta data
    - sanitized digital audio
    - transcription (Katakana/Alphabet)
    - translation (Ja/En)
    - gloss
    - footnotes
3. establish a working procedure
4. costs estimation, evaluation
5. open source access
6. develop learning materials

final product not yet available...

Ainu Language Archiving Project (2014-15) Agency for Cultural Affairs

focus preservation of language

objectives

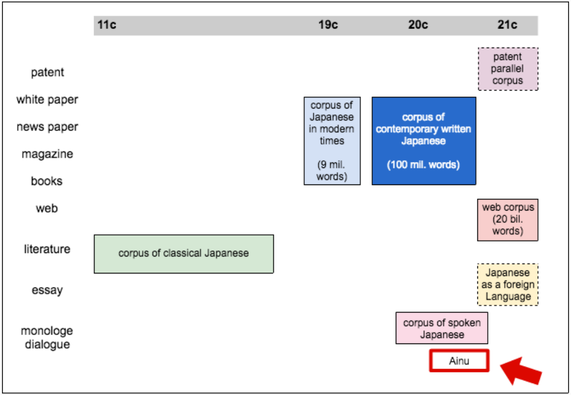
1. survey of primary (audio) resources in existence
2. digitalization of the NIBUTANI collection
    - meta data
    - sanitized digital audio
    - transcription (Katakana/Alphabet)
    - translation (Ja/En)
    - gloss
    - footnotes
3. establish a working procedure
4. costs estimation, evaluation
5. open source access
6. develop learning materials

final product not yet available...

Corpora Project National Institute for Japanese Language and Linguistics

focus NLP research

coverage



11c
19c
20c
21c
patent
white paper
news paper
magazine
books
web
literature
essay
monologe
dialogue
patent parallel corpus
corpus of Japanese in modern times (9 mil. words)
corpus of contemporary written Japanese (100 mil. words)
web corpus (20 bil. words)
corpus of classical Japanese
Japanese as a foreign Language
corpus of spoken Japanese
Ainu

tools fulltext search, cross-corpus search, multilingual UI, etc.

license open for research purpose

Outcomes

1. Topical Dictionary of Conversational Ainu [1]
   - released in March 2015
   - 3500 headwords
2. Glossed Audio Corpus of Ainu Folklore [2]
   - released in March 2016 new!
   - **1800** sentences, **15000** words
   - **959** possible tag patterns in train set

NLP Applications

so far:

1. PoS Tagging Ptaszynski et al. 2012 [4]
   - **Look-up:** 96.26 %
   - **HMM:** 83.64 %
2. Syntactic Parsing Ptaszynski et al. 2013 [3]
   - straightforward projection of syntactic tree from Japanese onto Ainu
   - rough idea only

NLP Applications

so far:

1. **PoS Tagging** Ptaszynski et al. 2012 [4]
   - **Look-up:** 96.26 %
   - **HMM:** 83.64 %
2. **Syntactic Parsing** Ptaszynski et al. 2013 [3]
   - straightforward projection of syntactic tree from Japanese onto Ainu
   - rough idea only

short-term: digitalization

NLP Applications

so far:

1. **PoS Tagging** Ptaszynski et al. 2012 [4]
   - **Look-up:** 96.26 %
   - **HMM:** 83.64 %
2. **Syntactic Parsing** Ptaszynski et al. 2013 [3]
   - straightforward projection of syntactic tree from Japanese onto Ainu
   - rough idea only

short-term: digitalization

long-term:

- speech recognition and speech synthesis
- annotation support system (like spell checker)
- Ainu → Ja translation support system

Thank You!

References

- National Institute for Japanese Language and Linguistics.
  A topical dictionary of conversational ainu, 2015.

- National Institute for Japanese Language and Linguistics.
  A glossed audio corpus of ainu folklore, 2016.

- M. Ptaszynski, M. Kazuki, and Y. Momouchi.
  Nlp for endangered languages: Morphology analysis, translation support and shallow parsing of ainu language, 2013.

- M. Ptaszynski and Y. Momouchi.
  Part-of-speech tagger for ainu language based on higher order hidden markov model.
  Expert Systems with Applications, 39(14):11576–11582, 2012.