Improving Tokenization, Transcription Normalization and Part-of-speech Tagging of Ainu Language through Merging Multiple Dictionaries

Karol Nowakowski Michal Ptaszynski Fumito Masui

KITAMI Institute of Technology

The Ainu people

* Native inhabitants of Hokkaidō.

* Estimated size of Ainu population in Hokkaidō – around 16 thousand people (Hokkaidō regional government, 2013).

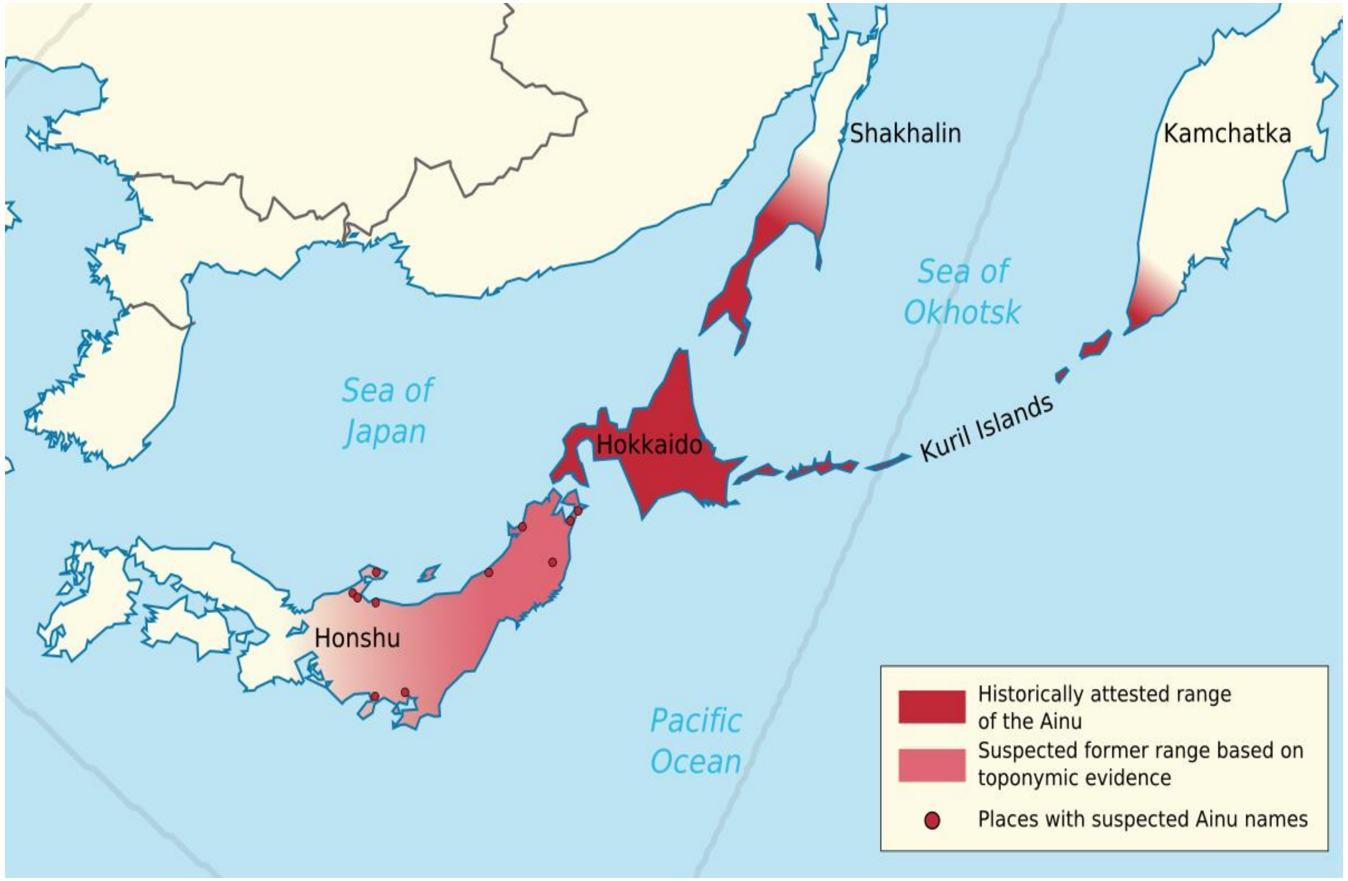

Shakhalin
Kamchatka
Sea of Okhotsk
Sea of Japan
Hokkaido
Kuril Islands
Honshu
Pacific Ocean
🟥 Historically attested range of the Ainu
🟥 Suspected former range based on toponymic evidence
● Places with suspected Ainu names

Image source: https://commons.wikimedia.org

Ainu language

* Language isolate (no confirmed relation to any other language)

* SOV (Subject-Object-Verb) word order (same as Japanese)

* Polysynthetic (especially classical language, such as in *yukar* stories)

Example:

Iramante oruspe ka aeukoisoytak

Meaning: “We can also talk about hunting”

Source: https://www.academia.edu/13753728/Polysynthesis_in_Ainu._In_M._Fortescue_M._Mithun_and_N._Evans_eds_Handbook_of_Polysynthesis._Oxford_OUP._Draft_._forthcoming_

Current situation

* Only 7.2% of Ainu people are able to communicate in the Ainu language (survey by Hokkaidō regional government conducted in 2013, with 586 respondents)

* Status: Critically endangered / nearly extinct

Example:

Iramante oruspe ka aeukoisoytak

Meaning: “We can also talk about hunting”

Source: https://www.academia.edu/13753728/Polysynthesis_in_Ainu._In_M._Fortescue_M._Mithun_and_N._Evans_eds_Handbook_of_Polysynthesis._Oxford_OUP._Draft_._forthcoming_

Ainu language preservation and revitalisation:

* Ainu language classes

* radio course (STV Radio, Sapporo)

* annual Ainu language speech contest (held by The Foundation for Research and Promotion of Ainu Culture),

* “The Ainu Times” (published quarterly)

* music groups singing in the Ainu language ("Oki", "Dub Ainu Band")

http://www.tonkori.com

Aims of this research

- create language analysis toolkit for the Ainu language
- facilitate analysis of the Ainu language by linguists and researchers of the Ainu literature
- contribute to the process of preservation and reviving of the Ainu language

Previous work – POST-AL

- In 2012 Ptaszynski and Momouchi created POST-AL (“Part of Speech Tagger for the Ainu Language).
- POST-AL performs the following tasks:

1. Transcription normalization – modificaton of parts of text that do not conform to modern rules of transcription (e.g. *kamui* -> *kamuy*).

Example:

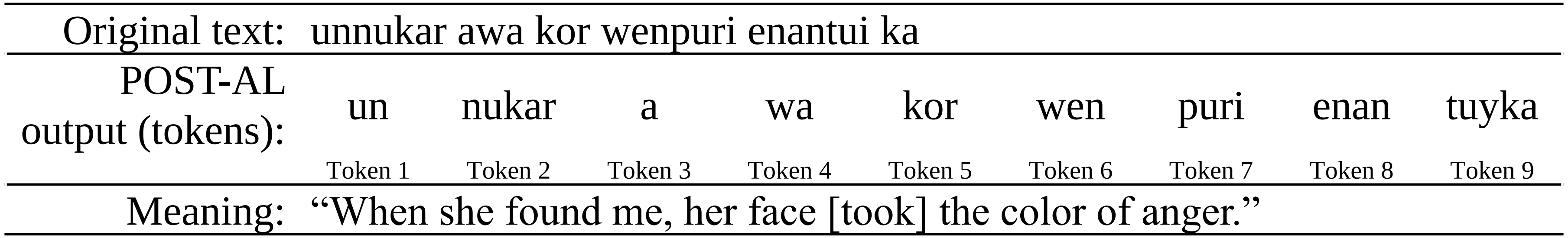
Original text:	Shineantota petetok un shinotash kushu payeash awa
Normalized transcription:	Sineantota petetok un sinotas kusu payeas awa
Meaning:	“One day when I went for a trip up the river”

Previous work – POST-AL

- In 2012 Ptaszynski and Momouchi created POST-AL (“Part of Speech Tagger for the Ainu Language).
- POST-AL performs the following tasks:

1. Transcription normalization – modificaton of parts of text that do not conform to modern rules of transcription (e.g. *kamui* -> *kamuy*).
2. Word segmentation (tokenization) – a process in which the text is separated into tokens (words, punctuation marks, etc.), which become the basic unit for further analysis.

Example:



Original text:	unnukar awa kor wenpuri enantui ka								
POST-AL output (tokens):	un	nukar	a	wa	kor	wen	puri	enan	tuyka
	Token 1	Token 2	Token 3	Token 4	Token 5	Token 6	Token 7	Token 8	Token 9
Meaning:	“When she found me, her face [took] the color of anger.”								

Previous work – POST-AL

- In 2012 Ptaszynski and Momouchi created POST-AL (“Part of Speech Tagger for the Ainu Language).
- POST-AL performs the following tasks:

1. Transcription normalization – modificaton of parts of text that do not conform to modern rules of transcription (e.g. *kamui* -> *kamuy*).
2. Word segmentation (tokenization) – a process in which the text is separated into tokens (words, punctuation marks, etc.), which become the basic unit for further analysis.
3. Part-of-speech tagging – assigning a part-of-speech marker to each token.

Example:

POST-AL tagger iyosno ku hosipire kusne na
output:【副】【人接】【他】【助動】【終助】

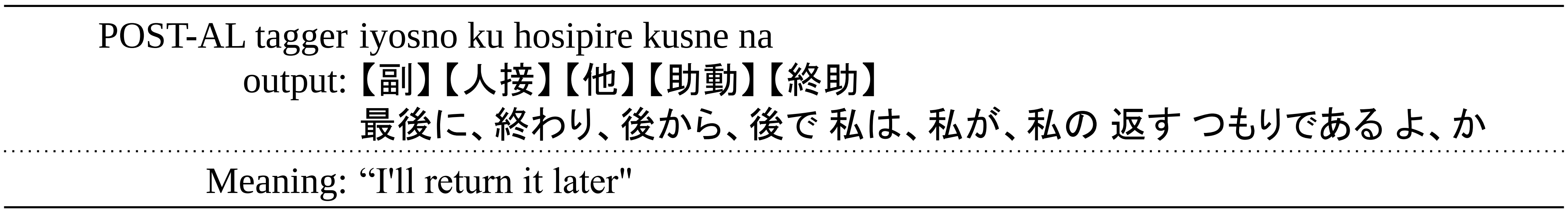
Meaning: “I'll return it later"

Previous work – POST-AL

- In 2012 Ptaszynski and Momouchi created POST-AL (“Part of Speech Tagger for the Ainu Language).
- POST-AL performs the following tasks:

1. Transcription normalization – modificaton of parts of text that do not conform to modern rules of transcription (e.g. *kamui* -> *kamuy*).
2. Word segmentation (tokenization) – a process in which the text is separated into tokens (words, punctuation marks, etc.), which become the basic unit for further analysis.
3. Part-of-speech tagging – assigning a part-of-speech marker to each token.
4. Word-to-word translation (into Japanese).

Example:



POST-AL tagger	iyosno ku hosipire kusne na
output:	【副】【人接】【他】【助動】【終助】
	最後に、終わり、後から、後で 私は、私が、私の 返す つもりである よ、か
Meaning:	“I'll return it later"

POST-AL’s dictionary base

- Originally, it contained one dictionary: *Ainu shin-yoshu jiten* (lexicon to Yukie Chiri’s *Ainu Shin-yoshu* (“Ainu Songs of Gods”)) by Kirikae (2003)
- 2,019 entries
- The dictionary has been transformed to XML format
- Each entry contains:

1. Token (word, morpheme, etc.)
2. Part of speech
3. Meaning (in Japanese)
4. Usage examples (not for all entries)
5. Reference to yukar story it appears in (not for all entries)

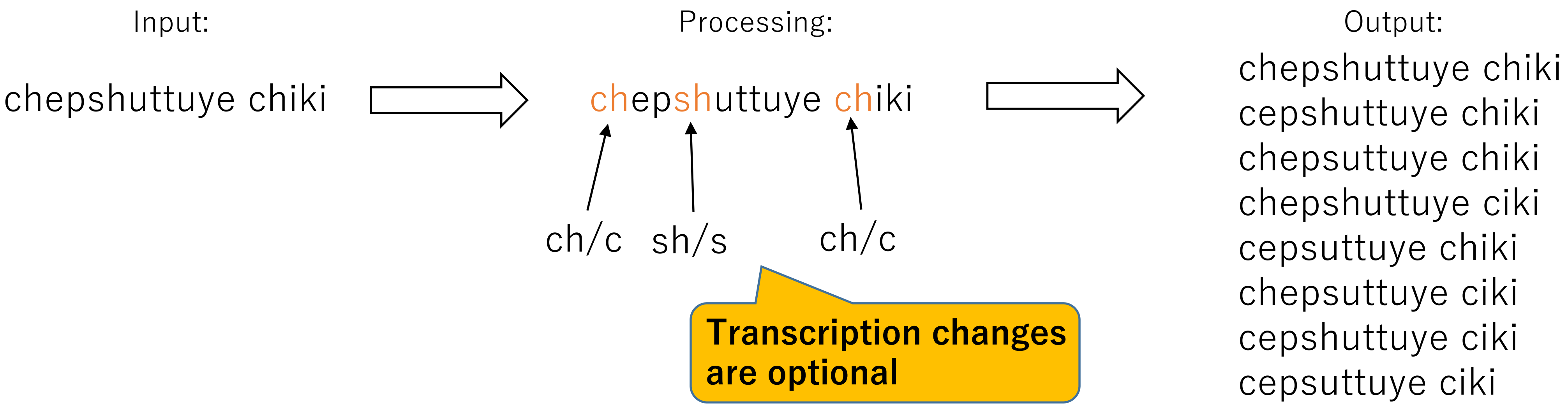
Sample entry:

```
<word>aep</word>
<morph>a$^{2}$-e$^{1}$-p$^{1}$</morph>
<pos>名詞</pos>
<tr>食べ物</tr>
<ref>aep'omuken</ref>
```

Improving transcription normalization

Transcription change rules:

Original transcription													
ch	sh(i)	ai	ui	ei	oi	au	iu	eu	ou	b	g	d	m
c	s	ay	uy	ey	oy	aw	iw	ew	ow	p	k	t	n
Modern transcription standard													



Input:

chepshuttuye chiki

Processing:

chepshuttuye chiki

ch/c sh/s ch/c

Transcription changes are optional

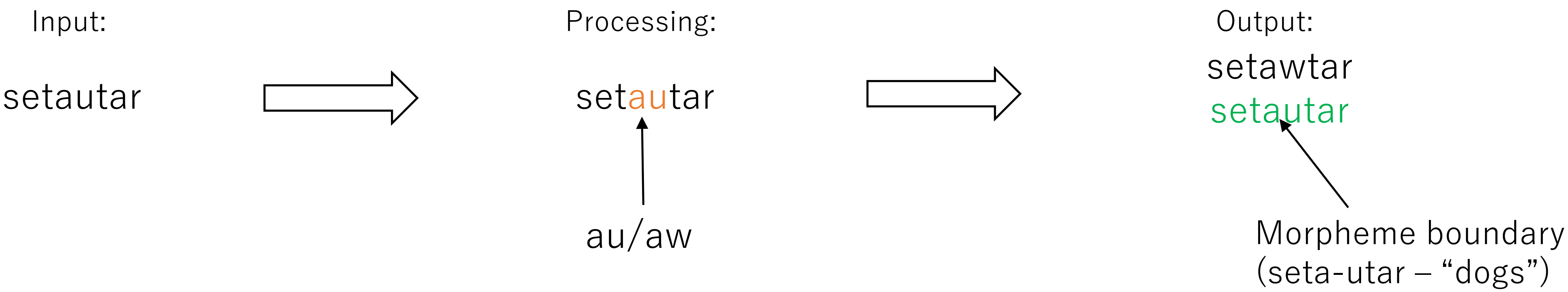
Output:

chepshuttuye chiki
cepshuttuye chiki
chepsuttuye chiki
chepshuttuye ciki
cepsuttuye chiki
chepsuttuye ciki
cepshuttuye ciki
cepsuttuye ciki

Improving transcription normalization

Transcription change rules:

Original transcription													
ch	sh(i)	ai	ui	ei	oi	au	iu	eu	ou	b	g	d	m
c	s	ay	uy	ey	oy	aw	iw	ew	ow	p	k	t	n
Modern transcription standard													



Input:

setautar

Processing:

setautar

au/aw

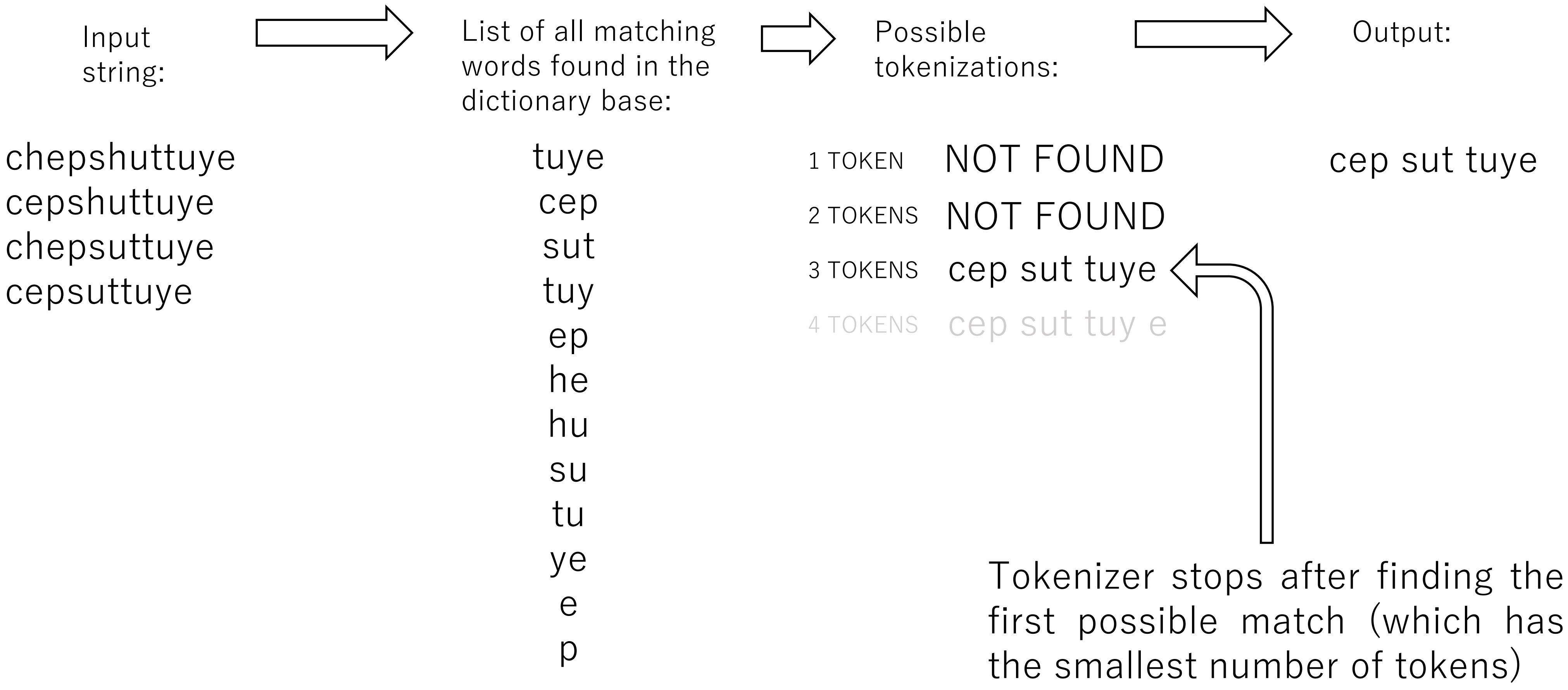
Output:

setawtar

setautar

Morpheme boundary
(seta-utar – “dogs”)

Improving tokenizer



Input string:

chepshuttuye
cepshuttuye
chepsuttuye
cepsuttuye

List of all matching words found in the dictionary base:

tuye
cep
sut
tuy
ep
he
hu
su
tu
ye
e
p

Possible tokenizations:

1 TOKEN	NOT FOUND
2 TOKENS	NOT FOUND
3 TOKENS	cep sut tuye
4 TOKENS	cep sut tuy e

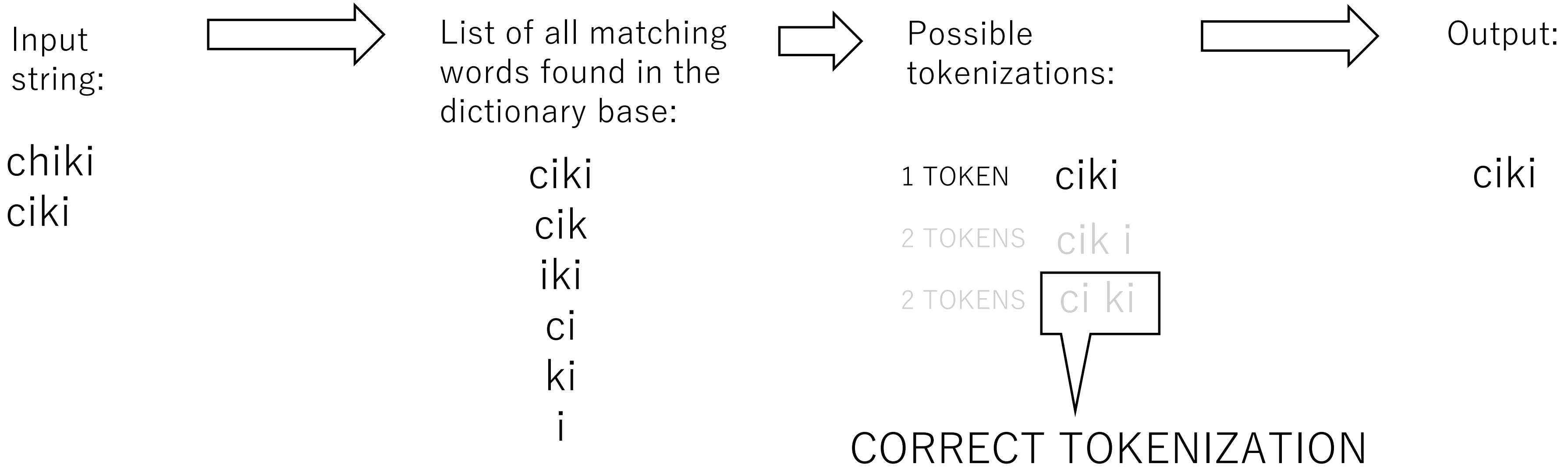
Output:

cep sut tuye

Tokenizer stops after finding the first possible match (which has the smallest number of tokens)

Improving tokenizer

PROBLEM: This tokenization algorithm always prefers long words over shorter ones.



Input string:

chiki
ciki

List of all matching words found in the dictionary base:

ciki
cik
iki
ci
ki
i

Possible tokenizations:

1 TOKEN ciki
2 TOKENS cik i
2 TOKENS ci ki

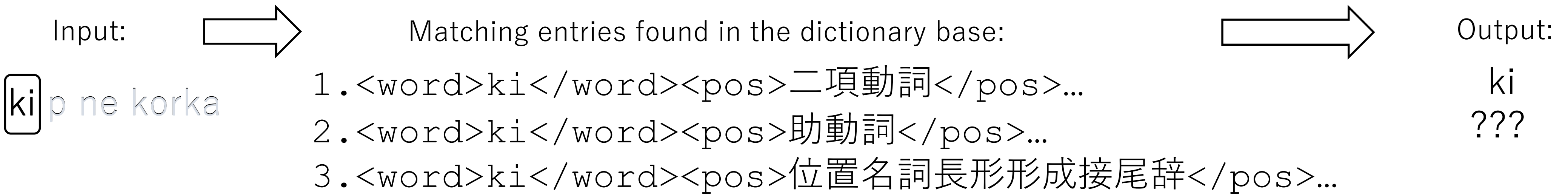
CORRECT TOKENIZATION

Output:

ciki

Improving part-of-speech tagger

“Tagging is a disambiguation task” (some words have more than one possible part-of-speech) (Jurafsky and Martin, 2016. *Speech and Language Processing*)



Input:

ki p ne korka

Matching entries found in the dictionary base:

1.<word>ki</word><pos>二項動詞</pos>…

2.<word>ki</word><pos>助動詞</pos>…

3.<word>ki</word><pos>位置名詞長形形成接尾辞</pos>…

Output:

ki
???

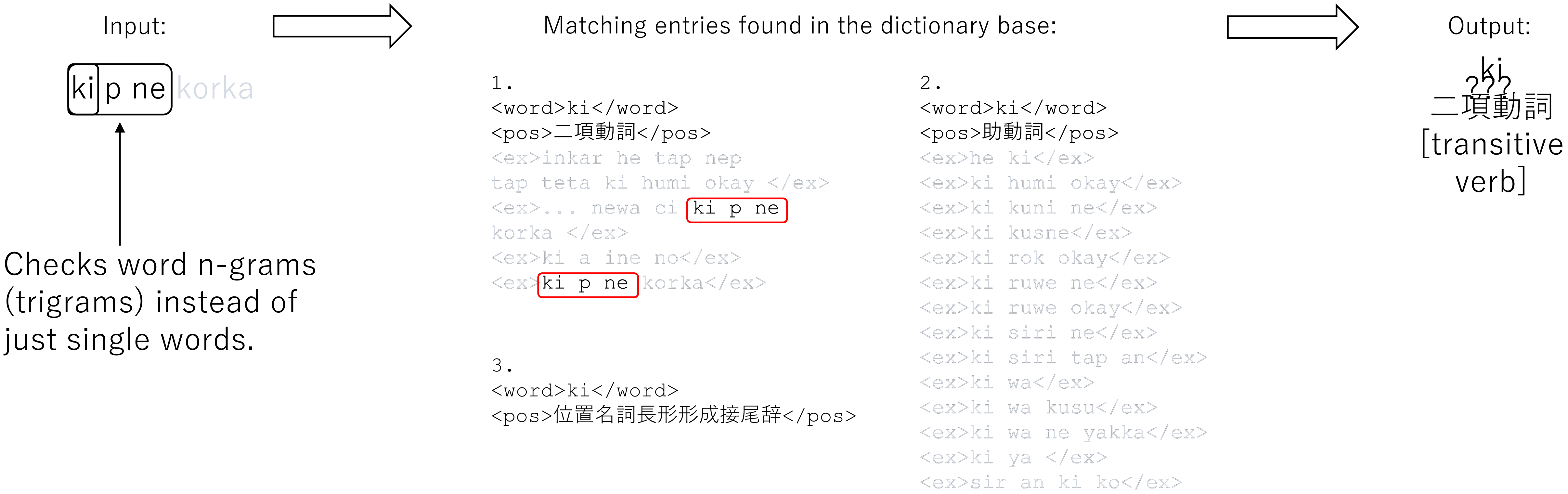
Two methods of POS disambiguation applied in POST-AL:

1. N-gram based POS disambiguation
2. Term Frequency (TF) based POS disambiguation

Improving part-of-speech tagger

N-gram based POS disambiguation:

* Uses sample sentences included in the dictionary base for determining the correct POS tag



Input:

ki p ne korka

Checks word n-grams (trigrams) instead of just single words.

Matching entries found in the dictionary base:

```
1.
<word>ki</word>
<pos>二項動詞</pos>
<ex>inkar he tap nep
tap teta ki humi okay </ex>
<ex>... newa ci ki p ne
korka </ex>
<ex>ki a ine no</ex>
<ex>ki p ne korka</ex>

3.
<word>ki</word>
<pos>位置名詞長形形成接尾辞</pos>
```

```
2.
<word>ki</word>
<pos>助動詞</pos>
<ex>he ki</ex>
<ex>ki humi okay</ex>
<ex>ki kuni ne</ex>
<ex>ki kusne</ex>
<ex>ki rok okay</ex>
<ex>ki ruwe ne</ex>
<ex>ki ruwe okay</ex>
<ex>ki siri ne</ex>
<ex>ki siri tap an</ex>
<ex>ki wa</ex>
<ex>ki wa kusu</ex>
<ex>ki wa ne yakka</ex>
<ex>ki ya </ex>
<ex>sir an ki ko</ex>
```

Output:

ki
???
二項動詞
[transitive verb]

Improving part-of-speech tagger

TF based POS disambiguation:

* Checks term frequency of each candidate word (= number of sample sentences included in the dictionary base) for determining the correct POS tag

Input:

ki p ne korka

Matching entries found in the dictionary base:

```
1.
<word>ki</word>
<pos>二項動詞</pos>
<ex>inkar he tap nep
tap teta ki humi okay </ex>
<ex>... newa ci ki p ne
korka </ex>
<ex>ki a ine no</ex>
<ex>ki p ne korka</ex>
```

4

```
2.
<word>ki</word>
<pos>助動詞</pos>
<ex>he ki</ex>
<ex>ki humi okay</ex>
<ex>ki kuni ne</ex>
<ex>ki kusne</ex>
<ex>ki rok okay</ex>
<ex>ki ruwe ne</ex>
<ex>ki ruwe okay</ex>
<ex>ki siri ne</ex>
<ex>ki siri tap an</ex>
<ex>ki wa</ex>
<ex>ki wa kusu</ex>
<ex>ki wa ne yakka</ex>
<ex>ki ya </ex>
<ex>sir an ki ko</ex>
```

14

```
3.
<word>ki</word>
<pos>位置名詞長形形成接尾辞</pos>
```

0

Output:

ki

???

助動詞

[auxiliary verb]

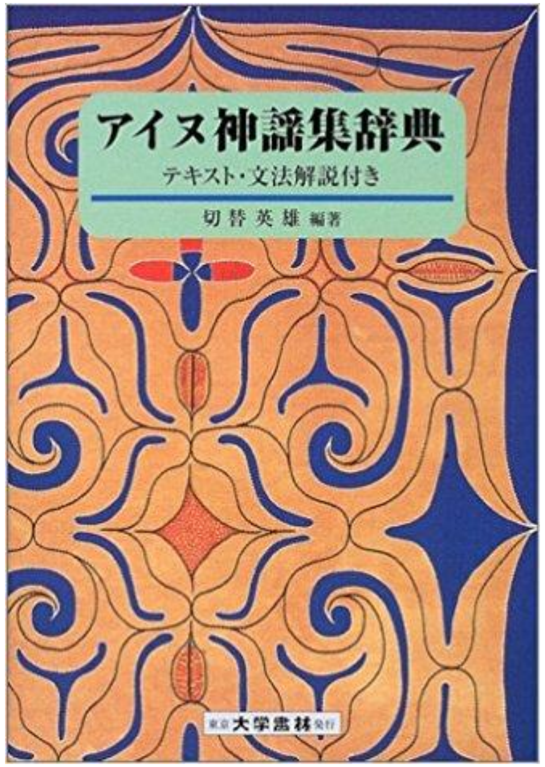
Improving part-of-speech tagger

- Word n-grams are more reliable as a method for POS disambiguation
- On the other hand, for many cases there are no relevant usage examples in the dictionary base
- To compensate for that, we created a modified tagging algorithm, which in such cases also takes into account the Term Frequency

Expanding POST-AL’s dictionary base

Dictionaries used:

1. *Ainu shin-yōshū jiten* (Kirikae, 2003) – based on classical Ainu language (*yukar* epics). The dictionary contains 2,019 entries.



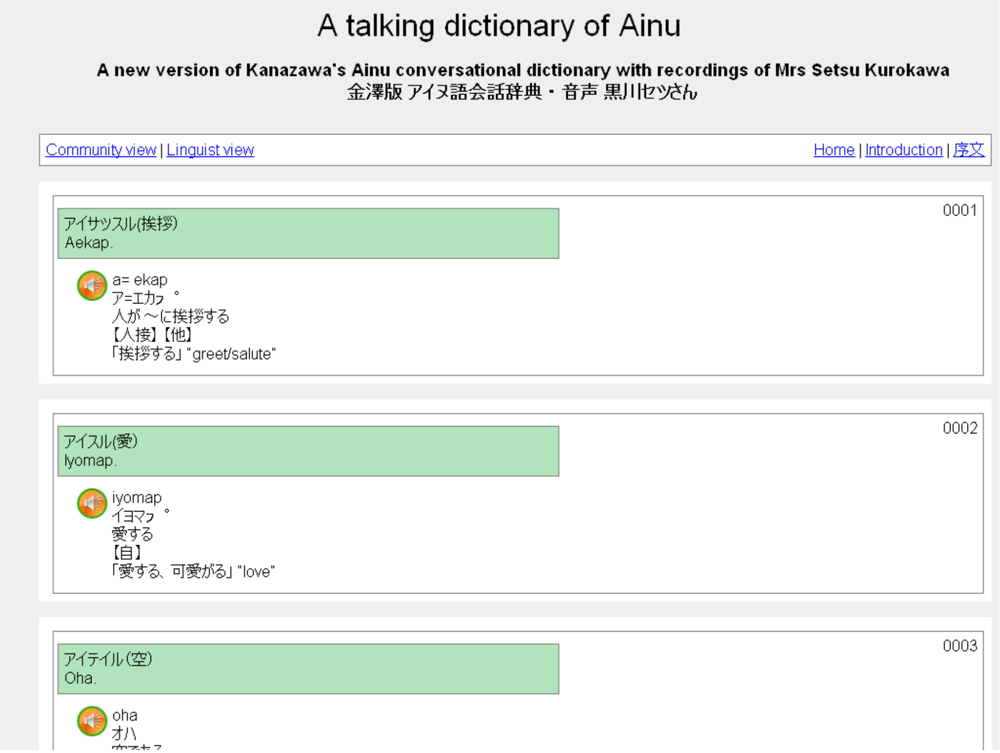
Sample entry:

```
<word>aep</word>
<morph>a$^{2}$-e$^{1}$-p$^{1}$</morph>
<pos>名詞</pos>
<tr>食べ物</tr>
<ref>aep'omuken</ref>
```

Expanding POST-AL’s dictionary base

Dictionaries used:

2. A Talking Dictionary of Ainu: A New Version of Kanazawa‘s Ainu Conversational dictionary (Bugaeva and Endō, 2010) – an online dictionary, based on the *Ainugo kaiwa jiten* (Jinbō and Kanazawa, 1898). Original dictionary contains 3,847 entries.



A talking dictionary of Ainu

A new version of Kanazawa's Ainu conversational dictionary with recordings of Mrs Setsu Kurokawa
金澤版 アイヌ語会話辞典・音声 黒川セツさん

Community view | Linguist view | Home | Introduction | 序文

0001
アイサツスル(挨拶)
Aekap.

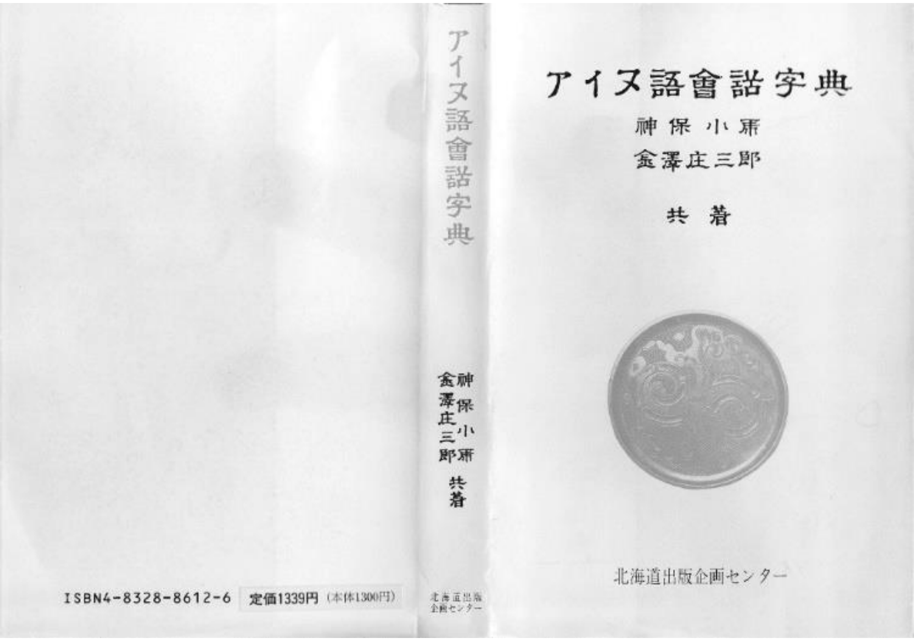
a= ekap
ア=エカㇷ゚
人が ～に挨拶する
【人接】【他】
「挨拶する」"greet/salute"

0002
アイスル(愛)
Iyomap.

iyomap
イヨマㇷ゚
愛する
【自】
「愛する、可愛がる」"love"

0003
アイテイル(空)
Oha.

oha
オハ

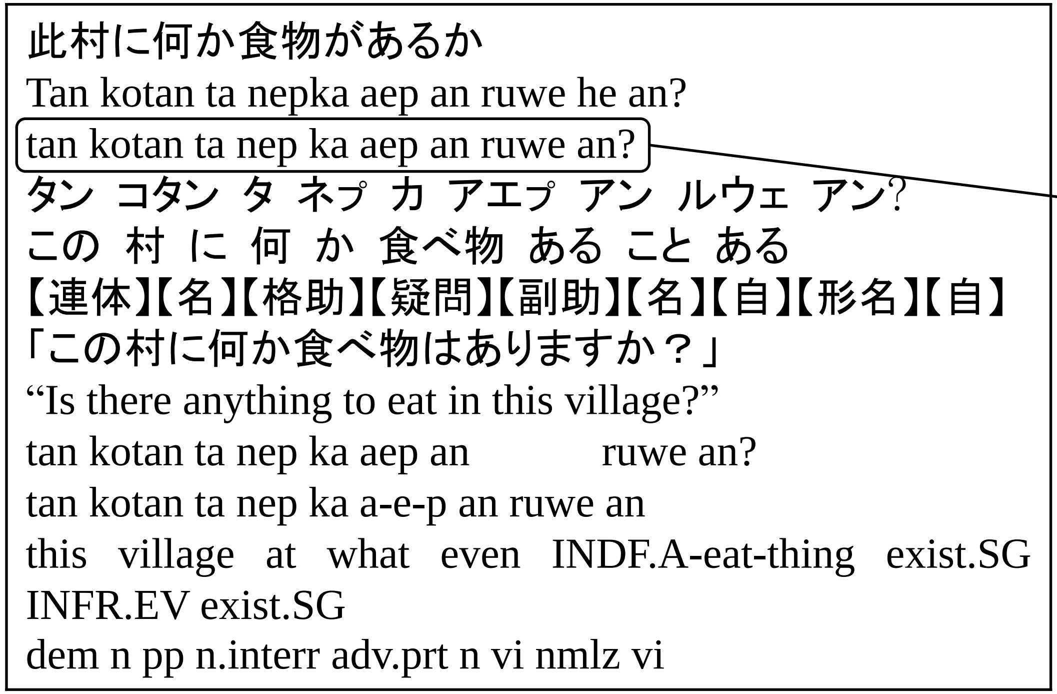


Expanding POST-AL’s dictionary base

Dictionaries used:

2. A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational dictionary (Bugaeva and Endō, 2010) – an online dictionary, based on the *Ainugo kaiwa jiten* (Jinbō and Kanazawa, 1898). Original dictionary contains 3,847 entries.

Sample entry (original):



此村に何か食物があるか
Tan kotan ta nepka aep an ruwe he an?
tan kotan ta nep ka aep an ruwe an?
タン コタン タ ネㇷ゚ カ アエㇷ゚ アン ルウェ アン?
この 村 に 何 か 食べ物 ある こと ある
【連体】【名】【格助】【疑問】【副助】【名】【自】【形名】【自】
「この村に何か食べ物はありますか？」
“Is there anything to eat in this village?”
tan kotan ta nep ka aep an ruwe an?
tan kotan ta nep ka a-e-p an ruwe an
this village at what even INDF.A-eat-thing exist.SG INFR.EV exist.SG
dem n pp n.interr adv.prt n vi nmlz vi

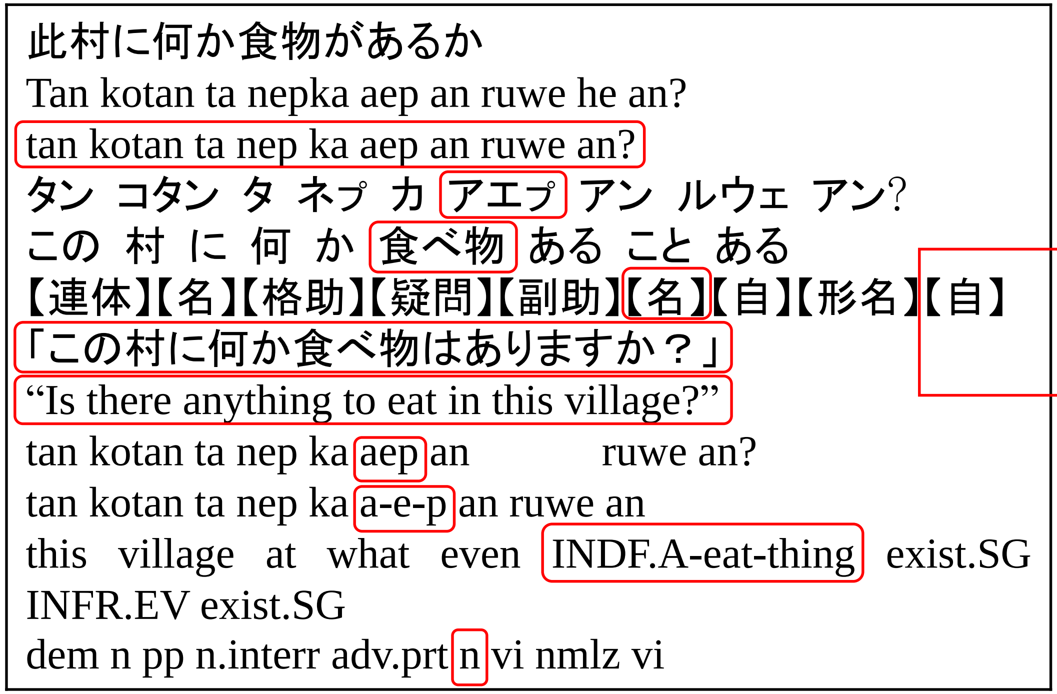
Original entries often consist of more than one word (multiple words or phrases)

Expanding POST-AL’s dictionary base

Dictionaries used:

2. A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational dictionary (Bugaeva and Endō, 2010) – an online dictionary, based on the *Ainugo kaiwa jiten* (Jinbō and Kanazawa, 1898). Original dictionary contains 3,847 entries.

Sample entry (original):



此村に何か食物があるか
Tan kotan ta nepka aep an ruwe he an?
tan kotan ta nep ka aep an ruwe an?
タン コタン タ ネㇷ゚ カ アエㇷ゚ アン ルウェ アン?
この 村 に 何 か 食べ物 ある こと ある
【連体】【名】【格助】【疑問】【副助】【名】【自】【形名】【自】
「この村に何か食べ物はありますか？」
“Is there anything to eat in this village?”
tan kotan ta nep ka aep an ruwe an?
tan kotan ta nep ka a-e-p an ruwe an
this village at what even INDF.A-eat-thing exist.SG INFR.EV exist.SG
dem n pp n.interr adv.prt n vi nmlz vi

Sample entry (modified dictionary):

```
<word>aep</word><kana>アエプ</kana>
<morph>a-e-p</morph><pos>【名】
</pos>
<pos_en>n</pos_en>
<tr>食べ物</tr><tr_en>food</tr_en>
<ge>INDF.A-eat-thing</ge>
<ex>tan kotan ta nep ka aep an ruwe
an?</ex>
<ex_jp>この村に何か食べ物はありますか？
</ex_jp>
<ex_en>Is there anything to eat in this
village?</ex_en>
```

Expanding POST-AL's dictionary base

Dictionaries used:

3. Combined dictionary (1+2).

A) extracted entries containing words listed in both dictionaries

B) automatically unified duplicate entries, basing on their Japanese translations (at least one kanji character in common)

Entry from *Ainu shin-yōshū jiten*

```
<word>aep</word>
<morph>a$^{2}$-e$^{1}$-p$^{1}$</morph>
<pos>名詞</pos>
<tr>食べ物</tr>
<ref>aep'omuken</ref>
```

Entry from Ainu Conversational Dictionary

```
<word>aep</word><kana>アエプ</kana>
<morph>a-e-p</morph><pos>【名】
</pos>
<pos_en>n</pos_en>
<tr>食べ物</tr><tr_en>food</tr_en>
<ge>INDF.A-eat-thing</ge>
<ex>tan kotan ta nep ka aep an ruwe
an?</ex>
<ex_jp>この村に何か食べ物はありますか？
</ex_jp>
<ex_en>Is there anything to eat in this
village?</ex_en>
```

Expanding POST-AL’s dictionary base

Dictionaries used:

3. Combined dictionary (1+2).

A) extracted entries containing words listed in both dictionaries

B) automatically unified duplicate entries, basing on their Japanese translations (at least one kanji character in common)

C) that resulted in a dictionary containing 4,161 entries.

```
<word>aep</word><kana>アエプ</kana>
<morph_kk>a$^{2}$-e$^{1}$-p$^{1}$</morph_kk>
<morph_jk>a-e-p</morph_jk>
<pos_jk>【名】</pos_jk>
<pos_kk>名詞</pos_kk>
<pos_en>n</pos_en>
<tr>食べ物</tr><tr_en>food</tr_en>
<ex>tan kotan ta nep ka aep an ruwe an?</ex>
<ex_jp>この村に何か食べ物はありますか？</ex_jp>
<ex_en>Is there anything to eat in this village?</ex_en>
<ge>INDF.A-eat-thing</ge>
<ref> aep'omuken</ref>
```

Entry from combined dictionary

Evaluation experiments

Transcription normalization results:

		Avg. result (F-score)
DICTIONARY	1. *Ainu shin-yōshū jiten* (Kirikae)	91.85%
	2. Ainu Conversational Dictionary (Jinbō and Kanazawa)	87.96%
	3. Combined dictionary (1+2)	92.48%

Tokenization results:

		Avg. result (F-score)
DICTIONARY	1. *Ainu shin-yōshū jiten* (Kirikae)	86.73%
	2. Ainu Conversational Dictionary (Jinbō and Kanazawa)	69.93%
	3. Combined dictionary (1+2)	87.73%

Evaluation experiments

POS tagging results:

		Avg. result (F-score)	Tagging algorithm version:	
		Avg.	N-grams	TF
DICTIONARY	1. *Ainu shin-yōshū jiten* (Kirikae)	72.16%	NO	YES
		71.71%	YES	NO
		74.72%	YES	YES
	2. Ainu Conversational Dictionary (Jinbō and Kanazawa)	80.01%	NO	YES
		77.28%	YES	NO
		81.55%	YES	YES
	3. Combined dictionary (1+2)	90.62%	NO	YES
		90.27%	YES	NO
		92.82%	YES	YES

Conclusions

1. Improved the following functions of POST-AL:

- Transcription normalization
- Tokenizer
- POS tagger

2. Expanded POST-AL's dictionary base by combining 2 dictionaries:

- found out that the combination improved overall performance of the system

Thank you for your attention!

Evaluation experiments

Applied datasets:

- Yukar (9-13) from *Ainu shin-yōshū* (“Ainu Songs of Gods”)
- JK dictionary sample sentences
- Sample text from Masayoshi Shibatani’s *The Languages of Japan*
- Mukawa dialect samples (by Chiba University)

アイヌ民譚集

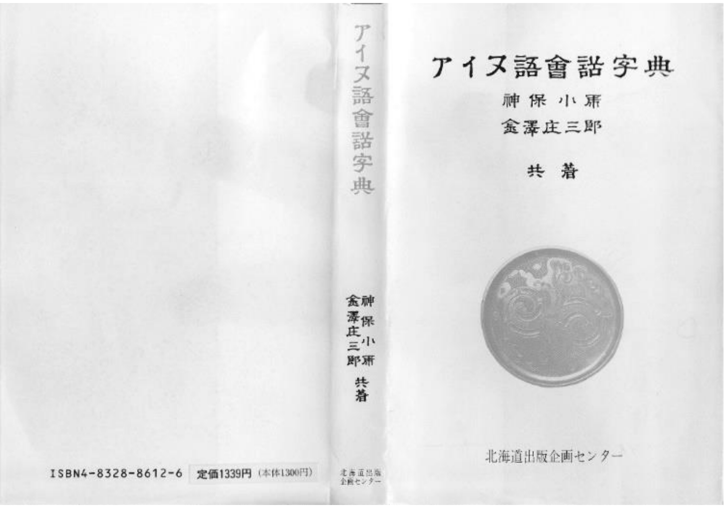
付，えぞおばけ列伝

知里真志保編訳

アイヌに伝わる数々の昔話のうち，とくに「パナンペ説話」を選んで訳出したもの。よろずに巧智を発揮して成功するパナンペをみたペナンペは，それをうらやんで真似するがかえってひどい目にあうばかり。野趣に富む痛快な物語15篇を収める。併せて，人間くさいおばけが次々と登場する「えぞおばけ列伝」を付した。(解説＝関 敬吾)

赤 81-1

岩波文庫



アイヌ語會話字典

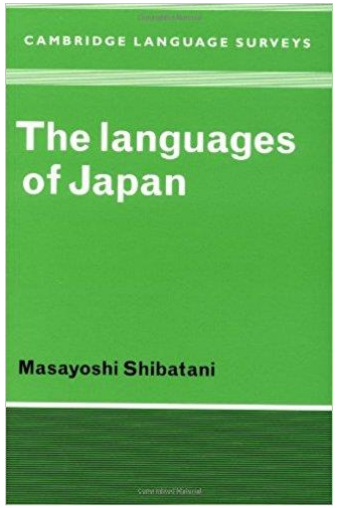
神保小虎

金澤庄三郎

共著

北海道出版企画センター

ISBN4-8328-8612-6 定価1339円（本体1300円）



CAMBRIDGE LANGUAGE SURVEYS

The languages of Japan

Masayoshi Shibatani

アイヌ語鵡川方言

日本語ーアイヌ語辞典

アイヌ語鵡川方言

日本語ーアイヌ語辞典

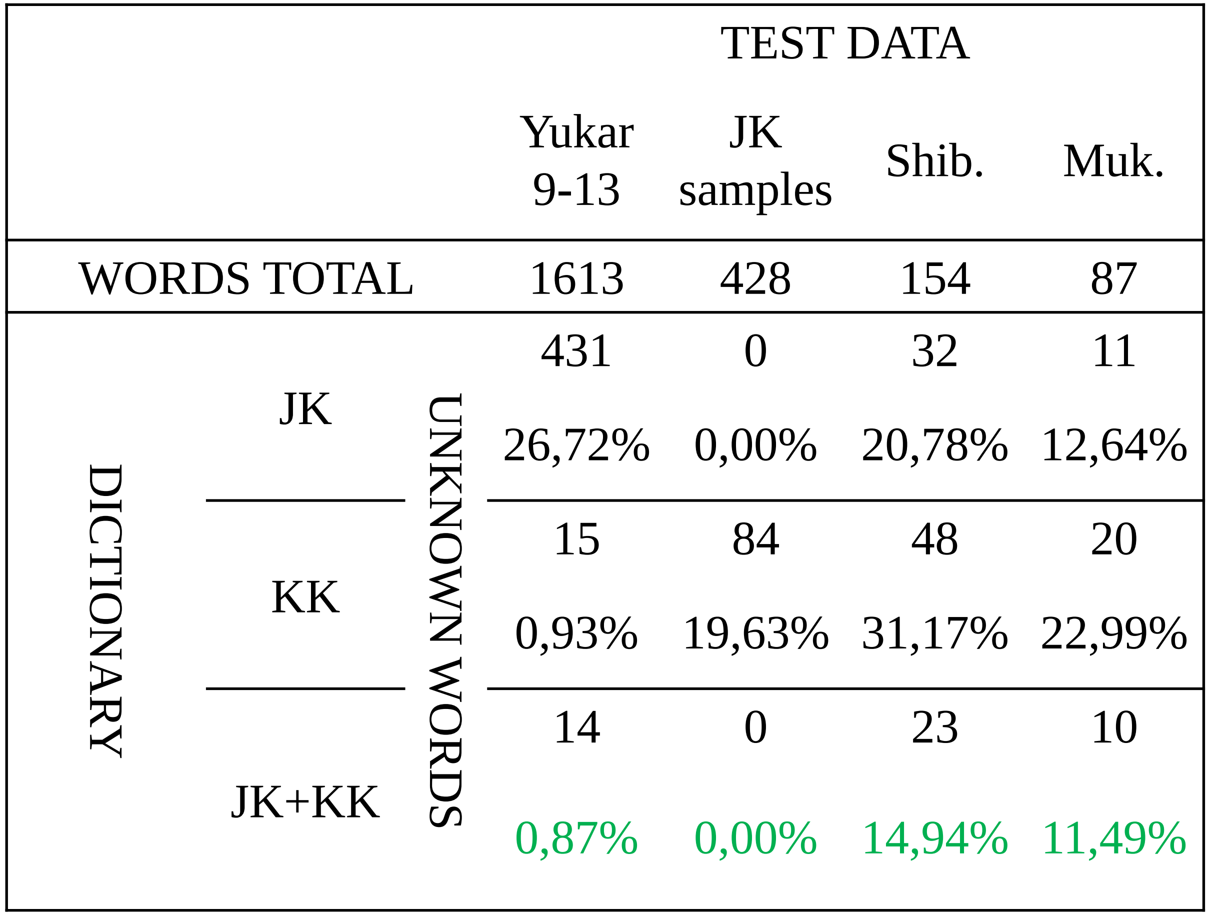
下のボックスに調べたい単語を入れて検索し

索引：【選択して下さい】

範囲：□見出し語 □内容 □和訳 □音声ファイル名

Evaluation experiments

Statistics of unknown words:



			TEST DATA			
			Yukar 9-13	JK samples	Shib.	Muk.
WORDS TOTAL			1613	428	154	87
DICTIONARY	JK	UNKNOWN WORDS	431	0	32	11
			26,72%	0,00%	20,78%	12,64%
	KK		15	84	48	20
			0,93%	19,63%	31,17%	22,99%
	JK+KK		14	0	23	10
			0,87%	0,00%	14,94%	11,49%

Evaluation experiments

Transcription normalizat

		Yukar 9	Yukar 10					
DICTIONARY	JK	92.34%	91.35%					
	KK	97.18%	98.55%					
	JK+KK	96.43%	97.11%	94.80%	91.41%	96.79%	78.32%	92.48%

Relatively low results for sample sentences from JK dictionary.
Explanation:
We decided not to apply some of the transcription change rules observed only in that dictionary (such as ‘ra’→’r’ (e.g. *arapa→arpa*) or ‘ei’→’e’ (e.g. *reihei→rehe*)), as initial tests indicated that including them in the algorithm can cause errors with processing yukars and other texts.

Evaluation experiments

Tokenization experiment results (F-score):

		TEST DATA								
		Yukar 9	Yukar 10	Yukar 11	Yukar 12	Yukar 13	JK samples	Shibatani	Mukawa	Avg.
DICTIONARY	JK	66.53%	64.40%	67.33%	64.13%	67.80%	87.07%	72.80%	74.40%	69.93%
	KK	89.23%	92.30%	93.07%	85.63%	92.73%	74.80%	68.60%	76.20%	86.73%
	JK+KK	85.37%	91.40%	90.03%	84.63%	91.87%	87.10%	79.90%	79.80%	87.73%

Evaluation experiments

F

The gap between the results of tagging Yukar 10 and samples from JK dictionary can be partially explained by differences in part of speech classification of certain words between the two dictionaries applied in the system and the annotations (gold standard) provided by Momouchi (2008). For example, Momouchi annotated the word *ne* („to be”) as ‘auxiliary verb’, whereas in the dictionary base it is listed as ‘transitive verb’.

DICTIONARY						
	JK+KK	85	6.00%	90.62%	NO	YES
		86.3	94.20%	90.27%	YES	NO
		87.95%	97.70%	92.82%	YES	YES

Future tasks

1. Develop a tokenization algorithm based on word n-grams rather than single words.
2. Enlarge the dictionary base by adding other dictionaries, such as the *Ainu-Japanese Dictionary: Saru Dialect* by Suzuko Tamura.
3. Expand the dictionary base with the information about alternative transcription methods appearing in older texts (in order to improve the normalization of transcription in such texts).
4. Build a statistical model of the Ainu language, reflecting probability distribution over different sequences (bigrams or trigrams) of parts of speech, and use it to improve POS tagging performance.