NLP for Endangered Languages: Morphology Analysis, Translation Support and Shallow Parsing of Ainu Language

Michal Ptaszynski † Mukaichi Kazuki ‡ Yoshio Momouchi ‡
† Department of Computer Science, Kitami Institute of Technology
ptaszynski@cs.kitami-it.ac.jp
‡ Department of Electronics and Information Engineering,
Faculty of Engineering, Hokkai-Gakuen University
momouchi@eli.hokkai-s-u.ac.jp

Abstract

This paper describes our research on computer processing of Ainu language with the use of various NLP techniques. Ainu is an endangered language close to extinction. At present linguists and anthropologists make a great effort to preserve the language by analyzing and understanding it. However, most of the work in this matter is done manually, which makes it an uphill task. Previously we have presented POST-AL, a part-of-speech tagger for Ainu language. This paper describes recent improvements to the system as well as other enhancements made with an aim to help Ainu language researchers. In particular, we have enhanced the POS tagger with analysis of morphological information. We have also added a translation support tool for Ainu language translators and made a first step toward deeper syntactical analysis of Ainu language by creating a simple shallow parser.

1 Introduction

Ainu language is a language of Ainu people, mostly living on Hokkaido, a northern island of Japan. Ainu are the native inhabitants of northern parts of Japan and Sakhalin, however, their exact origin is unknown. Similarly, the origins of Ainu language have also been a mystery for linguists. Till the present day no proof has been proposed showing similarity of Ainu language to any other known world language. This makes Ainu language a unique language isolate [1]. It is also one of the most critically endangered world languages [2].

By this research we most of all wish to contribute to the task of reviving the Ainu language. At present linguists and anthropologists join their efforts to preserve the language. One way to approach the full understanding of the language is by performing a thorough analysis of Ainu language artifacts, such as myths and stories. However, most of the work in this matter is done manually, which makes it an uphill task. Previously we have presented POST-AL, a ***P**art-**O**f-**S**peech **T**agger for **A**inu **L**anguage* [3, 4]. This paper describes recent improvements to the system as well as other enhancements made with an aim to help Ainu language researchers. In particular, we have enhanced the POS tagger with analysis of morphological information. We have also added a translation support tool for Ainu language translators and made a first step toward deeper syntactical analysis of Ainu language by creating a simple shallow parser.

The paper outline is as follows. In section 2 we present previous research done in Ainu language with focus on linguistic and NLP-related research. Section 3 presents the description of system components (tokenization, POS tagging, translation and parsing). Section 4 contains conclusions and points out some of the future work plans.

2 Previous Research

Some of the first research on Ainu language are dated on the end of 19th century. It was performed by Bronisław Piłsudski, a Polish cultural anthropologist. Piłsudski studied Ainu culture and language, and prepared some of the first glossaries [5]. A few years later Batchelor [6] published his *Ainu-English-Japanese Dictionary*. Among linguistic research done in modern times, most consist of collections of Ainu epic stories and myths [5, 7], dictionaries and lexicons [8, 9, 10], and grammar descriptions [11, 13, 14]. As for the research in NLP, Momouchi and colleagues took an attempt to create a machine translation system for Ainu (to Japanese). Within this research Azumi and Momouchi [16, 17] prepared ground for analysis and retrieval of hierarchical Ainu-Japanese translations. Momouchi et al. [18] began a process of annotating Ainu “yukar” stories for the need of machine translation system. Lastly, Momouchi and Kobayashi [19] began creation of a system for translation of Ainu place names.

Except the above research we have previously presented POST-AL, a part-of-speech tagger for Ainu language [3, 4]. The system performs tokenization, POS tagging and translation of Ainu language tokens to Japanese language. In the next section we describe the present state of the system and improvements to the previous version.

3 System Description

The system as described in our previous research [3, 4] performs tokenization, POS tagging and token translation. In the below description we will abbreviate the previously described parts and focus on improvements.

3.1 Dictionary

As the base dictionary for POST-AL we used *Ainu shin-yōshū jiten* (Lexicon to Yukie Chiri's Ainu Shin-yōsyū (Ainu Songs of Gods)) by written by Kirikae [10]. It is one of the newest Ainu language dictionaries with a firm part-of-speech classification developed especially to reflect the differences between Ainu parts of speech model to models of other languages. The dictionary by Kirikae [10] was transformed into an XML database using dictionary source files provided by the author of the dictionary himself. The original text of the dictionary contains different types of information, all of which were used in the system:

1. token (word, morpheme, etc.)
2. part of speech
3. translation/meaning (in Japanese)
4. reference to the story it appears in (not for all cases)
5. usage examples (not for all cases)

3.2 Tokenizer

Tokenization is a process in which the text is separated into tokens. In general tokens consist of words and punctuation marks. Texts in Ainu language, which usually include stories and narratives, most often appear in their printed form either undivided, or with chunks of text separated with a caesura (pause in recitation within one line of a poem). Therefore we needed to apply a tokenization method to be able to perform POS tagging of untokenized texts. We applied a standard approach to tokenization, namely dictionary lookup (DL). In the **DL-LSM** method (*Dictionary Lookup with Longest String Matching*) the input text is firstly glued together disregarding any other potential separations. Then the dictionary lookup is performed according to the Longest Match Principle, which assumes that the matching is done beginning with the longest words in the lexicon ending on the shortest ones.

3.3 POS-Tagger

Initially we developed and compared two methods for part-of-speech tagging. The first one (S-POST), based on statistics of parts of speech in the lexicon. The second one (CON-POST), based on a higher order HMM, using n-grams as contextual information for the processed word. As the CON-POST method showed better much performance we describe it shortly below.

CON-POST: (*Contextual Part of Speech Tagging*) This method uses an approach to POS tagging based on a higher order Hidden-Markov Model (HMM). HMM is a model in which a given word is analyzed with respect to the word preceding or succeeding it (bigrams). A higher order HMM is taking into account not one, but two or more succeeding words (trigrams and longer). We trained the HMM model on the examples that appear in the original dictionary on which the system is based.

3.4 Morphological Analysis

Morphological analysis is one of the main improvements to the original system. The task of morphological analysis in linguistic terms refers to the analysis of linguistic units (morphemes, root words, prefixes, etc.) within the structure of language. This task requires different approaches depending on the type of analyzed language. In English for example, being analytic language (with small degree of inflection), it is sufficient to focus on isolated morphemes. Agglutinative languages, like Japanese, join word stems and morphemes. On the other hand, in polysynthetic languages, like Ainu language, meaning of one word can consist of a number of compound morphemes. Therefore in the task of morphological analysis of Ainu language it is necessary to separate all morphemes a word is made of. Moreover, since one morpheme can have many meanings it is necessary to specify the exact meaning the morpheme is used in. For example, a word *eramesinne* ("to feel relieved to see/because of something [expressed previously]"), used as a single transitive verb can be analysed as follows:

e	supplement prefix, "by the means of; because of",
rame	variant of *ram*,
∟ *ram*	noun, "heart, mind",
sinne	intransitive verb, "to feel calm, relieved".

The word *eramesinne* is composed of three other words (*e, rame, sinne*). Moreover, one of the words, *rame*, is a variant of word *ram*, which information also needs to be included in the analysis. This way each word needs to be analyzed recursively until the original meaning of a single morpheme is reached. The result of such analysis is the meaning structure of the original word ("to feel relieved because of something") shown as a sum of smaller compound meanings ("becaufe of" [something] + "heart" + "feels calm, relieved").

The dictionary by Kirikae, on which POST-AL is based, contains both the compound words as well as separate morphemes. Moreover, in situations where the meaning of morpheme is ambiguous (more than one meaning) Kirikae added annotations of references to the original morphemes. Therefore using Kirikae's dictionary and analyzing each compound word recursively we were able to perform basic morphological analysis for Ainu language. One example of the analysis is represented in Figure 1.

3.5 Token Translation

Translation of tokens annotated with POS is an additional feature we included in the system. The translations of tokens (to Japanese) are selected from the lexicon. Previously we compared two methods for selecting the translations: random and contextual. The contextual method achieved better results, thus we describe it shortly below.

CON-ToT: (*Contextual Token Translation*) This method is the extension of CON-POST. The translation is selected specifically for the word selected in the contextual part-of-speech tagging, based on Hidden Markov Model trained on the dictionary examples.

3.6 Translation Support

Translation support is another improvement to the previous system. The translations of tokens in Kirikae's dictionary contain as much information as possible. For example, the word *eramesinne* is translated as *∼ga∼wo mite anshin suru* ("∼(subj.)∼(obj.) see and become relieved"). This means that when the word is used there is always a (specified) subject which is relieved because of a (specified) object[1]. However, in the process of translation the information included in the full token translation becomes redundant, which could hinder the translation. Therefore to help Ainu language researchers and translators we added an option to simplify token translations. When the translation simplification option is enabled, all redundant particles are deleted from the front of words and all brackets with additional information as well as other characters, like ∼ are deleted from the translations. An example of this function is represented in Figure 1. The translations in the simplified form could also be more useful in training a machine translation system for Ainu language [16, 17, 19].

3.7 Shallow Parsing

Although POS tagging provides information on parts of speech for each word, the analysis of the language is incomplete without a deeper analysis of sentence structure. Usually such analysis (parsing) is performed with a grammar parser. Parsing a sentence divides the sentence into separate chunks and provides analysis of relationships between those chunks in a tree-like output. It shows how words are related to each other in an overall sentence structure. Ainu language, being a polysynthetic language, has a sophisticated grammar. Many words relate to each other and constitute different meanings depending on context. There have been several attempts to describe the grammar structure of Ainu language [11, 12, 13, 14]. However, no computational model of Ainu language grammar has been proposed till now. Therefore it is still difficult to create a fully functional parser for Ainu language. However, the sentence structure of Ainu language is generally similar to Japanese. Therefore we assumed some general rules for sentence chunking should also work for the Ainu language. We applied the POS information obtained from the main system to create a simple shallow parser, or chunker. The chunker firstly divides the sentence into clauses containing the longest possible string of nouns consecutive verbs and closely related morphemes (particles, prefixes, suffixes). Next it divides the clauses into noun phrases (NP) and verb phrases (VP). This simplified parser does not provide all information about the sentence, however it could become useful in creating an actual parser in the future.

3.8 System Output Options

Except the part of speech naming model used by Kirikae, there are two others. The first one, more general by Nakagawa [8], attempts to minimize the number of POS names. The second one on the other hand, by Tamura [9], is more robust and proposes many separate part of speech names. The three POS naming standards can be considered the most influential. It is not in our competence to arbitrarily decide which standard is the best, thus we included all of them in POST-AL. As the default one we use the one by Kirikae. However, the output can be changed by adding an additional parameter in the command line ("-t" - for Tamura and "-n" for Nakagawa). This shifting of POS naming standards was possible due to the fact that Kirikae follows Tamura in the naming sophistication (with some simplifications), and Nakagawa in the naming form. For example, a verb *ahun* "to enter [the house]" appears as "intransitive verb" in Tamura, but as "type-1 verb" in Nakagawa and Kirikae. Except the main three we also added a more general simplified POS classification as well as translation of all POS names to English. An example showing outputs in different standards is represented in figure 1. As an option we also added two possible versions of output. The first one, vertical, typical for POS taggers, and the second one, horizontal, more readable and familiar to language anthropologists studying Ainu language. The two types of output are represented in figure 1.

4 Conclusions and Future Work

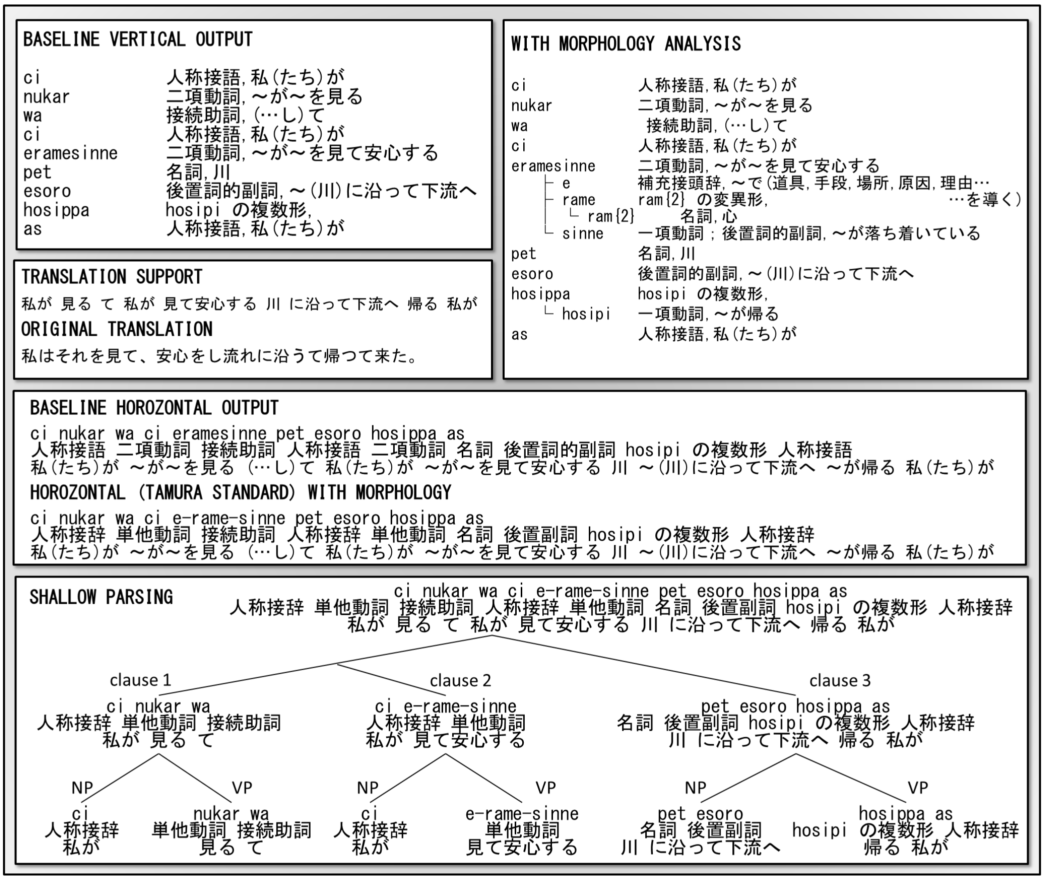
In this paper we described our research on processing of Ainu language. We described recent improvements to the system as well as other enhancements made with an aim to help Ainu language researchers. In particular, we have enhanced the POS tagger with analysis of morphological information. We have also added a translation support tool for Ainu language translators and made a first step toward deeper syntactical analysis of Ainu language by creating a simple shallow parser. All of those components still need to be evaluated in the future. At present thorough evaluation is difficult due to the lack of appropriate datasets. In the near future we plan to create datasets for evaluation of shallow parser as well as apply the translation support tool in training a machine translation system.

Acknowledgments

This research was supported by (JSPS) KAKENHI Grant-in-Aid for Scientific Research (Project Number: 24600001).

The authors express their gratitude to Associate Professor Hideo Kirikae of Hokkai-Gakuen University for his contribution of the dictionary source files and numerous insightful comments on the desirable structure and usability of the system.

[1] There is also a different word which has similar meaning, but is used without specifying an object, *yaykahumsu*, which translates as "to feel relieved; heave a sigh of relief; stroke one's chest with relief".



```
BASELINE VERTICAL OUTPUT

ci            人称接語,私(たち)が
nukar         二項動詞,～が～を見る
wa            接続助詞,(…し)て
ci            人称接語,私(たち)が
eramesinne    二項動詞,～が～を見て安心する
pet           名詞,川
esoro         後置詞的副詞,～(川)に沿って下流へ
hosippa       hosipi の複数形,
as            人称接語,私(たち)が
```

```
TRANSLATION SUPPORT
私が 見る て 私が 見て安心する 川 に沿って下流へ 帰る 私が
ORIGINAL TRANSLATION
私はそれを見て、安心をし流れに沿うて帰つて来た。
```

```
WITH MORPHOLOGY ANALYSIS

ci            人称接語,私(たち)が
nukar         二項動詞,～が～を見る
wa             接続助詞,(…し)て
ci            人称接語,私(たち)が
eramesinne    二項動詞,～が～を見て安心する
   ├ e        補充接頭辞,～で(道具,手段,場所,原因,理由…
   ├ rame     ram{2} の変異形,                    …を導く)
   │ └ ram{2}     名詞,心
   └ sinne    一項動詞；後置詞的副詞,～が落ち着いている
pet           名詞,川
esoro         後置詞的副詞,～(川)に沿って下流へ
hosippa       hosipi の複数形,
   └ hosipi   一項動詞,～が帰る
as            人称接語,私(たち)が
```

```
BASELINE HOROZONTAL OUTPUT
ci nukar wa ci eramesinne pet esoro hosippa as
人称接語 二項動詞 接続助詞 人称接語 二項動詞 名詞 後置詞的副詞 hosipi の複数形 人称接語
私(たち)が ～が～を見る (…し)て 私(たち)が ～が～を見て安心する 川 ～(川)に沿って下流へ ～が帰る 私(たち)が
HOROZONTAL (TAMURA STANDARD) WITH MORPHOLOGY
ci nukar wa ci e-rame-sinne pet esoro hosippa as
人称接辞 単他動詞 接続助詞 人称接辞 単他動詞 名詞 後置副詞 hosipi の複数形 人称接辞
私(たち)が ～が～を見る (…し)て 私(たち)が ～が～を見て安心する 川 ～(川)に沿って下流へ ～が帰る 私(たち)が
```

Figure 1: Examples of different outputs.

References

[1] Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge university Press, London.

[2] Christopher Moseley (ed.). 2010. *Atlas of the Worlds Languages in Danger*, 3rd ed. Paris, UNESCO Publishing. Online version: http://www.unesco.org/culture/languages-atlas/

[3] Michal Ptaszynski and Yoshio Momouchi, "POST-AL: Part-of-Speech Tagger for Ainu Language", In *Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, pp. 763-766, 2012.

[4] Michal Ptaszynski and Yoshio Momouchi, "Part-of-Speech Tagger for Ainu Language Based on Higher Order Hidden Markov Model", *Expert Systems With Applications*, Vol. 39, Issue 14 (2012), pp. 1157611582, Elsevier, 2012.

[5] Bronisław Piłsudski (Author), Alfred F. Majewicz (Editor). 2004. *The Collected Works of Bronislaw Pilsudski: Materials for the Study of the Ainu Language and Folklore*, v.3, Pt. 2: Materials for the Study of the Ainu, (Trends in Linguistics: Documentation). Mouton de Gruyter (Oct 2004)

[6] John Batchelor. 1905. *An Ainu-English-Japanese dictionary (including a grammar of the Ainu language)*. Tokyo Methodist Pub. House.

[7] Yukie Chiri. 1978. *Ainu shin-yōshū*. Tokyo, Iwanami Shoten.

[8] Hiroshi Nakagawa. 1995. *Ainugo Chitose Hōgen Jiten: The Ainu-Japanese Dictionary: Chitose Dialect*. Sōfūkan.

[9] Suzuko Tamura. 1998. *Ainugo Chitose Hōgen Jiten: The Ainu-Japanese Dictionary: Saru Dialect* [In Japanese]. Sōfūkan.

[10] Hideo Kirikae. 2003. *Ainu shin-yōshū jiten: tekisuto bumpō kaisetsu tsuki* (Lexicon to Yukie Chiri's Ainu Shin-yōsyū (Ainu Songs of Gods) with Text and Grammatital Notes) [In Japanese]. Sapporo: Hokkaidō Daigaku Bungakubu Gengōgaku.

[11] Kyōko Murasaki. 1979. *Karafuto ainugo. Bunpō-hen* (Sakhalin Ainu. Grammar volume) [In Japanese]. Tokyo, Kokushokan-kōkai.

[12] Kyōsuke Kindaichi. 1993. Ainu yūkara gohō tekiyō (An outline grammar of Ainu epic stories) [In Japanese]. In *Ainugogaku kōgi* 2 (Lectures on Ainu studies 2). *Kindaichi Kyōsuke zenshū. Ainugo I*, v. 5, 145-366. Tokyo, Sanseidoo.

[13] Tomomi Satō. 2008. Ainugo bunpō no kiso (The basics of Ainu grammar) [In Japanese]. Tokyo, Daigakushorin.

[14] Suzuko Tamura. 2000. *The Ainu Language*. Tokyo, Sanseido.

[15] Anna Bugaeva. 2010. Internet Applications for Endangered Languages: A Talking Dictionary of Ainu. *Waseda Institute for Advanced Study Research Bulletin*, No.3, pp. 73-81.

[16] Yasunori Azumi and Yoshio Momouchi. 2009a. Development of Analysis Tool for Hierarchical Ainu-Japanese Translation Data [In Japanese]. *Bulletin of the Faculty of Engineering at Hokkai-Gakuen University*, No.36, pp.175-193.

[17] Yasunori Azumi and Yoshio Momouchi. 2009b. Development of Tools for retrieving and analyzing Ainu-Japanese translation data and their applications to Ainu-Japanese machine translation system [In Japanese]. *Engineering Research: The Bulletin of Graduate School of Engineering at Hokkai-Gakuen University*, No.9, pp.37-58.

[18] Yoshio Momouchi, Yasunori Azumi and Yukio Kadoya. 2008. Research Note: Construction and Utilization of Electronic Data for "Ainu Shin-yōsyū" [In Japanese]. Bulletin of the Faculty of Engineering at Hokkai-Gakuen University, No. 35, pp. 159-171.

[19] Yoshio Momouchi and Ryosuke Kobayashi. 2010. Dictionaries and Analysis Tools for the Componential Analysis of Ainu Place Name [In Japanese]. *Engineering Research: The Bulletin of Graduate School of Engineering at Hokkai-Gakuen University*, No.10, pp.39-49.