information

MDPI

Article

Improving Basic Natural Language Processing Tools for the Ainu Language

Karol Nowakowski [1,*], Michal Ptaszynski [1], Fumito Masui [1] and Yoshio Momouchi [2]

[1] Department of Computer Science, Kitami Institute of Technology, 165 Koen-cho, Kitami, Hokkaido 090-8507, Japan; ptaszynski@cs.kitami-it.ac.jp (M.P.); f-masui@mail.kitami-it.ac.jp (F.M.)

[2] Department of Electronics and Information Engineering, Faculty of Engineering, Hokkai Gakuen University, 1-1, Nishi 11-chome, Minami 26-jo, Chuo-ku, Sapporo, Hokkaido 064-0926, Japan

* Correspondence: karol_nowakowski@interia.pl

Received: 27 August 2019; Accepted: 20 October 2019; Published: 24 October 2019

Abstract: Ainu is a critically endangered language spoken by the native inhabitants of northern Japan. This paper describes our research aimed at the development of technology for automatic processing of text in Ainu. In particular, we improved the existing tools for normalizing old transcriptions, word segmentation, and part-of-speech tagging. In the experiments we applied two Ainu language dictionaries from different domains (literary and colloquial) and created a new data set by combining them. The experiments revealed that expanding the lexicon had a positive impact on the overall performance of our tools, especially with test data unrelated to any of the training sets used.

Keywords: Ainu language; endangered languages; normalization; word segmentation; part-of-speech tagging

1. Introduction

UNESCO estimates that at least half of the languages currently used around the world are losing speakers and about 90% of them may be replaced by dominant languages by the end of the 21st century [1]. Technologies being developed within the fields of natural language processing (NLP) and computational linguistics have a great potential to support the urgent tasks of documenting, analysing and revitalizing endangered languages. At the same time, the rapid development and the spread of language technologies observed in recent decades may result in creating a technological gap between smaller languages and majority languages—which in turn would threaten the survival of the former group—if linguistic minorities are not provided with equal access to said technologies.

For these reasons, multiple research initiatives have been undertaken in recent years with the aim of developing linguistic resources (such as lexicons [2] and annotated corpora [3]) and speech or text processing technologies [4,5] for under-resourced and endangered languages. Abney and Bird [6] advocated for the construction of a multi-lingual corpus in a consistent format allowing for cross-linguistic automatic processing and the study of universal linguistics. Bird and Chiang [7] discussed the potential role of machine translation in language documentation. Blokland et al. [8,9] and Gerstenberger et al. [10,11] proposed the application of proven natural language processing approaches as a method to facilitate language documentation efforts, in particular to automate the process of corpus annotation and to support the integration of legacy linguistic materials in contemporary documentation projects. The "Digital Language Survival Kit" [12], published as a part of the Digital Language Diversity Project, lists some of the basic resources and technologies (such as spell checkers, part-of-speech taggers, and speech synthesis and recognition tools) necessary to improve the digital vitality of minority languages.

The aim of this research is to develop technologies for automatic processing of Ainu—a language isolate that is native to northern parts of Japan, which is currently recognized as nearly extinct (e.g., by Lewis et al. [13]).

In particular, we aimed at improving the part-of-speech tagger for the Ainu language (POST-AL), a tool for computer-supported linguistic analysis of the Ainu language, initially developed by Ptaszynski and Momouchi [14].

The task of developing NLP tools for Ainu poses several challenges. Firstly, large-scale digital language resources required for many NLP tasks (such as annotated corpora) are not available for the Ainu language. In this paper we describe our attempt to solve this problem by merging two different digitized dictionaries into one data set. Secondly, there exists no single standard for transcription and word segmentation of the Ainu language, especially in texts collected in earlier years. To address that problem, POST-AL has been equipped with the functions of transcription normalization and word segmentation. In this paper we describe in detail the proposed methodology including recent improvements. Another functionality of POST-AL is part-of-speech (POS) tagging. To improve this accuracy we developed a hybrid method of POS disambiguation, combining lexical n-grams and term frequency. The results of evaluation experiments presented in this paper show that there are differences in part-of-speech classification of certain forms between authors of different dictionaries and text annotations, which creates yet another challenge, to be tackled in the future.

The remainder of this paper is organized as follows. In Section 2 we briefly describe the characteristics and the current status of the Ainu language. In Section 3 we provide an overview of some of the previous studies on the Ainu language, including the few existing research projects in the field of natural language processing. Section 4 presents our algorithms for normalization, word segmentation and part-of-speech tagging. In Sections 5 and 6 we introduce the training data (dictionaries) and test data used in this research. Section 7 summarizes the evaluation methods we applied. In Section 8 we present the results of the evaluation experiments. Finally, Section 9 contains conclusions and some ideas for future improvements.

2. The Ainu Language

The Ainu language is the language of the Ainu people, the native inhabitants of Japan's northernmost island of Hokkaidō. Historically however, the Ainu inhabited a vast territory stretching from the southern part of the Kamchatka Peninsula in the north throughout the Kurile Archipelago, Sakhalin, Hokkaidō, down to the Tōhoku region in northern Honshū [15].

Although numerous attempts have been made to relate Ainu to Paleo–Asiatic, Ural–Altaic, or Malayo–Polynesian languages, to individual languages spoken in the same region, such as Japanese and Gilyak, or even to such remote groups of languages as Semitic and Indo–European (see [15]), until the present day none of these hypotheses have been proven or gained wider acceptance. Thus, Ainu is most often regarded as a language isolate. In terms of typology, Ainu is an agglutinating, polysynthetic language with SOV (subject-object-verb) word order. Ainu verbs are obligatorily marked with pronominal affixes (different for intransitive and transitive verbs) indicating person and number of the subject and the object [16]. Polysynthetic characteristics (such as incorporation and concentration of various morphemes in the verbal complex) are stronger in classical Ainu (the language of the traditional Ainu epics) than in colloquial language [17]. The first of the following examples demonstrates noun incorporation in Ainu. In the second one, a similar meaning is expressed without incorporation.

(A) ku-kamuy-panakte [18]

1st.person.singular.subject-god(s)-punish

"I was punished by the gods"

(B) kamuy en-panakte [18]

god(s) 1st.person.singular.object-punish

"The gods punished me"

Phonemic inventory of the Ainu language consists of five vowel phonemes: /i, e, a, o, u/, and eleven consonant phonemes: /p, t, k, c, s, h, r, m, n, y, w/. Some experts (e.g., Shirō Hattori [19]) treat the glottal plosive [ʔ] as an additional consonant phoneme. In Hokkaidō Ainu, there are the following types of syllables: V, CV, VC, CVC (C = consonant, V = vowel). In Sakhalin Ainu, there are two additional possibilities: VV and CVV, where VV represents a long vowel [20] (nevertheless, all resources used in this research belong to Hokkaidō dialects). A variety of phonological and morphophonological (i.e., only applying to certain morphemes) alternations can be observed at syllable boundaries. In the following example from Shibatani [17], syllable-final /r/ followed by /n/ is realized as /n/:

/akon nispa/

a-kor nispa

we-have rich.man

"Our chief"

For the majority of its history, the Ainu language did not have a written form, but instead had a rich tradition of oral literature, transmitted from generation to generation. One of the best known examples of the Ainu literary forms are the *yukar*, narrative poems about gods and heroes. Most written documents in the Ainu language are transcribed using the Latin alphabet and/or Japanese *katakana* script (all textual data in Ainu applied in this research is written in Latin script).

Current Situation

While the exact number of Ainu language speakers (as well as the size of the population of Ainu people) is difficult to determine, in a survey conducted in 2013 by the Hokkaidō regional government [21], only 7.2% out of 586 respondents answered that they were able to communicate using the Ainu language. This situation is a consequence of the language shift from Ainu to Japanese which started in the 19th century and resulted in the mother tongue of the Ainu people no longer being transmitted to next generations [20].

That being said, in the last few decades the Ainu people have started to regain pride in their culture, especially after the Japanese Government enacted the "Act for Promotion of Ainu Culture, Dissemination of Knowledge and Educational Campaign on Ainu Traditions" (English translation is available at: http://hrlibrary.umn.edu/instree/law-ainu.html) in 1997 and officially recognized the Ainu as indigenous people of Hokkaidō in 2008, refuting the deep-rooted myth of the Japanese being a homogeneous nation. One of the effects of the re-awakening of Ainu identity was the increase of interest in the Ainu language. A number of Ainu language courses are offered throughout Hokkaido, but also in other regions of Japan (e.g., in Tokyo). The Foundation for the Research and Promotion of Ainu Culture (FRPAC) holds an annual Ainu language speech contest and collaborates with the STV Radio in Sapporo in broadcasting a series of Ainu language courses. A magazine in the Ainu language, "Ainu Taimuzu" [Ainu Times] (http://www.geocities.jp/otarunay/taimuzu.html), has been published since 1997. There are also musicians singing in the Ainu language, such as the "Dub Ainu Band" (http://www.tonkori.com/).

3. History of Ainu Language Research

The earliest sources on the Ainu language date back to the 17th century, most of them being small wordlists compiled by travellers, missionaries, or official Japanese interpreters [20]. The 19th century brought the first dictionary publications by Japanese authors, such as Uehara-Abe [22] and Jinbō-Kanazawa [23] (Ainu–Japanese), as well as by Europeans: Dobrotvorskij [24] (Ainu–Russian), Batchelor [25] (Ainu–English–Japanese), Radliński [26] (Ainu–Polish–Latin) and others. Further development of the Ainu lexicography in Japan occurred in the second half of the 20th century and resulted in the publishing of some of the most comprehensive dictionaries of the Ainu language (all of them compiled as Ainu–Japanese bilingual dictionaries), such as the ones by Mashiho Chiri [27–29], Shirō Hattori [19], Hiroshi Nakagawa [30], Suzuko Tamura [31], Shigeru Kayano [32], and Hideo Kirikae [33].

Another important branch of Ainu language studies is the documentation and study of the Ainu people's oral literature. One of the pioneers in this field was a Polish anthropologist, Bronisław Piłsudski, who spent several years in Sakhalin between 1886 and 1905, studying Ainu language and culture, and in 1912 published a collection of 27 Ainu texts with English translations and comments. He also produced the earliest known sound recordings of the Ainu language, dating from 1902–1903 [34].

A latter example, and probably one of the best known studies concerning the Ainu oral tradition are the works of Kyōsuke Kindaichi (e.g., [35,36]), who devoted his research to translating and analysing the *yukar* epics.

Similar studies were also undertaken by Yukie Chiri (native Ainu who compiled a collection of 13 *yukar* stories: the *Ainu shin-yōshū* [37], first published in 1923) and Shigeru Kayano [38], among others.

Natural Language Processing for Ainu

As for Ainu language studies involving modern digital technologies, there is a considerable number of research projects focused on creating online dictionaries and repositories of materials in Ainu (texts with translations, as well as voice and video recordings), such as the ones by the National Institute for Japanese Language and Linguistics [39], Chiba University Graduate School of Humanities and Social Sciences [40], and The Ainu Museum [41]. However, in the field of natural language processing, little attention is paid to endangered languages and Ainu is no exception. From 2002 Momouchi and colleagues have been working on preparing ground for an Ainu–Japanese machine translation system. In the first stage of their research they tried to develop methods for automatic extraction of word translations based on a small parallel corpus [42–44]. Later Azumi and Momouchi [45,46] started developing tools for analysis and retrieval of hierarchical Ainu–Japanese translations. Momouchi, Azumi and Kadoya [47] annotated one of the *yukar* stories included in the *Ainu shin-yōshū* (namely *Pon Okikirmuy yayeyukar "kutnisa kutunkutun"*) with information such as parts of speech, Japanese translations and normalized transcription, with the intent of using it for the development of a machine translation system. Lastly, Momouchi and Kobayashi [48] compiled a dictionary of Ainu place names and used it to create a system for the analysis of Ainu topological names. A more recent project, by Senuma and Aizawa [49,50], aims at creating a small dependency treebank in the scheme of Universal Dependencies. However, their research is still in the initial phase.

POST-AL—Natural Language Processing Tool for the Ainu Language

In addition to the research described above, in 2012 Ptaszynski and Momouchi started developing POST-AL (part-of-speech tagger for the Ainu language), a tool for computer-aided processing of the Ainu language. In its present form, POST-AL performs five tasks:

- Transcription normalization: modification of parts of text that do not conform to modern rules of transcription (e.g., *kamui*→*kamuy*);

- Word segmentation (tokenization): a process in which the text is divided into basic meaningful units (referred to as *tokens*). For writing systems using explicit word delimiters, tokenization is relatively simple. However, in some languages (such as Chinese) word boundaries are not indicated in the surface form, or orthographic words are too coarse-grained and need to be further analyzed—which is the case for many texts written in Ainu;
- Part-of-speech tagging: assigning a part-of-speech marker to each token;
- Morphological analysis (see Ptaszynski et al. [51]);
- Word-to-word translation (into Japanese).

One of the core elements of the POST-AL system is its dictionary base. Originally, it contained one dictionary, namely the *Ainu shin-yōshū jiten* by Kirikae [33]. In 2016 Ptaszynski, Nowakowski, Momouchi and Masui investigated the possibilities of improving the part-of-speech tagging function of the system by testing it with four different dictionaries. In this paper we present the results of further tests comparing two wide-coverage dictionaries of the Ainu language: the *Ainu shin-yōshū jiten* and the *A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational dictionary* by Bugaeva et al. [52]. Moreover, we describe our attempt to combine both dictionaries into one database, in order to improve the system's overall performance.

Until recently there were no commonly accepted rules for transcribing the Ainu language (for detailed analyses of notation methods employed by different authors and how they changed with time, please refer to Kirikae [53], Nakagawa [54] and Endō [55]). At the same time, since the decline of the Ainu language community had already started in the 19th century, texts collected in earlier years became important records of the language. The modernization of such texts involves two tasks:

- Modifying character representations of certain sounds, such as 'ch'→'c', 'sh'→'s', 'ui'→'uy', 'au'→'aw';
- Correcting word segmentation. Authors of older transcriptions tended to use less word delimiters (texts were divided according to poetic recitation rules, into chunks often containing multiple syntactic words). In the context of our research, it leads to an increase in the proportion of forms not covered in dictionaries. Thus, it is necessary to split some of the orthographic words into smaller units.

In order to facilitate the analysis and processing of such documents, Ptaszynski and Momouchi [14] developed a maximum matching algorithm-based tokenizer, including several heuristic rules for normalization of transcription in older texts. Ptaszynski et al. [56] investigated the possibility of adapting the tokenizer algorithm included in the Natural Language Toolkit (http://www.nltk.org/) for segmenting Ainu, but it did not perform as well as the dedicated tokenizer. In the present research we improved the dictionary lookup algorithm applied in the tokenizer and enhanced the normalization rules.

Another functionality of POST-AL which we aimed to improve in the research presented here, is part-of-speech tagging. To achieve that, we modified the tagging algorithm by including the information about term frequency in the process of part-of-speech disambiguation.

4. System Description

In this section we present the technical details of modified algorithms for transcription normalization, word segmentation, and part-of-speech tagging.

4.1. Transcription Normalization Algorithm

The role of this part of the program is to detect all substrings of a given input string that are equal to the upper part of any of the transcription change rules shown in Table 1, and generate a list of all possible transcriptions, where each of such substrings is either substituted with the lower part of the corresponding change rule or retained without modification. Given an input string with n substrings to be potentially modified, a list of 2^n strings will be generated. An example is shown in Table 2. Unlike

in the original version of POST-AL, transcription change rules are optional—the decision as to which of them should be applied (which of the possible transcriptions of the input string to select) is made in the next step by the segmentation algorithm.

Table 1. Transcription change rules applied in part-of-speech tagger for the Ainu language (POST-AL).

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

Table 2. A fragment of text before and after processing with the normalization algorithm.

Input String	List of Output Strings	Meaning
chepshuttuye	cepsuttuye cepshuttuye chepsuttuye chepshuttuye	"to exterminate fish"

4.2. Tokenization Algorithm

4.2.1. Input

The type of input for the tokenizer depends on whether we want to apply transcription normalization or not:

- With transcription normalization: if the text has to be corrected in terms of transcription, then the word segmentation algorithm takes a list of all possible transcriptions of a given input string, which has been generated in the previous stage;
- Without transcription normalization: if there is no need for transcription normalization, the input only includes one string (the one that has been provided by the user).

4.2.2. Word Segmentation Process

Instead of applying a maximum matching algorithm, the new tokenizer performs a dictionary lookup in order to find a single token or the shortest possible sequence of tokens from the lexicon, such that after concatenation is equal to the input token. If the current processing pipeline includes transcription normalization, the tokenizer iterates through all variants of the input string and selects the one that allows for a complete match with the shortest sequence of lexicon items (an example is shown in Table 3). If more than one variant can be matched with a sequence containing a certain number of tokens, priority is given to the variants following the modern transcription rules listed in the lower part of Table 1 (the matching algorithm iterates through the strings with modernized transcription, before proceeding to the ones where change rules were not applied).

Table 3. A fragment of text before and after processing with the tokenization algorithm.

Input	Shortest Sequence of Tokens to Match
cepsuttuye cepshuttuye chepsuttuye chepshuttuye	cep sut tuye

There are two reasons why the transcription normalization process is finalized at this stage and not earlier: firstly, in the two dictionaries applied as the training data in this research (see Section 5)

there are more than one hundred items containing character sequences included in Table 1 as obsolete transcription rules (which means that there are exceptions to those rules). Secondly, older texts often include space-delimited units written as multiple segments in modern transcriptions, which are subject to processing with the word segmentation algorithm. This results in character combinations corresponding to one of the transcription change rules, often occurring at token boundaries (an example is shown in Table 4), in which case no modification should be applied, but that becomes clear only after word segmentation has been performed. This means that the dictionary lookup algorithm described here not only detects word boundaries but also performs disambiguation of transcription change rules.

Table 4. Example of a situation where the transcription change rules should not be applied.

Input Token	Strings Generated by the Transcription Normalization Algorithm	Gold Standard Transcription and Word Segmentation	Meaning
setautar	setawtar setautar	seta utar	"dogs"

4.3. Part-of-Speech Tagger

The part-of-speech tagger proposed by Ptaszynski and Momouchi [14] performs part-of-speech disambiguation based on sample sentences (i.e., lexical n-grams) included in the dictionary. Namely, for each ambiguous token found in the input text it extracts 2- and 3-grams containing that token, searches for them in the dictionary and returns the part-of-speech tag of the candidate entry with the highest number of matches. However, at this point the database only includes two dictionaries, therefore for many cases there exist few or no relevant usage examples. To compensate for that, we created a modified tagging algorithm, which in such cases also takes into account the term frequency (number of occurrences in the database) of each candidate term and returns the POS tag assigned to the item with the highest value. For instance, the form *sak* used as transitive verb (meaning 'to lack; not to have') appears 14 times in our dictionary, whereas the noun *sak* ('summer') has three occurrences, which means that according to the proposed disambiguation method "transitive verb" should be selected as the POS tag for the token *sak*.

In the present research, in order to verify the performance of different part-of-speech disambiguation methods, we prepared three variants of the tagging algorithm:

- With n-gram based POS disambiguation (as in the original POST-AL system);
- With TF (term frequency) based POS disambiguation;
- N-grams + TF (TF based disambiguation is only applied to cases where n-gram based disambiguation is insufficient).

5. Dictionaries

5.1. Ainu shin-yōshū jiten

The base dictionary originally used in POST-AL was a digital version of the *Ainu shin-yōshū jiten*, a lexicon to Yukie Chiri's *Ainu shin-yōshū* (a collection of thirteen mythic epics), developed by Kirikae [33]. The dictionary comprises 2019 entries, each of them containing the following types of information: form (word or morpheme), morphological analysis, part of speech (POS), translation into Japanese, reference to the story it appears in, and usage examples (not for all cases) [57]. In the following sections we will refer to this dictionary as "KK". Listing 1 shows a sample entry.

Listing 1: An entry from the *Ainu shin-yōshū jiten* by Kirikae, converted to XML format (KK dictionary).

```
<word>aep</word>
<morph>a{2}-e{1}-p{1}</morph>
<pos>名詞 [noun]</pos>
<tr>食べ物 [food]</tr>
<ref>aep’omuken</ref>
```

5.2. Ainu Conversational Dictionary

The *Ainugo kaiwa jiten* (“Ainu conversational dictionary”) [23] was one of the first dictionaries of the Ainu language, compiled by a Japanese researcher, Shōzaburō Kanazawa, who visited Hokkaidō several times between 1895 and 1897. He published the dictionary in 1898 under the supervision of professor Kotora Jinbō. The original dictionary contains 3847 entries, most of which presumably belong to the Saru dialect [58].

In 2010, Bugaeva et al. [52] released the *A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational Dictionary*, which is an online dictionary of Ainu, based on the original *Ainugo kaiwa jiten*. Apart from the original content (Ainu words and phrases and their Japanese translations), the dictionary provides additional information, including corrected Latin transcription, modern Japanese translations, part-of-speech classification, and English translations. Furthermore, with the help of a native speaker of Ainu (Setsu Kurokawa) mistakes and misinterpretations found in the original dictionary were corrected [58]. A sample entry is presented in Listing 2.

In 2015, a revised version of the above mentioned online dictionary has been released under the name of *A Topical Dictionary of Conversational Ainu* [39].

Listing 2: An entry from A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational Dictionary.

```
此村に何か食物があるか  [Japanese translation as in the original lexicon [23]]
Tan kotan ta nepka aep an ruwe he an?  [Latin transcription as in the original lexicon [23]]
tan kotan ta nep ka aep an ruwe an?  [modernized Latin transcription]
タン コタン タ ネプ カ アエプ アン ルウェ アン?  [transcription in kana syllabary]
この 村 に 何 か 食べ物 ある こと ある  [word-to-word translation to Japanese]
【連体】【名】【格助】【疑問】【副助】【名】【自】【形名】【自】  [part-of-speech annotation]
dem n pp n.interr adv.prt n vi nmlz vi  [part-of-speech annotation]
この村に何か食べ物はありますか？  [modern Japanese translation]
Is there anything to eat in this village?  [English translation]
tan kotan ta nep ka a-e-p an ruwe an  [morphemes]
this village at what even INDF.A-eat-thing exist.SG INFR.EV exist.SG  [gloss]
```

In the present research we use *A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational Dictionary*. Original entries often consist of more than one word (multiple words or phrases). Therefore, in order to apply the dictionary in POST-AL we modified it, dividing such entries into separate single-word entries. The original entries that consist of more than one word have been added to the modified dictionary as usage examples, which POST-AL uses for part-of-speech disambiguation. Finally, we performed automatic unification of duplicate entries (entries containing words appearing in multiple entries of the original dictionary), using the translations provided by Bugaeva et al. to determine whether each homonym should be treated as a duplicate or a separate entry. These modifications resulted in a dictionary containing 2555 single-word entries. The basic format of the entries has been adjusted to conform to the dictionary format required by POST-AL (the same format which is also used in the *Ainu shin-yōshū jiten*). Each entry contains the following information: headword, morphological boundaries, part(s) of speech (in Japanese and English), Japanese and English translation and morpheme-to-morpheme interpretation (explanations of meaning or function of each morpheme). Furthermore, 1496 entries contain usage examples with Japanese and English translations. The last modification we performed was excluding 62 usage examples (428 words) from

the dictionary, in order to use them as test data in evaluation experiments (see Section 6). Total number of usage examples in the final version of the dictionary is 12,513 (including duplicates). In the following sections we will refer to this dictionary as "JK". A fragment is presented in Listing 3.

Listing 3: An entry from the JK dictionary.

```
<word>aep</word><kana>アエプ</kana>
<morph>a-e-p</morph><pos>名詞</pos>
<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>
```

5.3. Combined Dictionary

In order to increase the POST-AL system's versatility, we decided to combine the two dictionaries described in Sections 5.1 and 5.2 into one dictionary base. To achieve this, we extracted entries containing items listed in both dictionaries and automatically unified them, based on their Japanese translations (namely, homonymous entries with at least one *kanji* character in common have been unified and the rest was retained as separate entries). That resulted in a dictionary containing 4161 entries. In the following sections we will refer to this dictionary as "JK+KK". Listing 4 shows an entry from this resource.

Listing 4: An entry from the JK+KK dictionary.

```
<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>名詞</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>
```

6. Test Data and Gold Standard

For evaluation of the proposed system we used four different datasets:

- Yukar epics: Five out of thirteen *yukar* stories (no. 9–13) from the *Ainu shin-yōshū* [37]. Apart from the original version by Chiri, we also used the variants edited by Kirikae, who manually corrected their transcription and word segmentation according to modern linguistic conventions, and included them in the *Ainu shin-yōshū jiten* [33]. The modernized version comprises a total of 1608 tokens. Later we refer to this dataset as "Y9–13". In the experiment with POS tagging, we only used a subset of the data in question, namely the story no. 10: *Pon Okikirmuy yayeyukar "kutnisa kutunkutun"* ["Kutnisa kutunkutun"—a song Pon Okikirmuy sang], which has 189 tokens—later it will be abbreviated to "Y10". A fragment is presented in Table 5.

Table 5. A fragment from the Y9–13; original text by Chiri (top) and postprocessed by Kirikae (middle).

Shineantota petetok un shinotash kushu payeash awa
Sine an to ta petetok un sinot as kusu paye as a wa
Meaning: "One day when I went for a trip up the river"

- Samples from the Ainu Conversational Dictionary: Sixty two sentences (428 tokens) from the *A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational Dictionary*, which were excluded from the training data (see Section 5.2). Apart from the original text by Jinbō and Kanazawa [23], we also used the modernized version by Bugaeva et al. [52]. Later we refer to this dataset as "JK samples". A fragment is shown in Table 6.

Table 6. A sentence from the JK samples; original version (top) and with transcription normalized by Bugaeva et al. (middle).

Tambe makanak an chiki pirika?
Tanpe makanak an ciki pirka?
Meaning: "What should I do about it?"

- Shibatani's colloquial text samples: Both datasets mentioned above are either obtained directly from one of the dictionaries applied as the training data for our system (Ainu Conversational Dictionary) or from the compilation of *yukar* stories on which one of these dictionaries was based (*Ainu shin-yōshū*). To investigate the performance with texts unrelated to the system's training data, we decided to apply other datasets as well. As the first one we used a colloquial text sample included in *The Languages of Japan* [17], namely a fragment (154 tokens) of Kura Sunasawa's memoirs written in the Ainu language, *Ku sukup oruspe* ("My life story") [59], transcribed according to modern linguistic rules. Later we refer to this dataset as "Shib.";
- Mukawa dialect samples: We also used a sample (11 sentences, 87 tokens) from the Japanese–Ainu Dictionary for the Mukawa Dialect of Ainu [40]. It is a transcribed version of audio materials Tatsumine Katayama recorded between 1996 and 2002 with two native speakers of the Mukawa dialect of Ainu: Seino Araida and Fuyuko Yoshimura, containing 6284 entries. Later we refer to this dataset as "Muk."

Below we describe the variants of the test data applied in testing each element of our system.

6.1. Test Data for Transcription Normalization

To test the transcription normalization performance we used two datasets: Y9–13 and JK samples. Each of them was prepared in two versions:

- Original ("O"): Original texts by Chiri or Jinbō and Kanazawa, without any modifications;
- Original, with spaces removed ("O-SR"): Original texts by Chiri and Jinbō and Kanazawa, preprocessed by removing any word segmentation (whitespaces) from each line.

As the gold standard data, we used the modernized versions of both texts [33,52].

6.2. Test Data for Tokenization

For evaluation of the tokenizer, we prepared two different versions of the test data:

- Modern transcription, spaces removed ("M-SR"): The first variant includes all four datasets. In the case of Y9–13 and JK samples, modernized versions by Kirikae and Bugaeva et al. were used. Each line of text was preprocessed by removing whitespaces;
- Original spaces, modernized transcription ("O/M"): In this variant, only two datasets were used: the Y9–13 and JK samples. We retained the word segmentation (usage of whitespaces or lack thereof) of the original texts [23,37]. However, in order to prevent differences between

transcription rules applied in original and modernized texts from affecting word segmentation experiment results, the texts were preprocessed by unifying their transcription with the modern (gold standard) versions. Table 7 shows to what extent the word segmentation of original texts is consistent with modernized versions by Kirikae and Bugaeva et al. (the evaluation method is explained in Section 7).

Table 7. Results of the evaluation of word segmentation in original texts by Chiri or Jinbō and Kanazawa against modern versions.

	Y9–13	JK Samples	Overall
Precision	0.998	0.949	0.983
Recall	0.609	0.918	0.674
F-score	0.756	0.933	0.800

The reason for performing the experiment on two versions of the test data was to verify which approach is more effective: retaining whitespaces even if in some cases it will hinder proper segmentation, or removing any word segmentation used in the original text. To illustrate the problem, below is a fragment of text from Y9–13 in original transcription by Chiri [37], the modern transcription by Kirikae [33], and two versions prepared for the experiment:

- Original transcription: unnukar awa kor wenpuri enantui ka;
- Modern transcription (gold standard): un nukar a wa kor wen puri enan tuyka;
- Modern transcription, spaces removed: unnukarawakorwenpurienantuyka;
- Modern transcription, original word segmentation: unnukar awa kor wenpuri enantuy ka;
- Meaning: "When she found me, her face [took] the color of anger."

Modern transcriptions of all four texts [17,33,40,52] were used as the gold standard.

6.3. Test Data for POS Tagging

To evaluate part-of-speech tagging performance we used two texts: Y10 and JK samples, both of them in modern transcription [33,52]. Gold standard POS annotations were provided by Momouchi et al. [47] and Bugaeva et al. [52], respectively.

Table 8 presents the statistics of all four datasets used for evaluation, including their different variants.

Table 8. Statistics of the samples used for testing.

Data		Variant	Characters (Excluding Spaces)	Tokens
Ainu shin-yōshū	**Y9–13**	O* O-SR**	6883	1076 N/A
		M-SR*** O/M****	6501	N/A 1076
		Kirikae [33]	6501	1608
	Y10	Kirikae [33]	822	189
Ainugo Kaiwa Jiten/ A Talking Dictionary of Ainu... **(JK samples)**		O O-SR	1742	418 N/A
		M-SR O/M	1617	N/A 416
		Bugaeva et al. [52]	1617	428
Shibatani's colloquial text samples **(Shib.)**		Shibatani [17]	583	154
Mukawa dialect samples **(Muk.)**		Chiba University... [40]	341	87

* Original transcription; ** Original transcription with spaces removed; *** Modern transcription with spaces removed; **** Modern transcription with original spaces.

7. Evaluation Methods

All experimental results were calculated with the means of precision (P), recall (R), and balanced F-score (F). In this section we provide definitions of the evaluation metrics for each type of experiments and describe the evaluation methodologies.

7.1. Balanced F-Score

Balanced F-score is the harmonic mean of precision and recall and its calculation method is the same across all experiments:

$$F = 2\frac{PR}{(P+R)}. \tag{1}$$

7.2. Evaluation of Transcription Normalization

In the case of transcription normalization, precision is calculated as the percentage of correct single-character edits (deletions, insertions or substitutions) within all edits performed by the system, and recall as the percentage of correct edits performed by the system within all edits needed to normalize the transcription of a given text. The total number of edits performed is equal to the Levenshtein distance between the input and the output, and the total number of edits needed, to the Levenshtein distance between the input and the gold standard. To calculate the number of correct edits we used a combination of the Levenshtein distance between the input and the output, and between the output and the gold standard (for each edit performed by the system we checked if it reduced the edit distance to the gold standard).

$$P = \frac{correct\ \ edits}{all\ \ returned\ \ edits} \tag{2}$$

$$R = \frac{correct\ \ edits}{all\ \ gold\ \ standard\ \ edits}. \tag{3}$$

As was explained in Section 4.1, the process of transcription normalization is finalized at the stage of tokenization. Therefore, after processing all texts with the transcription normalization algorithm, they were also processed with the tokenizer. In order to prevent tokenization errors from affecting the evaluation results, whitespaces were removed from both the output texts and the gold standard texts.

7.3. Evaluation of Tokenization

In the case of tokenization, precision is calculated as the proportion of correct separations (spaces) within all separations returned by the system, whereas recall is the number of correct spaces the system returned divided by the number of spaces in the gold standard.

$$P = \frac{correctly\ \ predicted\ \ spaces}{all\ \ returned\ \ spaces} \tag{4}$$

$$R = \frac{correctly\ \ predicted\ \ spaces}{all\ \ gold\ \ standard\ \ spaces}. \tag{5}$$

7.4. Evaluation of Part-of-Speech Tagging

In this case, precision is calculated as the percentage of correct annotations within all annotations made by the system. Recall is the percentage indicating how many correct annotations the system returned compared to the gold standard.

$$P = \frac{correct\ \ annotations}{all\ \ system's\ \ annotations} \tag{6}$$

$$R = \frac{correct\ \ annotations}{all\ \ gold\ \ standard\ \ annotations}. \quad (7)$$

As was shown in Section 5, the KK dictionary and JK dictionary differ in terms of part-of-speech classification. Furthermore, Momouchi et al. [47] annotated Y10 according to yet another part-of-speech classification standard, introduced by Tamura [31]. Therefore, in order to evaluate the POS tagging experiment results we used part-of-speech conversion tables built in POST-AL (see [14]). Specifically, we converted all annotations to the part-of-speech classification standard used by Nakagawa [30]. The conversion method is shown in Table 9. As an alternative method for the evaluation of tagging results, we used a simplified POS standard (Table 10), where subclasses of general word classes are not differentiated (e.g., intransitive and transitive verbs are both considered the same class: "verb"), thus in Section 8.3 two different results are given for each experiment, depending on the part-of-speech conversion table used ("Nakagawa" and "simplified").

Table 9. Table for conversion of other Ainu part-of-speech standards into Nakagawa's standard.

JK	KK	Tamura (1996)		Nakagawa (1995)
完全動詞 (complete verb)	ゼロ項動詞 (complete verb)	完全動詞 (complete verb)	→	0 項動詞 (complete verb)
自動詞 (intransitive verb)	一項動詞 (intransitive verb)	自動詞 (intransitive verb)	→	1 項動詞 (intransitive verb)
他動詞 (transitive verb)	二項動詞 (transitive verb)	単他動詞 (transitive verb)	→	2 項動詞 (transitive verb)
複他動詞 (ditransitive verb)	三項動詞 (ditransitive verb)	複他動詞 (ditransitive verb)	→	3 項動詞 (ditransitive verb)
	人称代名詞 (personal pronoun)	人称代名詞 (personal pronoun)	→	代名詞 (pronoun)
	指示代名詞 (demonstrative pronoun)		→	代名詞 (pronoun)
	疑問不定代名詞 (interrogative indefinite pronoun)	疑問代名詞 (interrogative pronoun)	→	疑問詞 (interrogative)
	疑問不定副詞 (interrogative indefinite adverb)	疑問副詞 (interrogative adverb)	→	疑問詞 (interrogative)
	指示副詞 (demonstrative adverb)		→	副詞 (adverb)
後置副詞 (postpositive adverb)	後置詞的副詞 (postpositive adverb)	後置副詞 (postpositive adverb)	→	副詞 (adverb)
	指示連体詞 (demonstrative prenoun adjectival)		→	連体詞 (prenoun adjectival)
	後置詞 (postposition)		→	格助詞 (case particle)
	名詞的助詞 (nominal particle)		→	名詞 (noun)

Table 10. POS conversion table, simplified standard.

Nakagawa (1995)		Simplified Standard
0 項動詞 / 1 項動詞 / 2 項動詞 / 3 項動詞 (complete verb / intransitive verb / transitive verb / ditransitive verb)	→	動詞 (verb)
固有名詞 / 代名詞 / 位置名詞 / 形式名詞 (proper noun / pronoun / locative noun / expletive noun)	→	名詞 (noun)
格助詞 / 接続助詞 / 副助詞 / 終助詞 (case particle / conjunctive particle / adverbial particle / final particle)	→	助詞 (particle)
人称接辞 / 接頭辞 / 接尾辞 (personal affix / prefix / suffix)	→	接辞 (affix)

8. Results and Discussion

8.1. Transcription Normalization

Table 11 shows the results of transcription normalization experiments. Transcription normalization based on Kirikae's lexicon achieved the highest scores for the Y9–13 dataset, which is not surprising, since the dictionary is based on *yukar* epics. In the case of JK samples, however, performance with the combined dictionary (JK+KK) was as good as with the JK dictionary only. Furthermore, the combined dictionary achieved the best overall results. In all test configurations the results for texts with original word segmentation retained were slightly better. Relatively low values of recall for normalization in JK samples, observed across all combinations of dictionaries and input text versions, can be explained by a high occurrence of forms transcribed according to non-standard rules modified by Bugaeva et al. in the modernized version of the dictionary, but not included in the list of universal transcription change rules applied in this research, such as 'ra'→'r' (e.g., *arapa*→*arpa*), 'ri'→'r' (e.g., *pirika*→*pirka*), 'ru'→'r' (e.g., *kuru*→*kur*), 'ro'→'r' (e.g., *koro*→*kor*) or 'ei'→'e' (e.g., *reihei*→*rehe*). This is due to the fact that these rules are so far only observed in the dictionary of Jinbō and Kanazawa and more importantly, initial tests performed during the development of the algorithm showed that including them in the algorithm can cause errors when processing yukars and other texts.

Table 11. Transcription normalization experiment results (best results in bold).

			Y9–13	JK Samples	Overall	Input Text Version:
DICTIONARY	JK	Precision	0.871	0.942	0.885	O-SR
		Recall	0.897	0.658	0.833	
		F-score	0.884	0.775	0.859	
		Precision	0.890	**0.956**	0.903	O
		Recall	0.897	**0.658**	0.833	
		F-score	0.893	**0.780**	0.867	
	KK	Precision	0.967	0.899	0.954	O-SR
		Recall	**0.966**	0.628	0.876	
		F-score	0.966	0.740	0.913	
		Precision	**0.980**	0.926	**0.969**	O
		Recall	0.958	0.628	0.871	
		F-score	**0.969**	0.749	0.917	
	JK+KK	Precision	0.953	0.942	0.951	O-SR
		Recall	0.964	0.658	**0.883**	
		F-score	0.958	0.775	0.916	
		Precision	0.971	**0.956**	0.968	O
		Recall	0.958	**0.658**	0.879	
		F-score	0.964	**0.780**	**0.921**	

8.2. Tokenization

The results of tokenization experiments are shown in Table 12. Table 13 shows a fragment from Y9–13 (M-SR) before and after segmentation. Similarly to transcription normalization, the tokenization algorithm also performed the best for *yukar* stories (Y9–13) when coupled with the *Ainu shin-yōshū jiten* (KK). Analogically, for JK samples, the JK dictionary was the best. It shows a weak point of the presented segmentation algorithm: while adding new forms to the lexicon improves its versatility (ability to process texts from different domains), it also increases the number of possible mistakes the tokenizer can make with texts for which the original lexicon had been (nearly) optimal. The combined dictionary performed better than the other two dictionaries on test data unrelated to the training data (Shib. and Muk.), and also achieved the best overall results (F-score). On the other hand, overall recall was higher with the KK dictionary. To some extent this might be explained by the differences in word segmentation between the two dictionaries applied in this research: many expressions (e.g., *oro wa*, 'from' or *pet turasi*, 'to go upstream') written as two separate segments by Kirikae (both in the lexicon part of the *Ainu shin-yōshū jiten*, as well as in his modernized transcriptions of the *yukar* stories, which we use as the gold standard data), are transcribed as a single unit (*orowa*, *petturasi*) by Bugaeva et al. Once these forms are added to the lexicon, the word segmentation algorithm, which prefers long tokens over shorter ones, stops applying segmentation to the tokens *orowa* and *petturasi* (and that causes recall to drop). This phenomenon occurs in the opposite direction as well: The only two types of tokenization errors made in the JK samples (O/M) when the combined dictionary was used, but not with the JK dictionary, were both of this type—the expressions transcribed by Bugaeva et al. as *somo ki* ('do not') and *te ta* ('here') are listed as *somoki* and *teta* in the *Ainu shin-yōshū jiten*. Scores achieved by the tokenizer on texts with original word boundaries retained (Y9–13 (O/M) and JK samples (O/M)) were higher than with spaces removed. This means that the original word segmentation, even if it causes some errors (as with the word *tuyka*—see Section 6.2), still supports tokenization rather than hindering it.

Table 12. Tokenization experiment results (best results in bold).

			Y9–13	JK Samples	Y9–13 + JK Samples	Shib. + Muk.	Overall	Input Text Version:
DICTIONARY	JK	Precision	0.575	**0.935**	0.634	0.742	0.644	M-SR
		Recall	0.772	0.907	0.801	0.808	0.801	
		F-score	0.659	0.921	0.708	0.774	0.714	
		Precision	0.652	0.933	0.700	n/a	n/a	O/M
		Recall	0.894	**0.984**	0.913	n/a	n/a	
		F-score	0.754	**0.957**	0.792	n/a	n/a	
	KK	Precision	0.921	0.703	0.867	0.649	0.838	M-SR
		Recall	0.889	0.842	0.879	0.822	**0.873**	
		F-score	0.905	0.766	0.873	0.726	0.855	
		Precision	**0.950**	0.772	0.904	n/a	n/a	O/M
		Recall	**0.944**	0.981	**0.952**	n/a	n/a	
		F-score	**0.947**	0.864	0.928	n/a	n/a	
	JK+KK	Precision	0.905	0.943	0.913	**0.776**	**0.896**	M-SR
		Recall	0.854	0.896	0.863	**0.860**	0.863	
		F-score	0.879	0.919	0.887	**0.816**	**0.879**	
		Precision	0.939	0.932	**0.937**	n/a	n/a	O/M
		Recall	0.919	0.975	0.931	n/a	n/a	
		F-score	0.929	0.953	**0.934**	n/a	n/a	

Table 13. A fragment from Y9–13 (M-SR) before and after tokenization.

Input:	kekehetakcepsuttuyecikikusnena
Tokenizer output:	keke hetak cep sut tuye ciki kusne na
Gold standard:	keke hetak cep sut tuye ci ki kusne na
Meaning:	"Now I'm going to show you how to make fish extinct" [60]

8.3. Part-of-Speech Tagging

The results of part-of-speech tagging experiments are presented in Table 14. Table 15 shows a fragment from the JK dictionary analyzed with POST-AL (with both POS annotations and word-to-word translation into Japanese). The results indicate that as long as the tagger is trained with language data belonging to the same type as the test data (i.e., classical Ainu of the *yukar* epics, also covered in the KK dictionary, and colloquial language of the JK dictionary), part-of-speech disambiguation based on lexical n-grams is more accurate than the method using term frequency. But it also shows that combining both approaches (with priority given to n-grams) provides the best performance in each case. While to a certain extent the most frequent tag approach compensates for the shortcomings of lexical context-based disambiguation method in our low-data conditions, it is far from perfect. Firstly, it ignores the context in which the given token appears, and secondly, in our case the frequency of each tag is calculated from usage examples included in the dictionary, which by no means can be regarded as a balanced representation of the language. As a result, the tagger still makes a considerable amount of disambiguation errors (see Table 16). One of the important tasks for the future is the compilation of a part-of-speech annotated corpus, which will allow us to build more robust disambiguation models. The tagger presented in this paper, while imperfect, can be useful in the process of creating such corpus, e.g., by applying it in an active learning scenario [61].

We also found out that although the JK dictionary and KK dictionary belong to different domains (colloquial and classical language), combining them both improved overall POS tagging performance and in the case of the Y10 dataset yielded the best results of all combinations.

There is a gap between the results of tagging Y10 and JK samples, which can be partially explained by differences in part of speech classification of certain items between the two dictionaries applied in the system and the annotations (gold standard) provided by Momouchi et al. [47]. For example, Momouchi et al. annotated *ne* ("to be") as "auxiliary verb", whereas in the training data it is listed as "transitive verb". In the experiment with Y10, the JK+KK dictionary, and the n-gram+TF based tagging algorithm (the combination that yielded the best result for Y10), this token and other errors of this type accounted for 61% of all incorrect predictions (see Table 16). We hope that to some extent this problem can be solved in the future by adding another dictionaries (such as Tamura's and Nakagawa's dictionaries) with the information about alternative part-of-speech classification standards, to the dictionary base of POST-AL.

Table 14. Part-of-speech tagging experiment results (best results in bold).

			Test Data						Tagging Algorithm Version:	
			Y10		JK Samples		Average			
Part-of-speech Standard:			Nakagawa	Simplified	Nakagawa	Simplified	Nakagawa	Simplified	N-Grams	Term Frequency
Dictionary	JK	Precision	0.771	0.786	0.965	0.974	0.868	0.880	NO	YES
		Recall	0.540	0.551	0.965	0.974	0.753	0.763		
		F-score	0.635	0.648	0.965	0.974	0.800	0.811		
		Precision	0.702	0.718	0.967	0.972	0.835	0.845	YES	NO
		Recall	0.492	0.503	0.967	0.972	0.730	0.738		
		F-score	0.579	0.592	0.967	0.972	0.773	0.782		
		Precision	0.794	0.809	**0.977**	**0.981**	0.886	0.895	YES	YES
		Recall	0.556	0.567	**0.977**	**0.981**	0.767	0.774		
		F-score	0.654	0.667	**0.977**	**0.981**	0.816	0.824		
	KK	Precision	0.821	0.859	0.713	0.763	0.767	0.811	NO	YES
		Recall	0.807	0.845	0.563	0.603	0.685	0.724		
		F-score	0.814	0.852	0.629	0.674	0.722	0.763		
		Precision	0.853	0.886	0.666	0.737	0.760	0.812	YES	NO
		Recall	0.840	0.872	0.526	0.582	0.683	0.727		
		F-score	0.847	0.879	0.588	0.650	0.717	0.765		
		Precision	0.859	0.891	0.728	0.790	0.794	0.841	YES	YES
		Recall	0.845	0.877	0.575	0.624	0.710	0.751		
		F-score	0.852	0.884	0.643	0.697	0.747	0.791		
	JK+KK	Precision	0.855	0.876	0.960	0.970	0.908	0.923	NO	YES
		Recall	0.850	0.872	0.960	0.970	0.905	0.921		
		F-score	0.853	0.874	0.960	0.970	0.906	0.922		
		Precision	0.866	0.892	0.942	0.949	0.904	0.921	YES	NO
		Recall	0.861	0.888	0.942	0.949	0.902	0.919		
		F-score	0.864	0.890	0.942	0.949	0.903	0.920		
		Precision	**0.882**	**0.903**	**0.977**	**0.981**	**0.930**	**0.942**	YES	YES
		Recall	**0.877**	**0.898**	**0.977**	**0.981**	**0.927**	**0.940**		
		F-score	**0.880**	**0.901**	**0.977**	**0.981**	**0.928**	**0.941**		

Table 15. A sentence from the JK dictionary processed by POST-AL, with POS annotations (second line) and word-to-word translation into Japanese (fourth line).

POST-AL output:	iyosno ku hosipire kusne na 副詞　人称接辞　他動詞　助動詞　終助詞 [Adverb　Personal affix　Transitive verb　Aux. verb　Final particle] 最後に/後で　私は/私が/私の　返す　つもりである　よ/か ['the end'/'later'　'I'/'my'　'return'　'intend'　EMPHASIS]
Meaning:	"I'll return it later"

Table 16. Statistics of POS tagging errors in the experiment with Y10, the JK+KK dictionary, and the n-gram+TF based tagger.

Type of Error	Count
Tagger (disambiguation error)	8 (35%)
Dictionary (out-of-vocabulary item)	1 (4%)
POS classification (the same word, but different tag)	14 (61%)

9. Conclusions and Future Work

In this paper we presented our research in improving POST-AL, a tool for computer-aided processing of the critically endangered Ainu language. In addition to improving the algorithms for transcription normalization, word segmentation, and part-of-speech tagging, we also expanded the system's dictionary base by combining two comprehensive Ainu language dictionaries. We found out that the combination improved overall performance of our tools, especially with objective samples unrelated to the training data.

In the future we will enlarge the dictionary base by adding other dictionaries such as the ones by Nakagawa [30] or Tamura [31], and expand it with information about alternative transcription methods appearing in older texts, in order to normalize transcription in such texts more effectively. We also plan to use our system and/or state-of-the-art tools developed for other languages, such as the SVMTool [62] and the Stanford Log-Linear Tagger [63], to build a part-of-speech annotated corpus of Ainu. Having such a resource, we will be able to build statistical language models, which will allow the development of further language technologies, such as speech recognition and statistical machine translation systems. We believe that it will also be useful for linguists, as well as learners and instructors of the Ainu language. Furthermore, we will use the corpus to generate new dictionaries and release them as mobile applications. Other tasks for the near future include the development of a speech synthesizer and a morphological analyzer.

Author Contributions: Conceptualization, M.P. and K.N.; methodology, M.P., Y.M. and K.N.; software, M.P. and K.N.; validation, F.M.; formal analysis, K.N.; investigation, K.N.; resources, F.M. and Y.M.; data curation, M.P., Y.M. and K.N.; writing—original draft preparation, K.N.; writing—review and editing, M.P., F.M. and Y.M.; visualization, K.N.; supervision, F.M. and M.P.; project administration, F.M. and M.P.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive feedback. We also thank Christopher Bozek for diligent proofreading of the manuscript. We are also grateful to Jagna Nieuważny for useful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UNESCO ad Hoc Expert Group on Endangered Languages. Language Vitality and Endangerment. Available online: http://www.unesco.org/culture/ich/doc/src/00120-EN.pdf (accessed on 23 December 2017).

2. Kazeminejad, G.; Cowell, A.; Hulden, M. Creating lexical resources for polysynthetic languages—The case of Arapaho. In Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, Honolulu, HI, USA, 2–5 March 2017; Association for Computational Linguistics: Honolulu, HI, USA, 2017; pp. 10–18.
3. Anastasopoulos, A.; Lekakou, M.; Quer, J.; Zimianiti, E.; DeBenedetto, J.; Chiang, D. Part-of-Speech Tagging on an Endangered Language: A Parallel Griko-Italian Resource. *arXiv* **2018**, arXiv:1806.03757.
4. Besacier, L.; Barnard, E.; Karpov, A.; Schultz, T. Automatic speech recognition for under-resourced languages: A survey. *Speech Commun.* **2014**, *56*, 85–100. [CrossRef]
5. Ruokolainen, T.; Kohonen, O.; Virpioja, S.; Kurimo, M. Supervised Morphological Segmentation in a Low-Resource Learning Setting using Conditional Random Fields. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, 8–9 August 2013; Association for Computational Linguistics: Sofia, Bulgaria, 2013; pp. 29–37.
6. Abney, S.; Bird, S. The Human Language Project: Building a Universal Corpus of the World's Languages. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 88–97.
7. Bird, S.; Chiang, D. Machine Translation for Language Preservation. In Proceedings of the COLING 2012: Posters, Mumbai, India, 8–15 December 2012; The COLING 2012 Organizing Committee: Mumbai, India, 2012; pp. 125–134.
8. Blokland, R.; Fedina, M.; Gerstenberger, C.; Partanen, N.; Rießler, M.; Wilbur, J.D. Language Documentation meets Language Technology. In Proceedings of the 1st International Workshop in Computational Linguistics for Uralic Languages (IWCLUL 2015), Tromsø, Norway, 16 January 2015.
9. Blokland, R.; Partanen, N.; Rießler, M.; Wilbur, J. Using Computational Approaches to Integrate Endangered Language Legacy Data into Documentation Corpora: Past Experiences and Challenges Ahead. In Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages, Honolulu, HI, USA, 26–27 February 2019; Volume 2, pp. 24–30.
10. Gerstenberger, C.; Partanen, N.; Rießler, M.; Wilbur, J. Utilizing Language Technology in the Documentation of Endangered Uralic Languages. *North. Eur. J. Lang. Technol.* **2016**, *4*, 29–47. [CrossRef]
11. Gerstenberger, C.; Partanen, N.; Rießler, M.; Wilbur, J. Instant Annotations—Applying NLP Methods to the Annotation of Spoken Language Documentation Corpora. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*; Association for Computational Linguistics: St. Petersburg, Russia, 2017; pp. 25–36.
12. Berger, K.C.; Hernaiz, A.G.; Baroni, P.; Hicks, D.; Kruse, E.; Quochi, V.; Russo, I.; Salonen, T.; Sarhimaa, A.; Soria, C. *Digital Language Survival Kit: The DLDP Recommendations to Improve Digital Vitality*; 2018. Available online: http://wp.dldp.eu/wp-content/uploads/2018/09/Digital-Language-Survival-Kit.pdf (accessed on 19 October 2019).
13. Lewis, M.; Simons, G.; Fennig, C. (Eds.) *Ethnologue: Languages of the World*, 19th ed.; SIL International: Dallas, TX, USA, 2016.
14. Ptaszynski, M.; Momouchi, Y. Part-of-Speech Tagger for Ainu Language Based on Higher Order Hidden Markov Model. *Expert Syst. Appl.* **2012**, *39*, 11576–11582. [CrossRef]
15. Majewicz, A. *Ajnu. Lud, jego język i tradycja ustna*; [Ainu. The people, its language and oral tradition]; Wydawnictwo Naukowe UAM: Poznań, Poland, 1984.
16. Bugaeva, A. Southern Hokkaido Ainu. In *The languages of Japan and Korea*; Tranter, N., Ed.; Routledge: London, UK, 2012; pp. 461–509.
17. Shibatani, M. *The languages of Japan*; Cambridge University Press: London, UK, 1990.
18. Satō, T. Ainugo Chitose hōgen ni okeru meishi-hōgō: Sono shurui to kanren sho-kisoku [Noun incorporation in the Chitose dialect of Ainu: Types and related rules]. *Hokkaidō-Ritsu Ainu Minzoku Bunka Kenkyū Sentā Kenkyū Kiyō/Bulletin of the Hokkaido Ainu Culture Research Center* **2012**, *18*, 1–32.
19. Hattori, S. *Ainugo hōgen jiten*; [Dictionary of Ainu dialects]; Iwanami Shoten: Tōkyō, Japan, 1964.
20. Refsing, K. *The Ainu Language. The Morphology and Syntax of the Shizunai Dialect*; Aarhus University Press: Aarhus, Denmark, 1986.
21. Hokkaidō Government, Environment and Lifestyle Section. *Hokkaidō Ainu seikatsu jittai chōsa hōkokusho*; [Report of the Survey on the Hokkaidō Ainu actual living conditions]; 2013. Available online: http://www.pref.hokkaido.lg.jp/ks/ass/ainu_living_conditions_survey.pdf (accessed on 23 October 2019).

22. Uehara, K.; Abe, C. *Ezo hōgen moshiogusa*; [Ezo dialect dictionary]; 1804.
23. Jinbō, K.; Kanazawa, S. *Ainugo kaiwa jiten*; [Ainu conversational dictionary]; Kinkōdō Shoseki: Tōkyō, Japan, 1898.
24. Dobrotvorskij, M. *Ainsko-russkij Slovar*; [Ainu-Russian Dictionary]; V Universitetskoj tipografii: Kazan, Russia, 1875.
25. Batchelor, J. *E-wa-ei santsui jisho. An Ainu–English–Japanese Dictionary and Grammar*; Hokkaidō-chō: Sapporo, Japan, 1889.
26. Radliński, I. Słownik narzecza Ainów zamieszkujących wyspę Szumszu w łańcuchu Kurylskim przy Kamczatce, ze zbiórow Prof. B. Dybowskiego [A dictionary of the dialect of Ainu inhabiting the Shumshu Island in the Kurile Archipelago near Kamchatka, collected by prof. B.Dybowski]. In *Słowniki Narzeczy ludów Kamczackich*; Nakładem Akademii Umiejętności: Cracow, 1891; Volume 1.
27. Chiri, M. *Bunrui Ainu-go jiten. Dai-ikkan: Shokubutsu-hen*; [Dictionary of Ainu, vol. I: Plants]; Nihon Jōmin Bunka Kenkyūsho: Tōkyō, Japan, 1953.
28. Chiri, M. *Bunrui Ainu-go jiten. Dai-sankan: Ningen-hen*; [Dictionary of Ainu, vol. III: People]; Nihon Jōmin Bunka Kenkyūsho: Tōkyō, Japan, 1954.
29. Chiri, M. *Bunrui Ainu-go jiten. Dai-nikan: Dōbutsu-hen*; [Dictionary of Ainu, vol. II: Animals]; Nihon Jōmin Bunka Kenkyūsho: Tōkyō, Japan, 1962.
30. Nakagawa, H. *Ainugo Chitose Hōgen Jiten*; [Dictionary of the Chitose Dialect of Ainu]; Sōfūkan: Tōkyō, Japan, 1995.
31. Tamura, S. *Ainugo jiten: Saru hōgen. The Ainu-Japanese Dictionary: Saru dialect*; Sōfūkan: Tōkyō, Japan, 1996.
32. Kayano, S. *Kayano Shigeru no Ainugo jiten*; [Shigeru Kayano's Ainu dictionary]; Sanseidō: Tōkyō, Japan, 1996.
33. Kirikae, H. *Ainu shin-yōshū jiten: Tekisuto, bumpō kaisetsu tsuki*; [Lexicon to Yukie Chiri's Ainu Shin-yōshū with text and grammatical notes]; Daigaku Shorin: Tōkyō, Japan, 2003.
34. Majewicz, A. Ed. *The Collected Works of Bronisław Piłsudski*; Mouton de Gruyter: Berlin, Germany, 1998–2004; Volumes 1–3.
35. Kindaichi, K. *Ainu Jojishi, Yūkara no Kenkyū*; [Studies of yukar, the Ainu epics]; Tōyō Bunko: Tōkyō, Japan, 1931.
36. Kindaichi, K.; Kannari, M. *Ainu Jojishi Yūkara-shū*; [Collection of *yukar*, the Ainu epics], vols. 1-8; Sanseidō: Tōkyō, Japan, 1959–1968.
37. Chiri, Y. *Ainu shin-yōshū*; [Collection of Ainu mythic epics]; Kyōdo Kenkyūsha: Tōkyō, Japan, 1923.
38. Kayano, S. *Kayano Shigeru no Ainu shinwa shūsei*; [A collection of Ainu myths by Shigeru Kayano] (vols. 1–10); Heibonsha: Tōkyō, Japan, 1998.
39. National Institute for Japanese Language and Linguistics. A Topical Dictionary of Conversational Ainu. Available online: http://ainutopic.ninjal.ac.jp (accessed on 25 August 2017).
40. Chiba University Graduate School of Humanities and Social Sciences. Ainugo Mukawa Hōgen Nihongo—Ainugo Jiten [Japanese—Ainu Dictionary for the Mukawa Dialect of Ainu]. Available online: http://cas-chiba.net/Ainu-archives/index.html (accessed on 25 February 2017).
41. The Ainu Museum. Ainu-go Ākaibu [Ainu Language Archive]. Available online: http://ainugo.ainu-museum.or.jp/ (accessed on 25 August 2018).
42. Katō, D.; Echizen'ya, H.; Araki, K.; Momouchi, Y.; Tochinai, K. Automatic Construction of the Bilingual Words Dictionary for Ainu-to-Japanese Using Recursive Chain-link-type Learning. In Proceedings of the 1st Forum on Information Technology, Tokyo, Japan, 25–28 September 2002; pp. 179–180.
43. Momouchi, Y. Incremental Direct Translation of Noun Phrases of the Ainu Language to Japanese. *IPSJ SIG Tech. Rep.* **2002**, *162*, 79–86.
44. Echizen'ya, H.; Araki, K.; Momouchi, Y. Automatic extraction of bilingual word pairs using Local Focus-based Learning from an Ainu-Japanese parallel corpus. *Bull. Fac. Eng. Hokkai-Gakuen Univ.* **2005**, *32*, 41–63.
45. Azumi, Y.; Momouchi, Y. Development of analysis tool for hierarchical Ainu-Japanese translation data. *Bull. Fac. Eng. Hokkai-Gakuen Univ.* **2009**, *36*, 175–193.
46. Azumi, Y.; Momouchi, Y. Development of tools for retrieving and analyzing Ainu-Japanese translation data and their applications to Ainu-Japanese machine translation system. *Eng. Res. Bull. Grad. Sch. Eng. Hokkai-Gakuen Univ.* **2009**, *9*, 37–58.

47. Momouchi, Y.; Azumi, Y.; Kadoya, Y. Research note: Construction and utilization of electronic data for "Ainu shin-yōsyū". *Bull. Fac. Eng. Hokkai-Gakuen Univ.* **2008**, *35*, 159–171.
48. Momouchi, Y.; Kobayashi, R. Dictionaries and analysis tools for the componential analysis of ainu place name. *Eng. Res. Bull. Grad. Sch. Eng. Hokkai-Gakuen Univ.* **2010**, *10*, 39–49.
49. Senuma, H.; Aizawa, A. Toward Universal Dependencies for Ainu. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), Gothenburg, Sweden, 22 May 2017; pp. 133–139.
50. Senuma, H.; Aizawa, A. Universal Dependencies for Ainu. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 2354–2358.
51. Ptaszynski, M.; Mukaichi, K.; Momouchi, Y. NLP for Endangered Languages: Morphology Analysis, Translation Support and Shallow Parsing of Ainu Language. In Proceedings of the 19th Annual Meeting of The Association for Natural Language Processing, Nagoya, Japan, 12–15 March 2013; pp. 418–421.
52. Bugaeva, A.; Endō, S.; Kurokawa, S.; Nathan, D. A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational Dictionary. Available online: http://lah.soas.ac.uk/projects/ainu/ (accessed on 25 November 2015).
53. Kirikae, H. Ainu ni yoru Ainugo hyōki [transcription of the Ainu language by Ainu people]. *Koku-bungaku kaishaku to kanshō* **1997**, *62*, 99–107.
54. Nakagawa, H. Ainu-jin ni yoru Ainugo hyōki e no torikumi [efforts to transcribe the Ainu language by Ainu people]. In *Hyōki no shūkan no nai gengo no hyōki = Writing Unwritten Languages*; Shiohara, A., Kodama, S., Eds.; Tōkyō gaikoku-go daigaku, Ajia/Afurika gengo bunka kenkyūjo [The Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies]: Tōkyō, Japan, 2006; pp. 1–44.
55. Endō, S. Nabesawa Motozō ni yoru Ainugo no kana hyōki taikei: Kokuritsu Minzoku-gaku Hakubutsukan shozō hitsu-roku nōto kara [Ainu language notation method used by Motozō Nabesawa: From the written notes held by the National Museum of Ethnology]. *Kokuritsu Minzoku-Gaku Hakubutsukan Chōsa Hōkoku* **2016**, *134*, 41–66.
56. Ptaszynski, M.; Ito, Y.; Nowakowski, K.; Honma, H.; Nakajima, Y.; Masui, F. Combining Multiple Dictionaries to Improve Tokenization of Ainu Language. In Proceedings of the 31st Annual Conference of the Japanese Society for Artificial Intelligence, Nagoya City, Japan, 23–26 May 2017.
57. Ptaszynski, M.; Nowakowski, K.; Momouchi, Y.; Masui, F. Comparing Multiple Dictionaries to Improve Part-of-Speech Tagging of Ainu Language. In Proceedings of the 22nd Annual Meeting of The Association for Natural Language Processing, Sendai, Japan, 7–11 March 2016; pp. 973–976.
58. Bugaeva, A. Internet applications for endangered languages: A talking dictionary of Ainu. *Waseda Inst. Adv. Study Res. Bull.* **2011**, *52*, 73–81.
59. Sunasawa, K. *Ku Sukup Oruspe*; [My life story]; Miyama Shobō: Sapporo, Japan, 1983.
60. Peterson, B. Project Okikirmui. The Complete Ainu Legends of Chiri Yukie, in English. Available online: http://www.okikirmui.com/ (accessed on 17 September 2017).
61. Ringger, E.; McClanahan, P.; Haertel, R.; Busby, G.; Carmen, M.; Carroll, J.; Seppi, K.; Lonsdale, D. Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation. In *Proceedings of the Linguistic Annotation Workshop*; Association for Computational Linguistics: Prague, Czech Republic, 28–29 June 2007; pp. 101–108.
62. Giménez, J.; Márquez, L. SVMTool: A general POS tagger generator based on Support Vector Machines. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), European Language Resources Association (ELRA), Lisbon, Portugal, 26–28 May 2004.
63. Toutanova, K.; Klein, D.; Manning, C.; Singer, Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), Edmonton, Canada, 27 May–1 June 2003; pp. 252–259.