A proposal for a unified corpus of the Ainu language

KAROL NOWAKOWSKI[†1] MICHAL PTASZYNSKI[†1]
FUMITO MASUI[†1]

Abstract: Ainu is an endangered language of the Ainu people, native inhabitants of northern Japan. It has been the subject of many studies, but most scholars work on small amounts of language data. We propose a corpus of Ainu covering a wide range of documents, in a consistent structure that will enable large-scale linguistic analysis and support the development of NLP technologies for Ainu, contributing to the process of its revitalization. The corpus contains parallel text in Ainu and Japanese. Its subset includes POS annotations produced by Ainu language experts. For the remaining parts, annotations will be generated automatically. At present, resources collected for the corpus comprise 1.86M characters (410K tokens) of text in Ainu. Their utility for NLP applications has been verified by applying them in a tokenization system, which achieved average F-score above 95%.

Keywords: Ainu language, endangered language, parallel corpus, annotated corpus

1. Introduction

Rapid development of language technologies that we have experienced in last decades, would be impossible without the existence of large-scale digital language corpora. Apart from being crucial for the field of NLP, they are also powerful resources for linguistic studies. However, such resources are not available for the majority of 7000 existing languages[a]. One of them is Ainu – a critically endangered language isolate spoken by the native inhabitants of northern parts of Japan. Due to its unique characteristics (such as noun incorporation or the usage of affixes – instead of pronouns – to express grammatical person), it has been the subject of a number of linguistic and anthropolinguistic studies. However, most researchers, even if they apply modern digital technologies, only work on small amounts of language data. Moreover, there has been no general agreement on such matters as word classes existing in the Ainu language, thus different conventions for linguistic description have been used among experts[b]. To initiate further development of Ainu language studies, we propose a unified corpus of the Ainu language. The corpus covers a wide range of existing documents, in a consistent structure that will enable large-scale linguistic analysis. It will also support the development of language technologies for the Ainu language, contributing to the ongoing process of Ainu language revitalization.

The remainder of this paper is organized as follows. In section 2 we review some related work in that area of language corpora. In section 3 we introduce the Ainu language resources included in our corpus, as well as the ones we plan to add to it in the near future. Section 4 presents the structure of the proposed corpus. In section 5 we explain our approach to the task of annotating the corpus with parts of speech. In section 6 we describe an experiment where the proposed corpus was applied as a training corpus for an n-gram based word segmentation algorithm. Finally, section 7 contains conclusions and ideas for future improvements.

2. Background and related work

For English and other major languages, digital corpora have been compiled since 1960s, one of the pioneering works in the field being the Brown Corpus [5]. As of today, some of the largest corpora available for these languages include hundreds of millions or even billions of words (e.g. [6][7]). In recent years, corpora started to be developed for some of the under-resourced or endangered languages, as well (e.g. [8]). In the case of Ainu, there is only one research project whose name includes the word 'corpus': A Glossed Audio Corpus of Ainu Folklore [9], and the size of that corpus is somewhat modest, at 22,559 tokens. In addition to that, there exist several online repositories of texts in the Ainu language [10][11][12][13]. We will describe them in detail in the next section.

Apart from studies focused on a particular language (or a language pair, in the case of parallel corpora), there is also an ongoing initiative, started by Steven Abney and Steven Bird [14], to build a large-scale multi-lingual corpus in a consistent format allowing for automatic processing. The idea has been picked up by Emerson *et al.* [15], who released the SeedLing, a resource containing small amounts of data from 1451 languages, including 15 sentences in Ainu.

3. Selection of texts for the corpus

Today, the Ainu language is not used as a language of everyday communication, nor as an official language in any institutions. As a result, the amount of available texts in Ainu (in particular, digitized texts) is limited. For that reason, rather than selecting a sample of documents for the corpus, we want to collect and include as many existing materials as possible. That being said, in order to maintain high linguistic quality of our resource, some sort of selection criteria need to be defined. A first intuition would be not to accept texts by non-native speakers, but it must be remembered that Ainu people – including some of those acting as informants in language documentation projects – have not been using Ainu as their first

[†1] Kitami Institute of Technology

[a] According to *Ethnologue* [1], there are 7,097 living languages.

[b] As an example, in each of the three modern dictionaries featuring parts of speech – by Nakagawa [2], Tamura [3] and Kirikae [4] – different part-of-speech classification is used.

language of communication for several decades, therefore the question of what level of Ainu language proficiency should be regarded as sufficient is not trivial. For instance, the *Ainu Times*[c], a magazine in Ainu existing since 1997, is published by the Ainu Language Pen Club led by an ethnic Japanese, Satoshi Hamada. On the other hand, its value as a rare example of Ainu being used in the context of contemporary affairs is indisputable and in the future, we shall consider including its contents in the corpus. Moreover, as we explain in section 4.1, each text in the corpus is accompanied by metadata including detailed information about its authors – using that information the users will be able to narrow down their search to texts that satisfy their requirements.

3.1 Materials included so far

Below we introduce the Ainu language resources included in the corpus so far. The majority of them are products of language documentation projects. Altogether, they comprise a total of 1.86 million characters (410 thousand tokens).

1) *Ainu shin-yōshu* [16]

A collection of 13 epics (*yukar*) compiled by Yukie Chiri. For our corpus we used a version with modernized transcription, published by Hideo Kirikae in 2003 [4].

2) A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational dictionary [10]

An online dictionary based on the *Ainugo kaiwa jiten*, a dictionary compiled by Shōzaburō Kanazawa and Kotora Jinbō, and published in 1898 [17]. It contains 3,847 entries, each of them consisting of a single word, multiple words (synonyms) or a sentence. For our corpus we used the modernized transcription provided by Bugaeva and Endō. Apart from the Ainu text and Japanese and English translations, the dictionary contains information about morphology as well as part-of-speech annotations.

3) Glossed Audio Corpus of Ainu Folklore [9]

A digital collection of 10 Ainu folktales with glosses (morphological annotation) and translations into Japanese and English.

4) Dictionary of Ainu place names [18]

A dictionary of Ainu place names in the form of a database. It contains a total of 3,152 topological names, along with an analysis of their components (including part-of-speech annotation) and Japanese translations.

5) Dictionary of the Mukawa dialect of Ainu [11]

An online resource developed on the basis of 150 hours of speech in the Ainu language (Mukawa dialect), recorded by Tatsumine Katayama between 1996 and 2002 with two native speakers: Seino Araida and Fuyuko Yoshimura. It contains 6,284 entries, comprising a total of 64,656 tokens of text in Ainu.

6) Collection of Ainu Oral Literature [12]

A digital collection of 98 texts in Ainu: 53 prose tales (*uwepeker*), 29 mythic epics (*kamuy yukar*), 11 heroic epics (*yukar*) and 5 texts of other types.

7) Ainu Language Archive – Materials [13]

An online archive of texts in Ainu, based upon voice and video recordings obtained from three native speakers: Matsuko Kawakami, Toshi Ueda and Tatsujirō Kuzuno.

Text collection	Characters total (excluding whitespaces)	Tokens total
Ainu shin-yōshu (revised by Kirikae [4])	36,780	8,786
A Talking Dictionary of Ainu… [10]	55,655	12,978
Glossed Audio Corpus of Ainu Folklore [9]	86,214	22,559
Dictionary of Ainu place names [18]	26,872	9,246
Dictionary of the Mukawa dialect of Ainu [11]	324,022	64,656
Collection of Ainu Oral Literature [12]	479,461	106,257
Ainu Language Archive – Materials [13]	856,555	185,919
Total	1,865,559	410,401

Table 1 Statistics of the Ainu language materials included in the corpus

3.2 Materials to include in the future

In the near future, we plan to expand our corpus with the following materials available online.

1) Ainu Language Material Release Project [19]

An online repository of voice recordings obtained from two native speakers of Ainu: Matsuko Kawakami and Suteno Orita, accompanied by transcriptions (in *kana* and Latin alphabet) and translations into Japanese, released by the Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies. It contains 17 documents (mostly *uwepeker* – prose tales), comprising a total of 39,398 tokens of text in Ainu.

2) Ainu textbooks by FRPAC[d]

A series of 24 textbooks for eight different dialects of Ainu (seven dialects spoken in Hokkaido and Sakhalin Ainu), published by the Foundation for Research and Promotion of Ainu Culture.

3) Ainu Language Radio Course textbooks[e]

A series of textbooks for the “Ainu-go Rajio Kōza” (“Ainu Language Radio Course”), broadcasted by the STV Radio in Sapporo since 1998.

4) The *New Testament*

Translation of the *New Testament* into Ainu by a British missionary, John Batchelor [20], who spent more than sixty years among the Ainu people [21]. Batchelor had received no linguistic training and his works related to the Ainu language have been criticized by some experts (e.g. by Chiri Mashiho [21]). However, due to the relatively large size of the resource, as well as the uniqueness of its contents, we will definitely consider adding it to our corpus.

Furthermore, we plan to digitize the following printed materials.

1) *Ku sukup oruspe* [22]

Memoirs of a speaker of the Ishikari dialect of Ainu – Kura Sunasawa – written in Ainu and Japanese, and edited by Hideo Kirikae. It contains ca. 10,000 words of text in Ainu [23].

[c] http://www.geocities.jp/otarunay/taimuzu.html

[d] https://www.frpac.or.jp/web/learn/language/dialect.html

[e] https://www.stv.jp/radio/ainugo/index.html

2) *Akor Itak* [24]

The first standard textbook of the Ainu language, published in 1994 by the Hokkaido Utari Association (now Hokkaido Ainu Association).

3.3 Representativeness and balance

An important problem in corpus design is the selection of a sample of texts that is representative for the language in question [25]. However, due to the fact that we intend to include as many texts as possible in the corpus, that problem is irrelevant for our project. Of course, it does not mean that our corpus constitutes a balanced representation of the Ainu language – most existing documents are transcriptions of oral literature (such texts amount for nearly 80% of the materials already added to the corpus, listed above).

4. Elements and structure of the corpus

At present, the proposed corpus consists of 215 documents. For each document, we created from 2 to 3 separate XML files: one for the metadata, one for Ainu-Japanese (and English, if available) parallel text, and one for the tokenized Ainu text with linguistic annotations, if such annotations were available. File names of all files pertaining to the same document share a common prefix (document ID), which consists of a 4-letter code for the collection a given document belongs to, and document number – for example, the ID of the first document from the Collection of Ainu Oral Literature [12] is "NIBU1".

4.1 Metadata

The following information about each document is stored in a separate file ("[document ID]-header.xml"):

- Collection name (e.g. "Ainu shin-yoshu");
- Document title, if available;
- Author(s) of the text (i.e. informants), if known;
- Author(s) of transcription;
- Author(s) of translations, if available;
- Copyright information;
- Year of creation, if available;
- Year of publication, if available;
- Date of obtaining;
- Source of the material (bibliographic reference or URL);
- Type of language. Available options: "literary" (i.e. oral literature, prayers), "conversational", "other", "undefined";
- Genre, e.g. "yukar" (heroic epics), "kamuy yukar" (mythic epics), "menoko yukar" (women's epics), "uwepeker" (prose tales), "upaskuma" (teachings of the ancestors), "isoytak" (personal narrative), "yaysama" (improvised songs), "upopo" (sitting songs), "inonno itak" (prayers), "daily conversation", "memoirs", "press", "blog", "other", "undefined";
- Dialect of Ainu, if known.

In addition, there are separate metadata files for text collections (used for storing general information about each collection and copyright information) and for authors of texts (Ainu speakers acting as informants), explaining from which region each of them came and describing their backgrounds.

4.2 Parallel text

The base type of information included in each document is the text in Ainu, transcribed in Latin alphabet. With the exception of two resources ([11] and [18]) and some fragments of documents from other collections, document-level translations into Japanese are available. That allowed us to create an Ainu-Japanese parallel corpus, rather than just a monolingual corpus of Ainu. Moreover, English translations are present in [9] and [10] and Peterson [26] released translations of the *Ainu shin-yōshu* – we also included these materials in our parallel corpus.

The majority of texts used in the corpus are transcripts of oral literature, which was traditionally recited in verses, not necessarily corresponding to sentences. Verse boundaries are reflected (by line breaks) both in transcription of the original speech and in Japanese translations. Knowing that the process of sentence alignment would be time- and labor-consuming, we decided to use that inherent structure of the source material, and established verse as the unit of bitext alignment in those documents. Parallel text in this format is stored in XML files sharing the common suffix "-verses" in their file names. An example is shown in Figure 1. On the other hand, the text (and its Japanese and English translations) in [9] and [10] is generally split into sentences, therefore we store it in files with the suffix "-sentences".

```
<verse id="v3">
  <text lang="ain">sine an to ta pet esoro sinot=as kor</text>
  <text lang="jpn">ある日川に沿って遊びながら</text>
  <text lang="eng">One day, when I went for a swim along the stream,</text>
</verse>
```

Figure 1 Fragment from the parallel corpus, aligned at the level of verses. Collection: *Ainu shin-yōshu*. File: "SYOS12-verses.xml". English text by Peterson [26].

4.3 Annotations

For materials, where linguistic annotations are available, we created additional files, sharing the common suffix "-annotations". At present, our corpus contains the following types of annotation:

- **Part-of-speech (POS) annotation** – available for [10], [18] and a subset (4 out of 13 documents) of [4]. For details see the next section. An example of POS annotated document is shown in Figure 2.
- **Word-level translation (into Japanese)** – available for [18], a subset (4 documents) of [4] and a subset (25 documents) of [13].
- **Morpheme-level glosses** – available for [9], [10] and a subset (18 documents) of [13].

The structure of our corpus as described above was inspired by the data format proposed by Steven Abney and Steven Bird [27]. However, it was designed as a tentative model and in the future we shall consider converting the corpus into one of the standard formats used in corpus building, such as XCES[f] or TEI[g].

```
<verse id="v3">
  <token id="t13">
    <text lang="ain">sine</text>
    <pos lang="jpn" tagset="tamura">連体詞</pos>
  </token>
  <token id="t14">
    <text lang="ain">an</text>
    <pos lang="jpn" tagset="tamura">自動詞</pos>
  </token>
  <token id="t15">
    <text lang="ain">to</text>
    <pos lang="jpn" tagset="tamura">名詞</pos>
  </token>
```

Figure 2 A fragment of POS annotated text. Collection: *Ainu shin-yōshu*. File: "SYOS12-annotations.xml".

```
<verse id="v16">
  <token id="t87">
    <text lang="ain">rapokke</text>
    <tr lang="jpn">そのうちに</tr>
    <morph id="m98">
      <text lang="ain">rapok</text>
      <gloss lang="jpn">間</gloss>
    </morph>
    <morph id="m99">
      <text lang="ain">ke</text>
      <gloss lang="jpn">〜の</gloss>
    </morph>
  </token>
  <token id="t88">
    <text lang="ain">a=</text>
    <tr lang="jpn">（私の・）</tr>
    <morph id="m100">
      <text lang="ain">a=</text>
      <gloss lang="jpn">4.(他主)=</gloss>
    </morph>
  </token>
```

Figure 3 A fragment of text annotated with word-level translations and morpheme-level glosses. Collection: Ainu Language Archive – Materials. File: "AASI20-annotations.xml".

5. Part-of-speech annotations

Apart from its unprecedented size, the biggest added value of our corpus will be the linguistic annotations. One of the basic type of annotation included in many language corpora is part-of-speech (POS) annotation. At this point, a subset of our corpus includes POS annotations produced manually by experts, namely Bugaeva and Endō [10], Momouchi and Kobayashi [18], and Momouchi (for a part of the *Ainu shin-yōshu*). However, that covers less than 10% of all texts in the corpus. On the other hand, manual annotation is a time-consuming process, therefore our plan is to apply bootstrapping techniques (as described e.g. in [28]) in order to perform POS annotations for the remaining part of the corpus automatically. A dedicated POS tagger for Ainu (POST-AL) was already developed by Ptaszynski and Momouchi in 2012 [29] and can be used for the annotation task. Furthermore, in the near future we plan to adapt and test some of the state-of-the-art taggers developed for other languages, such as the SVMTool [30] and the Stanford Log-linear Tagger [31]. In the final stage, the annotations generated using automatic tools will be checked by Ainu language experts.

Another challenge related to POS annotations is the fact that different part-of-speech classifications are used in each of the existing part-of-speech annotated resources. To solve that problem, we converted all annotations to the part-of-speech classification standard used by Nakagawa [2] and added the result to the corpus as alternative annotations (with the "tagset" attribute set to "nakagawa"). The conversion standard we used is shown in Table 1.

Bugaeva and Endō [10]	Momouchi and Kobayashi [18]		Nakagawa [2]
完全動詞 (complete verb)	完全動詞 (complete verb)	>	0 項動詞 (complete verb)
自動詞 (intransitive verb)	自動詞 (intransitive verb)	>	1 項動詞 (intransitive verb)
他動詞 (transitive verb)	単他動詞 (transitive verb)	>	2 項動詞 (transitive verb)
複他動詞 (ditransitive verb)	複他動詞 (ditransitive verb)	>	3 項動詞 (ditransitive verb)
	人称代名詞 (personal pronoun)	>	代名詞 (pronoun)
	疑問代名詞 (interrogative pronoun)	>	疑問詞 (interrogative)
	疑問副詞 (interrogative adverb)	>	疑問詞 (interrogative)
後置副詞 (postpositive adverb)	後置副詞 (postpositive adverb)	>	副詞 (adverb)

Table 1 Table for the conversion of other Ainu part-of-speech standards into Nakagawa's standard

6. Application of the corpus in automatic word segmentation

In order to verify the utility of materials included in the corpus for NLP applications, we applied it as a training corpus in an experiment with automatic word segmentation algorithm developed by Nowakowski *et al.* [32].

In the Ainu language in its written form, there are no clear guidelines regarding correct word segmentation. The problem is especially remarkable in older texts, as their authors tended to

f) http://www.xces.org/
g) http://www.tei-c.org/

use less word segmentation (sentences were divided not into single words, but into chunks containing multiple words). As a consequence, automatic processing of such texts is impossible without a mechanism for word boundary identification.

Original transcription [16]:	Nenkatausa wakka unkure
Modern transcription [4]:	Nen ka ta usa wakka un kure
Meaning:	Someone, please give me water

Table 2 Example of different word segmentation in the original and modernized transcription of the *Ainu shin-yōshū*

To address that problem, Ptaszynski and Momouchi [29] and Nowakowski *et al.* [33] developed dictionary based word segmentation algorithms, but their effectiveness was limited due to the fact, that they relied solely on a lexicon, without using any form of contextual or statistical information. As a solution, Nowakowski *et al.* [32] proposed a word n-gram based word segmentation algorithm. To test both the algorithm and the proposed corpus, we extracted n-gram data from the texts included in the corpus and used it as the training data for the system. Details of the experiments can be found in [32] – here we will only note, that after applying all 7 text collections mentioned in section 3.1, the segmentation algorithm achieved results (F-score) between 91.4% and 99.6% (95.9% on average) on held out data, whereas with only one text collection used (namely, the *Ainu shin-yōshu*), the average F-score was 88.2%. This shows that creating a large-scale corpus of Ainu can effectively support the development of NLP technologies for that language.

7. Conclusions and future work

In this paper, we described the scope and structure of the newly developed corpus of the Ainu language. At present, our corpus is still relatively small, but results of experiments with n-gram based word segmentation algorithm – where it was applied as the training data – indicate that it allows for substantial performance improvement in NLP applications, as compared to experiments using small, homogeneous datasets.

Important tasks for the future include, apart from expanding the corpus with new materials, automatic generation of part-of-speech annotations for all documents. We also plan to adapt or develop a browser tool which will let users search the contents of the corpus (e.g. concordances) and release it in the form of a Web service. Furthermore, if we are able to receive necessary permissions from the parties holding copyrights for the materials used, we intend to make the entire resource publicly available.

References

[1] Gary F. Simons and Charles D. Fennig (eds.). (2018). *Ethnologue: Languages of the World, Twenty-first edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

[2] Hiroshi Nakagawa. (1995). *Ainugo Chitose Hōgen Jiten* [Dictionary of the Chitose dialect of Ainu]. Sōfūkan, Tokyo.

[3] Suzuko Tamura. (1996). *Ainugo jiten: Saru hōgen. The Ainu-Japanese Dictionary: Saru dialect*. Sōfūkan, Tokyo.

[4] Hideo Kirikae. (2003). *Ainu shin-yōshu jiten: tekisuto, bumpō kaisetsu tsuki* [Lexicon to Yukie Chiri's Ainu shin-yōshu with text and grammatical notes], Daigaku Shorin, Tokyo.

[5] Henry Kučera and W. Nelson Francis. (1967). *Computational Analysis of Present-day American English*. Brown University Press.

[6] Mark Davies. (2008-). The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available online at https://corpus.byu.edu/coca/ (accessed 2018-08-20).

[7] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. (2012). "YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information". In *Proceedings of The AISB/IACAP World Congress 2012 in Honour of Alan Turing*, 2nd Symposium on Linguistic and Cognitive Approaches To Dialog Agents (LaCATODA 2012), pp. 40-49, 2-6 July 2012, University of Birmingham, Birmingham, UK.

[8] Joel Martin, Howard Johnson, Benoit Farley and Anna Maclachlan. (2003). Aligning and using an English-Inuktitut parallel corpus. In Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond - Volume 3, pp. 115-118.

[9] Hiroshi Nakagawa, Anna Bugaeva and Miki Kobayashi. (2016). A Glossed Audio Corpus of Ainu Folklore. NINJAL. Available online at http://ainucorpus.ninjal.ac.jp (accessed 2018-08-20).

[10] Anna Bugaeva and Shiho Endō (eds.). (2010). A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational dictionary. Retrieved November 25, 2015 from http://lah.soas.ac.uk/projects/ainu/

[11] Chiba University Graduate School of Humanities and Social Sciences. (2014). Ainugo Mukawa Hōgen Nihongo – Ainugo Jiten [Japanese – Ainu Dictionary for the Mukawa Dialect of Ainu]. Retrieved February 25, 2017 from http://cas-chiba.net/Ainu-archives/index.html

[12] Nibutani Ainu Culture Museum. (n.d.). Collection of Ainu Oral Literature. Retrieved December 11, 2017 from http://www.town.biratori.hokkaido.jp/biratori/nibutani/culture/language/

[13] The Ainu Museum. (2017-2018). Ainu-go Ākaibu [Ainu Language Archive]. Retrieved December 15, 2017 from http://ainugo.ainu-museum.or.jp/

[14] Steven Abney and Steven Bird. (2010). 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*, pp. 88–97, Uppsala, Sweden, 11-16 July 2010.

[15] Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer and Michaela Regneri. (2014). SeedLing: Building and using a seed corpus for the Human Language Project. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 77–85.

[16] Yukie Chiri. (1923). *Ainu shin-yōshu* [Ainu songs of gods]. Kyōdo Kenkyūsha, Tokyo.

[17] K. Jinbō and S. Kanazawa. (1898). Ainugo kaiwa jiten [Ainu conversational dictionary]. Kinkōdō Shoseki, Tokyo.

[18] Yoshio Momouchi and Ryosuke Kobayashi. (2010). Ainu-go chimei kōsei yōso kaiseki no tame no jisho to kaiseki tsūru no kōsei. Dictionary and Analysis Tools for the Componential Analysis of Ainu Place Names.

[19] Information Resources Center, Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign

Studies. (n.d.). AA-ken Ainu-go shiryō kōkai purojekuto [Ainu Language Material Release Project]. Retrieved July 21, 2018 from http://ainugo.aa-ken.jp/

[20] *Chikoro Utarpa ne Yesu Kiristo Ashiri Aeuitaknup Oma Kambi. The New Testament of our Lord and Saviour Jesus Christ in Ainu*. Translated by John Batchelor. Printed for the Bible Society's committee for Japan by the Yokohama bunsha. 1897.

[21] Kirsten Refsing. (1986). *The Ainu language. The morphology and syntax of the Shizunai dialect*. Aarhus University Press, Aarhus.

[22] Kura Sunasawa. (1983). *Ku sukup oruspe* [My life story]. Miyama Shobō, Sapporo.

[23] Masayoshi Shibatani. (1990). The languages of Japan. Cambridge University Press, London.

[24] Hokkaidō Utari Kyōkai [Hokkaido Ainu Association]. (1994). *Akor Itak* [Our Language]. Sapporo.

[25] Douglas Biber. (1993). Representativeness in Corpus Design. In *Literary and Linguistic Computing*, Volume 8, Issue 4, 1 January 1993, pp. 243–257.

[26] Benjamin Peterson. (2013). Project Okikirmui The complete Ainu legends of Chiri Yukie, in English. Available online at http://www.okikirmui.com/ (accessed 2018-08-20).

[27] Steven Abney and Steven Bird. (2011). Towards a Data Model for the Universal Corpus. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pp. 120-127.

[28] Stephen Clark, James R. Curran and Miles Osborne. (2003). Bootstrapping POS taggers using Unlabelled Data. School of Informatics. University of Edinburgh.

[29] Michal Ptaszynski and Yoshio Momouchi. (2012). Part-of-Speech Tagger for Ainu Language Based on Higher Order Hidden Markov Model. In: *Expert Systems With Applications* 39 (2012), pp. 11576–11582. Issue14.

[30] J. Giménez and L. Márquez. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC’04).

[31] K. Toutanova, D. Klein, C. Manning and Y. Singer. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

[32] Karol Nowakowski, Michal Ptaszynski and Fumito Masui. (2018). *Word n-gram based tokenization for the Ainu language*. Manuscript submitted for publication.

[33] Karol Nowakowski, Michal Ptaszynski and Fumito Masui. (2017). Improving Tokenization, Transcription Normalization and Part-of-speech Tagging of Ainu Language through Merging Multiple Dictionaries. In: *Proceedings of the 8th Language & Technology Conference* (LTC'17), pp. 317-321.