Universal Dependencies for Ainu

Hajime Senuma†‡, Akiko Aizawa‡†
†University of Tokyo, ‡National Institute of Informatics
{senuma, aizawa}@nii.ac.jp

Abstract

This paper reports an on-going effort to create a dependency tree bank for the Ainu language in the scheme of Universal Dependencies (UD). The task is crucial both language-internally (language revitalization) and language-externally (providing sources for new features and insights to UD). Since the language shows many of the representative phenomena of a type of languages called polysynthetic languages, an annotation schema to Ainu can be used as a basis to extend the current specification of UD. Our language resource comprises an annotation guideline, dependency bank based on UD, and a mini-lexicon. Although the size of the dependency bank will be small and contain only around 10,000 word tokens, it can serve as a base annotation for the next step. Our mini-lexicon is encoded under the W3C OntoLex specification with UD and UniMorph (UM) features with the system-friendly JSON-LD format and thus bearable to future extensions. We also provide a brief description of dependency relations and local features used in the bank such as pronominal cross-indexing and alienability.

1. Introduction

The project of Universal Dependencies (UD) (Nivre et al., 2016) marks a milestone in the history of natural language processing (NLP), as it unifies syntactic annotation schemes across languages/corpora and enables cross-lingual processing. At the CoNLL Shared Task 2017 for UD, 33 teams participated in the task, proving that the community is quickly thriving (Zeman et al., 2017).

We are currently working on the UD annotation scheme and corpus of Ainu, a language spoken by the Ainu people, an ethnic minority in Japan. A part of the work was presented elsewhere (Senuma and Aizawa, 2017). The task aims to tackle two purposes:

1. to promote the language revitalization of Ainu, as it is a highly endangered language but has a rich amount of the records of oral literature, and

2. to serve as a basis to test if universal specifications in NLP such as UD and UniMorph (UM) (Sylak-Glassman, 2016) can encode the world's languages.

The first issue is urgent because Ethnologue classifies the languages as nearly extinct (Lewis et al., 2016), and scholars have been alerting the status of language usages (DeChicchis, 1995; Sato, 2012). To mitigate the situation, Bugaeva (2011) created a freely-accessible Ainu dictionary for daily conversation based on the old dictionary published in 1898 by Kotora Jimbō and Shōzaburo Kanazawa. Our work will be in line of these movements.

Whereas the first motivation is language-internal, the second motivation is external, as it contributes to our deep understanding of natural languages. Being a member of polysynthetic languages (Baker, 1996), the language exhibits many peculiar properties from the view point of the speakers of major languages. Complex verb formation rules called noun incorporation (Mithun, 1984) may be the hallmark of such properties. Other properties (some of which will be discussed later) include pluractionality, possessed case, alienability, and multiple pronominal markers. These "peculiar" properties are, however, actually very common in world's languages, if we look at typological data such as WALS Online (Dryer and Haspelmath, 2013). Through the examination of annotating Ainu, it may be possible to be used to extend the current specifications of NLP including UD.

2. Dataset and System Description

We are currently working on the annotation of a collection of traditional Ainu songs *Ainu Shin'yōshū*, transcribed by Yukie Chiri into latin scripts, including Japanese translations by herself (Chiri, 1923). We also consulted the work of (Kirikae, 2003), which retranscribed Yukie Chiri's writing into one close to modern orthographic system and appended a lexicon for all words appeared in the text. The work consists of thirteen mythological songs in the style of *kamuyyukar* type poetry, typical in the Ainu oral literature. Since the work is the most famous Ainu literature, distributing resources to read it in the open format will be useful for language revitalization. Furthermore, Ainu poetry is usually told by *atomte itak* "Adorned Speech", a variant of Ainu which has more polysynthetic nature than *yayan itak* "Common Speech", so it serves as a good resource to expand the inventory of UD and UM.

At this stage, the annotation process has been mainly done by the first author alone, since it is experimental and requires the quality of linguistic efforts rather than the quantity of human power. Although the annotation process is behind schedule, but we plan to release the data set until camera-ready. We estimate the size of the resulting corpus will be around 10,000. It is very small in comparison with other corpora, but still it is enough to create a sound documentation for annotating the language.

Although Ainu did not have orthographic systems in ancient times, two writing systems have been used in these 200 years: one based on latin scripts and one based on Japanese katakana. Latin-based one reflects the phonology of Ainu more accurately, while katakana-based system is more friendly for elder people and young children in Japan. the Foundation of Research and Promotion of Ainu Culture (FRPAC) provides instructions for both systems.

Since the original Ainu text was written in the outdated version of latin-based orthography, we retranscribed it into a

latin-based modern orthographic system , almost equivalent to FRPAC. We, however, adopted two orthography rules not seen in FRPAC, following Tamura (2000):

1. irregular accent positions are marked by an acute, as in *húre*, and
2. irregular positions of glottal stops are marked by apostrophe, as in *yay'eyukar*.

Our data set contains three items:

1. UD annotation guideline, in the form of Markdown texts, the official format for UD documentation,
2. dependency tree bank formatted in JSON, and
3. mini-lexicon formatted in JSON-LD, following the W3C OntoLex/lemon inventory.

3. Data Format

This section briefly discusses the design of data format and system used in our annotation.

3.1. Dependency bank

Our dependency bank is written in JSON format. It has a similar style to sd-parse, the official UD annotation format for documentation, although it is still JSON, maintaining easiness to be used in systems.
By using a converter, we translate data into CoNLL-U, the official format of UD. We also developed a dependency bank viewer for our dataset (Figure 1). A tooltip pops up to show the contents of our lexicon by pointing to a word that the reader does not know. The viewer will serve as a handy tool to learn the Ainu language.

3.2. Lexicon

Each entry of our lexicon contains basic information such as word forms, pronunciations, and the concise definitions of its senses. It also contains bibliographic references to three dictionaries: Kirikae (2003; Nakagawa (1995; Tamura (1996). It is formatted in JSON. For example:

```
{
  "@context":
    "...",
  "@id": "san_1",
  "@type": "Word",
  "canonicalForm": {
    "latn": "san",
    "kana": "",
    "ipa": "san"
  },
  "lexicalForm": [
    {
      "latn": "sap",
      "kana": "",
      "ipa": "sap",
      "feature": [
        "um:intr",
        "x:pluract"
      ]
    }
  ],
  "pos": "verb",
  "feature": ["um:intr"],
  "sense": [
    {
      "@id": "san_1_1",
      "reference": {
        "prefLabel":
          "to go downstream"
      },
      "usage": "...",
      "bibliography": [
        {
          "bib:key": "bib:Kirikae2003",
          "bib:loc": "p.˜378"
        },
        {
          "bibkey": "bib:Nakagawa1995",
          "bib:loc": "p.˜203"
        },
        {
          "bibkey": "bib:Tamura1996",
          "bib:loc": "pp.˜602-603"
        }
      ]
    },
    {
      "@id": "san_1_2",
      "reference": {
        "prefLabel": "to go toward ..."
      },
      "bibliography": [
        {
          "bib:key": "bib:Tamura1996",
          "bib:loc": "pp.˜602-603"
        }
      ]
    }
  ]
}
```

In reality it is in W3C JSON-LD (Sporny et al., 2017), the specification for linking data. By using a context file it is possible to expand each element into the RDF triples, enabling ontological data exchange. For example, in our context file, there are mappings like the following.

```
"ontolex": "http://www.w3.org/ns/lemon/ontolex#",
"latn": {
  "@id": "ontolex:writtenRep",
  "@language": "ain-Latn"
},
```

Then we can convert it into expanded triples as

```
_:b0
http://www.w3.org/ns/lemon/ontolex#writtenRep
"san"@ain-latn .
```

This way we can link our lexicon to the vocabulary of OntoLex/Lemon (W3C Ontology-Lexicon Community, 2017), bridging gaps.
Our lexicon has features associated to the inventory of UM (Sylak-Glassman, 2016; Cotterell et al., 2017) which also allows morphosyntactic descriptions.
Although not seen in the above example, our lexicon can also contain word compositions and synonym/antonym relations. Through OntoLex it may be possible to connect to WordNet in the future.

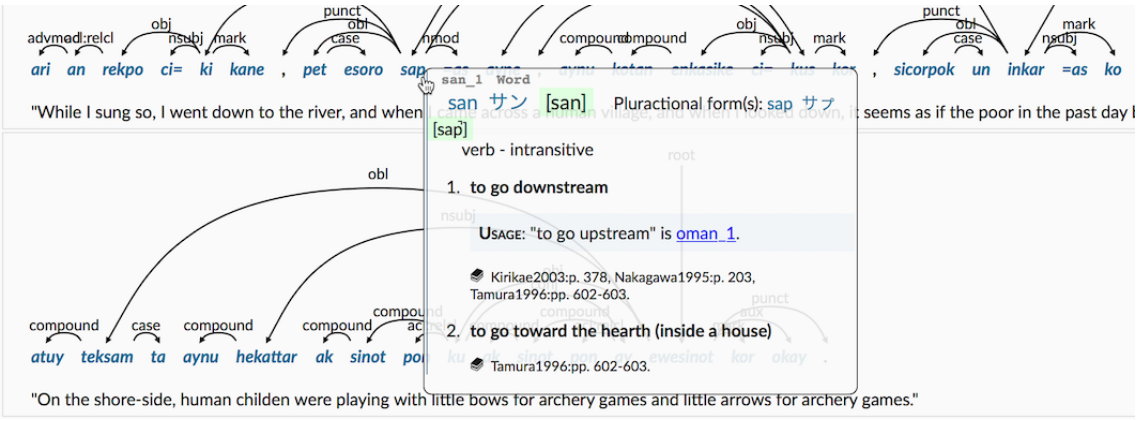


Figure 1: Screen shot of the dependency bank viewer with linking to the lexicon

4. Linguistic Properties

Ainu is abundant for linguistic phenomena not seen in major languages. In this extended abstract we mention two of such properties.

In this section, the style of interlinear glossings are roughly based on the Leipzig Glossing Rules (Comrie et al., 2008).

4.1. Pronominal cross-indexing

The first one is related to pronominal agreement (or pronominal *indexing* in the functionalist terminology (Croft, 2003)). The Ainu language has two sets of pronominal markers: clitics (such as *e*=, "you") and pronouns (such as *eani* "you"). Nevertheless, in almost all cases, pronouns do not occur in texts, and only clitics are realized.

horkew *e=* *ne*
wolf 2= COP

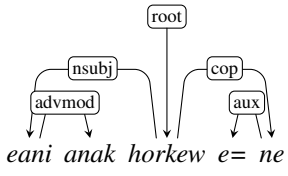
"You are a wolf."

However, in some cases (especially if we want to emphasize who involved a topic), pronouns can be optionally used, although clitics are obligatory.

eani *anak* *horkew* *e=* *ne*
2SG.S INT wolf 2SG.S= COP

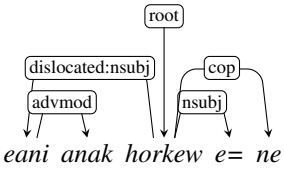
"(As for you), you a wolf."

This kind of phenomena prohibits us to annotate even for simple declarative sentences, because we do not know which tokens should be counted as subject. In previous work (Senuma and Aizawa, 2017), we annotated this phenomenon in the following strategy.



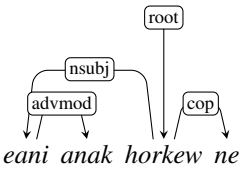
The problem of this approach is that the pronominal clitic *e*= is treated as an auxiliary, rather than an argument. This approach contradicts to formalists and functionalists in orthodox linguistics. Some formalists in the Chomskyan tradition claim that clitics (such as *e*=) are true arguments to predicates in some polysynthetic languages and pronouns are mere adjuncts (Jelinek, 1984; Baker, 1996). On the other hand, functionalists reject the idea of arguments/adjuncts dichotomy in the first place; Haspelmath (2013) called pronouns/nominals as *cominals* and claimed that in these *cross-indexing* constructions both are real arguments . At any rate, a clitic *e*= must be treated as an argument, for cross-lingual comparisons, although at the same time we should also adhere to UD's approach that there should be no more than one `nsubj` (nominal subject).

A solution suggested by the UD community is the usage of `dislocated`, a relation originally used to encode dislocated pronouns, commonly seen in informal French. Descriptively, the approach is also justified by the fact that in an Ainu grammar published in 1936, Mashiho Chiri "likens the use of the Ainu personal pronouns to those of Latin and French, and contends that the following expressions are parallel in their use of the over pronoun (Ainu *kuani* 'I', Latin *ego*, and French *moi*)" (Shibatani, 1990, p. 30). However, unlike the inherent informality of French, Ainu systematically uses this system. According to Haspelmath (2013, p. 8), French dislocation is not counted as conominals, because pronominal indexes (such as *e*= in Ainu) and conominals must be in the same narrow phrases, while French dislocation occurs outside these phrases. We thus introduce language-specific features `dislocated:nsubj`, `dislocated:obj`, and `dislocated:iobj` to indicate that they are conominals with valid status as subjects and objects, without violating UD's restriction that a predicate must not have more than one argument per role.



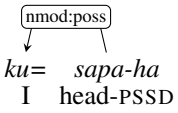
In some dialects of Ainu, for informal speech, it is re-

ported that pronominal clitics were dropped, possibly due to the influence of Japanese (Izutsu, 2006). Since UD has the realization-first approach, in these cases conominals are promoted to common arguments, as in:



4.2. Alienability and possessed case

In most major languages, syntactic information is marked on its dependent, e.g., "my head" (rather than "I head-have"). The Ainu language, however, marks it on the head (*possessum* (Croft, 2003) or *possessee*), as in *ku=sapa-ha* "I=head-PSSD" in some circumstances.

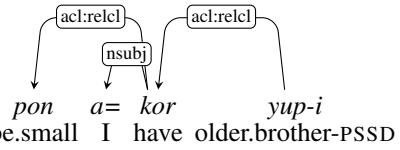


"my head"

Such phenomenon is called the *possessed case* (Sylak-Glassman, 2016). In reality, out of 236 languages recorded in the WALS Online, the number of languages with dependent-marking is 98 (42%) and that with head-marking possession is 78 (33%), and therefore it is by no means a minor construction.

The UD corpus for a Uralic language Hungarian, being an exceptional language in Europe which has head-marking possession, utilizes UD's "layered feature" (e.g., `Person[psor]=3` "it is possessed by 3rd person") to annotate possessed cases. Ainu possessees do not, however, inflect on number/person, and its meaning is only realized on a clitic such as *ku=*, thereby prohibiting the usage of these systems. We thus borrowed PSSD (possessed case) from the UniMorph inventory, and used a language-specific feature `Case=Pssd`.

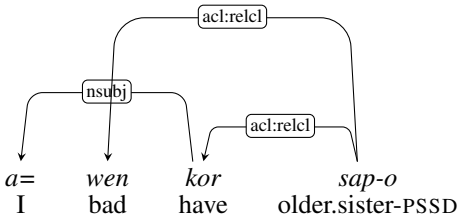
In addition, Ainu has another possessive construction with a relative clause (the phrase was taken from Shibatani (1990, p. 44)).



"my young older brother" (lit. "older brother that is young and that I=have")

In the above case, the form *a=kor yupi* ("older brother I have") is optional, and in common speech the form *a=yupihi* is preferred. But not all nouns have possessed forms. Only two classes of nouns called inalienable nouns (such as body parts and kins) and locative nouns (such as *or* "place" and *enka* "higher place (above)") have possessed cases. If nouns are alienable, relative clause based constructions are mandatory. We use language-specific lexical features `Alienability=Naln` (inalienable) and `Alienability=Aln` (alienable), borrowed from the inventory of UniMorph, and `Locativity=Yes` to annotate these words.

It is interesting to see that the pronominal clitic can reside in anywhere in the relative-clause construction, exhibiting crossing dependencies in some cases (the phrase was taken from Shibatani (1990, p. 44), too).



"my dear older sister" (lit. "older sister I=that is bad and (...) have")

5. Conclusion

In this paper, we presented an attempt to construct UD bank and lexicon for Ainu, including dataset description, data format description, and explanations for some linguistic properties. Compared with previous work published from the UD Workshop, we created a system and viewer to annotate Ainu more easily, and a lexicon which can be also used as a bridge to the community of morphology and the community of ontology/lexicon. We furthermore refined our annotation scheme so that it meets the standards of both linguistic typology and UD. We plan to publish the dataset in permissible open licenses such as CC-BY and MIT in the near future.

6. Acknowledgements

This work was supported by CREST, Japan Science and Technology Agency. We are also grateful to the following researchers for valuable information and suggestions: William Croft, Teresa Lynn, Hiroshi Nakagawa (Chiba University), Joakim Nivre, Sebastian Schuster, and Francis Tyers.

7. Bibliographical References

Baker, M. C. (1996). *The Polysynthesis Parameter*. Oxford University Press.

Bugaeva, A. (2011). Internet Applications for Endangered Languages: A Talking Dictionary of Ainu. *Waseda Institute for Advanced Study Research Bulletin*, 3:73–81.

Chiri, Y. (1923). *Ainu Shin'yōshū [Collection of Ainu Kamuyyukar]*. Kyodo Kenkyusha.

Comrie, B., Haspelmath, M., and Bickel, B. (2008). The Leipzig Glossing Rules. Technical report, Max Planck Institute for Evolutionary Anthropology and University of Leipzig.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2017). CoNLL-SIGMORPHON 2017 Shared Task: Universal

Morphological Reinflection in 52 Languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Croft, W. (2003). *Typology and Universals, 2nd ed.* Cambridge University Press.

DeChicchis, J. (1995). The current state of the Ainu language. *Journal of Multilingual and Multicultural Development*, 16(1-2):103–124, January.

Dryer, M. S. and Haspelmath, M. (2013). WALS Online. Technical report, Max Planck Institute for Evolutionary Anthropology.

Haspelmath, M. (2013). Argument indexing: a conceptual framework for the syntactic status of bound person forms. In Dik Bakker et al., editors, *Languages Across Boundaries: Studies in Memory of Anna Siewierska*, pages 197–226. Walter de Gruyter.

Izutsu, K. (2006). Ainugo no Hinshi Bunrui Saikō: Iwayuru Ninshō Daimeishi o Megutte [Ainu Parts of Speech Revisited: with Special Reference to So-called Personal Pronouns]. *Journal of Hokkaido University of Education: Humanities and Social Sciences*, 56(2):13–27.

Jelinek, E. (1984). Empty Categories, Case, and Configurationality. *Natural Language & Linguistic Theory*, 2(1):39–76, January.

Kirikae, H. (2003). *Ainu Shin'yōshu Jiten [Lexicon to Ainu Shin'yōshū]*. Daigaku Shorin.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2016). *Ethnologue: Languages of the World, Nineteenth edition*. SIL International.

Mithun, M. (1984). The Evolution of Noun Incorporation. *Language*, 60(4):847–894.

Nakagawa, H. (1995). *Ainugo Chitose Hōgen Jiten [The AinuJapanese Dictionary: Chitose Dialect]*. Sōfūkan.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Sato, T. (2012). Ainugo no Genjō to Fukkō [The Present Situation of the Ainu Language and Its Revitalization]. *Gengo Kenkyu*, 142:29–44.

Senuma, H. and Aizawa, A. (2017). Toward Universal Dependencies for Ainu. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*.

Shibatani, M. (1990). *The Languages of Japan*. Cambridge University Press.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., and Lindström, N. (2017). JSON-LD 1.1: A JSON-based Serialization for Linked Data, Draft Community Group Report 12. Technical report, W3C.

Sylak-Glassman, J. (2016). The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema), Working Draft, v. 2. Technical report.

Tamura, S. (1996). *Ainugo Saru Hōgen Jiten [The Ainu-Japanese Dictionary: Saru Dialect]*. Sōfūkan.

Tamura, S. (2000). *The Ainu Language*. ICHEL Linguistic Studies. Sanseidō.

W3C Ontology-Lexicon Community. (2017). Lexicon Model for Ontologies. Technical report.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinková, S., Hlaváčová, J., Kettnerová, V., Nka Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., De Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., De Paiva, V., Droganova, K., Alonso, H. M., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *CoNLL*, (1):1–19.