Making a Phonological Corpus of Nanai Language

Introduction

Nanai is a severely endangered language from the Upper Amur River region with fewer than 200 speakers. This paper intends to start a publically-available phonological corpus of Nanai language that can be used for further enquiries into its phonological system, especially into the nature of its vowel harmony.

Nanai People

Nanai people are a Tungusic people who have traditionally inhabited the banks of the Amur/Heilongjiang, Sunggari/Songhuajiang and Ussuri/Wusuli rivers, currently residing around the Chinese-Russian border. Among the Tungusic peoples Nanai are the 5th biggest with approximately 18,000 people; they are preceded by Manchu (10 million), Xibe (191,000), Ewenki (69,000) and Ewens (22,000).

Nanai ethnogenesis is a subject of ongoing research although it is believed that Tungus peoples mixed with a Paleosiberian substrate which for Nanai specifically was possibly Nivkh (Larin et al., 2003). The Tungusic motherland is likely located in Southern Manchuria and North Korea, and Nanai spread to the Amur basin between 2,000 to 1,000 years ago (Janhunen, 2005). The population of this region has relied on fishing as their main food source since at least 6th millenium BCE. Nanai have historically been fishers as well, although in the 1950s the economic programmes of the Soviet government forced Nanai people to seek employment in agriculture, industry and elsewhere (Larin et al., 2003).

The biggest issue for Nanai people is rampant unemployment (Larin et al., 2003).

Nanai Language

Nanai (Nani, Hezhe; ISO 639-3 gld, Glottolog nana1257) is a severely endangered Tungusic language spoken in the Upper Amur River region, in Russian Far East (Khabarovsk and Primorsky krai) and in the Heilongjiang province of China.

Orok and especially Ulch languages are very close to Nanai. Its position relative to the other Tungusic languages remains a controversial subject, with some linguists proposing a dialect continuum view of the whole family, others dividing it into the Northern and Southern branches and including Nanai into the latter, some offering a 3-branch classification and others offering a 4-way split into Ewenic, Udegheic, Nanaic and Jurchenic branches (Whaley et al., 1999; Whaley, 2012).

The prestige dialect of the Najkhin village has the most speakers (Sem, 1997), but many others exist: Daerga, Dada, Kukan, Achan etc. It is proposed that the Kilen dialect spoken in China is a language in its own right (Zhang, 2013). The same might be true for Kili dialect spoken in Russia (Janhunen, 2005). Both Kilen and Kili are extremely close to Nanai in terms of morphology and syntax but Kili sound system and vocabulary is closer to Ewenki while Kilen is believed to be close to Udeghe and Orok (Janhunen, 2005; Zhang, 2013).

Russian Nanai speakers use modified Cyrillic alphabet proposed in 1936 while Chinese Nanai remains an unwritten language (Zhang, 2013). It is worth mentioning that the first Nanai orthography was proposed in 1928 and then changed 3 times (Janhunen, 2005). At the same time, Soviet linguists created an artificial literary standard for Nanai that was not accepted by many dialect speakers, and began printing textbooks that did not gain popularity amongst the speakers (Janhunen, 2005).

No publically available corpora of Nanai exist as of 2020. Most published audio and video materials are recordings of fairy-tales collected in the 1980s by Soviet folklorists (Kile et al., 2018) and more recent data collected by linguist and language revitalisation activist Vasily Kharitonov from 2017 onwards (Kharitonov, Xisangoru).

Sociolinguistic Situation

Nanai is severely endangered with around 180 L1 speakers (Campbell et al., 2017), not more than twenty speakers remain in China (Moseley, 2010). In 1970s many Soviet children from indigenous peoples of the North, Siberia and the Far East were sent away to boarding schools that were often located far from their hometown; the same was true for Nanai children. This practice disrupted the traditional system of language transmission and contributed to the decline of Nanai (Larin et al., 2003).

In Russia, it is taught to children in Nanai primary schools as a second language for 2-6 hours per week, yet no students acquire the ability to speak it (Kharitonov, 2013). Adults can study Nanai in Nikolajevsky-na-Amure teaching college, Far Eastern State University of Humanities and Herzen State Pedagogical University of Russia in Saint-Petersburg (Sulyandziga, 2003). As of 2010s, there are no L1 speakers younger than 50 (Ko & Yurn, 2011; Kharitonov, 2013).

Nanai speakers have a positive attitude about revitalization efforts but it is not actively happening due to problems in communication between language activists, teachers, organisations and potential speakers (Kharitonov, 2013).

Grammar

Nanai sound system is relatively well-researched. Most researchers report 17 or 18 phonemic consonants and 6 phonemic vowels arranged in a double triangular system.

	–back	**+back**	
		–rounded	+rounded
–RTR	i iː	ə əː	u uː
+RTR	ɪ ɪː	a aː	o oː

All vowels can be short or long:

- пиктэ /piktə/ ‘child’
- пиктэ /piːktə/ ‘nettle’

Word-final /n/ is not realised, nasalising the preceding vowel instead:

- би /bi/ [bi] ‘to live’
- бин /bin/ [bĩ] ‘life’

Possible syllable structure is (CC)V(C):

- эситул /ə.si.tul/ ‘immediately’
- эрчэн /ər.cən/ ‘lower part of the roof’

Nanai has vowel harmony, distinguishing two classes of vowels: /i ə u/ and /ɪ a o/. The nature of the alteration between them is unclear. Soviet researchers believed it to be a high-low type harmony but several Korean researchers propose a [+RTR]/[−RTR] contrast (Yun et al., 2016). Moreover, since /i/ and /ɪ/ are not distinguished orthographically, and reports that /ɪ/ is weakened in non-initial syllables (Ko & Yurn, 2011), there is no universally accepted description of this alteration.

	Labial	Lamino-alveolar	Dorso-palatal	Dorso-velar
Plosive	p b	t d	c ɟ	k g
Fricative		s		x
Nasal	m	n	(ɲ)	ŋ
Trill		r		
Approximant	w	l	j	

17 or 18 consonantal phonemes are usually reported; Ko & Yurn argue that /ɲ/ should not be considered a phoneme since there are no minimal pairs with /ni/ and /ɲi/ contrasting, hence it is not possible to prove that they are in fact independent (2011).

Corpus Linguistics and Corpus Phonology

Corpus linguistics is one of the most widely used research methods that can be traced back to 18th century although it had only taken its current form in 1960s. With the advent of the Internet, corpora started to be used even wider, for an array of studies including morphosyntax, language change and variation, et cetera (Durand et al., 2014).

Corpus phonology is an emerging subfield of corpus linguistic that uses corpora to perform analysis on phonological phenomena, distributional patterns and variation (Cole, 2012).

A phonological corpus allows performing many types of analysis over the same data: for example, using a thoroughly marked corpus researchers can quickly collect possible realisations of phonemes and their environments or calculate how long a long vowel is compared to a short vowel, and whether this difference changes between stressed and unstressed syllables. If a corpus has additional data, such as an additional layer segmented by syllables, it can provide information about suprasegmental properties of the language: intonation, tone, speech tempo et cetera.

Despite the fact that Nanai is relatively well-researched, it remains endangered and is expected to become extinct in less than 50 years. Given the circumstances, it is important to preserve as much linguistic materials in Nanai as possible. Annotation and segmentation ease the usage of linguistic materials primarily because it is impossible to seek out specific items (sounds, parts of speech, intonation patterns) in raw data.

It is crucial for corpora to be publically available as open access to scientific research increases both readership and citations. It also enables scientists from low-income countries and researchers who are not affiliated with any specific organisation to conduct high-quality analysis, adding to the overall sum of

human knowledge (Wynne, 2005). Consequentially, a phonological corpus of Nanai would be useful for linguists interested in Tungusic languages and Nanai language teachers alike.

Methodology

The first step in corpus creation is acquiring some audio or video data. Often the recordings are made by the corpus linguists themselves since using the data collected by other researchers requires permissions.

The second step is transcription and segmentation, which is done either manually or via scripts. The most popular applications for speech analysis are Praat and ELAN. The researcher must be familiar with the sound system of the language; otherwise they will have no point of reference for assigning individual sounds to phonemes.

The third step is analysis of individual phenomena. I chose the nature of the alteration between /i/ and /ɪ/ to be my research question, intending to see if my data proves either theory about Nanai vowel harmony.

Ladefoged and Maddieson offer two ways of deducing [ATR]/[RTR] contrast:

- in some African languages [+ATR] is associated with lowered F3;
- in many African languages lowered F1 can be found in [+ATR] vowels.

On the other hand, Yun et al. (2016) use the relationship between F1 and F2 to show the differences in /i/ and /ɪ/ quality.

Collecting formants for each instance of /i/ and /ɪ/ can be done with the Praat script language. The instrument to conduct further analysis of two categorical independent variables (type of vowel) on one continuous dependent variable (formant frequency) is ANOVA or two-way analysis of variance. Its results assess the main effect of each independent variable or lack thereof. If the resulting p-value is smaller than 0.05 then the null hypothesis can be rejected, in other words, there is a statistically significant difference between the formant frequencies of /i/ and /ɪ/.

If one of the methodologies returns a p-value that is larger than 0.05, it means that the contrast between the two formants is negligible and can occur due to chance alone.

Data Collection

Linguist 310 is not intended to include fieldwork, thus I had to find Nanai recordings made by other researchers. Vasily Kharitonov offered his recording of a Nanai fairy tale “Mergen ningman” read by Raisa Alekseevna Beldy who spent her childhood in Dada village but acquired the prestige Najkhin dialect later. Kharitonov suggested this data as it was used in a cartoon made for a Russian government-sponsored project Gora Samotsvetov, which is publically available on YouTube (Beldy et al., 2017). It is worth noting that at the time of the recording, Raisa Alekseevna spent 4 years speaking mostly Russian.

The audio quality in the cartoon is not ideal: speech is sometimes distorted by background music.

I extracted the audio from the cartoon with ffmpeg (Ffmpeg) and converted it into .wave format to be used in Praat.

Transription

Transcription was made in Praat (Boersma & Weenink), a free and open-source cross-platform application developed by Paul Boersma and David Weenink of the University of Amsterdam.

Well-made corpora have to pertain to a certain set of standards, for example, the raw data must be accessible for further manipulation; the corpus must include documentation mentioning the instruments used in its making, annotation scheme, remarks about the quality of the annotation et cetera (Wynne, 2005). This information will be added to the corpus and provided in Appendix 4.

At first, I separated the audio stream into words using the text of the fairy-tale, which revealed several inconsistencies stemming from slips of the tongue. Kharitonov helped correcting these fragments.

After that, I created another tier "Phonemes" where words were split into phonemes, then copied the resulting intervals to the third tier "Sounds". On this tier I marked some allophonic variations, such as:

- /k/ and /g/ are realised as [q] [ɢ] before low back vowels /a/ and /o/
- /s/ is realised as [ɕ] before /i/ and /ɪ/
- /x/ is realised as [χ] before back vowels /ə, u, a, o/
- /n/ in the word-final position is realised as nasalisation on the preceding vowel.

The phonemic/word tiers do include vowel length but it was added after completion of the research project.

The boundaries of individual sounds were chosen according to the illustrations in Ladefoged & Ferrari Disner (2012). Segments that are distorted by background music are left blank.

One of the challenging aspects of sound segmentation was defining boundaries for word-initial plosives since their articulation starts with a period of silence. Another one was segmenting sequences of vowels since they can either belong to a diphthong or exist independently.

The resulting textgrid file will be uploaded to the open access repository Figshare (Figshare).

Analysis

Since /i/ and /ɪ/ are both represented by the letter 'и' in writing, it is not immediately evident if 'и' participates in vowel harmony, or if it is a 'neutral' vowel, akin to e/i vowels in Finnish that can occur in word with either series of harmonising vowels. Furthermore, Ko & Yurn claim that /ɪ/ weakens to /i/ after the last [+RTR] vowel (2011).

Ladefoged and Maddieson provide two ways of measuring the advanced tongue root variation: by comparing the relations between F3 and F4 and by comparing the relations between F2−F1 and F1 (1996). Also, Yun et al. (2016) measured the relationship between F1 and F2, so these measurements were also used.

I extracted the F1-F4 formant data for all vowels using the script from Appendix 1 that I wrote. Praat output was converted to JSON format using TextGrid package (Cesine).

After that, I processed the resulting values with C# code that I wrote for this task (Appendix 2). I picked a selection of measurements of /i/ or /ɪ/ that belong to the first syllable of the word since this is the only position where Ko & Yurn could find reliable contrast.

From this data, I created several charts in Excel and assessed them visually. Neither one of the resulting charts shows significant correlation within groups; the spread appears to be roughly equal.

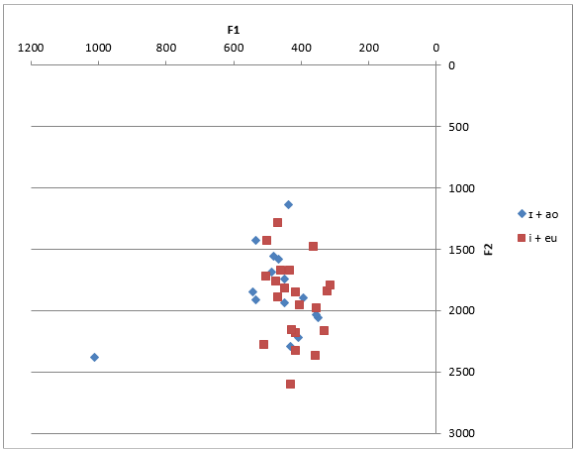

Figure 1. F1/F2 relationships for first /i/ and /ɪ/

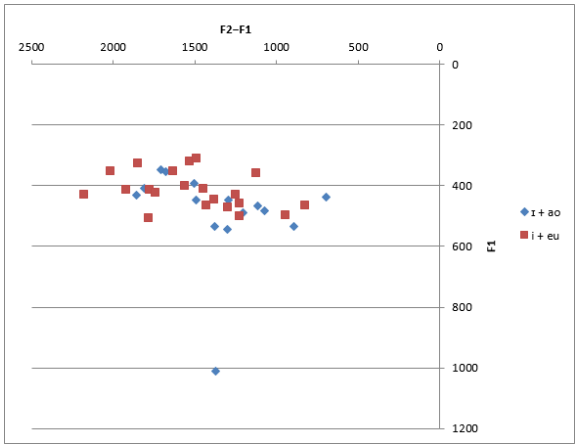

Figure 2. F2−F1/F2 relationships for first /i/ and /ɪ/

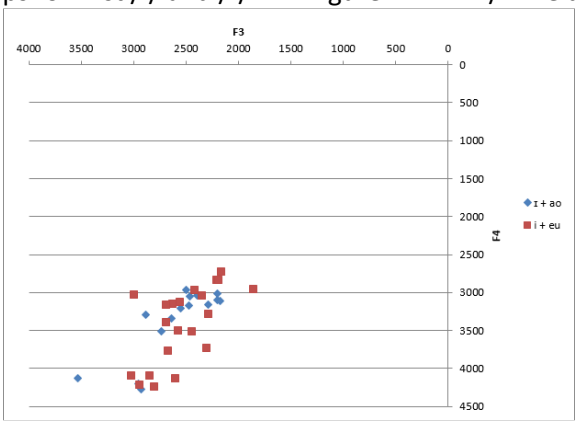

Figure 3. F3/F4 relationships for first /i/ and /ɪ/

The mean values of all formants are:

	F1	F2	F3	F4
ɪ	451	1808.573	2527.755	3316.799
i	418	1921.487	2529.841	3427.362

The set of measurements in Yun et al. (2016), provided for reference:

	F1	F2	F3	F4
ɪ	429	2251	3017	-
i	335	2352	3106	-

Statistically, the observed difference is not sufficient to prove that the formants for /i/ and /ɪ/ are truly different, i. e. the difference it is not due to chance alone. A two-factor ANOVA (analysis of variance) with replication is able prove or disprove the null hypothesis that there are no significant differences in mean formant values for /i/ and /ɪ/. I removed one outlier during the data preparation per ANOVA testing requirement.

The analysis does not allow us to reject the null hypothesis as all of the p-values are larger than 0.05:

- p-value(F1/F2) = 0.325569983
- p-value(F2−F1/F2) = 0.284086285
- p-value(F3/F4) = 0.99658067.

Thus, it is not possible to conclude that the observed formants for /i/ and /ɪ/ are truly different.

These findings directly contradict Yun et al. (2016) and most other researchers, but the underlying reason for this is not immediately clear. Most likely, this is the result of language attrition; other contributing factors are dialectal variation (no studies of Dada dialect exist; if the Dada vowel harmony is different, 'и' might be a neutral vowel there) or distortion from background music. Moreover, the data used by Yun et al. is not a relatively free speech flow but individual word utterances; the contrast might be weakened in different circumstances.

Conclusion

Language corpora have a multitude of practical and research applications including phonological analysis of phonological features and their variation. Making phonological corpora for endangered languages is even more important since there is a possibility that there would not be any native speakers left in the future. The corpus is supplied with documentation and an audio file with raw data.

Creating a phonological corpus involves data collection, manipulation and transcription. Prepared corpus can be used to perform all kinds of analysis. One example of such analysis is measuring the variation in formant frequencies F1-F4 of /i/ and /ɪ/ phonemes that I performed.

The results show that there is no statistically significant difference in median formant frequencies of these phonemes. It might be attributed to a number of factors, such as language attrition, suboptimal quality of data, the nature of the recording, and dialectal variation.

Appendices

Appendix 1

The Praat script used to extract vowel formants:

```
sound$ = selected$("Sound")
textGrid$ = selected$("TextGrid")

select TextGrid 'textGrid$'
numberOfPhonemes = Get number of intervals: 3
appendInfoLine: "Number of segments: ", numberOfPhonemes

select Sound 'sound$'
To Formant (burg)... 0 5 5000 0.025 50

output$ = "formants.csv"
writeFileLine: "'output$'", "time,phoneme,F1,F2,F3,F4"

for thisInterval from 1 to numberOfPhonemes
    #appendInfoLine: thisInterval

    select TextGrid 'textGrid$'
    phoneme$ = Get label of interval: 3, thisInterval
    #appendInfoLine: phoneme$

    phonemeStartTime = Get start point: 3, thisInterval
    phonemeEndTime   = Get end point:   3, thisInterval
    duration = phonemeEndTime - phonemeStartTime
    midpoint = phonemeStartTime + duration/2

    select Formant 'sound$'
    f1 = Get value at time... 1 midpoint Hertz Linear
    f2 = Get value at time... 2 midpoint Hertz Linear
    f3 = Get value at time... 3 midpoint Hertz Linear
    f4 = Get value at time... 4 midpoint Hertz Linear

    appendFileLine: "'output$'",
                      ...midpoint, ",",
                      ...phoneme$, ",",
                      ...f1, ",",
                      ...f2, ",",
                      ...f3, ",",
                      ...f4
endfor
appendInfoLine: newline$, newline$, "End"
```

Appendix 2

C# code used to extract i/ɪ formants:

```
using System;
using System.Collections.Generic;
using System.IO;
using System.Linq;
using System.Threading.Tasks;
using Newtonsoft.Json.Linq;

namespace PraatFormants
{
    class Program
    {
        static async Task Main(string[] args)
        {
            var allPhonemes = await
ReadFormantsAsync(@"formants.csv");

            var json = JObject.Parse(File.ReadAllText(@"data.json"));
            var words = json["items"][0]["intervals"];

            await WriteFormantsAsync(allPhonemes, words, "i", "data-
i.csv");
            await WriteFormantsAsync(allPhonemes, words, "ɪ", "data-
ɪ.csv");
        }

        private static async Task WriteFormantsAsync(FormantsRow[]
allPhonemes, JToken words, string targetPhoneme, string path)
        {
            using var sw = File.CreateText(path);

            foreach (var word in words)
            {
                if (word.Value<string>("text") == "") continue;

                var min = double.Parse(word.Value<string>("xmin"));
                var max = double.Parse(word.Value<string>("xmax"));

                var phonemes = allPhonemes
                    //.Where(p => p.Phoneme != "")
                    .Where(p => p.Timestamp >= min && p.Timestamp <
max)
                    .ToArray();

                var lastVowelIndex = phonemes.Select((ph, i) => new
{ ph.Phoneme, Index = i }).LastOrDefault(x =>
IsVowel(x.Phoneme))?.Index;
                if (lastVowelIndex == null)
```

```
                    continue;

                for (var i = 0; i < lastVowelIndex; i++)
                {
                    if (phonemes[i].Phoneme != targetPhoneme)
continue;

                    if (i > 0 && IsVowel(phonemes[i - 1].Phoneme))
continue;
                    if (IsVowel(phonemes[i + 1].Phoneme)) continue;

                    await
sw.WriteLineAsync(WritePhoneme(phonemes[i]));
                }
            }
        }

        static string WritePhoneme(FormantsRow row)
        {
            return string.Join(",", row.Formants.Select(f =>
double.IsNaN(f) ? "" : f.ToString()));
        }

        static readonly string[] Vowels = { "i", "ĩ", "ə", "əˆ", "u",
"ũ", "ɪ", "a", "ã", "o", "õ" };
        static bool IsVowel(string phoneme) => Vowels.Any(v =>
phoneme.StartsWith(v));

        static async Task<FormantsRow[]> ReadFormantsAsync(string
path)
        {
            using var sw = File.OpenText(path);
            await sw.ReadLineAsync();
            var list = new List<FormantsRow>();

            for (; ; )
            {
                var line = await sw.ReadLineAsync();
                if (line == null) break;

                var parts = line.Split(',');
                list.Add(new FormantsRow
                {
                    Timestamp = double.Parse(parts[0]),
                    Phoneme = parts[1],
                    Formants = parts.Skip(2).Select(x =>
double.TryParse(x, out var res) ? res : double.NaN).ToArray(),
                });
            }
```

```
            return list.ToArray();
        }
    }

    class FormantsRow
    {
        public double Timestamp { get; set; }
        public string Phoneme { get; set; }
        public double[] Formants { get; set; }

        public override string ToString() => Phoneme;
    }
}
```

Appendix 3

Words tier

Буэ балдипу Россияду, эй буэ тӯмпу, эй буэ килдэмпу. Россиядумэ боа ялодоани чӯ эгди мангбосал. Эй чу даи эрдэнгэ России мангбони - Амур. Нанид… хэсэдиэни Мангбо. Нёани Хабаровскай краеду хэейни, Мангбо киradоани нāнисал балдичи. Мангбола найсал хадёнгои, сиагои бāричи. Балана нāнисал лахама, корима, серома дёгду балдихачи. Тактова уйлэ гогда мо оялани ангой бӣчичи Туй тами бэюндэ, сингэрэдэ таоси мокчамари мутэси бӣчин А нāни найдоани тōкпон би ни Дё хадёни хэм хачин илгаку бӣчин. Тэй илгасал амбасалди этӯри бичин. Пиктэвэри кандёми ангой бӣчичи. Аминасални пиктэи эмуэкэндуэни этӯри сэвэкэсэлбэ лōричи. Тул-тул сонгой, энуси пиктэвэ сāвори, хадода сэвэкэсэлбэ лочивачи. Пиктэвэри улэсимэри, сонгаси осигоани нингмамба гусэрэйчи. Хай нингман осини, тэй нингман бōбой дёло Туй осихани кэтэ горо бӣэси.

Дуэнтэ киradоани эм нāни халани балдихачи. Амини, энини, гучи нӯчи хусэ пиктэчи. Гэ туй очини хони пурэн бэюнсэлду Мэргэн балдилохани. Нёандоачи дёбон эгди осихани. Пиктэвэ сиавамбори, омиамбори хупиуривэ гэлини. Манга маси сампар Мэргэн урэхэни. Эси мэнэ бэюнсэлбэ бэлэчими тэпчиухэни Мэргэн эрдэнгэ бāхани, мурчийни. Хаоси хэм энэхэчи? Хай дяка осихани? Гэ тэй бэюнсэл гусэрэхэчи мэргэнчи хони нгэлэпси гōгда амбан нёанчианчи хукчухэмбэни. Кама валиаха, туй би дяка. Аминаи гэлэндэгуми энэми ая. Мэргэн хай гойдами баргичигоани хадён-да нёани анā. Туй бӣди энэхэни Гойдами энэмиэ, энэмиэ тэингуй мурчихэни ядахани. Тэни сиами тэпчухэни хай дяка? Нгэлэпси амбāн дяка, боко няронду бӣ дяка. Паталан сиасисини, морайни, гудем гэлини. «Хориосу» морайни, «бэлэчиусу!» Хай макикаси бала хориро! Ичэру, элэ-элэ будемби! Кэсиэ бāха паталан, ундини: ми синчи баняламби наондёан най. Ми синди эдили-дэ аясии симби улэсилухэмби. Хай асигой гэлэдемби, ми корпиасимби! Тотара Мэргэн гусэрэхэни аминаи, гōгда амбāн хони-да эрдэлэхэмбэни. Тэй гōгда амбамба ми сāрии, нёани мэпи гэрбиэсини дуэнтэ эдени. Нёандиани соридой минчи ичэхэри Ми симбивэ бэлэчидемби.

Таванкидиа Мэргэн энэлупсинкини сапси кирачиани исихани. Ичейни, огда чаду бӣ. Улэн огда, боя-да анā. Эй хай дяка? Ата-та гучэ ихэни адоличи! Тэнг дāи! Пиктэкудэ! Хэсэкудэ! Хаоси сӣ най энэйси? Мэргэн гӯчэнчи гусэрэхэни: «Гōгда амбāмба гэлэмэчии» Нёани мэпи дуэнтэ эдениэм хэсини. Мӣ сāрамби тэй «эдембэ», Амимбаси мангбо дōлани, пэгилэ дюлденди маси уихэни Хэм согдатава нёандоани таонгоани. Энимбэси нā дōлани тэвухэни. Эгди айнаня нёанчи, адолива ангогоани. нгэгден боава эчиэ ичейчи Сӣ мимбивэ бэлэчихэндуэси мӣ поктова ичуэндэмби. Гэ туй нёанчи энэми тэпчиухэчи, гӯчэн дюлэси муэвэ энэй. Тэй хамиалани Мэргэн огдади гиолими энэй. Каодярару исихапу. Эси ичэндугуивэ, хони бӣни чаду амиси. Гэ чаду бӣ. Муэ дочани ими ая. Эчиэ такоани пиктэи. Хаду айнганяду сихэни. Гэ улэн, эси тэни туй бӣ дяка. Гэ «дуэнтэ эдени» бӣни боачи исихачи. Аорини Гōгда амбāн, эчиэ сэнэни. Тэй чаду онголоду энимбэси Гōгда амбāн дяпачини. Гэ, эй мӣ синчи! сидямби! Садячи, хонида найди хополамборива Эй гаса! Дэгдэйни Дяпу, морайни, эди хаморира! Гэ эси чаду хэмтуни агдахачи Хоня улэн Гōгда амбāмба хэтэхэчи. Гōгда амбāмба тагохачи, писачихачи, нянгā нюлэхэчи, диливани-рагда хамаси эчиэ нэкуэчи. Туй дуэнтэду хайду-да биэ тэни. Гэ хоня улэн вездеход осихани, чади дёкчи дидюхэчи. Эси дэм балдичи улэн, ая кэсику.

Phonemes tier

bu rəsijədu əi uə tu:mpu əi buə kildəmpu rəsijədum boa jalodoanɪ cu: əgdi maŋbosal əi cu da:ɪ ərdəŋgə rəsi maŋboɲɪ amur nanɪd xəsəɬiəni maŋbo ɲoanɪ xabar sk kraɪdu xəjəni maŋbo kɪradoanɪ na:nɪsal balɬɪcɪ maŋbola najsal xaɬongoɪ sɪagoɪ ba:rgɪcɪ balana na:nɪsal laxama korɪma sɪroma ɬogdu balɬɪxacɪ taktowa ujlə gogda mo ojalanɪ aŋgoɪ bi:cici tujtamɪ bəjundə siŋgərədə taosɪ mokcamarɪ mutəsi: bi:cin a na:nɪ naɪdoantanɪ to:kpon bi:ni ɬoxaɬonɪ xəm xacɪn ɪlgaku bi:cin təj ɪlgasal ambasalɬɪ ətu:ri bicin piktəwəri kaŋɬomɪ aŋgoj bi:cici amɪnasalnɪ piktəi əmuəkənduəni ətu:ri səwəsəlbə lo:rɪcɪ tultul soŋgoj ənusi piktəwə sa:worɪ xaloda səwəkəsəlbə locɪwaci piktəwəri ulə:siməri soŋgasɪ osɪgoanɪ nɪŋmamba gusərici xaj nɪŋman osɪnɪ təj nɪŋman bo:boj ɬolo tuj osɪxanɪ kə:tə goro bi:ə:si

duəntə kɪradoanɪ əm na:nɪ xalanɪ balɬɪxacɪ amɪnɪ ənini guci ŋuci xusə piktəci gə tuj ocɪnɪ xonɪ purən bəjunsəldu mərgən balɬɪloxanɪ ɲoandoacɪ ɬobon əgɬi osixanɪ piktəwə sɪawamborɪ omɪmborɪ xupiuwə gəlini maŋga masɪ sampa mərgən urəxəni əsi mənə bəjunsəlbə bələcimi təpciuxəni mərgən ərdəŋgə ba:xanɪ murcini xaosɪ xəm ənəxəci xaj ɬaka osɪxanɪ gə təj bəjunsəl gusərəxəci mərgəɲci xonɪ ŋələpsi go:gda amban ɲoaɲcɪancɪ xukcuxəmbəni kama walɪaxa tuj bi ɬaka amɪnaɪ gəlndəgumi ənəmi aja mərgən xaj gojdamɪ bargcɪorɪ xaɬon da ɲoanɪ ana: tuj bi:ɬi ənəxəni gojdamɪ ənəmiə ənəmiə təinguj murcixəɲi jadaxanɪ təni sɪamɪ təpcuxəni xaj ɬaka ŋələpsi amba:n ɬaka boko ɲarondu bi: ɬaka patalan sɪasɪnɪ morajnɪ guɬəm gəlini xorɪosu moraɪnɪ bələciusu xaj makɪkasɪ bala xorɪro icəru ələələ budəmbi kəsiə ba:xa patalan unɬini mi siɲci baɲalambɪ naonɬoan naj mi sinɬi əɬiləi də ajasɪɪ simbi uləsiluxəmbi xaj asɪgoj gələɬəmbi mi korpɪasɪmbɪ totara mərgən gusərəxəni amɪnaɪ go:gda amba:n xonɪ da ərdələxəmbəni təj go:gda ambamba mi sa:rɪjɪ ɲoanɪ məpi gərbiəsini duəntə əɬəni ɲoanɬaɲɪ sorɪdoj minci icəxəri mi simbiwə bəlcɬəmbi

tawankɪɬɪa mərgən ənəlupsinkini sapsɪ kɪracianɪ ɪsɪxanɪ icəjni ogda cadu bi: ulən ogda boja da ana: əj xaj ɬaka ata ta gucə ixəni adolɪcɪ təŋ da:i piktəkudə xəsəkudə xaosɪ si: naj ənəjsi mərgən gu:cə:ɲci gusərəxəni go:gda amba:mba gələməciji ɲoanɪ məpi duəntə əɬəniəm xəsini mi: sa:rambɪ təj əɬəmbə amɪmbasɪ maŋbo do:lanɪ pəgilə ɬulɬənɬi masɪ uixəni xəm sogdatawa ɲoandoanɪ taoŋgoanɪ ənimbəsi na: do:lanɪ təwuxəni əgɬi ajŋaɲa ɲoancɪ adolɪwa aŋgogwanɪ ŋəgɬiən boawa əciə icəci si: mimbiwə bələcixənduəsi mi: poktowa icuənɬəmbi gə tuj ɲoancɪ ənəmiə təpciuxəci gu:cə:n ɬuləsi muəwə ənəj təj xamɪalanɪ mərgən ogdaɬɪ gɪolɪmɪ ənəj kaoɬararu ɪsɪxapu əsi icənduguiwə xonɪ bi:ni cadu amɪsɪ gə cadu bi: muə docanɪ imi aja əcə takoanɪ piktəi xadu ajŋaɲadu sixəni gə ulə:n əsi təni tuj bi: ɬaka gə duəntə əɬəni bini boacɪ ɪsɪxacɪ aorɪnɪ go:gda amba:n əciə sənəni təj cadu oŋgolodu ənimbəsi go:gda amba:n ɬapacɪnɪ gə əj mi: sinci sɪɬamɪ saɬacɪ xonɪda najɬɪ xopolamborɪwa əj gasa dəgdəjni ɬapu morajnɪ əɬi xamorɪra gə əsi cadu xəmtuni agdaxacɪ xoɲa ulən go:gda amba:mba xətəxəci go:gda amba:mba tagoxacɪ pɪsacɪxacɪ ɲaŋga: ɲuləxəci ɬɪlɪwanɪ ragda xamasɪ əciə nəkuci tuj duəntədu xajdu da biə təɲi gə xoɲa ulə:n osɪxaɲɪ cadɪ ɬokcɪ ɬiɬuxəci əsi dəm balɬɪcɪ ulə:n aja kəsiku

Sounds tier

rəɕijədu əi uə tumpu əi buə kildəmpu rəɕijədum bwa jalodwaɲɪ cu əgdi mãbosal əi cu daɪ ərdə̃ŋgə rəɕi maŋboɲɪ amur naɲɪd... xəsəɬiəɲi maŋbo ɲoaɲɪ χabarsq qraɪdu χəjəɲi mãbo kɪradwaɲɪ naɲɪsal balɬɪcɪ mãbola najsal xaɬongoɪ ɕɪagoɪ bargɪcɪ balana naɲɪsal laxama qorɪma ɕɪroma ɬogdu balɬɪxacɪ taqtoʊa ujlə gogda mo ojalaɲɪ aŋgoɪ bicici tuj tamɪ bəjundə ɕiŋgərədə taoɕɪ moqcamarɪ mutəɕi bicĩ a naɲɪ naɪdwantaɲɪ toqpõ biɲi ɟo χaɬoɲɪ χəm χacĩ ɪlɢaku bicĩ təj ɪlgasal ambasalɬɪ əturi bicin piktəʊəri qaɲɬomɪ aŋgoj bicici amɪnasalɲɪ piktəi əmuəkənduəɲi əturi səʊəsəlbə lorɪcɪ tultul sõŋɢoj ənuɕi piktəʊə saworɪ xaloda səʊəkəsəlbə locɪwaci piktəʊəri uləɕiməri sõŋgwaɕɪ oɕɪgwaɲɪ ɲɪŋmamba gusərici ɲɪŋmã oɕɪɲ təj a o ɟolo tuj oɕɪχaɲɪ kətə ɢoro biəɕi

duəntə kɪradwaɲɪ əm naɲɪ xalaɲɪ balɬɪxacɪ amɪɲɪ əniɲi guci ŋuci xusə piktəci gə tuj ocɪɲɪ xoɲɪ purən bəjũsəldu mərgən balɬxaloxaɲɪ ɲoandwacɪ ɟobon əgɬi oɕχaɲɪ piktəʊə ɕɪaʊamborɪ omɪmborɪ xupiuʊə gə maŋga maɕɪ sampa mərgən urəxəɲi əɕi mənə bəjunsəlbə bələcimi təpciuχəɲi mərgən ərdəŋgə baχaɲɪ murciɲi xaoɕɪ χəm ənəxəci χaj ɟaqa oɕχaɲɪ gə təj bəjũsəl gusərəxəci mərgəɲci xoɲɪ ŋələpɕi gogda ambã ɲoaɲcaɲcɪ xuqcxəmbəɲi qama ʊalɪaχa tuj bi ɟaqa amɪnaɪ ɢəlndəgumi ənəmi aja mərgən χaj gojdamɪ bargcɪorɪ xaɬon da ɲoaɲɪ ana tuj biɟi ənəxəɲi gojdamɪ ənəmiə ənəmiə təĩguj murciχəɲi jadaxaɲɪ təɲi ɕɪamɪ təpcuxəɲi xaj ɟaqa ŋələpɕi amban ɟaqa boqo ɲarondu bi ɟaqa patalan ɕɪaɕɪɲɪ morajɲɪ guɟəm gəliɲi χorɪosu moraɪɲɪ bələciusu xaj maqqaɕɪ bala χorɪro icəru ələ ələ budəmbi kəɕiə baχa patalan uɲɬiɲi mi ɕiɲci baɲalambɪ naoɲɟoan naj mi ɕiɲɟi əɟiləi də ajaɕɪɪ ɕimbi uləɕiluxəmbi xaj aɕɪgoj ɢələɟəmbi mi qorpɪaɕɪmbɪ totara mərgən gusərəxəɲi amɪnaɪ ɢogda ambam xoɲɪ da ərdələxəmbəɲi təj ɢogda ambamba mi sarɪɪ ɲoaɲɪ məpi gərbiəɕiɲi duəntə əɟəɲi ɲoaɲɟaɲɪ sorɪdoj miɲci icəχəri mi ɕimbiʊə bəlcɟəmbi

tawankɪɟɪa mərgən ənəlupɕiɲkiɲi sapɕɪ kɪraciaɲɪ ɪɕχaɲɪ icəjɲi ogda cadu bi ulə:n ogda boja da ana əj χaj ɟaqa ata ta gucən ixəɲi adolɪcɪ tən dajɪ pi t udə χəsəku xaoɕɪ ɕi naj ənəjɕi mərgən gucəɲci gusərəxəɲi ɢogda ambamba gələməciji ɲoaɲɪ məpi duəntə əɟəɲiəm xəɕiɲi mi sarambɪ tə əɟəmbə amɪmbaɕɪ mãŋbo dolaɲɪ pəgilə ɟulɟəɲɟi maɕɪ uixəɲi χəm sogdataʊa ɲoandwaɲɪ taoŋgwaɲɪ əɲimbəɕi nadolaɲɪ təwuxəɲi əgɬi jaŋaɲa ɲoaɲcɪ adolɪʊa aŋɢogwaɲɪ ŋəgɬiən boawa əciə icəci ɕi mimbiʊə bələcxənduəɕi mi pqktoʊa icuəɲɟəmbi gə tuj ɲoaɲcɪ ənəmiə təpcuχəci gucə ɟuləɕi muəwə ənəj təj χamɪalaɲɪ mərgən ogdaɟɪ gɪolɪmɪ ənəj qaoɟararu ɪɕɪχapu əɕi icəndguiwə χoɲɪ biɲi cadu amɪɕɪ gə cadu bi muə docaɲɪ imi aja əcə takoaɲɪ piktəi χadu ajŋaɲadu ɕixəɲi gə ulən əɕi təɲi tuj bi ɟaqa gə duəntə əɟəɲi biɲi bwacɪ ɪɕɪχacɪ aorɪɲɪ ɢogda amban əciə ənəɲi təj cadu oŋɢolodu əɲimbəɕi gogda amban ɟapcɪɲɪ gə əj mi ɕĩci ɕɪɟamɪ saɟacɪ xoɲɪda najɟɪ χopolamborɪʊa əj gasa dəgdəjɲi ɟapu morajɲɪ əɟi xamorɪra gə əɕi cadu χəmtuɲi agdaxacɪ hoɲa ulən ɢogda ambamba xətəχəci gogda a amba ta χacɪ pɪsacχacɪ aŋga ɲuləxəci ɟɪlɪuaɲɪ ragda xamaɕɪ əciə nəkuci tuj duəntədu χajdu da biə təɲi gə χoɲa ulən ʊ χ ɕχa cadɪ ɟoqcɪ ɟiɟuxəci əɕi dəm balɟɪcɪ ulən aja kəɕiku

Appendix 4

Corpus documentation

People involved

This corpus was created by Aidan Winberry in 2020 from a recording of Raisa Alekseevna Beldy reading the text of a Nanai fairy-tale ”Mergen ningman”. The recording was made by Vasily Kharitonov.

Annotation scheme

The information about Nanai phonemes is taken from (Ko & Yurn, 2011).

Coding scheme

The sounds and phonemes are represented by their IPA symbols in Unicode.

The text of the fairy-tale is provided in Cyrilic Nanai orthography. Nanai writing system is nearly-phonemic so a Latin transcription layer would simply copy the phonemic tier in IPA.

In several segments the sound is corrupted by background music; in this case the annotation on phonemic and phonetic level is omitted.

Annotation quality

Annotations were made without consulting with the dictionaries. All phonemes and allophonic variants are marked aurally.

Diphthongs are not thoroughly marked.

The “Words” tier follows the text of the fairy-tale while the “Phonemes” and “Sounds” tiers represent what is actually being said instead. Several utterances end with an ellipsis which marks correcting slips of tongue.

Bibliography

1. Beldy, Raisa Alekseevna et al., *Mergen ningman*. 2017 https://www.youtube.com/watch?v=HCJqav2LFhE
2. Boersma, Paul and Weenink, David. *Praat*. http://www.praat.org/
3. Campbell, Lyle et al. *The Catalogue of Endangered Languages (ElCat)*. Endangered Languages Project, 2017, http://endangeredlanguages.com/userquery/download/
4. Cesine. Textgrid package. *Npm.js*. https://www.npmjs.com/package/textgrid
5. Cole, Jennifer et al. "Corpora, Databases, and Internet Resources: Corpus Phonology with Speech Resources Using The Internet For Collecting Phonological Data Speech Manipulation, Synthesis, and Automatic Recognition in Laboratory Phonology Phonotactic Patterns in Lexical Corpora". *The Oxford Handbook of Laboratory Phonology*, edited by Cohn, Abigail C., et al. Oxford University Press, 2012.
6. Durand, Jacques, et al., editors. *The Oxford Handbook of Corpus Phonology*. 1st ed, Oxford University Press, 2014.
7. Ffmpeg 4.2, https://ffmpeg.org/
8. Figshare University of Auckland. https://auckland.figshare.com/
9. Gries, Stefan Thomas. *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge, 2009.
10. Janhunen, Juha. 'Tungusic: An Endangered Language Family in Northeast Asia'. International Journal of the Sociology of Language, vol. 2005, no. 173, Jan. 2005. DOI.org (Crossref), doi:10.1515/ijsl.2005.2005.173.37.
11. Kharitonov, Vasily. 'Some thoughts on the revitalization of Nanai language'. CAES, vol. 3, no. 1, March 2017
12. Kharitonov, Vasily. *Xisangoru* (Хисангору). http://xisango.ru
13. Kile et al., *Nanajskij folklor. Ningman, Siokhor, Telungu* (Нанайский фольклор. Нингман, Сиохор, Тэлунгу). Nauka, [1996] 2018, https://elibrary.ngonb.ru/catalog/524/15116/
14. Ko, Dongho, and Gyudong Yurn. *A Description of Najkhin Nanai*. Seoul National University Press, 2011.
15. Ladefoged, Peter, and Maddieson, Ian. *The Sounds of the World's Languages*. Blackwell Publishers, 1996.
16. Ladefoged, Peter, and Ferrari Disner, Sandra. *Vowels and consonants*. 3rd ed, Wiley-Blackwell, 2012.
17. Larin, V. L., et al., editors. *Istorija I kultura nanajtsev* (История и культура нанайцев). Nauka, 2003
18. Moseley, Christopher (ed.). *Atlas of the World's Languages in Danger*, 3rd ed. UNESCO Publishing, 2010, http://www.unesco.org/culture/en/endangeredlanguages/atlas
19. Sem, Lidija Ivanovna. 'Nanajskij jazyk (Нанайский язык) '. *Entsiklopedija. Yazyki mira*. Russian Academy of Sciences, 1997, pp. 173–188.
20. Sulyandziga R. V. et al. *Korennyje malochislennyje narody Severa, Sibiri i Dalnego Vostoka Rossijskoj Federatsii. Obzor sovremennogo polozheniya* (Коренные малочисленные народы Севера, Сибири и Дальнего Востока Российской Федерации. Обзор современного положения). Moscow, 2003, http://www.raipon.info/peoples/nanai/nanai.php
21. Whaley, Lindsay J., et al. 'Revisiting Tungusic Classification from the Bottom up: A Comparison of Ewenki and Oroqen'. Language, vol. 75, no. 2, June 1999, p. 286. DOI.org (Crossref), doi:10.2307/417262.

22. Whaley, Lindsay. ‘Deriving Insights about Tungusic Classification from Derivational Morphology’. Johanson, Lars, and Martine Irma Robbeets, editors. *Copies versus Cognates in Bound Morphology*. Brill, 2012.
23. Wynne, Martin (ed.). Developing Linguistic Corpora: a Guide to Good Practice. AHDS Guides to Good Practice, 2005, http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm
24. Yun Jiwon, et al. ‘A Phonetic Study of Nanai Vowels: Using Automated Post-Transcriptional Processing Techniques’. ALTAI HAKPO, no. 26, June 2016, pp. 29–44. DOI.org (Crossref), doi:10.15816/ask.2016..26.003.
25. Zhang, Paiyu, *The Kilen language of Manchuria. Grammar of a Moribund Tungusic Language*. University of Hong Kong, 2013, http://hdl.handle.net/10722/181880