A tundrai nyenyec (egynyelvű) korpusz munkálatai: kihívások, módszerek, eredmények[1]

Mus Nikolett, Metzger Réka

Nyelvtudományi Kutatóközpont

This paper reviews work on the Tundra Nenets corpus currently under way in the Hungarian Research Centre for Linguistics. Its most important aim is to provide methodological underpinnings for future initiatives of creating online corpora from the data of languages whose sociolinguistic background and level of digital elaboration are similar to those of Tundra Nenets. This paper reviews the stages of our work: the selection of texts for the corpus, their digital processing (web scraping, OCR, transcription), the uniformization of texts, and the creation of corpus files. We summarise our results so far, and provide guidance concerning possibilities of search in the corpus. We show at what levels the data are currently annotated, and outline our plans for the future.

Keywords: Tundra Nenets corpus, digital processing of written texts, uniformization of characters, the NoSketchEngine corpus management system, annotation

Kulcsszavak: tundrai nyenyec korpusz, írott szöveg digitális feldolgozása, karakterek egységesítése, NoSketchEngine korpuszkezelő, annotálás

1. Bevezetés

Tanulmányunk a „Nyelvjárási variáció és nyelvi változás elméleti és kísérletes megközelítése a tundrai nyenyec nyelvben” elnevezésű (NKFIH FK_129235 számú) kutatási projekt keretében zajló tundrai nyenyec (szamojéd, uráli) nyelvi korpusz munkálatának módszertanát foglalja össze és lépéseit részletezi.[2] Legfontosabb célunk az, hogy az alkalmazott módszerek részletes bemutatásával a tundrai nyenyec nyelvhez hasonló dokumentáltságú és digitális feldolgozottságú nyelvek jövőbeni korpuszépítési munkáit támogassuk, és bemutassuk az eddig elért eredményeinket.

[1] A tanulmány szerzői köszönetüket fejezik ki bírálóiknak a tanulmány korábbi változatához fűzött értékes észrevételeikért, továbbá a folyóirat szerkesztőinek a tanulmány összeállításában nyújtott segítségükért, tanácsaikért és megjegyzéseikért.

[2] A projekt céljai és eredményei ezen a linken találhatók: http://www.nytud.hu/oszt/elmnyelv/thea/index.html, a tundrai nyenyec korpusz pedig a következő linken érhető el: https://tundranenetsdata.nytud.hu/bonito.

DOI: 10.15776/NyK.2021.117.5

A tundrai nyenyec nyelvi korpuszunk létrehozásának oka az volt, hogy szerettük volna a mostanáig kizárólag nyomtatásban elérhető, valamint a saját gyűjtésű, még ki nem adott tundrai nyenyec nyelvű szövegeket archiválni, digitálisan feldolgozni, és a tudományos közösség számára elérhetővé tenni. Fontos szempont volt munkánk során, hogy egy kereshető szöveggyűjteményt alakítsunk ki, mivel a jelenleg digitálisan elérhető tundrai nyenyec anyagok elsősorban arra szolgálnak, hogy illusztrálják a nyelvet (ezekről bővebben lásd a **2**. részt), de nem lehet bennük adatokat keresni. Munkánkkal így ezt a hiányt szeretnénk betölteni. Korpuszunkkal ösztönözni szeretnénk a tundrai nyenyec nyelv kutatását, leírását, nyelvészeti megközelítésű vizsgálatát célzó munkákat.

Tanulmányunk a következőképpen épül fel. A **2.** részben bemutatjuk a tundrai nyenyec nyelvet és az elérhető digitális anyagokat, a nyelvre fejlesztett nyelvtechnológiai eszközöket. Ezt követően ismertetjük a korpuszépítési munkánk lépéseit a **3.** részben. A **4.** rész az online korpuszt mutatja be. A jelenleg folyó munkákat és jövőbeli terveinket az **5.** részben foglaljuk össze, majd a **6.** részben összegezzük az eredményeinket és a tanulmány lényeges mondanivalóit.

2. A tundrai nyenyec nyelv bemutatása

A tundrai nyenyec nyelv egyike az Oroszországi Föderáció területén beszélt számos őshonos kisebbségi nyelvnek. A nyelvet Északkelet-Európában és Északnyugat-Szibériában beszélik a Nyenyec Autonóm Körzetben (Arhangelszki terület), a Jamal-Nyenyec Autonóm Körzetben, valamint a Tajmiri Dolgan-Nyenyec járásban (Krasznojarszki határterület). Ezen felül tundrai nyenyec beszélők további csoportjai találhatók a Hanti-Manysi Autonóm Körzetben (Tyumenyi terület), a Komi Köztársaság területén és a Murmanszki területen. Az alábbi térkép ábrázolja ezeket a területeket.[3]

[3] A térképet Gulyás Zoltán készítette, és letölthető az alábbi oldalról: http://www.nytud.hu/oszt/elmnyelv/thea/tools.html#maps

1. térkép: A tundrai nyenyec beszélők élőhelye

A tundrai nyenyec nyelv veszélyeztetett, EGIDS (Expanded Graded Intergenerational Disruption Scale) értéke 6b. Ez a számérték azt mutatja, hogy (i) a tundrai nyenyec beszélt nyelvi változatát a mindennapi kommunikáció során még minden generáció használja, (ii) a nyelvátadás a generációk között még (többnyire) folytatólagos, de (iii) a beszélők száma évről évre csökken (Trevilla 2009).

A 2010-es oroszországi népszámlálás adatai alapján 21 926 nyenyec beszélő van. Ez megközelítőleg a fele azoknak, akik nyenyec etnikumúnak vallották magukat (ezek pontos száma 43 777). A népszámlálás során nem tettek különbséget a tundrai és az erdei nyenyec nyelvek között. Utóbbi a tundrai nyenyec legközelebbi rokona, amelyet sokáig a nyenyec egyik nyelvjárásának tartottak. Az újabb irodalom viszont a köztük megfigyelhető különbségekre alapozva két külön nyelvként kezeli az erdei és a tundrai nyenyecet (vö. Sammallahti 1974; Popova 1978; Salminen 1998; Koskarjova 2005; Nikolaeva 2014). Az erdei nyenyecek számát nagyjából 2 000 főre becsülik (Toulouze 2003; Koskarjova 2005; Vol-

zhanina 2007). Következésképpen megközelítőleg 20 000 tundrai nyenyec beszélőt találunk még.

A tundrai nyenyecnek három nyelvjárási csoportja van, ezek további nyelvjárási egységekre tagolódnak (vö. Hajdú 1968; Tyerescsenko 1993; Salminen 1998; lásd 1. táblázat).

1. táblázat: A tundrai nyenyec nyelvjárások

Nyelvjárási csoport	**Nyelvjárási egység**
Nyugati csoport	kolgujevi
	kanini
	timani
	malaja-zemljai
Középső csoport	bolsaja-zemljai
Keleti csoport	obi/uráli
	jamali
	tazi
	nadimi
	tajmiri

A nyelvjárások közötti különbségekről alig áll rendelkezésünkre adat. A szakirodalomban főként fonológiai, morfológiai és szókészlettani különbségekről olvashatunk (Salminen 1998; Ackerman – Salminen 2006). Szisztematikus összehasonlító mondattani kutatások ebben a témában tudomásunk szerint még nem történtek.[4]

Az Oroszországi Föderáció hivatalos nyelve az orosz, amely többek között az oktatás nyelve is. Az orosz mellett más (őshonos) kisebbségi nyelvek beszélői is élnek a tundrai nyenyecek lakta területeken. A gazdag etnikai és nyelvi diverzitás következtében napjainkban alig találunk egynyelvű tundrai nyenyec beszélőt (Pakendorf 2010; Kasten – de Graaf 2013; Vakhtin 2015). A hagyományos tundrai nyenyec életmód a vándorló réntartáshoz kapcsolódik. Az utóbbi évtizedekben egyre több és több közösség kényszerült letelepedni, és hagyományos életmódját meg-

[4] Kutatásunk egyik célja éppen az, hogy kérdéstípusok mondattani és prozódiai különbségeit vizsgálja két nyelvjárási egységben, a jamali és a tajmiri nyelvjárásban.

változtatni. Ez a folyamat negatív hatást gyakorolt a nyelvhasználatra és a generációk közötti nyelvátadásra (Dudeck 2013; Laptander 2013; Liarskaya 2013). A különböző etnikumok közötti vegyes házasságok is negatív módon befolyásolják a nyelvhasználatot (Vagramenko 2017). Az utóbbi években viszont számos olyan kezdeményezést figyelhetünk meg a térségben, amely a nyelvvesztési folyamatot lassítja (pl. a tundrai iskolákról lásd Laptander 2013).

Annak ellenére, hogy már a 17–18. századból találunk nyenyec nyelvű gyűjtéseket (lásd pl. Daniel Gottlieb Messerschmidt és Philip Johann Strahlenberg adatait), a nyenyec írásbeliség viszonylag fiatalnak számít. Egy egységes nyenyec írásbeliség létrehozásának ötlete az 1920-as – 1930-as években fogalmazódott meg először (Touluze 1999: 53). Az írott sztenderd és egységes irodalmi nyelv alapjául a tundrai nyenyec jamali nyelvjárását választották, amely még ma is a legelterjedtebb írott változat. Újságokat (pl. *Няръяна вындер* 'Vörös tundra'), online forrásokat (pl. wikipedia szócikkek), tv- és rádiófelvételek lejegyzéseit találjuk meg a jamali nyelvjárásban.[5] Írott sztenderdet és egységes írásrendszert viszont a mai napig nem sikerült kialakítani. Ezért az elérhető írott források között jelentős különbségek vannak, s ez befolyásolja a nyelv digitális feldolgozását, a korpuszépítési munkát.

A tundrai nyenyec írásbeliség a cirill ábécén alapul, hasonlóan más nyugat-szibériai nyelvekéhez. A cirill ábécé további karakterekkel egészült ki a tundrai nyenyecben, ezek a veláris nazális /ŋ/ jelölésére használatos <ӈ>, valamint a gégezárhang /ʔ/ jelölésére szolgáló <”> és <’>. A cirill ábécé nem teljesen alkalmas a nyelv fonémarendszerének reprezentálására. Többek között például a magánhangzók hosszúságát sok forrásban egyáltalán nem jelölik, valamint bizonyos esetekben ugyanaz a karakter jelöl különböző fonémákat, pl a cirill <э> használatos a tundrai nyenyec /e/ és /æ/ fonémák lejegyzésére. Egységes és következetes latin átírás sem létezik a tundrai nyenyec nyelvre. Az elérhető átírások közötti különbségek elsősorban a magánhangzórendszer eltérő értelmezéséből fakadnak (Hajdú 1968; Salminen 1993, 1998; Staroverov 2006; Kavitskaya – Staroverov 2008).

Vannak online elérhető tundrai nyenyec nyelvű szöveggyűjtemények és adatbázisok. Itt a teljesség igénye nélkül említhetjük meg például az „Endangered Languages and Cultures of Siberia” c. projekt oldalán talál-

[5] Ezek a források az alábbi linken érhetők el: http://nvinder.ru/rubric/yalumd, https://incubator.wikimedia.org/wiki/Special:PrefixIndex/Wp/yrk/, http://yamal-region.tv

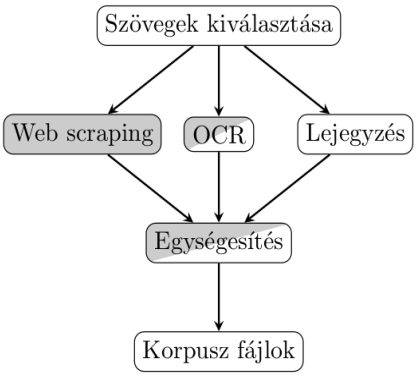
ható audiofelvételeket és azok morfológiailag annotált, oroszra fordított lejegyzéseit.[6] Itt a tundrai nyenyec anyagokon kívül találunk even, erdei enyec, kolimai jukagir, északi hanti, tundrai jukagir és udihe nyelvi gyűjteséket is. A veszélyeztetett nyelvek online archívuma (ELAR) is tartalmaz tundrai nyenyec anyagokat, amelyek csak regisztrált felhasználók számára hozzáférhetőek.[7] Az obi-ugor és szamojéd nyelvek tagadó szerkezeteit vizsgáló tipológiai projekt (Typology of Negation in Ob-Ugric and Samoyedic Languages / NOS) oldalán is találunk annotált szövegmutatványokat és példamondatokat.[8] Emellett „Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben" elnevezésű projekt adatbázisa is tartalmaz tundrai nyenyec anyagokat.[9] A felsorolt adatbázisok elsősorban abból a célból készültek, hogy megmutassák a nyelv általános tipológiai jellemzőit, vagy illusztráljanak bizonyos nyelvi jelenségeket. Az adatbázisokban keresésre csak manuálisan van lehetőség.

Ami a nyelv digitális támogatottságát illeti, szó- és karakter n-gram gyakorisági listákat találunk IPA átírásban az An Crúbadán projekt honlapján.[10] Emellett szövegelemző, paradigma- és szógenerátor, illetve digitális szótár érhető el a Giellatekno oldalán.[11]

3. A korpuszépítés folyamata és lépései

Az alábbi ábra mutatja a tundrai nyenyec nyelvi korpusz munkálatai során kidolgozott és követett lépéseket. Az ábrán fehér színnel jelöltük azokat a munkaszakaszokat, amelyeket manuálisan végeztünk, míg szürke szín jelöli az automatizáltakat. Két olyan munkaszakasz volt, amely egyaránt kívánt automatizált és manuális munkát (OCR és egységesítés).

6 http://www.siberianlanguages.surrey.ac.uk/about/
7 https://elar.soas.ac.uk/Collection/MPI120925
8 https://www.univie.ac.at/negation/sprachen/nenetsa.html
9 http://www.nytud.hu/oszt/elmnyelv/urali/adatbazisok_tundrainyenyec.html
10 http://crubadan.org/languages/yrk-x-tundra-acad
11 http://giellatekno.uit.no/cgi/index.yrk.eng.html



1. ábra: A tundrai nyenyec nyelvi korpusz munkálatainak lépései

3.1. A korpusz szövegeinek kiválasztása

Korpuszunkban két forrástípusból származnak szövegek. Az egyik forrást a már publikált – nyomtatásban megjelent vagy online elérhető – szövegek alkotják. A másik csoportot pedig a projektünk keretében általunk gyűjtött anyagok szolgáltatják, illetve fogják szolgáltatni.

A nyomtatásban elérhető (tehát nem általunk gyűjtött) tundrai nyenyec szövegek kiválasztása során egy olyan szöveggyűjteményt szerettünk volna összeállítani, amely – a lehetőségeinkhez mérten – nagy számú szövegszót, valamint megbízható és természetes nyelvi adatot tartalmaz (vö. pl. Himmelmann 1998; McEnery – Hardie 2011). Így a következő tényezőket vettük figyelembe: a szövegek típusa (írott / írásban rögzített beszélt / beszélt), a szövegek nyelvjárás(i csoportj)a, az anyanyelvi adatközlő / informáns neve, neme, életkora a lejegyzés idején. A fenti szempontok egy része magától értetődő, de vannak kategóriák, amelyek némiképp magyarázatra szorulnak. Ilyen a szövegek típusára vonatkozó három kategória (írott / írásban rögzített beszélt / beszélt), amelyeket Schneider (2005) figyelembevételével alakítottunk ki. Vannak ugyanis olyan publikált tundrai nyenyec szövegek, amelyek egy valós beszédhelyzetben elhangzott szöveg írásban történt rögzítései. Ilyenek tipikusan a korábbi terepmunkák során nagy mennyiségben gyűjtött és publikált folklórszövegek, illetve újságokban megjelent interjúk szövegei. Ezeket nem tekinthetjük pusztán beszélt nyelvi szövegeknek, mivel az eredeti felvételek sok esetben nem érhetők el, így nem tudjuk, hogy a

szöveg szerkesztői milyen mértékben módosították az eredetileg elhangzott szövegeket. Az alábbi táblázat a korpuszunkban megtalálható szövegek típusát foglalja össze.[12]

2. táblázat: A tundrai nyenyec korpusz szövegeinek típusa

Szövegek forrása	Szövegek típusa	Szövegek műfaja	Szavak száma	Szövegek aránya a korpuszban
nyomtatott/ publikált	írott	újságcikkek	307 263	65,76%
		társalgási kézikönyvek	2 759	0,59%
		módszertani kézikönyvek	3 093	0,66%
	írásban rögzített beszélt	folklórszövegek	146 830	31,43%
		interjúk	6 021	1,29%
saját gyűjtés	beszélt	elbeszélés	1 246	0,27%
Összesen			467 212	

Az volt az eredeti szándékunk, hogy ezek a tényezők nagyjából egyforma szövegszó mennyiségben legyenek a korpuszunkban reprezentálva. Az anyanyelvi adatközlők / informánsok egyértelmű beazonosítására azonban csak 113 676 szót tartalmazó szöveg esetében volt lehetőség, ez a teljes szövegállomány (467 212) 24,33%-a. A szövegek lejegyzésének az idejét 319 422 szót tartalmazó szöveg (a korpusz 68,37%-a) esetében tudtuk csak meghatározni. Következésképpen a korpuszunk pillanatnyilag nem tekinthető kiegyensúlyozottnak, és bizonyos paraméterek szerint reprezentatívnak sem. A szövegeket azonban a fenti és további kritériumok (pl. a szövegek forrása, kiadási éve, szerkesztő személye) szerint rendszereztük, és a hozzáférhető adatokat egy katalógusban rögzítettük. A katalógus online változatát, amely egyelőre nem tartalmazza az adatközlőkre / informánsokra vonatkozó személyes adatokat, a projektünk

[12] A szövegek írott és beszélt nyelvi típus alapján való megkülönböztetése azért is indokolt, mert korábbi szintaktikai vizsgálatok arra a következtetésre jutottak, hogy a tundrai nyenyec nyelv írott és beszélt nyelvi változata között jelentős szerkezeti eltérések tapasztalhatók (lásd pl. Asztalos és mtsai 2017).

honlapján közzétettük.[13] Ez a katalógus szolgáltatja az alapját további gyűjtéseinknek.

A fenti paramétereket emellett metaadatként hozzárendeltük a szövegekhez, így az online korpuszban ezen kritériumok alapján létrehozhatók alkorpuszok. Ezek segítségével különböző szociolingvisztikai adatok és szövegtípusok alapján összehasonlító vizsgálatokat is végezhetünk a korpuszban.

3.2. A szövegek digitális feldolgozása

A szövegek kiválasztása és katalogizálása után azok digitális feldolgozása következett. Három szövegtípust különböztettünk meg egymástól: (i) online elérhető szövegek; (ii) nyomtatásban megjelent szövegek; (iii) hangfelvételek. Ezeket a típusokat eltérő módszerekkel kellett feldolgoznunk, az alábbiak szerint:

(i) csoport szövegei	⇒	web scraping
(ii) csoport szövegei	⇒	(szkennelés és) OCR
(iii) csoport szövegei	⇒	lejegyzés

Ebben a munkaszakaszban az volt a célunk, hogy a nyers adatokat egységes formátumúvá, UTF8 karakterkódolású .txt fájlokká alakítsuk.

3.2.1. Web scraping[14]

Az online elérhető szövegek forrása a korábban említett *Няръяна вындер* 'Vörös tundra' nevet viselő újság. Itt több száz tundrai nyenyec cikk érhető el, melyekhez 2013 februárja óta orosz fordítás vagy kisebb összegzés is társul. Ekkora mennyiségű szöveget hosszadalmas lett volna manuálisan menteni, ezért automatizáltuk a folyamatot egy web scraper implementálásával. Ez a script első lépésként minden egyes cikk url linkjét kigyűjti, majd ezeket egyesével végigjárva reguláris kifejezések segítségével kinyeri a számunkra lényeges metaadatokat és magát a szöveget a HTML címkék közül. Ennek kimenetele 793 tundrai nyenyec szöveges fájl lett, amely 286 166 tokent tartalmaz. Az egyes cikkekhez tartozó orosz fordításokat és összegzéseket is elmentettük, mivel azok lesznek az

[13] A katalógus az alábbi linken érhető el: https://tundranenetsdata.nytud.hu/index.html#corpus

[14] A web scraping kifejezés egy webes felületen található tartalom összegyűjtését és lemásolását jelenti.

alapjai a későbbi orosz–tundrai nyenyec párhuzamos korpuszunknak (a tervekről bővebben lásd az **5**. részt).

3.2.2. OCR[15]

A nyomtatásban hozzáférhető szövegeket beszkenneltük, majd a kapott .pdf kiterjesztésű fájlokat .txt fájlokká konvertáltuk.[16] Ehhez az Abbyy FineReader optikai karakterfelismerő programot használtuk.[17] Az OCR program kimenetét első lépésben manuálisan ellenőriztük, összehasonlítva az eredeti .pdf fájlokat az OCR program kimenetével. E folyamat során 179 800 szóból álló szöveget kellett manuálisan ellenőriznünk. Az első javítást követően szúrópróbaszerűen kiválasztottunk szövegeket, és azokat újra ellenőriztük. Az ellenőrzés során azt tapasztaltuk, hogy a kimeneti fájlok továbbra is tartalmaznak karakterfelismerési hibákat. Ezért szükségesnek találtuk egy második manuális ellenőrzés beiktatását. A második ellenőrzés során azonban nem hasonlítottuk össze újra a már ellenőrzött szövegeket, hanem más módszert alkalmaztunk. Ezt azért választottuk, mert az első körös manuális ellenőrzés nagyon időigényes volt. A második ellenőrzési kör során listát készítettünk a korpuszban előforduló összes szóról és olyan karakterláncokat kerestünk, amelyek megsértik a következő tundrai nyenyec fonotaktikai szabályt:

> **[Sz:]** A tundrai nyenyecben magánhangzóval, mássalhangzó torlódással, zöngétlen/lazán ejtett zár(jellegű)hanggal, affrikátával, pergőhanggal és gégezárhanggal nem kezdődhet szó (vö. Hajdú 1968).

A fenti szabályt megsértő karakterláncokat kilistáztuk, manuálisan visszakerestük és javítottuk a végleges fájlokban. Ez a második OCR ellenőrzés tehát nem teljes körű ellenőrzés volt.

3.2.3. Lejegyzés

2017-ben öt rövid, összesen 1246 szót tartalmazó tundrai nyenyec szöveget rögzítettünk audio (.wma) formátumban Khadry Okotetto tundrai

[15] Az OCR az *optical character recognition* rövidítése, magyar jelentése optikai karakterfelismerés. Az optikai karakterfelismerés nyomtatott szövegek elektronikus átalakítása.

[16] A txt az angol *text file* rövidítése, magyarul szövegfájl. Ez a szöveges dokumentumformátum nem tartalmaz semmilyen formázást.

[17] https://pdf.abbyy.com/

nyenyec informánsunk közlésében. Ezeket a szövegeket az informánsunk lejegyezte. Ennek során azonban nem fonetikai lejegyzést használt, hanem a szövegeket az általa ismert helyesírási formában rögzítette.[18]

3.3. Karakterek egységesítése

Miután a nyers adatokat UTF8 karakterkódolású .txt fájlokká alakítottuk, szükséges volt a szövegek bizonyos mértékű egységesítése annak érdekében, hogy a korpuszban történő keresés során minden létező adatot kinyerjen a felhasználó.[19] Amint említettük, nincsen egységes tundrai nyenyec írott nyelv és helyesírás, ezért a szóalakokat nem sztenderdizáltuk, kizárólag csak a karaktereket. A szavak egységesítésére elsősorban azért nem vállalkoztunk, mert az eltérő szóalakok nyelvjárási különbségekre is utalhatnak, s ezt az információt nem akartuk elveszíteni a szövegek feldolgozása során. Így van ez például a mutató névmás esetében, amelynek alakja *тикы* a jamali nyelvjárásban, és *чикы* a tajmiri nyelvjárásban.

A karakterek egységesítése során két problématípust kellett megoldanunk:

1. Ugyanaz a karakter használatos eltérő (grammatikai) funkciók jelölésére.
2. Különböző karakterek használatosak ugyanannak a (grammatikai) funkciónak a jelölésére.

Tipikus példa az 1. pontban ismertetett problémára a kettős idézőjel (U+0022), amely egyaránt használatos idézetekben (tehát eredeti funkciójában), és jelöli a(z egyik) gégezárhangot. Azért, hogy meg tudjuk különböztetni egymástól az idézőjel két használatát, az idézetekre szolgáló karaktereket lecseréltük francia idézőjelekre (U+00AB, U+00BB). Ezt a lépést nem tudtuk automatizálni, így manuálisan végeztük a cserét.

A 2. pontban bemutatott karakterkódolási problémára példaként szolgál az aposztróf (U+0027) és a kettős idézőjel (U+0022) használata: mindkét karakter ugyanazt a fonémát, ti. a gégezárhangot /ʔ/ jelöli.

[18] Az eredeti audio fájlok az alábbi honlapon érhetők el: http://corpus.nytud.hu/people/eszter/uralic_preerc/yrk/uj_szovegek/Okotetto/audio/

[19] Az UTF8 az angol 8-bit Unicode Transformation Format angol kifejezés rövidítése, amelynek magyar jelentése 8 bites Unicode átalakítási formátum. Ez egy, a szövegek karaktereinek (betűk, számok, írásjelek stb.) egységes kódolási és használati szabványán alapuló megjelenítési módszer.

Salminen (1998) és Nikolaeva (2014) két gégezárhangot különböztet meg a tundrai nyenyecben. Az egyik gégezárhang nazálissal váltakozik, a másik pedig zéró fonémával bizonyos toldalékok előtt. Hajdú (1968) és Staroverov (2006) alapján azonban nem indokolt a tundrai nyenyecben kétféle gégezárhang megkülönböztetése a szinkrón nyelvleírásokban, mivel a feltételezett két gégezárhang semmilyen akusztikai tulajdonságban nem tér el egymástól. Annak ellenére, hogy egyetértünk azzal, hogy nincsen kétféle gégezárhang a tundrai nyenyecben, valamint a kétféle gégezárhang jelölése és megkülönböztetése nem következetes a szövegekben, megtartottuk a fenti két karaktert. Elsősorban azért jártunk így el, mert a beszélői közösséggel konzultálva azt tapasztaltuk, hogy a közösség használja és megkülönbözteti a két jelölést. Ennek magyarázata véleményünk szerint a preskriptív iskolai nyelvoktatásban keresendő.

Egy másik példa a 2. pontban szereplő problématípusra a veláris nazálisok előfordulása a szövegekben: három különböző graféma használatos a veláris nazális fonéma jelölésére (Ӈ = U+04C9, Ң = U+04A2, Ҋ = U+04C8). Mivel a három eltérő jelölésnek nyelvészeti magyarázatát nem találtuk, ezeket a karaktereket egységesítettük. Azt feltételezzük, hogy a grafémák különbözősége nyomdatechnikai okokkal magyarázható. A beszélői közösséggel konzultálva a virtuális billentyűzet applikációkban (pl. Gboard) használatos karakterre cseréltük le ezeket a grafémákat. A karakterek egységesítésének zárásaként az írásjeleket is egységesítettük.

3.4. Korpuszfájlok

Utolsó lépésként a karakterek egységesítése során kapott szövegeket elő kellett készíteni az általunk használt NoSketch Engine korpuszkezelő rendszer számára. A NoSketch Engine korpuszkezelő a Sketch Engine nyílt forráskódú változata (Trevilla 2009).

A korpuszkezelő kiválasztása során figyelembe kellett vennünk, hogy (i) szövegeink még nincsenek elemezve és annotálva. Emiatt egy olyan rendszerre volt szükségünk, mely hatékony keresést tesz lehetővé már ezen az elemzési szinten is. Ezt az igényünket a NoSkE maximálisan kielégíti. Kereshetünk egyszerűen szóalakokra, karakterekre vagy akár komplex keresési mintát is írhatunk reguláris kifejezések segítségével, melyekre például szavak különböző előfordulásai egyszerre tudnak illeszkedni. (ii) Távolabbi célunk között szerepel egy tundrai nyenyec–orosz kétnyelvű korpusz elkészítése. Ezt is figyelembe véve fontos szem-

pont volt, hogy a korpuszkezelő képes legyen különböző nyelvű szövegek kezelésére. A NoSkE alkalmas az eltérő nyelvű szövegek párhuzamosítására akkor is, ha nincsenek azonos szinten annotálva. (iii) A tundrai nyenyec digitálisan kevésbé támogatott nyelv. Olyan rendszert szerettünk volna, ahol ez nem jelent akadályt. (iv) Végül hangsúlyos szerepet játszott a felület letisztult, felhasználóbarát és személyre szabható jellege, ami megkönnyíti a használatát bárki számára.

A NoSkE bemenetéül két fájl szolgál. Az egyikhez először az összes szöveget át kellett alakítani egy-egy XML fájllá, ami a szöveget vertikálisan tartalmazza.[20] Ez azt jelenti, hogy minden sorban egy token szerepelhet és esetlegesen a hozzá tartozó metaadatok, mint a szótő (lemma) vagy grammatikai információ (POS-tag). A szövegeket strukturálhatjuk címkék segítségével, pl. bekezdésekre, mondatokra oszthatjuk, vagy jelölhetjük, ha egy írásjel tapad az előtte lévő vagy az azt követő szóhoz. Minden ilyen típusú fájlnak van egy gyökércímkéje, ahol definiálhatunk attribútumokat. Ezek alapján a felhasználó kereséskor szűrni tud az adatok között. Az általunk használt attribútumok: id (azonosító), az adat forrása, műfaja, az adatközlő neme és a dialektus. Az automatizált átalakítási folyamat után kapott XML fájlokat végül összefűztük egy nagy vertikális fájllá. A másik bemeneti feltétel a konfigurációs fájl, ahol a korpusz struktúráját, vázát adhatjuk meg. Itt kapnak helyet olyan információk is, mint a korpusz nyelve, karakterkódolása, rövid leírása stb. E kettő segítségével a NoSkE sikeresen összeállította a tundrai nyenyec egynyelvű korpuszt, ami 467 212 szóból áll.

4. Online korpusz

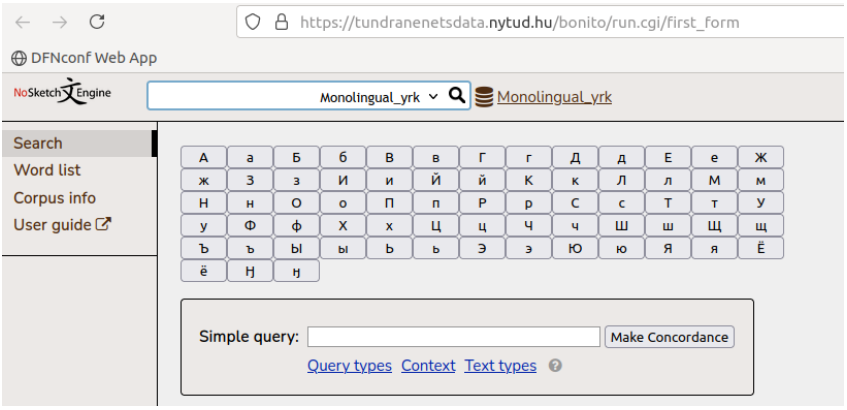
Az online tundrai nyenyec korpusz ideiglenesen elérhető a Nyelvtudományi Kutatóközpont egyik szerverén.[21] A korpusz jelenleg 467 212 szövegszót tartalmaz és egynyelvű, tehát csak tundrai nyenyec adatok kereshetők benne (a bővítési terveket a következő részben tárgyaljuk).

[20] Az XML az angol nyelvű *Extensible Markup Language* rövidítése, amely magyarul kiterjeszthető jelölőnyelvet jelent. Ez egy általános célú leíró nyelv, amely különböző adattípusok leírására képes, elsősorban az interneten keresztül történő strukturált szövegek és információ megosztására használják.

[21] https://tundranenetsdata.nytud.hu/bonito

4.1. A korpusz használata

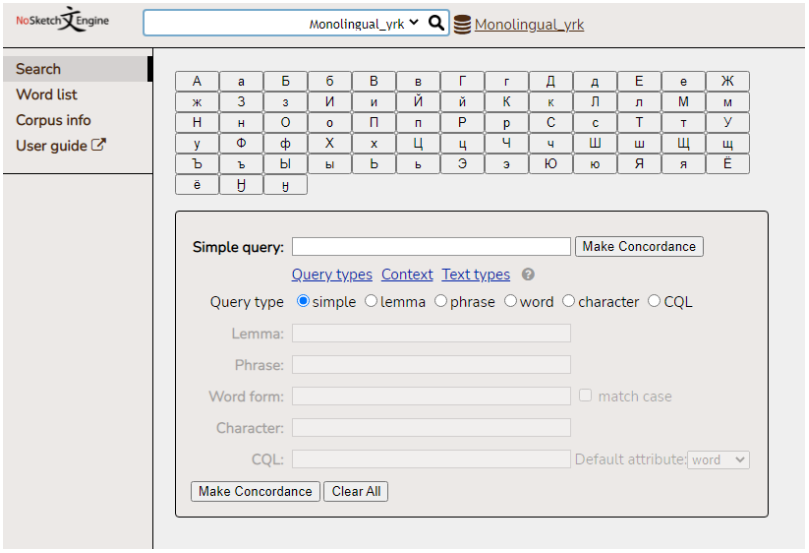
A korpusz jelenlegi formájában elsősorban arra használható, hogy karaktereket, karakterláncokat, morfémákat keressünk benne. Megfelelő tundrai nyenyec nyelvismeret esetében morfológiai és szintaktikai vizsgálatok is végezhetők a segítségével. A keresés egyszerűbbé tételének érdekében létrehoztunk egy cirill billentyűzetet a felületen (lásd 2. ábra). A felület jelenleg kizárólag angol nyelvű.



2. ábra: A tundrai nyenyec egynyelvű korpusz online keresőfelülete

A felhasználói felület megjelenítését módosítottuk, és olyan funkciókat állítottunk be, amelyek a szövegek mostani feldolgozottsági szintjén hasznosak lehetnek. Ennek során kikerült a menüpontok közül az ún. 'Find x' funkció, melynek segítségével részletes információt lehet szerezni arról, hogy például a keresett szó jellemzően milyen grammatikai kategóriában fordul elő. Korpuszunkban az egyszerű keresés mellett jelenleg lemmára, frázisokra, szóalakokra vagy karakterekre is kereshetünk. Továbbá elérhető olyan funkció is, ahol reguláris kifejezések segítségével alkothatunk komplex keresési mintákat a korpuszkereső nyelv használatával (CQL;[22] lásd 3. ábra)

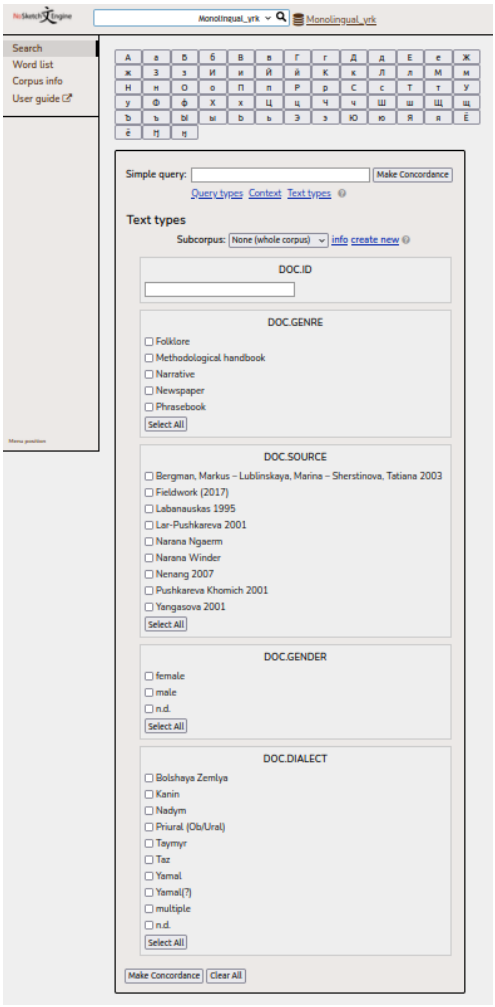
[22] https://www.sketchengine.eu/documentation/corpus-querying/



3. ábra: Keresési típusok

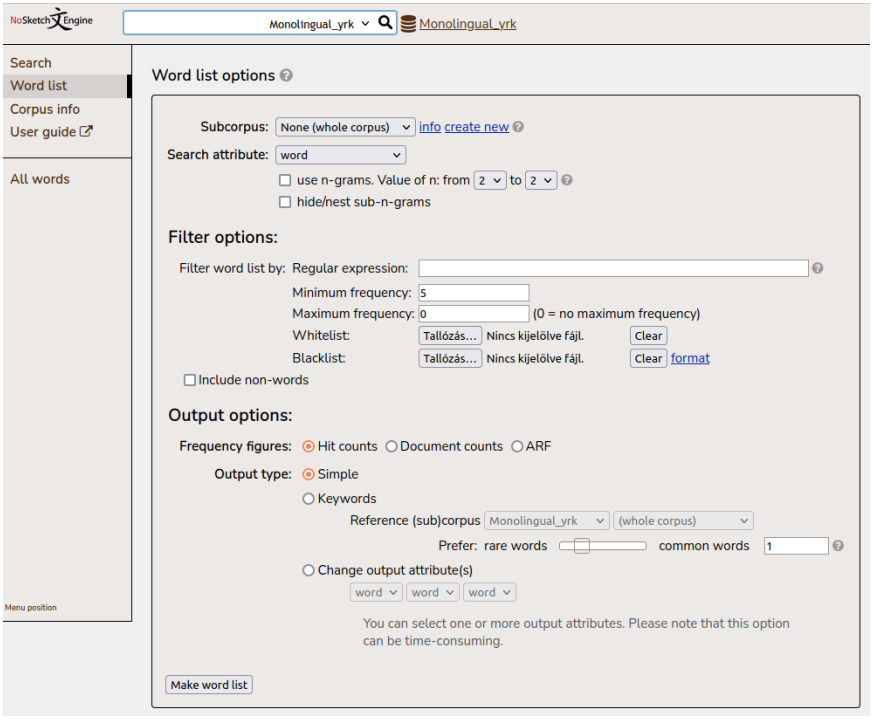
Létrehozhatunk alkorpuszokat is a felületen belül. Ahogy a korábbiakban szó volt róla, a szövegek gyűjtése és kiválasztása során azok metaadatait is rögzítettük abból a célból, hogy korpuszunk a jövőben megfeleljen a 3.1. részben tárgyalt főbb kritériumoknak. A következő adatokat gyűjtöttük össze: a szöveg forrása, a szöveg lejegyzésének éve, a szöveg típusa (írott / beszélt), (al)műfaja, tokenek száma, továbbá az adatközlő / informáns neve, neme, életkora, nyelvjárása. Minden egyes szöveghez hozzárendeltünk egy azonosító kódot is. Ezeket az adatokat egy .xlsx formátumú katalógusban tároljuk, és elérhetővé tettük a projekt honlapján. Ez konvertálható olyan formátummá, amelyet az általunk kiválasztott és a projekt során használni kívánt metaadatkezelő keretrendszer (IMDI/CMDI) megkövetel.[23] E keretrendszer használatával a nyelvi forrásról közölni kívánt metaadatok komponenseit egy profillá alakíthatjuk. Ezzel kívánjuk elősegíteni azt, hogy minél több kutatóhoz és nyelvhasználóhoz eljuthasson a korpuszunk. A metaadatok egy részét felhasználtuk arra is, hogy az online platformon segítségükkel alkorpuszt hozhassunk létre, és a keresést korlátozhassuk a szövegek azonosítójának, (al)műfajának, forrásának, az adatközlők nemének, valamint nyelvjárásának figyelembevételével (4. ábra).

[23] https://www.clarin.eu/content/component-metadata



4. ábra: Alkorpuszok

A NoSkE egy további beépített funkcióját is megtartottuk, így lehetőség van arra, hogy szavak gyakorisági listáit elkészíthessük a segítségével (lásd 5. ábra). Szűrhetünk a szavak között reguláris kifejezés, minimum és maximum előfordulás segítségével, vagy egyszerű szöveges fájl (.txt) feltöltésével megadhatunk szóalakokat, melyek mindenképp szerepeljenek, vagy éppen ne szerepeljenek az elkészülő listában.



5. ábra: Szavak gyakorisági listájának elkészítése

4.2. Annotálás

A digitális feldolgozást és a korpuszfájlok létrehozását követően a szöveg különböző szintjein szerettünk volna többletinformációt fűzni az adatokhoz. Az annotálási munkát azért kezdtük el, hogy szélesebb közönség is tudja használni a korpuszunkat. Megfelelő erőforrás és digitális eszközök híján azonban csak részleges szó- és szövegszintű annotálással tudtuk kezdeni ezt a munkát.

A szavak szintjén a személyes névmásokat láttuk el grammatikai információval. Azért esett a választásunk a személyes névmásokra, mert nincs nagy variabilitás az alakváltozataikban, így biztosan minden előfordulást ki tudtunk keresni és meg tudtunk jelölni. A NoSkE bemenetéül szolgáló vertikális fájl azért is van ilyen módon kialakítva, hogy a számítógép számára könnyen olvasható formában tudja tartalmazni az egyes tokenekhez tartozó információkat. Tabulátorokkal elválasztva szabadon megnevezhető metaadatot fűzhetünk a szavakhoz, melyeket a konfigurá-

ciós fájlban is definiálni kell a vertikális fájl struktúrájának megadásakor. A mi esetünkben szótővel (lemma) és POS-tag + morfológiai információval egészítettük ki a személyes névmásokat a szövegekben. A 3. táblázat tartalmazza a címkekészletet és a névmások gyakoriságát.

Az egyes alakokra a korábban említett CQL segítségével tudunk keresni. Az alábbi kifejezés közül (1) illeszkedni fog minden többes szám első személyű névmásra függetlenül annak esetétől, míg (2) az összes tárgyesetű személyes névmást fogja kilistázni.

(1) [pos="PPRON.1pl.*"]
(2) [pos=".*acc"]

A szöveg szintjén pedig a kérdéseket, illetve a rövidítéseket és a lábjegyzeteket jelöltük. Ezek mindegyikéhez szükség volt manuális feldolgozásra is. A kérdéseket azért szerettük volna megjelölni, mert a projekt, aminek keretében a korpusz készül, a kérdések szerkezetét vizsgálja. A kérdések megjelölése során először kilistáztuk azokat a mondatokat, amelyek végén kérdőjelet találtunk. A kérdések másik végpontjaként a kérdőjelet megelőző első írásjelet jelöltük meg. Így összesen 2557 db kérdést találtunk. Ellenőriztük a mondatokat és azok jelentését informáns bevonásával, majd külön címkével jelöltük az eldöntendő (1772 db) és a kiegészítendő (785 db) kérdéstípusokat. A rövidítések esetében kigyűjtöttük a lehetséges formákat négy karakter hosszúságig, melyekből 25 valódi rövidítést jelöltünk a szövegekben. A lábjegyzeteket már az OCR ellenőrzés során kapcsos zárójelekkel láttuk el, így ezek automatikus cseréjére volt csak szükség. Ezek jelöléséhez a NoSkE által használt XML formátumot tartottuk meg. A 4. táblázatban látható címkéket határoztuk meg.

Ezek a címkék jelenleg a szavak szintjén találhatók meg a szövegekben, keresésük is ehhez mérten lehetséges. A jövőben szeretnénk a keresési opciók közé könnyen kiválasztható módon felvenni.

3. táblázat: A tundrai nyenyec személyes névmások és címkéik a korpuszban

TOKEN	LEMMA	POS-tag	GYAKORISÁG
мань	мань	PPRON.1sg	2289
пыдар	пыдар	PPRON.2sg	534
пыда	пыда	PPRON.3sg	1848
мани'	мани'	PPRON.1du	180
пыдари'	пыдари'	PPRON.2du	20
пыди'	пыди'	PPRON.3du	186
маня"	маня"	PPRON.1pl	1547
пыдара"	пыдара"	PPRON.2pl	83
пыдо'	пыдо'	PPRON.3pl	838
си"ми	си"ми	PPRON.1sg.acc	587
сит	сит	PPRON.2sg.acc	469
сита	сита	PPRON.3sg.acc	105
сидни'	сидни'	PPRON.1du.acc	39
сидди'	сидди'	PPRON.2du.acc/3du.acc	30
сидади'	сидди'	PPRON.2du.acc/3du.acc	0
сидна"	сидна"	PPRON.1pl.acc	281
сидда"	сидда"	PPRON.2pl.acc	52
сидада"	сидда"	PPRON.2pl.acc	3
сиддо'	сиддо'	PPRON.3pl.acc	91
сидадо'	сиддо'	PPRON.3pl.acc	0

4. táblázat: Az annotálás során használt címkék

Kiegészítendő kérdés	<whq> … </whq>
Eldöntendő kérdés	<pq> … </pq>
Rövidítés	<abbrev> … </abbrev>
Lábjegyzet	<footnt> … </footnt>

5. Jelenleg folyó munkálatok, jövőbeli tervek

A korpuszunk pillanatnyilag tehát részlegesen (kis mértékben) annotált és egynyelvű. A munkánk még nem zárult le, a következő munkaszakaszban az alábbiak szerint tervezzük bővíteni és feldolgozni az adatokat, amennyiben rendelkezésre állnak majd az ehhez szükséges anyagi és emberi erőforrások:

– A korpusz anyagának bővítése további anyagokkal. Tervünk megvalósításához felvesszük a kapcsolatot olyan kutatóintézetekkel és archívumokkal, ahol még folyik a tundrai nyenyec nyelv kutatása. Továbbá a saját jövőbeli gyűjtéseinket is feldolgozzuk majd, és elérhetővé tesszük a korpuszban. Az adatgyűjtésünk során figyelembe fogjuk venni a katalógusunkat, és megpróbálunk olyan adatokat lekérdezni, amelyek jelenleg hiányoznak vagy alulreprezentáltak a korpuszban.
– A tundrai nyenyec szövegek fordítása. Első lépésként azt tervezzük, hogy egy tundrai nyenyec–orosz párhuzamos korpuszt hozunk létre, amely egyaránt kereshető mindkét nyelv oldaláról. Ehhez bizonyos források (mint pl. a folklórszövegek) esetében már készen vannak nyersfordítások, amelyeknek egy részét úgy adták ki nyomtatásban, hogy orosz fordításokat is mellékeltek hozzá. Ezek esetében szükség lesz a szövegek mondatszintű párhuzamosítására; ezt a feladatot anyanyelvi beszélők bevonásával tervezzük megvalósítani. Távlati tervünk között szerepel továbbá az is, hogy a szövegeket angolra is lefordítsuk, így létrehozva egy tundrai nyenyec–orosz–angol nyelvű párhuzamos korpuszt.

6. Összefoglalás

Tanulmányunkban bemutattuk a tundrai nyenyec nyelvi korpuszunk kialakításának folyamatát és lépéseit. Elsősorban az volt a célunk, hogy részletes útmutatót adjunk az általunk kidolgozott és követett módszerekről.

Az egyes munkaszakaszok a következők: a korpusz szövegeinek kiválasztása, a szövegek digitális feldolgozása (web scraping, OCR, lejegyzés), a karakterek egységesítése, korpuszfájlok létrehozása, online korpusz létrehozása. Ezek mellett bemutattuk, hogy hogyan lehet jelenleg használni a korpuszt, és beszámoltunk a korpuszban szereplő adatok részleges szó- és szövegszintű annotálásáról. Tanulmányunk végén a tervezett bővítési és módosítási lehetőségeinkről számoltunk be.

Irodalom

Ackerman, Farrell – Tapani Salminen (2006), Nenets. In: Brown, Keith (ed.), Encyclopaedia of language and linguistics, 2nd ed. Vol. 8. Elsevier, Amsterdam. 577–579.

Asztalos Erika – Gugán Katalin – Mus Nikolett (2017), Uráli VX szórend: nyenyec, hanti és udmurt mondatszerkezeti változatok. In: É. Kiss Katalin – Hegedűs Attila – Pintér Lilla (szerk.), Nyelvelmélet és diakrónia 3. PPKE BTK, Budapest. 30–62.

Dudeck, Stephan (2013), Challenging the state educational system in Western Siberia: taiga school by the Tiuitiakha River. In: Kasten – de Graaf (eds) (2013): 129–157.

Hajdú Péter (1968), Chrestomathia Samoiedica. Tankönyvkiadó, Budapest.

Himmelmann, Nikolaus P. (1998), Documentary and descriptive linguistics. Linguistics 36: 161–196.

Kasten, Erich – de Graaf, Tjeerd (eds) (2013), Sustaining indigenous knowledge: learning tools and community initiatives for preserving endangered languages and local cultural heritage. Kulturstiftung Sibirien, Fürstenberg.

Kavitskaya, Darya – Staroverov, Peter (2008), Opacity in Tundra Nenets. In: Abner, Natasha – Bishop, Jason (eds), WCCFL 27: Proceedings of the 27th West Coast Conference on Formal Linguistics. Cascadilla Proceedings Project, Somerville, MA. 274–282.

Koskarjova, N. B. [Кошкарева, Н. Б.] (2005), Очерки по синтаксису лесного диалекта ненецкого языка. Институт филологии СО РАН, Новосибирск.

Laptander, Roza (2013), Model for the tundra school in Yamal: a new education system for children from nomadic and semi-nomadic Nenets families. In: Kasten – de Graaf (eds) (2013): 181–194.

Liarskaya, Elena (2013), Boarding school on Yamal: History of development and current situation. In: Kasten – de Graaf (eds) (2013): 159–180.

McEnery, Tony – Hardie, Andrew (2011), Corpus linguistics: Method, theory and practice. Cambridge University Press, Cambridge.

Nikolaeva, Irina (2014), A Grammar of Tundra Nenets. Mouton de Gruyter, Berlin.

Pakendorf, Brigitte (2010), Contact and Siberian languages. In: Raymond Hickey (ed.), The handbook of language contact. Wiley-Blackwell, Malden. 714–737.

Popova, Ja. Ny. [Попова Я. Н.] (1978), Фонетические особенности лесного наречия ненецкого языка. Наука, Москва.

Salminen, Tapani (1993), On identifying basic vowel distinctions in Tundra Nenets. Finnisch-Ugrische Forschungen 51: 177–187.

Salminen, Tapani (1998), Nenets. In: Abondolo, Daniel (ed.), The Uralic languages. Routledge, London. 516–547.
Sammallahti, Pekka (1974), Material from Forest Nenets. Suomalais-Ugrilainen Seura, Helsinki.
Schneider, Edgar W. (2002), Investigating variation and change in written documents. In: Chambers, J. K. – Trudgill, Peter – Schilling-Estes, Natalie (eds), The handbook of language variation and change. Blackwell Publishing. 67–96.
Staroverov, Peter (2006), Vowel deletion and stress in Tundra Nenets. In: Gyuris, Beáta (ed.), Proceedings of the first Central European Student Conference in linguistics. http://www.nytud.hu/cescl/proceedings.html
Tyerescsenko, N. M. [Терещенко, Н. М.] (1993), Ненецкий язык. In: Ярцева, В. Н. – Елисеев, Ю. С. – Майтинская, К. Е. – Романова, О. И. (ред.), Языки мира: Уральские языки. Наука, Москва. 326–343.
Toulouze, Eva (1999), The development of a written culture by the indigenous peoples of Western Siberia. Arctic Studies 2: 53–85.
Toulouze, Eva (2003), The Forest Nenets as a double language minority. Pro Ethnologia 15: 95–108.
Trevilla, Lorena (2009), Ethnologue: languages of the world, SIL International. https://www.ethnologue.com
Vagramenko, Tatiana (2017), 'Blood' kinship and kinship in Christ's blood: nomadic evangelism in the Nenets Tundra. Journal of Ethnology and Folkloristics 11/1: 151–169.
Vakhtin, Nikolai (2015), Indigenous minorities of Siberia and Russian sociolinguistics of the 1920s: A life apart? Acta Borealia 32/2: 171–189.
Volzhanina, Elena A. (2007), The Forest Nenets: Habitat and population size in the 20th century, and the present demographic situation. Archaeology, Ethnology and Anthropology of Eurasia 30/2: 143–154.