DEBRECENI EGYETEM
FINNUGOR NYELVTUDOMÁNYI TANSZÉK
FOLIA URALICA DEBRECENIENSIA 30.
DEBRECEN, 2023

PermCorp: egy komi-permják korpusz létrehozása

NÉMETH Szilvia – SZABÓ Ditta – F. GULYÁS Nikolett

Eötvös Loránd Tudományegyetem, Budapest – nemeth.szilvia.viktoria@gmail.com
HUN-REN Nyelvtudományi Kutatóközpont, Budapest – szabo.ditta@nytud.hun-ren.hu
Eötvös Loránd Tudományegyetem, Budapest – nikolett.fgulyas@btk.elte.hu

1. Bevezetés

Jelen tanulmányban egy új komi-permják korpusz (PermCorp) létrehozásának első fázisát mutatjuk be.[1] Számos más finnugor nyelvvel összevetve a komi-permják nyelv dokumentáltsága alacsonynak tekinthető, ezen belül pedig az angol nyelven elérhető, így a nemzetközi kutatásban jobban hozzáférhető anyagok száma még ennél is csekélyebb. Projektünk elsődleges célja tehát a nyelv láthatóságának növelése mind az uralisztikában, mind az általános nyelvészetben nyelvtechnológiai eszközök létrehozásával.

A komi-permják az uráli nyelvcsalád permi ágához tartozik, legközelebbi rokon nyelvei a komi-zürjén és az udmurt. A komi-permjákok Oroszország európai részén, az Urál-hegység nyugati oldalán élnek és az oroszországi finnugor nyelvek beszélői körében egyedülálló módon 2005-ig többséget alkottak a kudimkari központú Komi-Permják Autonóm Körzetben. Ezt a közigazgatási területet aztán beolvasztották a Permi Területbe, így ők is kisebbségbe kerültek. A 2020-as oroszországi népszámlálási adatok szerint a komi-permjákok lélekszáma 55 785 fő volt, ami 41%-os csökkenést jelent a 2010-es adatokhoz viszonyítva (Pusztay 2022: 131).

Az uralisztikában a nyelvváltozat státusával kapcsolatban nincs egyetértés. Noha történetileg a komi-zürjének és a komi-permjákok két elkülönült csoportot alkotnak, utóbbiak nyelvváltozatát általában a komi egy nyelvjárásának tartják (Rédei 1978: 37; Bartens 2000: 9). E szerint a megközelítés

[1] Jelen tanulmány a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Fiatal kutatói kiválósági program (NKFIH FK 143242) keretében készült. Köszönjük a tanulmány névtelen lektorainak hasznos észrevételeit. Minden esetleges hiba a szerzők kizárólagos felelőssége.

DOI: 10.52401/fud/2023/11

szerint a komi nyelv három fő nyelvjárása az északon beszélt tulajdonképpeni komi (külső elnevezéssel zürjén), a délen használt permják (komi-permják), valamint a keleti, jazvai nyelvjárás (Hausenberg 1998: 305). A nem oroszországi kutatók általában ezt a megközelítést alkalmazzák (pl. Rédei 1978: 37–42; Bartens 2000: 9), mivel véleményük szerint a permják és a zürjén közti különbségek nem számottevők, elsősorban lexikális, kisebb részben fonológiai és morfoszintaktikai jellegűek, így a kölcsönös érthetőséget elviekben nem befolyásolják. Hasonló felfogást tükröznek a legújabb uralisztikai kézikönyvek vonatkozó fejezetei is (vö. Klumpp 2022: 471–474; Blokland 2023: 614–615), bemutatva a komi nyelvváltozatok osztályozásával kapcsolatos problémákat. Ezzel szemben az oroszországi, sok esetben anyanyelvi kutatók (vö. Batalova 1975, 2002; Ponomareva 2002, 2010; Lobanova 2017) azzal érvelnek a komi-permják önálló nyelvként való besorolása mellett, hogy a zürjének és permjákok egyrészt az oroszt használják a közös kommunikációban, másfelől az adott nyelvváltozatot annak beszélői is önálló nyelvnek tekintik. A komi-permják változat további variánsokra (nyelvjárásokra) tagolódik, ezek az északi, a déli és a keleti (izsvai, ld. Batalova 1975: 3; Bartens 2000: 31–32). Fontos még megemlíteni, hogy mind a zürjén, mind a permják változatnak létezik önálló irodalmi nyelve, utóbbi a Kudimkar környéki nyelvváltozaton alapul. A tanulmány további részében a komi-permják változatra nyelvként fogunk hivatkozni.

A nyelvi vitalitást tekintve a komi-permják 5-ös típusú, tehát fejlődő nyelv az EGIDS-skála szerint (Lewis et al. 2015). Ez azt jelenti, hogy hiába létezik sztenderdizált irodalmi nyelv, azt a többség nem használja, így a nyelv jövőbeni kilátásai is negatívak. A komi-permjáknak nincs hivatalos nyelvi státusa, ami tovább gyengíti a nyelvi vitalitást. Helimski (2003: 159) szerint a nyelv lazán kapcsolódik a Volga–Káma-vidéki nyelvi areához, ugyanakkor számottevő benne az orosz hatás is. Mindezeken túl fontos hangsúlyozni, hogy a nyelv dokumentáltsága még mindig nagyon alacsony például a komi-zürjénnel vagy az udmurttal összevetve. A WALS-ban (Dryer – Haspelmath 2013) például míg a komi-zürjénnel kapcsolatban 51, addig a komi-permjákra csak 24 szerkezeti jellemző szerepel, az ott található források egy része pedig már több mint 50 éves. Ezen túl a nyelvre vonatkozó, angol nyelven elérhető információk köre meglehetősen szűk.

Projektünk elsődleges motivációja tehát az volt, hogy nyelvtechnológiai eszközök segítségével hozzájáruljunk a komi-permják nyelv dokumentáltságának, valamint a nemzetközi szakirodalomban való láthatóságának növeléséhez. A projekt munkálatai 2022 decemberében kezdődtek, a kutatócsoport három főből áll, emellett több külső munkatárs is segíti munkánkat. A pro-

jekttagok közül ketten elsősorban a komi-permják nyelvi anyagok gyűjtéséért és feldolgozásáért felelnek, a harmadik résztvevő fő feladata a korpusz nyelvtechnológiai elveinek kidolgozása és a korpusz implementációja.[2] Az első munkaévben a korpusz felépítésének, a feldolgozandó szövegek körének, az annotáció fajtáinak a meghatározása, valamint kérdőíves anyaggyűjtés a fő feladata. Mindezeket természetesen nemzetközi kooperációban, komi-permják anyanyelvi szakértők segítségével terveztük elvégezni. Az ukrajnai háború miatt azonban a munkatervet részben módosítanunk kellett, összességében azonban mind a korpusz nyelvtechnológiai, mind a tartalmi részére vonatkozó elveket sikerült meghatároznunk. A továbbiakban ezek egy részét fogjuk részletesebben bemutatni. A tanulmány második részében ismertetjük a korpusz különféle definícióit, valamint a komi-permjákra eddig létrehozott korpuszokat és egyéb digitális eszközöket, azok előnyeit és a velük kapcsolatos kritikai észrevételeinket. A harmadik részben bemutatjuk korpuszunk tervezési munkálatait, a már említett elveket, valamint a szövegek kiválasztása során felmerült kérdéseket és az erre adott válaszainkat. Emellett kitérünk az annotációval kapcsolatos kérdésekre is. Végezetül összegezzük eddigi eredményeinket és vázoljuk a kutatás lehetséges további irányait.

2. A korpusz definíciójának különböző megközelítései és a jelenleg elérhető komi-permják források

Annak ellenére, hogy a komi-permják egyike az uráli nyelvcsalád legkevésbé dokumentált nyelveinek, létezik néhány interneten elérhető, kutatható korpusz is, ezek bemutatása előtt azonban fontos tisztázni, hogy a korpusz egy többjelentésű terminus a szakirodalomban (Gatto 2014: 8), nyelvészeti megközelítéstől függően eltérő feltételei vannak arra vonatkozóan, hogy mely szöveggyűjteményre alkalmazható. A definíciós különbségek ellenére a korpusz célja minden esetben az, hogy az alapján ismétlődő mintázatokat lehessen feltárni.

Gyakran használatos a terminus azokra a nyelvészeti kutatás céljára összeválogatott, változó méretű szöveggyűjteményekre, amelyek alapján egy adott kérdésre választ kaphatnak a kutatók. A kiválasztott szövegek ebben az esetben többnyire autentikusak, vagyis anyanyelvi beszélőktől származnak és nem, vagy nem elsődlegesen nyelvészeti vizsgálat céljára készültek (vagyis nem elicitált mondatok, hanem pl. folklórtörténetek, újságcikkek stb.). Ezek

[2] Vagyis a korpusz informatikai megvalósítása a közösen meghatározott elvek és specifikáció alapján.

a gyűjtemények többnyire kisebb méretűek, nem feltétlenül érhetők el digitális formátumban és annotációval sem rendelkeznek.[3]

A dokumentációs nyelvészek ennél szűkebben definiálják a korpusz fogalmát: digitális formátumú és valamilyen típusú annotációval ellátott autentikus szövegekből álló gyűjteményt értenek alatta, melyeken többféle kutatást is lehet végezni, vagyis nem egyetlen célra készülnek. Ezen korpuszok ritkán haladják meg a 150 000 tokenes terjedelmet és kis számú szövegtípus alkotja őket, mivel céljuk többnyire a nyelv vagy nyelvváltozat archaikus állapotának rögzítése. Azon nyelvek, melyek kisebbségi helyzetben és veszélyeztetett státuszban vannak, többnyire erősen forráshiányosak, így a korábban gyűjtött szövegek digitalizálása[4] és kutathatóvá tétele fontos lépés (Blokland et al. 2015), csakúgy, mint a lejegyzett, de a hanganyaggal együttesen elérhető beszédkorpuszok[5] létrehozása is.

A korpusznyelvészet ismét egy más megközelítésből határozza meg a korpusz fogalmát és jelentősen leszűkített jelentésben alkalmazza. Ebben az értelmezésben az autentikus szövegekből álló gyűjtemény célja, hogy egy nyelvet vagy nyelvváltozatot reprezentáljon (olyan precízen, ahogy csak lehet), ehhez pedig az alkotók súlyozzák (balanszolják) a különböző szövegtípusokat és a szövegek hosszának egységességére is figyelemmel vannak (Sinclair 1991: 171, McEnery – Wilson 2001: 29, Gatto 2014: 8). Az így létrejött korpuszok többnyire olyan nagy méretűek, hogy manuálisan nem dolgozhatók fel.[6]

Jelenleg komi-permják nyelven mindössze két szövegkorpusz létezik, melyek közül egy a dokumentációs nyelvészetre jellemző megközelítéssel készült, egy pedig vélhetően korpusznyelvészeti megközelítésből, de az aluldokumentáltsága miatt ez nehezen megítélhető. Ezen felül azonban csupán néhány kisebb szöveggyűjtemény érhető el az interneten.

[3] A korpusz ezen meghatározásának egy sor finnugor nyelveken történt kutatás megfelel, a komi-zürjént illetően pl. Baker (1986) és Klumpp (2009).

[4] Komi-zürjén nyelven például 2021-ben digitális korpusz formájában is elérhetővé vált (Partanen 2021) Vaszilij Litkin 1952-ben publikált szöveggyűjtése (Litkin 1952).

[5] Komi-zürjén nyelven például ebben a szellemben készült a Vászolyi Erik korábbi gyűjtéseit elérhetővé tevő korpusz (Blokland et al. 2021), az újabb keletű hangfelvételekből álló *IKDP* korpusz (Blokland – Chuprov 2019), valamint a média korpusz is, melyben az alkotók saját gyűjtései, illetve a médiában megjelent riportok hanganyagai is szerepelnek (Gerstenberger et al. 2016).

[6] Magyarországon a legismertebb ilyen korpusz a *Magyar Nemzeti Szövegtár* (Oravecz et al. 2014), de létezik komi-zürjén nyelvű, nagyobb méretű korpusz is, egy sajtónyelvi anyagokat tartalmazó, 1,75 millió tokenes (szóalakos), illetve egy közösségi médiában megjelent szövegekből álló, 1,85 millió tokenes korpusz (Arkhangelskiy 2019).

2.1. A komi-permják nyelvű korpuszok és szöveggyűjtemények

Az egyik komi-permják szövegkorpusz a *Korp* gyűjtemény alkorpuszaként érhető el (Borin et al. 2012). A komi-permjákon kívül a *Korp* komi-zürjén, udmurt, moksa, erza, hegyi mari és mezei mari alkorpuszokat egyaránt tartalmaz, melyek egy része további kisebb korpuszokra bontható. A komi-permják nyelv esetében nincsenek további alkorpuszok.

A permják főkorpusz kizárólag Wikipedia szövegeket tartalmaz (permják viszonylatban) rendkívül magas, 241 614 tokenszámmal.[7] Az adatokat alaposabban kifejtő részben láthatjuk, hogy ez 23 153 mondatot jelent. Jelen alkorpuszt 2020. december 9-én frissítették legutóbb, ami azt jelenti, hogy az azóta keletkezett komi-permják nyelvű Wikipedia-szócikkek itt már nem lesznek megtalálhatók.

A korpuszban a keresés csak cirill betűvel lehetséges. Az általános keresés esetében megadhatjuk, hogy a morféma vagy szó egy nyelvi elemen belül helyezkedjen-e el, vagy annak végén álljon, vagy esetleg esettől függetlenül keresse a rendszer. A keresőmotor időnként hibázik, például a *pö* (’állítólag’) partikula esetében nem mindig tudja megkülönböztetni a rendszer, hogy a morféma éppen egy nyelvi elem részeként vagy önálló szóként jelenik-e meg a mondatban. A keresés eredménye egy lista, mely 1-1 mondatot tartalmaz soronként, ezekben vastagon kiemelve látható az általunk keresett elem. Az egy oldalon megjelenő eredmények száma változtatható. Ugyancsak változtatható lenne az adott nyelvi elem kontextusának kiterjesztése jobbra vagy balra bővüléssel, ám a komi-permják esetén ez a funkció nem érhető el, pedig egyes nyelvtani elemek esetében elengedhetetlen lenne (például az evidencialitás vagy az aspektus vizsgálatában kimondottan permi nyelveknél, ahol egy grammatikai elem számos funkciót képes ellátni (Skribnik – Kehayov 2018: 539–543; Szabó 2022: 172). A keresés végrehajtása után lehetőségünk van a megjelenő mondatokban szereplő elemekről több információt megtudni. Egy-egy elemre kattintva a jobb oldalon felugró ablak tartalmazza például a szöveg keletkezésének dátumát és az adott nyelvi elem tulajdonságait, úgymint szófaj, grammatikai elemzés (az adott morféma lehetséges nyelvtani szerepének megadása), függőségi relációk és alapforma. A szófaj kategóriája nem mindig ismeri fel az adott elemet, ilyenkor ’unknown’ (ismeretlen) lesz az érték. A *pö* konzekvensen határozóként szerepel, a grammatikai elemzésnél is adverbiumként (Adv) jelenik meg, pedig a széles körben elfogadott nézet szerint ez egy partikula (Siegl 2004: 102). A függőségi

[7] A Korp rendszeren belül viszont ez a legkisebb tokenszámú alkorpusz, a mordvin és a mari alkorpuszok milliós nagyságrendűek.

viszonyokra vonatkozó adatok nem érhetők el minden esetben, azonban a mondat egy olyan eleménél, ahol a dependencia reláció X, vagyis ismeretlen, a dependencia-ábra megnyitásával (ahol a mondat teljes függőségi viszonyrendszerét láthatjuk) mégis megjelennek az információk. Az alapformák esetében is problémásak az eredmények: a rendszer általában vagy nem tudja meghatározni az alapformát, vagy meghatározza, de helytelenül, csupán az esetek kisebb részében tudja helyesen azonosítani az adott elem alapalakját. A kiterjesztett keresés során már nemcsak szóra, hanem szófajra, glosszára, alapalakra, vagy függőségi viszonyra is kereshetünk, de akár egynél több paramétert is beállíthatunk egy még összetettebb keresés érdekében. A keresőmotor ezen része jól működik, azonban a hibák, melyekkel az egyszerű keresés eredményeinél találkozunk, itt is előfordulnak. Tehát ha a szófaj kategóriájában a partikulákra keresünk rá, az eredmények ebben az esetben sem fogják tartalmazni a *pö* partikulát, mivel a rendszer azt adverbiumként ismeri. A korpusz egyik hasznos funkciója, hogy lehetőség van több keresést kombinálni, tehát rákereshetünk például két egymás után szereplő szóalakra, szófajra, vagy morfológiai elemre is. A keresési lehetőségek egy további fajtáját a haladóknak szánt CQP-val (corpus query processor)[8] történő lekérdezés képezi, melyhez egy útmutatót is találunk az oldalon.

A fenti hibák abból fakadnak, hogy a szövegeket automatikusan annotálták és glosszázták. A korpusz egyértelmű előnye, hogy nagy tokenszámú,[9] amivel nő a keresések sikerességének aránya. Hátránya viszont, hogy kizárólag az tudja használni, aki egyrészt ismeri a cirill betűket, másrészt, aki ért komi-permjákul, ugyanis angol (vagy bármely más nyelvű) fordítás nem érhető el egyik mondatnál sem. A grammatikai kategóriacímkék automatikusan történő megadása pedig sok esetben inkább félrevezető, semmint segítség a felhasználók számára.

A közelmúltban egy új komi-permják korpusz jelent meg a világhálón *Коми Кыв Корпус* [A komi nyelv korpusza][10] néven, amely összesen 6 196 963 tokent tartalmaz. Több szövegműfaj is megtalálható a gyűjteményben, melyek alapján szűrhetők a találatok, vagyis a szövegműfajok alkorpuszokat alkotnak, de akár egyszerre az összesben is kereshetünk. Találunk szépirodalmat,

[8] Ez a módszer úgynevezett reguláris kifejezésekkel működik, melyek segítségével definiálható, hogy a keresőmotor milyen szövegrészeket találjon meg úgy, hogy azt nem az annotációkban, hanem a korpuszt alkotó szövegek nyers változatában keresi.

[9] A korpusznyelvészetben a token egy, a szövegben található szóalakot jelent, vagyis a tokenek összesített száma megadja, hogy hány szóból áll a korpusz.

[10] Ld. http://perem.komicorpora.ru/

ismeretterjesztő irodalmat, drámát, bulvárt, tudományos irodalmat, költészetet, folklórt és egyéb más műfajokat is, de a korpuszépítés folyamata jelentősen aluldokumentált: nem találni információt arról, hogy milyen arányban szerepelnek benne az egyes szövegtípusok, illetve, hogy mi volt a szövegek kiválasztásának kritériumrendszere.

A keresési felület, mely maga a nyitóoldal is, átlátható. Itt elsősorban teljes szóra vagy szóalakra kereshetünk, ezúttal is csak cirill betűs írást alkalmazva, de a fejlesztők a cirill billentyűzet biztosításával segítik a munkát. Az adott elem pozícióját is kiválaszthatjuk: a megadott formára pontosan keressen a rendszer; vagy állhasson egy szó részeként (külön választható, hogy az elején, végén vagy a közepén); reguláris kifejezésként; az elem lehessen-e lemma vagy egy frázis része. Külön kereshetünk a korpuszban felhasznált források szerzőjére, a mű címére, a kiadás évére is. A keresés még tovább pontosítható a találatok rendszerezésének elvének megadásával és azok csökkenő vagy növekvő sorrendbe helyezésével. Az eddigieket összegezve, a Komi Kyv Korpus keresőmotorja elméletileg részletes és sokoldalú.

Maradva a *pö* partikulánál, egy egyszerű keresés esetén, mely során nem szűrtük a találatokat semmilyen paraméter szerint, a szöveggyűjteményben 1 662 találat van az általunk keresett elemre. A találatok itt is soronként jelennek meg. Minden egyes elem a rákattintás után tartalmaz(na) további információkat, úgymint morfológiai elemzést, lemma adatokat, valamint fordítást. De akárcsak a *Korp* esetén, ezúttal is azt tapasztalhatjuk, hogy egyrészt nem minden elemnél érhető el további adat, másrészt ahol van, ott is gyakran hiányos. A korpusz állományáról csak a keresés eredményén keresztül tudunk meg adatokat. Eszerint a legkorábbi szöveg 1921-ből származik, a legújabb szövegek 2022-ből valók. A szöveggyűjteményt a FU-Lab (The Finno-Ugric Laboratory for Support of the Electronic Representation of Regional Languages) nevű kutatócsoport hozta létre 2021-ben.[11]

Mindent összevetve ebben a korpuszban a keresés jól működik, akár szűrők használata mellett is. A keresőmotor nagyon részletes, azonban ezt elsőre nem tudja alkalmazni a felhasználó, hiszen keresés nélkül nem derül ki a korpusz szövegállománya sem a forrásokat, sem a számokat tekintve. Ezúttal is nagy hátránynak tekinthető, hogy a korpuszt csak a komi-permják nyelvet magas szinten ismerő kutatók tudják igazán használni, hiszen egyértelmű és hiánymentes glosszázás, valamint fordítás nem érhető el. További probléma, hogy a morfológiai elemekre való kereséskor nem lehet olyan lehetőséget megadni, ami jelezné az elem grammatikai kategóriáját, ezáltal a nagy elem-

[11] Elérhetősége: https://fu-lab.ru/

számú találatok közül manuálisan kell kiválogatni azokat az elemeket, amelyek ténylegesen toldalékmorfémák és azokat, amelyekben az elem az adott szó része. A kontextus jobbra és balra való kiterjesztése 1-1 mondattal funkció viszont tökéletesen alkalmazható.

Egy további ismert korpusz a Timofey Arkhangelskiy által 2018–2019-ben létrehozott *Corpora of Uralic Volga-Kama Languages* (Arkahgelskiy 2019), amely a szerkezetéből ítélve hét uráli nyelvű alkorpuszt tartalmazna, köztük komi-permják szövegekkel. A korpusz jelenleg öt nyelven (udmurt, komi-zürjén, erza, moksa, mezei mari) tartalmaz szövegeket, a komi-permják és a hegyi mari alkorpuszt eddig még nem publikálták, és a megjelenés várható idejéről nincs információnk.

A Niko Partanen és Jack Rueter által készített *Four Battles Corpus* négy nyelven (komi-zürjén, komi-permják, udmurt és hegyi mari) tartalmazza a (komi-permjákul) *Нёль воевöй случай* című történetet 1940-ből, mely a Finn Nemzeti Könyvtár Fenno-Ugrica nevű digitális gyűjteményében található nyers formában.[12] A korpusznak egyértelmű hátránya, hogy egy tömörített fájl letöltése után tekinthető csak meg a szöveggyűjtemény, így viszont informatikai ismeretek nélkül csak manuális keresés hajtható végre rajta. A szöveget mondatokra bontották, így például a komi-permják változatban 389 mondat található, melyek mindegyikében jelölt az elemszám. Az elemeknél külön számot kaptak a mondatban található írásjelek is. Minden egyes mondathoz tartozik orosz fordítás, ez egyértelműen pozitívuma a korpusznak. Glossza szórványosan érhető el az egyes elemek esetében, inkább szintaktikai szerepük, úgy mint a mondat alanya vagy tárgya kerültek feltüntetésre. A projekt célja egyértelműen érdekes és értékes, arról viszont semmit sem tudunk, hogy a jövőben a korpusz további szövegekkel, esetleg további nyelvekkel bővülni fog-e.

Hasonló típusú a Turkui Egyetem Volgai nyelvek kutatócsoportja[13] által a 2000-es években készített, párhuzamos szövegekből álló korpusz. A korpusz az interneten nem elérhető, használata engedélyköteles és regisztrációhoz kötött, de kutatási célokra bárki számára hozzáférhető.[14] A nyelvi anyag két párhuzamos korpuszból áll. Ezek egyike egy eredetileg orosz nyelvű elbeszélést, az ún. Pavlik Morozov-szöveget tartalmazza, amelyről tizenegy finnugor – köztük komi-permják – és két törökségi nyelvre készített fordítást

[12] Lásd még: Bradley et al. (2018).

[13] A kutatócsoportról ld. https://www.utu.fi/en/university/faculty-of-humanities/finnish-and-finno-ugric-languages/volgaic-languages

[14] Felhasználására ld. például F. Gulyás (2016).

egy-egy anyanyelvi adatközlő. A szöveg a 20. század elejéről származik, típusa szerint szépirodalmi, prózai szöveg, de néhol beszélt nyelvi jegyeket is mutat. A történet egy erősen propagandisztikus tanmese a pionírok hősiességéről és helyes magatartásáról. A lefordított szövegek az adott nyelv helyesírásának megfelelően vannak lejegyezve és egy kiegészítő szoftver segítségével a szövegek egyszerre, párhuzamosan olvashatók. A mondatok ugyan nem annotáltak, de külön sorokra vannak bontva és meg vannak számozva, ami segíti a könnyebb összehasonlítást. A komi-permják fordítás 1 608 db számozott mondatból áll, a tokenek száma 13 089. A korpusz kis mérete miatt nem alkalmas statisztikai elemzésre és figyelembe kell venni, hogy fordítás, tehát az orosznak mint forrásnyelvnek lehet nyelvi hatása. A másik alkorpusz egy kicsivel hosszabb, Finnországról és a finn természetről szóló, ismeretterjesztő szöveg, mely kevesebb, hét nyelven érhető csak el, a komi-permják nem szerepel közöttük. Összességében minden hátránya ellenére fontos hangsúlyozni, hogy a párhuzamos korpusz hiánypótló forrásnak számít.

2.2. A komi-permják nyelv egyéb digitális, nem korpusznak minősülő forrásai

Permják nyelvi anyagok a korpuszokon kívül is elérhetők digitális formában. Az előzőekben említett permják nyelvű Wikipedia egy folyamatosan bővülő, és ezzel egy mindig megújuló forrást jelent a nyelvvel foglalkozóknak. A permják Wikipedia 2009-ben indult (hivatalosan csak 2010-ben), mely ma már 3 461 szócikket tartalmaz, melyek között azonban vannak üresek, egy mondatot tartalmazók és 400 tokenesek is.

Kis számban, de léteznek komi-permják híroldalak, online publikált újságok, illetve találni VKontakte-oldalakat is, melyek elsősorban kulturális programokról számolnak be és amatőr költők műveit jelentetik meg. Két, mára már inaktív blog is elérhető. Sajnálatos módon internetes kommenteket alig találni.

Az ugyancsak említett Fenno-Ugrica digitális gyűjteményben nyelv szerint is tudunk keresést végezni, mely során láthatjuk, hogy komi-permják nyelven összesen 3 293 folyóirat-tétel érhető el, a könyv kategóriájából pedig 91. A szövegek szkennelve (esetleg .txt formátumban is) letölthetők, vagyis annotálás, nyelvtani elemzés nem része (de nem is célja) a Fenno-Ugrica szöveggyűjteményének. A nyelv szinkrón és diakrón vizsgálatához viszont egyaránt használhatók, ahogyan korpuszok forrásaiként is.

A Jack Rueter, Niko Partanen és Larisa Ponomareva által létrehozott *Comparative Permic Database* (Rueter et al. 2020) egy jó kezdeményezés, de jelen állapotában nem felhasználóbarát. Nincs online keresőfelülete, a .zip ki-

terjesztésű fájl letöltése és kicsomagolása után fér hozzá az érdeklődő az anyaghoz. Az adatbázis jelenleg 34 morfémát tartalmaz komi-zürjén és komi-permják nyelven, feltüntetve azok grammatikai jegyeit.

A nemrég megjelent Volga-vidéki finnugor nyelvek tipológiai adatbázisának (Havas et al. 2023) komi-permják adatai (vö. Asztalos et al. 2021) nem szövegforrásként, hanem a grammatikai elemek elemzésénél és a glosszázásnál segíthetik jelen korpusz kialakítását. Az adatbázis izolált példamondatokat tartalmaz, melyek mindegyike glosszázva és fordítással kerül publikálásra három nyelven: magyarul, angolul és oroszul.

3. A PermCorp

3.1. A PermCorp céljai és létrehozásának alapelvei

A fent ismertetett források alapján úgy véljük, hogy a komi-permják kutathatóságát jelentősen megkönnyítené egy korpusznyelvészeti szempontok alapján felépített új korpusz, mely azzal, hogy különböző kommunikációs szándékkal készült szövegekből áll, manuálisan ellenőrzött angol nyelvű címkézéssel és fordítással van ellátva (vagyis gold sztenderd minőségű), lehetővé tenné, hogy:

– általános következtetések legyenek levonhatók a permják nyelvről, hiszen egyetlen helyen lehetne vizsgálni több nyelvváltozatot is,
– a kutatási kérdések megválaszolása gyorsabbá váljon a keresési eredmények magas fokú pontosságával,
– a felhasználók a szövegek metaadatai alapján a kutatási kérdésüknek megfelelően összeválogathassák, leszűkíthessék a saját vizsgálati korpuszukat,
– a kutatók angol nyelven is hozzáférjenek a nyelvi anyagokhoz, valamint, hogy
– a manuális annotáció segítségével kiértékelhetővé váljanak a már létező permják morfológiai elemzők teljesítményei, és így lehetővé váljon azok továbbfejlesztése.

Egy ilyen korpusz létrehozása számos problémába ütközik még egy olyan forrásbőségű világnyelv esetében is, mint pl. az angol, de egy kis írásbeliséggel rendelkező veszélyeztetett nyelvnél további kihívások is fennállnak. Mielőtt bemutatnánk, hogy a PermCorp tervezése során milyen megoldások mellett döntöttünk, fontos áttekinteni a korpusznyelvészetben alkalmazott alapelveket.

A korpusznyelvészet alapvető kiindulópontja, hogy a nyelvet nem a nyelvészeti kutatás céljára előállított szövegek vagy elicitált példák alapján, hanem természetes (autentikus) adatokon keresztül figyeli meg (Gatto 2014: 9). Autentikus szöveg ebben az értelmezésben lehet fordítás vagy kódváltásokkal tarkított produktum is, sőt ha ezek jellegzetes formái a reprezentálni kívánt nyelvváltozat használatának, akkor fontos is, hogy megjelenjenek a korpuszban.

A korpusz ebben az értelmezésben egy olyan szöveggyűjtemény, mely arra törekszik, hogy reprezentáljon egy nyelvet vagy nyelvváltozatot (Gatto 2014: 10), hiszen csak akkor lehetséges általános érvényű következtetéseket levonni az alapján, ha a legtöbb olyan szövegtípus megjelenik az anyagban, ami előfordul a természetben. Ha például egy magyar nyelvű korpusz csak osztálytársak egymásnak írt leveleiből és a saját maguknak készített jegyzetekből áll, akkor ez alapján nem fogalmazható meg az egész magyar nyelvre vonatkozó állítás, hiszen számtalan jelenség (pl. a magázás) meg sem jelenik bennük. De ez a szöveggyűjtés nem reprezentálja a magyar diákok nyelvezetét sem, mert nem tartalmaz egy sor gyakran előforduló szövegtípust (pl. szünetben történő beszélgetéseket, dolgozatokat, a tanárral való kommunikációt stb.), ami komplexebben megvilágítaná, hogy a diákok hogyan használják a nyelvet.

A reprezentativitás biztosítására nem létezik egy minden esetben alkalmazható, objektív módszer, így ez minden esetben a korpuszépítők szubjektív becslésén alapul (Leech 2007: 143), de a tervezés során számos szempontot figyelembe kell venni. Először is, szükséges felmérni, hogy milyen szövegtípusok léteznek az adott nyelvváltozaton, ennek során azonban nem a szövegek tartalma, hanem azok kommunikációs funkciója a meghatározó (Wynne 2005). A kommunikáció témája sokkal szerteágazóbb, mint a funkciója, így lehetetlen feladat minden lehetséges téma szerepeltetése a korpuszban, az azonos funkciójú szövegek azonban sok szempontból hasonlítanak egymásra, még ha a szókincsük különbözik is. A téma ráadásul kevésbé meghatározó a nyelvi megformáltságban (a lexikont leszámítva), például a komi-permják nyelvtanról szóló általános iskolai tankönyv, Wikipedia-szócikk, illetve tudományos publikáció nyelvezete jobban eltér egymástól, mint pl. egy ugyanazon osztálynak szóló biológia és egy nyelvtan tankönyv nyelvezete.

Másodszor, a reprezentativitás biztosításához valamilyen módon szükséges megbecsülni a különböző szövegtípusok relatív gyakoriságát, és ehhez mérten kell meghatározni, hogy a korpusz darab- és tokenszámban mérve mennyit tartalmazzon az egyes szövegtípusokból, ez a balanszolás (Biber

1992: 174). Ez egy igen nehéz és szubjektív feladat, melynek a nyelvi helyzetet jól ismerő kutató intuicióján kell alapulnia (Wynne 2005). Jelentősen könnyebb a helyzet, ha a korpuszépítők pontosan ismerik annak a kutatásnak a célját, amire a korpusz szolgálni fog, de ha ez ismeretlen (például mert többféle kutatásra is alkalmas korpuszt kívánnak készíteni), akkor fontos biztosítani, hogy a jövőbeli felhasználók a szövegek metaadatai segítségével leválogathassák, összeállíthassák a saját vizsgálati korpuszukat (Wynne 2005). A korpuszépítés során érdemes alkorpuszokat kialakítani, melyek szövegei jól elkülöníthetők a többi alkorpuszéitól (pl. közvetlen, hivatalos, gyerekeknek szóló, tudományos stílusú stb.).

Harmadszor, egy nyelvváltozat reprezentálásánál nem hanyagolható el annak hatása sem, hogy a természetben előforduló szövegek hossza radikálisan különbözik, például a versek jelentősen rövidebbek, mint a regények. Ez problémát okoz a balanszolásban, mert választani kell, hogy a szövegek darabszámát (és szerzőinek számát) vagy a tokenjei számát tartja-e szem előtt a korpuszépítő. Ennek kiküszöbölésére használatos gyakori stratégia, az, hogy a korpuszépítők akkorára emelik a korpusz méretét, hogy ez ne jelentsen problémát, illetve hogy a hosszabb szövegekből csupán meghatározott hosszúságú részleteket emelnek be a korpuszba, így pl. több regény szövege is szerepelhet a korpuszban úgy, hogy tokenszerűen nem ez a szövegtípus fog dominálni. A részletek kiválasztásának (a mintavételnek) több módja is lehet, de szem előtt kell tartani, hogy a részlet vagy részletek minél jobban reprezentálják a szöveg egészét (Wynne 2005).

Számos kutatási célra (pl. szövegstatisztika, frazémák kinyerése stb.) csupán több millió tokenes korpuszok alkalmasak, más kérdések azonban kisebb korpuszon is vizsgálhatók, emiatt a korpusznyelvészeti megközelítésben az alkotók többnyire igyekeznek minél nagyobb gyűjteményt létrehozni. Ezek azonban többnyire semmilyen vagy automatikusan létrehozott címkézéssel rendelkeznek, így a manuális annotáció nem korlátozza a méretet.

A továbbiakban bemutatjuk, hogy a PermCorp tervezése során milyen nehézségekkel szembesültünk, és milyen megoldások mellett döntöttünk, hiszen a fentebb ismertetettek fényében látható, hogy a számos szubjektív döntés miatt kiemelt fontosságú a folyamat alapos dokumentálása a jövőbeli felhasználás szempontjából.

3.2. A PermCorp felépítése

A korpuszépítés során elsődleges célunk, hogy a komi-permják nyelv minél tágabb változatát reprezentáljuk, amivel kapcsolatban rögtön egy jelentős akadállyal szembesülünk. Bár a nyelv használata napjainkban is inkább a

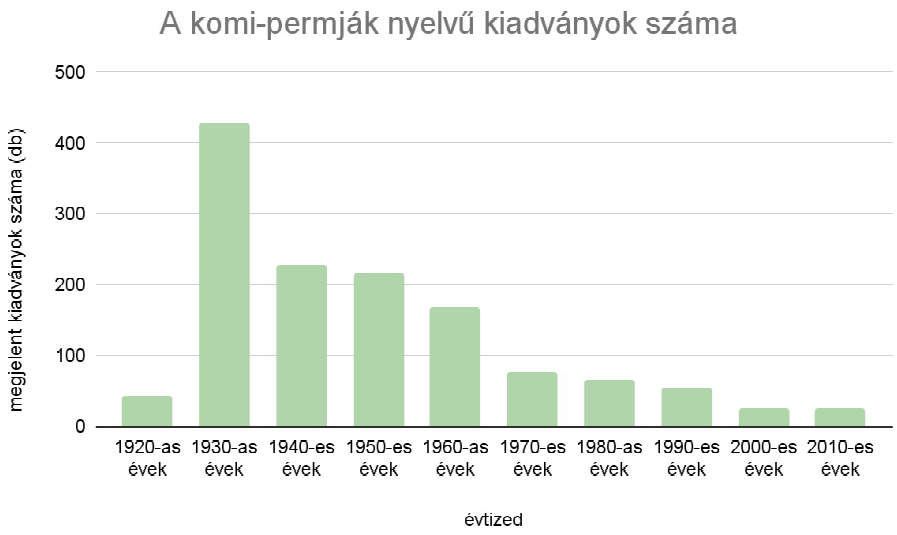
családi környezetre, mint a hivatali kommunikációra korlátozódott, ráadásul jellemzően inkább szóbeli, mint írásbeli formában, beszéd gyűjtésére nincs lehetőségünk. Azonban az elmúlt évszázad alatt több ezer oldalnyi szöveg került publikálásra, melyek hozzáférhetősége jelenleg erősen korlátozott, hiszen a kiadványok vagy eleve csak nyomtatásban (ráadásul kis példányszámban) jelentek meg, vagy nem kereshető formában vannak digitalizálva. Ezek többnyire irodalmi szövegek, folklórgyűjtések, újságcikkek, iskolai tankönyvek, de vannak szakszövegek és találni internetes műfajokat is. Közös bennük, hogy mindegyik a nyilvánosság számára készült (szemben pl. a magánlevelekkel, naplókkal, egyetemi vázlatokkal stb.), ezért azt a célt tűztük ki, hogy ezekből készítünk egy több alkorpuszból álló, szövegtípusonként súlyozott korpuszt, mely reprezentálni fogja a permják nyilvánosságnak szánt írott változatát.

Az alkorpuszok kialakításánál autentikusnak tartottuk a permják anyanyelvűek által készült szövegeket attól függetlenül, hogy milyen erős bennük az orosz hatás, azonban nem emeltünk be a korpuszba fordításokat, mivel ezek – elsősorban a korai időszakokban készültek – nem megfelelő minőségűek, kiforratlanságuk miatt még az orosz hatás milyenségéről sem informálnak.

Mivel a források mennyisége és a manuális annotáció erősen limitálta, hogy mekkora méretű korpuszt készíthetünk a projekt időszaka alatt, 300 000 tokennel tervezünk. Ez a dokumentációs nyelvészeti megközelítés alapján egy nagy, a korpusznyelvészeti megközelítés alapján azonban egy kis méretű korpuszt jelent, ezért annak érdekében, hogy ritkább nyelvi jelenségek is kutathatóak legyenek, úgy döntöttünk, hogy egy kisebb alkorpuszban nem autentikus (vagyis nyelvészeti kutatás céljára készült) szövegeket is elérhetővé teszünk, melyeket a felhasználó igény szerint kizárhat majd a vizsgálati korpuszából.

3.2.1. Az alkorpuszok kijelölése és a korpusz szövegeinek kiválasztása

A korpuszépítés első lépéseként felmértük, hogy milyen források (és milyen mennyiségben) jelentek meg komi-permják nyelven. A munkában segítségünkre volt két, a Wikipédián elérhető lista, melyek a permják kiadványokat tartják számon 1921 és 2016 között (Wikipedia 1, Wikipedia 2). Az ezekben található 1 328 kiadvány alapján látható, hogy az 1930-as évek forrásbősége után egyre csökkent a megjelenések száma, a ’70-es évektől pedig nem volt olyan évtized, amikor az új kiadványok száma elérte volna a 80-at (1. ábra).



1. ábra

Ahogy az a többi, Oroszország területén élő finnugor nyelv esetében is megfigyelhető, az 1920-as évek vége és a 30-as évek eleje felvirágzást hozott a komi-permják beszélőközösség kulturális fejlődésében. Ekkoriban indult meg a permják irodalmi norma kodifikálása a Kudimkar-Inyva környéki nyelvjárást alapul véve, ami később soha nem látott mennyiségű kiadvány megjelenésével járt. A korábbi, nem kodifikált cirill írásrendszert az 1920-as években először latin, majd ismét cirill betűs helyesírás létrehozása követte. Ebben az időszakban születtek az első általános iskolai tankönyvek (ábécéskönyvek, matematika, fizika, biológia, földrajz), az első ismeretterjesztő könyvek és füzetek (mezőgazdasági, háztartáshoz kapcsolódó, valamint néprajzi témákban), ekkortól léteztek folyóiratok, sőt megjelent már a permják szépirodalom képviselőinek első nemzedéke is, akik elsősorban lírai műfajokban alkottak (vö. Domokos 1985). A kezdeti évtizedekre jellemző volt egy többnyire fordításokat tartalmazó prózai szövegtípus, a propaganda célból készült tanmese, mely a későbbiek folyamán folytatás nélkül megszűnt.

A 30-as évek végén bekövetkező politikai megtorlások után 1945-től jelent meg a permják szépirodalom második nemzedéke, akik szintén verses műfajokban alkottak, bár publikáltak már néhány drámát is. E művek mellett folytatódtak a tankönyvkiadások, a propagandaszövegek, valamint a rendkívül gazdag permják folklór publikálása. Az 50-es években a korábbi szöveg-

típusokon kívül a permjáks szépirodalom harmadik generációjával egyre több regény, novella is kiadásra került. A 70-es évektől radikálisan csökkent az évente megjelenő kiadványok darabszáma, 1985-től (a peresztrojkával) pedig egyes szövegtípusok megszűntek. Ezt követően csupán az 2010-es évek törték meg az egyre ritkuló szövegközléseket, 2013-tól ugyanis számos, a permjáks nyelvet és kultúrát népszerűsítő VKontakte-oldal indult (melyeken a hosszabb-rövidebb posztok mellett igen gyakran amatőr szerzők versei is megjelennek), egyes folyóiratok online felületekre vándoroltak, a permjáks nyelvű tévéadásnak is lett saját weboldala, ahol a híreket tartalmazó videókat röviden írásban is összefoglalják, 2010-től pedig körvonalazódni kezdett a permjáks Wikipedia is. Csekély számú, jelenleg nem aktív blog[15] is elérhető már az interneten.

A felsorolt szövegtípusok kommunikációs célját felmérve a PermCorp tervezése során négy nagyobb alkorpuszt különítettünk el: az irodalmi, az információ átadására szolgáló, a gyerekeknek szánt, valamint a tudományos szövegeket tartalmazót. Következő lépésként megbecsültük az őket alkotó szövegtípusok relatív méretét, és ez alapján meghatároztuk, hogy az egyes alkorpuszok mekkora tokenméretűek lesznek (1. ábra).

Az irodalmi alkorpusz a szépirodalmi művek (lírai és prózai) mellett azokat a folklórszövegeket tartalmazza, melyek nem kimondottan gyerekeknek szólnak. Az információ átadására szolgáló alkorpuszba soroljuk az ismeretterjesztő szövegeket, a propagandaszövegeket, az újságcikkeket, hírösszefoglalókat, közösségi médiában megjelent tájékoztató jellegű bejegyzéseket, Wikipédia oldalakat. Speciális voltuk miatt külön alkorpuszba gyűjtjük a kisgyermekeknek készült szövegtípusokat (általános iskolai tankönyveket, meséket, mondókákat), illetve külön alkorpuszként határozzuk meg a tudományos szövegeket (szakkönyvek és egyetemi jegyzetek).

Mivel a korpusz mérete korpusznyelvészeti szempontból kicsi (300 000 token), úgy döntöttünk, hogy egy külön alkorpusz részeként elérhetővé teszünk olyan speciális forrásokat is, melyek nyelvészeti kutatások céljára készültek, vagyis nem számítanak autentikusnak. Erre azért van szükség véleményünk szerint, mert elképzelhető, hogy egyes fontos, de kimerítő jelleggel eddig le nem írt nyelvi jelenségek (pl. a különféle modalitások kifejezése és a különböző múlt idők és azok funkciói stb.) olyan ritkán jelennek meg a fentebb felsorolt szövegtípusokban, ami miatt korpuszalapú vizsgálatuk nem lenne megoldható. Ez az alkorpusz részben F. Gulyás Nikolett (2013, 2016) személytelen szerkezetek kutatása során gyűjtött elicitált mondataiból, illetve

[15] Például: http://permjak.blogspot.fi/

a projekt keretében permják anyanyelvű beszélőktől gyűjtött néhány oldalas irányított fogalmazásokból áll.

Mivel szándékunk szerint a korpusz a komi-permják teljes, nyilvánosságnak szánt írott változatát fogja reprezentálni, a korpusz szövegeit az 1920-as évektől napjainkig terjedő száz év kiadványaiból válogatjuk össze és két korszakba soroljuk. Korai nyelvmutatványoknak tekintjük a jelentős társadalmi változást hozó peresztrojkát megelőző, vagyis 1985 előtt született szövegeket, az ezt követően megjelent forrásokat pedig napjaink nyelvhasználati formájaként határozzuk meg.

A korai szövegek feldolgozásakor nehézséget jelent, hogy ezen kiadványok íráshagyománya sok esetben eltérő. Ez azt jelenti, hogy a ma is gyakorlatban lévő cirill írás mellett latin betűsre átírt szövegekkel is találkozhatunk, ezek viszont más és más átírási protokollt követnek. Nem ritka az sem, hogy a folklórműfajok esetén az azokat összegyűjtő kutató saját átírási rendszert alkalmazott. Érdemes itt megemlíteni, hogy az uráli nyelvek közül a magyar után a komi nyelvek rendelkeznek a legkorábbi nyelvemlékekkel. Írásbeliségük a 14. századra vezethető vissza, mely akkor még az úgynevezett abur ábécével történt. Az 1920-as években előfordult, hogy a latin és a cirill írást is alkalmaztak párhuzamosan.

Az általunk tervezett korpuszban ezt a fajta akadályt egy egységesített transzliterálás alkalmazásával szeretnénk kiküszöbölni, így a korpuszban való keresés rendkívül egyszerűen és gördülékenyen történhet, a különböző írásrendszerű szövegek a korpuszban közösen lesznek kereshetőek, miután egységes változatba konvertáljuk őket.

A korpusz felépítését az 1. ábrán olvasható arányok szerint tervezzük.

	%	token (db)
1. irodalmi szövegek	**40**	**120 000**
regények, novellák	30	36 000
versek	25	30 000
folklórszövegek	45	54 000
2. gyerekeknek készült szövegek	**30**	**90 000**
általános iskolai tankönyvek	50	45 000
mesék	40	36 000
gyermekversek, mondókák	10	9 000

3. internetes / információátadásra szánt szövegek	**20**	**60 000**
újságcikkek	60	36 000
ismeretterjesztő írások	20	12 000
Wikipedia-szócikkek	7	4 200
blogbejegyzések	6	3 600
közösségi média posztok	7	4 200
4. tudományos szövegek	**8**	**24 000**
publikációk	40	9 600
egyetemi jegyzetek	60	4 400
5. speciális szövegek	**2**	**6 000**
izolált mondatok	20	1 200
rövid fogalmazások	80	4 800

3.2.2. A korpusz szövegeinek annotálása

A digitalizálás, tisztítás, valamint a közös írásrendszerre alakítás után a kiválasztott szövegeket a projekt időtartama alatt többszintű címkézéssel látjuk el. A morfémaszintű glosszákon kívül szófaji annotációt (POS-tagging) is készítünk, illetve a mondatokat ellátjuk angol nyelvű fordítással is. A munkát a FLEx (FieldWorks Language Explorer) szoftver használatával végezzük majd.

A FLEx számos egyéb, dokumentációs nyelvészeti programmal (pl. az ELAN, EXMARaLDA) ellentétben nem igényel hanganyagot, vagyis gond nélkül címkézhetők általa írott szövegek is. Rendelkezik egy morfológiai elemzővel, mely egy kezdeti szótár, illetve egy nyelvtani vázlat segítségével hatékonyan előannotálja a szövegeket, megkönnyítve a címkézést végző nyelvészek dolgát.

Mivel a címkézést többen fogják végezni, az annotációs kézikönyvek összeállítása után tesztannotációt végzünk majd, vagyis a szövegek egy részét több nyelvész is fel fogja címkézni, így mérhetővé válik az egyetértési arány. Ez informálni fog arról, hogy a címkék fogalmai mennyire voltak pontosan, kimerítően definiálva, illetve láthatóvá válik majd az is, hogy mely kategóriák jelentenek problémát, melyek címkézése nehézkes. Amennyiben az egyetértési arány nem éri el az előzetesen meghatározott értéket, tovább pontosítjuk az annotációs sémát.

A komi-permják nyelvi helyzet ismeretében arra számítunk, hogy a szövegekben számos permják–orosz kódváltás fog megjelenni, a szövegek összeválogatásánál ezek kiküszöbölésére nem is törekszünk. A komik már a

11–13. században kereskedelmi kapcsolatban álltak a novgorodi fejedelemséggel, a 12. századtól területeik már Novgorodhoz tartoztak. A 14. századtól a Moszkvai Nagyfejedelemséghez csatolták a területet. Az azóta is fennálló szoros orosz kapcsolatok hatással voltak és vannak a permjákok (és más kisebb finnugor nyelvű népek) nyelvére (Hajdú – Domokos 1978: 190, Kuznetsov 2005: 58). Számos orosz lexikai és grammatikai elem honosult meg a permi nyelvekben az elmúlt évszázadok során, melyek közül egyes elemek hangtani, morfológiai és szintaktikai jellemzőikben igazodtak a permják nyelv grammatikai rendszeréhez, míg mások továbbra is az orosz mintát követik e tekintetben. Ezek között természetesen diakrón sorrend is felállítható, vagyis a komi-permják nyelvtani tulajdonságokat hordozó, orosz eredetű elemek értelemszerűen korábban jelentek meg a nyelvben, mint azok az orosz szavak, amelyek még tükrözik eredeti formájukat. A korpusz szöveganyagának szerkesztése során csupán azokat az orosz elemeket glosszázzuk majd, melyek fonetikai–fonológiai tulajdonságaik alapján ugyan orosz eredetűek, de a komi-permják nyelv hangtani tulajdonságaihoz már igazodtak, így például ragozásukat tekintve már komi-permják mintát mutatnak (*mijön bogatöś* [gazdag.pl], *sijön i radöś* 'minél gazdagabbak vagyunk, annál boldogabbak vagyunk'). Ugyanakkor nem glosszázzuk azokat az egyértelműen orosz elemeket és kódváltásszigeteket, melyek sem fonetikailag, sem morfológiailag nem igazodtak (még) a komi-permják nyelv struktúrájához.

4. Összefoglalás

Tanulmányunkban áttekintettük a PermCorp nevű komi-permják korpusz létrehozásának első fázisát, a munka során felmerült kérdéseket és az arra adott válaszainkat. Ismertettük, hogy a projekt létrehozását egyrészt a veszélyeztetett komi-permják nyelv alacsony dokumentáltsága, másfelől a nyelvre elérhető nyelvtechnológiai eszközök kis száma motiválta. Bemutattuk, hogy milyen korpuszok és egyéb digitális források, eszközök állnak most rendelkezésre a komi-permjákra vonatkozóan. Hangsúlyoztuk, hogy a jelenleg elérhető eszközökkel ellentétben a PermCorp egy manuálisan annotált, ún. gold sztenderd korpusz lesz, amelyet a különféle annotációkon (pl. interlineáris glosszák, POS-tagging) kívül angol nyelvű fordítással is ellátunk. A projekt jelenlegi szakaszában a korpuszba beépítendő szövegek körének kijelölése, valamint az annotálási elvek pontos meghatározása zajlik. Részletesen ismertettük a balanszálás kritériumait, valamint a felgyűjtött szövegek különféle típusait és azok arányát a korpuszon belül. Röviden kitértünk az annotáláshoz használni tervezett szoftver kiválasztására is. A projekt következő szakaszában a kijelölt szövegek egységes transzliterálását, valamint az irányított fo-

galmazások gyűjtését és feldolgozását fogjuk elvégezni. Reméljük, hogy a PermCorp a jövőben a finnugrisztikai és az általános nyelvészeti kutatások számára is hasznos eszköz lesz, amellyel hozzájárulhatunk a komi-permják szélesebb körű dokumentálásához.

Irodalom

ARKHANGELSKIY, TIMOFEY 2019: Corpora of social media in minority Uralic languages. In: Pirinen, Tommi A. – Kaalep, Heiki-Jaan – Tyers, Francis M. (eds), Proceedings of the fifth workshop on computational linguistics for Uralic languages, Tartu, January 7–8. University of Tartu, Tartu. 125–140.
http://komi-zyrian.web-corpora.net/index_en.html

ASZTALOS, ERIKA – F. GULYÁS, NIKOLETT – HORVÁTH, LAURA – TIMÁR, BOGÁTA 2021: New aspects in the study of Mari, Udmurt and Komi-Permyak: The typological database of the Volga Area Finno-Ugric languages. In: Szeverényi, Sándor (ed.), Uralic studies, languages, and researchers. Proceedings of the 5th Mikola Conference. Szeged, University of Szeged, Department of Altaic Studies. 255–274.
DOI: https://doi.org/10.14232/sua.2021.54.255-274

BAKER, ROBIN 1986: The role of animacy in Komi direct object marker selection. Ural-Altaische Jahrbücher. Neue Folge 6: 47–60.

BARTENS, RAIJA 2000: Permiläisten kielten rakenne ja kehitys. Mémoires de la Société Finno-Ougrienne 238. Suomalais-Ugrilainen Seura, Helsinki.

BATALOVA, R. M. [Баталова, Р. М.] 1975: Коми-пермяцкая диалектология. Наука, Москва.

BATALOVA, R. M. [Баталова, Р. М.] 2002: Кудымкарско-иньвенский диалект коми-пермяцкого языка. Mitteilungen der Societas Uralo-Altaica. Heft 23. Moskva – Groningen.

BIBER, DOUGLAS 1992: Representativeness in corpus design. In: Sampson, Geoffrey – Diana McCarthy (eds), Corpus linguistics: Readings in a widening perspective. Continuum, London. 174–197.

BLOKLAND, ROGIER 2023: Zyrian Komi. In: Abondolo, Daniel – Valijärvi, Riitta-Liisa (eds), The Uralic languages. Second edition. Routledge, London. 614–664.

BLOKLAND, ROGIER – CHUPROV, VASILIJ 2019: IKDP Spoken Komi Corpus, Korp Version (2019, January 01). [Dataset (Text corpus)]. Source: European language grid. https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0021_64F1_D

BLOKLAND, ROGIER – GERSTENBERGER, CIPRIAN – PARTANEN, NIKO – FEDINA, MARINA – RIEßLER, MICHAEL – WILBUR, JOSHUA 2015: Language documentation meets language technology. In: Pirinen, Tommi A. – Tyers, Francis M. – Trosterud, Trond (eds), Proceedings of the first international workshop on computational linguistics for Uralic languages. University of Tromsø, Tromsø. 8–18. DOI: https://doi.org/10.7557/5.3457

BLOKLAND, ROGIER – PARTANEN, NIKO – RIEßLER, MICHAEL 2021: langdoc/spoken-komi-corpus-vaszolyi: Spoken Komi Corpus: Erik Vászolyi (v0.1) [Data set]. Zenodo. https://zenodo.org/record/4591282

BORIN, LARS – FORSBERG, MARKUS – ROXENDAL, JOHAN 2012: Korp – the corpus infrastructure of Språkbanken. https://gtweb.uit.no/u_korp/?mode=koi#?lang=en

BRADLEY, JEREMY – KELLNER, ALEXANDRA – PARTANEN, NIKO 2018: Variation in word order in Permic and Mari varieties: a corpus-based investigation. Proceedings of the symposium ”Language contacts of the nations of Volga-Ural region”. Cheboksary, 21–24. 5. 2018.

DOMOKOS PÉTER 1985: A kisebb uráli népek irodalmának kialakulása. Akadémiai Kiadó, Budapest.

DRYER, MATTHEW S. – HASPELMATH, MARTIN (eds) 2013: WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

GATTO, MARISTELLA 2014: The web as corpus. Bloomsbury, London.

GERSTENBERGER, CIPRIAN – PARTANEN, NIKO – RIEßLER, MICHAEL – WILBUR, JOSHUA 2016: Utilizing language technology in the documentation of endangered Uralic languages. Northern European Journal of Language Technology 4: 29–47.

F. GULYÁS NIKOLETT 2013: Towards a classification of impersonal constructions in Komi: A functional-typological approach. In: Csepregi, Márta – Kubínyi, Kata – Sivonen, Jari (eds), Grammar and context: New approaches to the Uralic languages III. ELTE Finnugor Tanszék, Budapest. 31–49. https://edit.elte.hu/xmlui/handle/10831/9786

F. GULYÁS NIKOLETT 2016: Személytelen szerkezetek finnugor nyelvekben. Funkcionális és tipológiai megközelítés. ELTE Finnugor Tanszék, Budapest. (PhD-disszertáció, DOI: 10.15476/ELTE.2016.165)

HAJDÚ PÉTER – DOMOKOS PÉTER 1978: Uráli nyelvrokonaink. Tankönyvkiadó, Budapest.

HAUSENBERG, ANU-REET 1998: Komi. In: Abondolo, Daniel (ed.), The Uralic languages. Routledge, London – New York. 305–326.

HAVAS, FERENC – ASZTALOS, ERIKA – F. GULYÁS, NIKOLETT – HORVÁTH, LAURA – TIMÁR, BOGÁTA 2023: Typological database of the Volga area Finno-Ugric languages (VolgaTyp). ELTE Finnugor Tanszék, Budapest. volgatyp.elte.hu

HELIMSKI, EUGENE 2003: Areal groupings (Sprachbünde) within and across the borders of the Uralic language family: A survey. Nyelvtudományi Közlemények 100: 156–167.

KLUMPP, GERSON 2009: Variation in Komi object marking. In: Dufter, Andreas – Fleischer, Jürg – Seiler, Guido (eds), Describing and modelling variation in grammar. Mouton de Gruyter, Berlin – New York. 325–360.

KLUMPP, GERSON 2022: Permic: General introduction. In: Bakró-Nagy, Marianne – Laakso, Johanna – Skribnik, Elena (eds), The Oxford guide to the Uralic languages. Oxford University Press, Oxford. 471–486.

KUZNETSOV, NIKOLAJ 2005: A komik. In: Pusztay János (szerk.), A Volga–Káma-vidék finnugor népei. Szombathely. 55–72.

LEECH, GEOFFREY 2007: New resources or just better old ones? The Holy Grail of representativeness. In: Hundt, Marianne – Nesselhauf, Nadja – Biewer, Carolin (eds), Corpus linguistics and the web. Rodopi. 133–149.

LEWIS, PAUL M. – SIMONS, GARY F. – FENNIG, CHARLES D. (eds) 2015: Ethnologue: languages of the World. Eighteenth edition. SIL International, Dallas. http://www.ethnologue.com

LITKIN, VASZILIJ [Лыткин, Василий] 1952: Древнепермский язык: чтение текстов, грамматика, словарь. Издательство Академии Наук. СССР, Москва.

LITKIN, VASZILIJ [Лыткин, Василий] 1962: Коми-пермяцкий язык: Введение, фонетика, лексика и морфология. Коми-пермяцкое книжное издательство, Кудымкар.

LOBANOVA, ALEVITA [Лобанова, Алевита] 2017: Коми-пермяцкӧй кыв синтаксис. Кывтэчас да простӧй сёрникузя. ПГГПУ, Пермь.

MCENERY, TONY – WILSON, ANDREW 2001: Corpus linguistics. Edinburgh University Press, Edinburgh.

ORAVECZ, CSABA – VÁRADI, TAMÁS – SASS, BÁLINT 2014: The Hungarian Gigaword Corpus. Proceedings of the ninth international conference on language resources and evaluation (LREC' 14), Reykjavik, Iceland. European Language Resources Association (ELRA). 1719–1723.

PARTANEN, NIKO 2021: Old Komi: digital edition. Zenodo. https://github.com/langdoc/written-komi-corpus-old-komi

PONOMAREVA, L. G. [Пономарева, Л. Г.] 2002: Фонетика и морфология мысовско-лупьинского диалекта коми-пермяцкого языка. Удмуртский государственный университет, Ижевск. (Диссертация.)

PONOMAREVA, LARISA [Пономарева, Лариса] 2010: Komi-permják nyelvkönyv. Budapest. (Kézirat.)

PUSZTAY JÁNOS 2022: Az oroszországi 2020. évi népszámlálás uráli (finnugor) szempontból. Folia Uralica Debreceniensia 29: 129–138.

RÉDEI KÁROLY 1978: Chrestomathia Syrjaenica. Tankönyvkiadó, Budapest.

RUETER, JACK – PARTANEN, NIKO – PONOMAREVA, LARISA 2020: Comparative Permic Database (v1.0). Zenodo. https://doi.org/10.5281/zenodo.3596435

SIEGL, FLORIAN 2004: The 2nd past in the Permic languages. Form, function and a comparative analysis from a typological perspective. University of Tartu, Tartu. (MA Thesis.)

SINCLAIR, JOHN 1991: Corpus, concordance, collocation. Oxford University Press, Oxford.

SKRIBNIK, ELENA – KEHAYOV, PETAR 2018: Evidentials in Uralic languages. In: Aikhenvald, Alexandra Y. (ed.), The Oxford handbook of evidentiality. Oxford University Press, Oxford. 525–553.

SZABÓ, DITTA 2022: Az evidencialitás történeti-tipológiai vizsgálata az udmurtban. ELTE Finnugor Tanszék, Budapest. (PhD-disszertáció.)
WYNNE, MARTIN (ed.) 2005: Developing linguistic corpora: A guide to good practice. Oxbow Books, Oxford. https://users.ox.ac.uk/~martinw/dlc/

*

PermCorp: Towards the implementation of a Komi-Permyak corpus

In this paper we present the structure and the process of creating a new corpus of the Komi-Permyak language. The aim of the project is to collect the available literature in Permyak and to provide glossing and English translation for the texts. Since Komi-Permyak is an under-researched language, it is extremely important to produce a linguistically usable text collection that contributes to making the language more visible. The primary purpose of PermCorp is to represent the written version of the Permyak language in public use. In order to do this, we have collected texts of different genres, each of which will be represented by a sub corpus, and which have been balanced by text type. After the digitalization of the printed works and the transliteration, the texts will be labelled with multi-level tagging (glossing, POS tagging, English translation). During the project, we use the FLEx software, because its morphological analysis function with the dictionary and grammatical outline could help the researchers effectively in tagging.

Keywords: corpus, Komi-Permyak, FLEx, POS tagging, English translation

SZILVIA NÉMETH – DITTA SZABÓ – NIKOLETT F. GULYÁS