Egy komi-permják korpusz létrehozásának kihívásai: igék és melléknevek

Szabó Ditta,[1,2] F. Gulyás Nikolett,[1] Németh Szilvia[1]
[1]Eötvös Loránd Tudományegyetem, Finnugor Tanszék
[2]HUN-REN Nyelvtudományi Kutatóközpont

The paper reports on the current steps in the creation of a new corpus of written Komi-Permyak under grant number NKFIH FK 143242. The texts of the corpus are annotated by our research group using the FieldWorks Language Explorer (FLEx) software by pre-labelling sentences with the built-in general-purpose morphological parser of the software and manually checking the resulting machine-generated suggestions. As an output of the project, we will make the FLEx file available for other researchers to help them annotate their own texts. In order to use the general parser, we need to prepare it for the Komi-Permyak language system, i.e. we need to formalize Komi-Permyak morphology according to the needs of the software. After describing the basic principles of the parser, the paper presents some concrete examples of the challenges of this process in relation to adjectives and verbs.

Keywords: FieldWorks Language Explorer (FLEx), Komi-Permyak, corpus building, adjectives, verbs

Kulcsszavak: FieldWorks Language Explorer (FLEx), komi-permják, korpuszépítés, melléknevek, igék

1. Bevezetés

Tanulmányunk a *Komi-permják korpusz* elnevezésű, a komi-permják nyelv írott változatát tartalmazó korpusz létrehozásának jelenlegi fázisát mutatja be.[1] Írásunkban azokat a problémákat, általános, illetve specifikus kérdéseket tekintjük át, amelyek a projekt során, a nyelvtanépítéssel kapcsolatban merültek fel az annotáló rendszerben végzett munka közben. Mind a komi-permják nyelv általános dokumentáltsága, mind pedig az elérhető nyelvtechnológiai eszközök száma alacsonynak tekinthető több más finnugor nyelvvel (pl. az udmurttal vagy a mezei marival) összevetve. Kifejezetten

[1] Jelen tanulmány a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Fiatal kutatói kiválósági program (NKFIH FK 143242) keretében készült. A projektről további információ található a http://info.permcorp.elte.hu weboldalon. Köszönjük a tanulmány névtelen lektorainak hasznos észrevételeit. Minden esetleges hiba a szerzők kizárólagos felelőssége.

Nyelvtudományi Közlemények 120: 21–48.
DOI: 10.15776/NyK.2024.120.2

csekély az angol nyelven elérhető adatok mennyisége, ennek megfelelően a nyelv az általános nyelvészeti kutatásokban sem könnyen hozzáférhető. Projektünk így kettős célkitűzéssel rendelkezik: angol fordításokkal közölt szövegekből álló korpuszunk létrehozása egyrészt elősegíti, hogy a nemzetközi kutatásban jobban hozzáférhető legyen a komi-permják nyelv. Emellett pedig a projekt végén közzé fogjuk tenni az általunk használt nyelvtechnológiai eszköz (a FieldWorks Language Explorer, a továbbiakban FLEx) komi-permják nyelvre felkészített változatát is, amely reményeink szerint nemcsak a korpuszban nem szereplő komi-permják szövegeket glosszázni kívánó uralisták, hanem az anyanyelvi kutatók számára is hasznos lehet a jövőben.

A komi-permják az uráli nyelvcsalád permi ágához tartozik, Oroszország európai részén, a Permi és kisebb részben a Kirovi Területen beszélt, veszélyeztetett nyelv. Vitalitását tekintve az EGIDS-skálán 5-ös besorolású, ami azt jelenti, hogy a generációk közötti nyelvátadás nem minden esetben történik meg (Lewis et al. 2015). A legújabb, 2020-as (2021-ben publikált) oroszországi népszámlálás szerint a komi-permjákok lélekszáma 55 786 fő volt, ami 41%-os csökkenést jelent a 2010-es adatokhoz viszonyítva (Pusztay 2022: 131), ezen belül körülbelül 44 000 fő vallotta magát a nyelv anyanyelvi beszélőjének (PEREPIS 2020).

A nyelvváltozat státuszával kapcsolatban hosszú idő óta nincs egyetértés a kutatók körében. A komi-permjákot többen a komi nyelv egyik nyelvjárásának tekintik (Bartens 2000: 9; Rédei 1978: 37). Hausenberg (1998: 305) szerint a komi nyelvnek három nyelvjárása van: az északi (komi-zürjén), a déli (komi-permják), továbbá a keleti (jazvai) nyelvjárás. A nyelvjárásként való meghatározás melletti érvként gyakran említik, hogy a változatok közötti különbségek nem jelentősek, és elsősorban a lexikont és – kisebb részben – a fonológiát érintik (pl. Rédei 1978: 37–42). Az újabb szakirodalom is általában ehhez hasonló felfogást követ (ld. például Klumpp 2022: 471–474).

Ezzel szemben számos oroszországi, elsősorban anyanyelvi kutató szerint a komi-permják önálló nyelv (vö. Batalova 1975: 5, 2002: 7; Ponomarjova 2002, 2010), mivel a beszélők annak tekintik. A komi-permjákok és a komi-zürjének már a cári időktől kezdve adminisztratív szempontból elkülönülten éltek. 1925-ben létrehozták a Komi-Permják Autonóm Körzetet, amelyben a komi-permjákok az oroszországi finnugor nyelvek beszélői körében egyedülálló módon 60%-os többséget alkottak 2005-ig, amikor a területet beolvasztották a Permi Területbe. A zürjénnek és a permjáknak is van önálló irodalmi változata, noha a komi-permják irodalmi nyelv kialakításában nagy hatása volt a zürjénnek (Antal 2023: 45). Tanulmá-

nyunkban Batalova (1975) nyomán a komi-permjákra nyelvként hivatkozunk.

A komi-permják további nyelvjárásokra tagolódik, ezek az északi, a déli és a felső-kámai (Batalova 1975: 3; Bartens 2000: 31–32). Az irodalmi nyelv a déli, Kudimkar környéki nyelvváltozaton alapul. Az irodalmi változatba – a komi-zürjén mintájára – beemelték az ún. *v* ~ *l* morfológiai alternációt, ami eredetileg erre a nyelvváltozatra nem, csak az északira volt jellemző (Klumpp 2022: 473). Bár létezik sztenderdizált irodalmi nyelv, valamint van lehetőség a nyelv kis óraszámban történő tanulására (Zamyatin 2022: 79), a nyelvhasználat elsősorban a családi színtérre korlátozódik. Mivel a nyelvnek nincs hivatalos nyelvi státusza, a jövőbeni kilátásai kifejezetten negatívak.

A komi-permják nyelvű kiadványok száma alacsony (Zamyatin 2022: 86), a nyelvre elérhető nyelvtechnológiai eszközök száma pedig, ha növekszik is (Eberhard et al. 2024), még mindig nem túl magas. Ugyanez érvényes az udmurttal vagy a komi-zürjénnel való összevetésben is, például bár születtek már nagyobb szószámú szövegkorpuszok is, ezek pusztán automatikus annotációval vannak ellátva, ami (többek között a tanulmányban részletezett problémák miatt) azt eredményezi, hogy igen magas a hamis találatok aránya, jelentős korlátozásokkal használhatók. Az angolul is hozzáférhető anyagok köre pedig még ehhez képest is minimálisnak mondható. Jól érzékelteti a komi-permjákra vonatkozó, a nemzetközi kutatás számára angolul is hozzáférhető adatok szűkösségének mértékét az a tény, hogy a World Atlas of Language Structures nevű tipológiai adatbázisban csak 24 szerkezeti jellemzője szerepel, ez a szám a komi-zürjén esetében 51, a hantiéban 72, a magyarében pedig 155 (Dryer – Haspelmath 2013). Örvendetes, hogy a közelmúltban több olyan tipológiai adatbázis (Norvik et al. 2022; Havas et al. 2023; Skirgård et al. 2023) is napvilágot látott, amelyben a komi-permják adatok már nagyobb számban vannak képviselve.

Tanulmányunk a következőképpen épül fel: a bevezető rész (1) után ismertetjük a kis nyelvek korpuszépítésének gyakorlati problémáit, konkrétan az annotáláskor használandó morfológiai elemző kiválasztását (2. rész). A 3. részben bemutatjuk az általunk használt FLEx szoftver morfológiai elemzőjét, valamint a használata során felmerült kérdéseket. Ezek után két rövid esettanulmányt közlünk, hogy illusztráljuk a korpuszépítés során felmerülő egyedi és átfogóbb problémákat. A 4. részben a melléknévi, az 5. részben pedig az igei kategória morfológiai elemzésének problémáit ismertetjük. Végül pedig összegezzük eredményeinket és bemutatjuk a kutatás további irányait (6. rész).

2. A projekt során használt morfológiai elemző kiválasztása

Ugyan a komi-permják egyike az uráli nyelvcsalád legkevésbé dokumentált nyelveinek, az utóbbi években mégis egymással párhuzamosan egyszerre több szövegkorpusz is készült, hogy megkönnyítse a nyelv kutathatóságát. Ezek mindegyike elsődlegesen nem nyelvészeti vizsgálat céljára készült szövegekből álló digitális gyűjtemény (vö. Németh et al. 2023: 185–189), és mind tartalmaz valamilyen annotációt, melyeket elemzőszoftverek használatával automatikusan állítottak elő[2] (1. táblázat). Ez egy bevett gyakorlat a nagy beszélőközösséggel rendelkező nyelvekre készült nagy szószámú korpuszok esetében, mivel bár manuális ellenőrzés hiányában az eredmény elkerülhetetlenül nagy számú félreglosszázást és többértelmű címkéket is tartalmaz, ezt a korpusz nagy mérete mennyiségileg kompenzálja. A felhasználó ráadásul – akiről a korpuszalkotók implicite azt feltételezik, hogy magas szintű tudással rendelkezik az adott nyelvről – könnyedén kiszűrheti a hibás adatokat. A már létező komi-permják korpuszok csupán egy része tartalmaz mondatfordítást is a nyelvtani annotációkon, glosszákon kívül, és ezek minden esetben oroszul vannak, angol metanyelvvel nem találkozhatunk.

Kutatócsoportunk ezeknek a korpuszoknak az előnyeit szem előtt tartva – és tanulva a hibáikból – egy olyan új korpusz építésén dolgozik, mely kutatók szélesebb körének szól. Az automatikus annotációk manuális ellenőrzésével elérhető magasabb fokú pontosság, valamint az angol fordítások miatt reményeink szerint az alacsonyabb komi-permják (vagy orosz) nyelvi kompetenciával rendelkező nyelvészek is tudják majd értelmezni az adatokat, ezzel pedig ösztönözhetjük a komi-permják nyelv bevonását az Oroszországon kívüli nyelvoktatásba és a klasszikus uralisztikán kívül eső, például tipológiai célú vizsgálatokba is.[3]

A projektünk célja emellett az is, hogy egy olyan digitális eszközt adjunk más kutatók kezébe, amivel azok könnyebben dolgozhatják fel a saját maguk által gyűjtött vagy a kutatási céljaiknak a korpuszéinál jobban megfelelő komi-permják szövegeket, vagyis egy olyan szoftvert, ami automatikusan előannotálja a szövegeket, és amelynek kimenetét a felhasználó könnyedén javíthatja, pontosíthatja.

---

[2] Léteznek annotáció nélküli komi-permják szöveggyűjtemények is, például a Turku Komi-Permyak Corpus (https://finno-ugric-corpora.utu.fi/cqpweb/), vagy a Finn Nemzeti Könyvtár Fenno-Ugrica nevű digitális gyűjteménye (https://fenno-ugrica.kansalliskirjasto.fi/).

[3] A készülő korpusz szerkezetéről bővebben ld. Németh et al. (2023).

1. táblázat: A komi-permják nyelvű korpuszok

korpusz neve	a szövegek keletkezése	szövegek típusa	tokenszám	annotáció típusa és módja
Korp (Borin et al. 2012)	1999–2020	Wikipedia-szócikkek	241 614	szófaj, glossza, függőségi relációk (automatikus)
Коми Кыв Корпус – Перем Коми Юкӧн (2021)	1921–2022	szépirodalom (verses és prózai műfajok), folklór, ismeretterjesztő és tudományos írások, bulvár, egyéb	6 196 963	glossza (automatikus), orosz fordítás
Four Battles Corpus	1940	szépirodalom (prózai műfajok)	kb. 1900	glossza, mondattani szerepek (automatikus), orosz fordítás

Erre az eszközre a korpusz szövegeinek annotálása során nekünk is szükségünk van (hiszen 300 ezer tokent teljesen manuálisan annotálni[4] túl hosszú idő lenne, gépi előannotálással azonban jelentősen felgyorsítható a folyamat), létrehozásának pedig két lehetséges útja van: vagy készítünk egy saját morfológiai elemzőt (melyet felhasználóbarát felülettel ellátva publikálhatunk), vagy választunk egy hasonló célra készített, folyamatosan fejlesztés alatt álló szoftvert, melyet felkészítünk a komi-permják nyelvrendszerre (és ezt a beépített nyelvtant tartalmazó állományt publikáljuk). A kutatócsoport már a pályázat beadásakor az utóbbi megoldás mellett döntött, mert egy már kész program használatával 1) az adott szoftvert már amúgy is használó nyelvészek minimális energiabefektetéssel tudják alkalmazni az eszközünket, 2) az adott szoftvert még nem használó nyelvészeknek nem szükséges felhasználói kézikönyveket szerkeszteni, hiszen ezt a szoftver készítői elvégezték, 3) nem szükséges a morfológiai elemzőnkhöz felhasználói felületet programozni, 4) mentesülünk a komplexebb szoftverek fenntartásával járó fejlesztői munkálatok alól.

[4] A tanulmányban annotációnak nevezünk minden olyan címkézést, melyet a nyelvi anyaghoz társítunk, a készülő korpuszunk esetében ezek a glosszák, a szó- és mondatjelentések, valamint a szófaji kategóriák.

Az elmúlt évtizedekben számos olyan szoftver vált elérhetővé, mely aluldokumentált nyelvek szövegeinek feldolgozására készült, ezek többsége azonban egészen pontosan a nyelvi dokumentációt hivatott megkönnyíteni, így kizárólag hang- és/vagy videófelvételeket lehet szegmentálni, fordítani és glosszázni velük (ELAN, EXMARaLDA). Ezek a programok nem tartalmaznak általános célú morfológiai elemzőt, amit aktualizálhatnánk a komi-permják nyelvtanra, így bizonyos egyszerűbb automatizmusok kivételével kizárólag teljesen manuálisan lehet annotálni velük, ami azt jelenti, hogy az annotáló minden egyes szólakot manuálisan bont összetevőkre, és manuálisan adja meg (esetleg egy listából kiválasztva) a címkéket. Mivel a korpuszunk kizárólag eredetileg is írásban közölt szövegeket tartalmaz, és kifejezett igényünk volt morfológiai elemzőre is, ezeket a szoftvereket nem tudjuk használni, a FLEx azonban kiválóan megfelel a céljainknak.

A FLEx egy nonprofit alapítvány (SIL) által készített és jelenleg is aktívan fejlesztett szoftver, melynek elsődleges célja, hogy elősegítse a kis beszélőszámú nyelvek leírását, és kimondottan az új szótárak összeállításának és publikálásának folyamatát. A felületén keresztül két különböző morfológiai elemző (*parser*)[5] is használható, egy fonológiai alapokon dolgozó, még kísérleti fázisban álló, illetve egy hagyományos elemző. A kutatócsoport az utóbbit használja, így a tanulmány további részében kizárólag ezt értjük az elemző alatt. A szoftver újabb verziói már azt is lehetővé teszik, hogy több annotátor párhuzamosan dolgozzon ugyanazon az állományon.

A szövegek FLEx-ben történő feldolgozásának két alapvető megközelítésmódja van. Az egyik, hogy a felhasználó a bevitt szövegeket mondatról mondatra, szóalakról szóalakra haladva morfémákra szegmentálja és egyenként megadja a jelentéseiket, vagyis tulajdonképpen a glosszázással felépít egy szótárat, mely mind a szabad, mind a kötött morfémákat tartalmazza. A másik mód, hogy a felhasználó előzetesen importál a szótárba egy sor lexémát, emellett pedig többlépcsős szabályrendszerek segítségével felépíti a nyelvtant is, vagyis a kötött morfémáknak nem csupán a jelentését adja meg, hanem használatuknak szabályszerűségeit is (pl. azt,

[5] Bár a nyelvtechnológiában a *parser* terminus többnyire szintaktikai (vagy legalább a szöveg szavainak sorrendiségét is figyelembe vevő) elemző szoftvert jelent, a tanulmány során mégis ezt használjuk a FLEx morfológiai elemzőjére azon oknál fogva, hogy a FLEx dokumentációja és felülete is így hívja az eszközt. Úgy véljük, hogy ha nem is szerencsés a terminológia, ezzel nem zavarjuk össze a nyelvtechnológiai háttérrel nem rendelkező, de a FLEx iránt érdeklődő olvasókat.

hogy egy affixum milyen szófajú szóhoz kapcsolódhat, milyen változások következnek be a szótőben a hatására, milyen alakváltozatai vannak stb.). Utóbbi munkafolyamat előnye, hogy a megadott információk alapján az eredetileg általános célú, de ily módon az adott nyelvre aktualizált morfológiai elemző képes azonnal előannotálni az összes szöveget. Az előannotáció kimenetét a manuális ellenőrzés során jóvá lehet hagyni (el lehet fogadni), ki lehet javítani, a hiányzó elemzést lehet pótolni, illetve egy esetlegesen többjelentésű elemnél ki lehet választani az aktuálisan érvényes jelentést. A FLEx-ben történő automatikus elemzés egyik fontos korlátja, hogy ha a szótár nem tartalmaz egy lexémát, akkor a szoftver nem fog javaslatot tenni a toldalékolt alakjára, vagyis nem fogja leszegmentálni róla a szótárban esetleg már szereplő kötött morfémákat sem. Mivel azonban a FLEx nem csupán előzetesen felépített nyelvtannal használható, lehetőség van arra is, hogy bármikor újrafuttassuk az elemzőt, így a szótár bővülésével a parser egyre több lexéma toldalékolt alakját fogja felismerni a feldolgozott szövegek számának növekedésével párhuzamosan.

Mindez azt jelenti, hogy a komi-permják nyelvtan FLEx-ben való felépítésekor alkalmazkodnunk kell a szoftver morfológiai elemzőjéhez, és ennek megfelelően szükséges formalizálnunk a nyelvi rendszert. A formalizálás során alkalmazott elvek egy része bármilyen egyéb morfológiai elemző készítésekor megjelenne, néhány azonban a FLEx sajátossága. A tanulmány további részében nagy vonalakban áttekintjük a FLEx morfológiai elemzőjének fontosabb jellemzőit, majd az igék és a melléknevek kapcsán felmerült egyes problémák ismertetésén keresztül bemutatjuk, hogy mennyire komplex feladat ez.

3. A FLEx morfológiai elemzőjének alapelvei

Mivel a FLEx morfológiai elemzője arra készült, hogy mind agglutináló, mind flektáló nyelvekre használható legyen, így kezeli többek közt a ragozást, a szóképzést (képzőkkel és szóösszetételekkel is), az epentézist, a több elemből álló affixumokat (pl. cirkumfixumok), az infixációt és a reduplikációt is. Bár van lehetőség zéró morfémák megadására is, ezek lassíthatják az elemzést.

Ennek a flexibilitásnak az a hátránya, hogy a használati szabályok megadása egy bonyolult, többlépcsős struktúrán keresztül lehetséges még akkor is, ha az elemző szempontjából komplexebb problémák (pl. magánhangzó-harmónia) nem merülnek fel a komi-permjákkal kapcsolatban. Mivel azonban a felépített nyelvtan bármikor szerkeszthető, az esetleges hibák korrigálása az annotálás bármely pontján lehetséges. Az általunk FLEx-be épített formalizált nyelvtan minősége a projekt végeredményeként létrejövő

korpuszon nem fog látszani (hiszen manuálisan ellenőrzünk minden szóalakot), de mind a saját munkánk megkönnyítése, mind a projekt végén nyilvánosan elérhetővé váló FLEx állomány miatt a lehető legpontosabbnak kell lennie.

Az elemző működését és a nyelvtanépítés módját leíró legteljesebb dokumentáció a FLEx súgóján keresztül érhető el, ami példákkal illusztrálja néhány komplex probléma formalizálásának módját, vagyis a szabályok esetenként több szinten történő definiálását.

A FLEx parser célja megegyezik más morfológiai elemzőkével: 1) ellenőrizni, hogy a szóalak egy létező szó-e a szótár és a megadott szabályok alapján, 2) meghúzni a morfémák határát a szóalakon belül, és 3) jelentést vagy nyelvtani kategóriacímkét társítani az egyes morfémákhoz. Ennek érdekében az elemzőnek ismernie kell, 1) hogy az adott morfémának milyen alakjai elfogadhatóak (mi a szótöve, vannak-e helyesírási, nyelvjárási vagy egyéb változatai, amiket ugyanazon szóként szeretnénk azonosítani, valamint hogyan viselkedik szóösszetételekben), 2) a morfotaktikai szabályokat, vagyis hogy egy jólformált szóalakon mely kötött morfémák jelenhetnek meg. Amennyiben több toldalék is előfordulhat együtt, akkor egészen pontosan melyek ezek, és milyen relatív sorrend figyelhető meg a szótőhöz és a többi kötött morfémához viszonyítva.

A FLEx már az importálás során mondatokra bontja a szöveget a központozás alapján, de a parser az elemzés alatt csakis egy-egy szóalakot vizsgál: ellenőrzi, hogy található-e a szótárban olyan lexéma, melynek a megadott szabályokkal létrejöhet olyan alakja, amilyen a kérdéses szóalak. Amennyiben igen, szegmentálja a szóalakot, és egy legördülő menüben felsorolja az összes glosszázási lehetőséget. Jelentős gyakorlati korlátnak számít, hogy ha a szótár nem tartalmaz egy lexémát, akkor a szoftver nem fogja leválasztani róla a szótárban esetleg már szereplő kötött morfémákat sem, illetve hogy a több szóból álló szerkezeteket (pl. összetett múlt időket) nem képes felismerni.

Azt, hogy melyik szó milyen toldalékokat vehet fel, és ezek milyen sorrendben állhatnak, különféle szabályokkal lehet megadni. Ennek során egymásra épülő módokon külön-külön beállíthatók a nyelvekben meglévő betűk (a több karakterből állók is), szófajok, tőtípusok, speciális hangkörnyezetek (melyek pl. tőbelseji változásokat idéznek elő, vagy éppen a több karakterből álló betűk kettőzését definiálják), ragozási kategóriák (vagyis azok a grammatikai kategóriák, amelyek morfológiai eszközökkel kifejezhetők a nyelvben) és a szóösszetétel módja (a morfológiai és a szintaktikai összetételek szabályai).

A nyelvtan formalizálása során nem csupán a nyelvleírásban használt kategóriákat és azok rendszerét kell figyelembe venni, hanem időnként alternatív csoportosításra is szükség van. Például minden kötött morféma definiálásakor meg lehet adni, hogy az milyen kategóriájú szavakhoz járulhat. A kategória elsősorban azt jelenti, hogy milyen szófajhoz kapcsolódhat, és ha ezt megadjuk, akkor a parser jelentősen kevesebb szegmentálási hibát ejt és kevesebb téves elemzést ad. Például abban az esetben, ha a szótárban benne van a *fagy* főnév, a *fagy* ige és az *-Ás* képző anélkül, hogy szófaji kategóriához rendeltük volna, a parser a *fagyás* szóalakra felajánlja a főnév+képző glosszát is az ige+képzőn kívül, de ha megadjuk, hogy csak igéhez kapcsolódhat, akkor csak az ige+képző elemzést ajánlja fel. Noha a kategória elsősorban azt jelenti, hogy az adott morféma milyen szófajhoz kapcsolódhat, ez nem minden esetben igaz, ugyanis egy-egy morfémához csak egy kategória rendelhető, ezért a szófaji kategóriák hierarchiáját úgy kell összeállítani, hogy legyen egy olyan főkategória, amely megadásával az alatta található összes szófaj hozzárendelhető egy toldalékhoz. Emiatt a sajátosság miatt előfordulhat, hogy szavak egy csoportját nem a nyelvleírásban megszokott szófajok alapján szükséges elkülöníteni, hanem külön kvázi-szófajként kell beilleszteni a szófajok rendszerébe. Ilyen például, hogy a komi-permjákban (a magyarhoz hasonlóan) a melléknév számos főnévi végződést is felvehet, emiatt létre kell hozni egy olyan (a valóságban nem létező) felettes kategóriát, mely a melléknevet és a főnevet is magában foglalja (de pl. a mutató névmásokat nem, mert azok nem azonos elvek alapján toldalékolhatók). Ugyancsak ilyen probléma, amikor például az igék egy csoportja a többi igétől eltérő paradigmának megfelelően toldalékolódik (pl. történeti okokból), mert ilyenkor egy külön kategóriát (egy kvázi szófajt) kell létrehozni a nyelvtanépítés során és ehhez rendelni a kérdéses igéket, valamint a rajtuk megjelenő kötött morfémákat. Esetenként a kategorizációval sehogy sem tudjuk lefedni a valódi helyzetet, nem alakítható ki minden esetben jól működő hierarchia, ilyenkor értelemszerűen azt a részben működő kategóriarendszert érdemes választani, ami a legtöbb helyes elemzést eredményezi.

Egy szóalak jólformáltságának megállapításához azonban még távolról sem elég az affixumok szófaji kötöttségének megadása, hiszen egy meghatározott szófajú szótövön több toldalék is megjelenhet egyszerre, ám nem tetszőleges sorrendben. A FLEx-ben a morfotaktikai szabályokat az egyes szófajoknál lehet megadni sémák (template) formájában. Ehhez azonban az szükséges, hogy a kötött morfémák felvételekor megadjuk, hogy a végződés a jólformált szóalakon belül a szótőhöz (stem) képest hányadik morfémapozíciót (slot) töltheti be. Ennek megadását csak több lépcsőben

tudjuk kivitelezni. Először is előzetesen csoportosítanunk kell az affixumokat morfotaktikai viselkedésük alapján, és létre kell hozni a FLEx-ben is a csoportokat, körültekintően elnevezve őket, mivel az egyes affixumok szótárba vételekor csak a csoport nevét tudjuk megadni. A következő lépés, hogy amikor definiáljuk a szófajok jól formált szóalakjainak sorrendjét tartalmazó sémát, az egyes morfémapozícióknál megadjuk, hogy melyik morfémacsoport jelenhet meg az adott pozícióban. Ha több olyan morfémacsoport is van, ami megjelenhet például közvetlenül a szótő utáni első pozícióban, akkor több sémát szükséges létrehoznunk, mivel egy sémában morfémapozíciónként csupán egy csoport megadása lehetséges (erről a 4. és 5. pontban részletesebben is szó lesz).

A fentiekben bemutatottakhoz hasonló, egymásra épülő szabályrendszer egyszerre rugalmassá és bonyolulttá teszi a FLEx-es nyelvtan felépítését, és egyben rá is kényszeríti a korpuszépítőt, hogy olyan jelenségekkel is behatóbban kezdjen foglalkozni, amivel korábban a nyelvi rendszer leírásának szempontjából kielégítően, gyakorlati szempontból azonban kevésbé alaposan foglalkozott a szakirodalom. A tanulmány további részében néhány ilyen kérdéses pontot fogunk bemutatni a komi-permják melléknevekkel és igékkel kapcsolatban.

4. A melléknevekkel kapcsolatos kérdések a FLEx szempontjából

Egy, a korpuszépítés szempontjából számos kérdést felvető téma a komi-permják melléknevek kategoriális státusza volt. Bár nyelvészeti szempontból sokat megtudhatunk a melléknevekről a szakirodalom alapján (Bartens 2000: 130–141; Batalova 1975: 166–173; Rédei 1978: 80–84), ezek az információk gyakran nem elégségesek a korpuszépítés szempontjából. Az alábbiakban először áttekintjük, hogy a komi-permják mellékneveknek milyen, a szakirodalomban is tárgyalt grammatikai tulajdonságai vannak, mi szolgáltatja az alapot a FLEx nyelvtanának melléknevekre koncentráló részéhez.

Morfológiai értelemben a melléknévként jelölt szavak nem főnevek, tehát morfológiai viselkedésük eltér a főnevektől abban az értelemben, hogy nem jelölik az eset kategóriáját (Batalova 2002: 63–76; F. Gulyás 2023a). Emellett a melléknevek, akárcsak a főnevek, többes számba tehetők, esetükben a többes szám jele az *-öś* (1) (Batalova 1975: 166).

(1) *jonyďik-öś*
fehér-PL
'fehérek' (Batalova 1975: 166)

A melléknevek ugyanakkor jelzői pozícióban általában nem egyeztetődnek számban és esetben az általuk módosított főnévvel, de (feltehetően orosz

hatásra) elvétve találkozhatunk számban egyeztetett melléknévvel főnév mellett (2) (Bartens 2000: 130; F. Gulyás 2023b).

(2) *basök-öś* *źoriź-źez*
szép-PL virág-PL
'szép virágok' (Ponomarjova 2010: 52)

Állítmányként a melléknév mindig egyeztetődik számban az alannyal, tehát a többes szám jele tipikusan az állítmányként megjelenő mellékneveken tűnik fel (3) (Rédei 1978: 81).

(3) *Kyčöm nija* *basök-öś!*
milyen 3PL szép-PL
'Milyen szépek!' (Komi Kyv Korpus)

A komi-permják melléknév fokozható, mely során a középfokot egy szuffixummal, a felsőfokot egy partikulával fejezik ki. A középfok a *-žyk/-žyk* szuffixummal képezhető (4), ugyanakkor ez a toldalék nemcsak melléknevekhez, hanem igékhez, mennyiségjelzőkhöz és tagadószóhoz is járulhat (Batalova 1975: 167). A felsőfokot a *med* partikula fejezi ki, mely írásban megjelenhet egybeírva a melléknévvel (5) vagy különálló szóként is a melléknév előtt (6) (Bartens 2000: 138; Batalova 1975: 169).

(4) *pöriś-žyk*
öreg-COMP
'öregebb' (Batalova 1975: 167)

(5) *med-bur* *jort-tez*
SUPL-jó barát-PL
'legjobb barátok' (Komi Kyv Korpus)

(6) *med* *bur*
SUPL jó
'legjobb' (Rédei 1978: 82)

A melléknévből az *-a* szuffixum képez határozószót, ami közvetlenül a szótő után is állhat (7), de megelőzheti a komparatívusz jele is (8) (Batalova 1975: 167).

(7) *bur-a*
jó-ADV
'jól' (Bartens 2000: 140)

(8) *kuź-žyk-a*
hosszú-COMP-ADV
‘hosszabban’ (Batalova 1975: 167)

A fent említett grammatikai tulajdonságok ismerete mind elengedhetetlen a FLEx grammatikai fejlesztéséhez, ugyanakkor ennyi információ még nem elegendő ahhoz, hogy a FLEx képes legyen annotálni és glosszázni a mellékneveket. A szükséges adatok akár nyelvészeti kérdésként is felfoghatók és relevánsak lennének, mégsem kerülnek tárgyalásra a nyelvtanokban és kézikönyvekben. Itt nem másról, mint az említett morfémák lehetséges kombinációiról és azok sorrendjéről van szó. Amint láthattuk, a melléknevekkel megjelenő partikulákat és szuffixumokat külön-külön tárgyalják a grammatikák, az esetleges együttes előfordulásra csak a példákból következtethetünk.

A FLEx szempontjából ugyanakkor fontos információ, hogy a melléknéveken potenciálisan megjelenő affixumok milyen sorrendben és milyen szabályok szerint kombinálhatók a nyelvben. Ennek meghatározásához kimerítő nyelvleírás hiányában korpuszalapú vizsgálatra volt szükség, melynek eredményeit a 2. táblázat mutatja be. A *Komi Kyv Korpus – Perem Komi Jukön*[6] egy közel 7 millió tokent tartalmazó korpusz, mely jelentősen felülmúlja más működő komi-permják korpuszok anyagának nagyságát. Használhatóságát tekintve viszont komoly akadályokkal néz szembe a mindenkori kutató, hiszen a szövegek eredeti, cirill betűs képe nem rendelkezik egységesített átírással, sem pedig angol fordítással, orosz fordítás és glosszázás viszont helyenként elérhető (ld. a 2. fejezetben foglaltakat). A keresés is kizárólag morfémákra lehetséges, ezen kívül korpuszra történő szűrést és szerzői, valamint évszámbeli konkretizációt alkalmazhatunk, vagyis nyelvtani kategóriacímkékre és szófajra, vagy egyszerre több morfémára nem kereshetünk. Így a korpusz használata nehézkes és rendkívül lassú, ezért igazán nem is hatékony, és a nyelvészek jelentős része számára, alacsony komi-permják nyelvi kompetenciával vagy annak teljes hiányában, nem lehetséges.

A melléknevekhez kapcsolható affixumok esetében két keresési módot alkalmazhatunk. Az egyik során vagy csak a partikulára, vagy csak a szuffixumra keresünk, a másik megoldást követve egy-egy konkrét (lehetőleg gyakori) melléknév nominatívuszi vagy a potenciális toldalékokkal ellátott alakjaira kereshetünk, ám az utóbbi gyakorlat jelentősen csökkentheti a találatok számát.

[6] https://p.komicorpora.ru/

Az itt bemutatott eredmények az első típus alkalmazásával, vagyis meghatározott affixumok keresésével születtek. A találatokat úgy lehet pontosítani, hogy vagy egy konkrét szóalakra, vagy szókezdő, illetve szóvégi pozícióra keresünk. A korpuszban a következő affixumokat és azok kombinálhatóságát vizsgáltuk meg: a komparatívusz jelét (*-žyk/-žyk*), a melléknévi többes szám jelét (*-öś*), a szuperlatívusz jelét (*med(-)*), az adverbiális képzőjét (*-a*), valamint az egyéb ismert, de ritkábban előforduló képzők közül az intenzifikáló képzőt (*-öv*), a kicsinyítő (*-yńik*) és a becéző képzőt (*-ik*). Az egy szóalakon egyszerre megjelenő két toldalék előfordulási lehetőségeit a 2. táblázat szemlélteti. Egy adott melléknévi affixum egy szóalakon belül nem ismétlődhet, így ez az eshetőség a táblázatban mindenhol X-szel van jelölve.

2. táblázat: Két affixum kombinálhatósága a mellékneveken[7]

SLOT1 → SLOT2 ↓	COMP	AUG	DIM (*-yńik*)	DIM (*-ik*)	ADV	PL	SUPL
COMP	✗	✓	✗	✗	✓	✗	✗
AUG	✗	✗	✗	✗	✓	✗	✗
DIM (*-yńik*)	✗	✗	✗	✗	✗	✗	✗
DIM (*-ik*)	✗	✗	✗	✗	✗	✗	✗
ADV	✓	✗	✗	✗	✗	✗	✓
PL	✓	✓	✓	✓	✓	✗	✓
SUPL	✗	✗	✗	✗	✗	✗	✗

A FLEx szempontjából a melléknevekkel kapcsolatos fenti nyelvtani jelenségeket a következő lépésekben szükséges formalizálni. Mikor hozzáadjuk, a program a melléknevek szótári alakját elraktározza a lexikonban. Miután

[7] A SLOT1 a szótövet közvetlenül követő, a SLOT2 pedig az azt követő morfémapozíciót jelöli.

a FLEx számára definiáltuk, hogy a *bur* szó 'jó'-t jelent, a bevitt szövegekben fel fogja ismerni, és a glosszasorba kiírja a *bur* szó alá, hogy 'jó'. A FLEx ezen kívül már a szófaját is tudja, hiszen a lexikonba való felvitel során azt is megadtuk neki, hogy ez egy melléknév. Fentebb már említettük, hogy a mellékneveken megjelenő toldalékok sorrendjét és kombinálhatóságát is tudnunk kell a sikeres nyelvtan felépítéséhez, de még mielőtt elkezdenénk a FLEx-nek felsorolni ezeket a lehetséges, sorrend-specifikált párokat, meg kell határoznunk, hogy maximálisan hány darab toldalék jelenhet meg az adott POS-taggel, vagyis szófaji meghatározással ellátott szón. A POS-tagnek azért van itt kiemelten fontos szerepe, mert ahogy lejjebb látni fogjuk, a melléknevek képesek főnevesülni, s ilyenkor más toldalékokat is felvehetnek, mint melléknévi szerepükben. A FLEx-ben épített nyelvtanunk viszont továbbra is melléknévként fogja azonosítani ezeket a szavakat, s ebből a szempontból irreleváns, hogy egy adott mondatban a lexikonban szereplő melléknévi szófajba tartozó szó főnévi szerepet tölt be. Így a komi-permják esetében nemcsak a jelzői szerepű mellékneveken, hanem a főnevesült mellékneveken megjelenő toldalékokkal is számolnunk kell. A toldalékokat – ahogy az előző fejezetben utaltunk rá – csoportosítanunk kell attól függően, hogy a szótő (stem) utáni hányadik morfémapozíciót (slotot) tölthetik be, és milyen további toldalékok állhatnak előttük vagy utánuk. Az egy szón megjelenő maximális toldalékszám meghatározza a morfémapozíciók számát is.

Mivel a komi-permják tipikusan szuffixumokkal fejezi ki a grammatikai viszonyokat, a relációkat hordozó toldalékok a szótő után álló morfémapozíciókba kerülnek. Maga a szótő etalonként szolgál a többi morfémapozíció elhelyezéséhez, továbbá már kódolva van a lexikonban és szófajjelölést (POS-tag) is kapott. A szófaji meghatározás lehetővé teszi, hogy a szuffixumoknál fellépő homonímia esetén a FLEx csak a melléknevekhez járuló toldalékok elemzését ajánlja fel (és például az igeieket ne), ezen kívül azonban szükség van egy, a mellékneveket és főneveket tartalmazó felettes kategóriacímkére is, amit az egyes toldalékokhoz rendelhetünk hozzá azok szótári bejegyzésekor, mert ez biztosítja, hogy az elemző mind a főneveken, mind a mellékneveken felismerje az adott toldalékot.

A 2. táblázatban említett affixumok mindegyike megjelenhet egyedüliként a melléknév mellett (típustól függően előtte vagy utána), és egyik sem vonz kötelezően más affixumot. Ezeket kettesével társítva láthatjuk, hogy az első morfémapozícióban a többes szám jelén kívül bármi megjelenhet, így a komparatívusz, az intenzitás-, kicsinyítő, becéző képzők és a határozóképző is. A szuperlatívusz egy kivételes elem, mivel kizárólag az első morfémapozícióba kerülhet, ez esetben az első morfémapozíció megelőzi

a szótövet. A felsőfokú melléknév csak határozóképzőt (9) és többesjelet (10) vehet fel szuffixumként.

(9) *kin* *med-bur-a,* *med-basök-a,* *med-kužan-a*
aki SUPL-jó-ADV SUPL-szép-ADV SUPL-ügyes-ADV
tancujt-a-s *val's.*
táncol-FUT-3SG keringő
'... aki a legjobban, legszebben, legügyesebben fogja táncolni a keringőt.' (Komi Kyv Korpus)

(10) *bydsön* *ťihöj* *okean* *paśta* *med-bur-öś!*
egész Csendes óceán teljesen SUPL-jó-PL
'Az egész Csendes-óceán a legjobb (hely)!' (Komi Kyv Korpus)

Ahogy már említettük, a felsőfok jeleként szolgáló *med* partikula írásban a melléknévtől függetlenül is állhat (de csak közvetlenül előtte jelenhet meg). Ez a FLEx-ben nem okoz problémát, hiszen önálló elemként is bekerülhet a lexikonba, de ez esetben is megkaphatja a SUPL glosszát.

A komparatívusz a becéző- és a kicsinyítőképző mellett sosem jelenik meg, ugyanakkor az *-öv* augmentatív képző megjelenhet a középfok jele előtt az első morfémapozícióban (11), de fordított esetet nem mutatott a korpusz.

(11) *uľ-öv-žyk* *maťerik*
nedves-AUG-COMP talaj
'(jóval) nedvesebb talaj' (Komi Kyv Korpus)

Komparatívusz után csak a határozóképző (12) és a többes szám állhat (13); a határozóképző ugyancsak megengedi, hogy középfokjel kövesse egy kifejezésben (14). Ez utóbbi példa érdekes kérdéseket vethet fel nyelvészeti területen is, ti. van-e szemantikai különbség a szótő+COMP+ADV és a szótő+ADV+COMP szerkezetek között.

(12) *peryt-žyk-a*
gyors-COMP-ADV
'gyorsabban' (Komi Kyv Korpus)

(13) *vyna-žyk-öś*
erős-COMP-PL
'erősebbek' (Komi Kyv Korpus)

(14) *čök-a-žyk*
sűrű-ADV-COMP
'sűrűbben' (Komi Kyv Korpus)

A többes szám jele kétmorfémás szerkezet esetén első morfémapozícióba sosem kerülhet, viszont a második morfémapozícióban elhelyezkedve bármilyen toldalék megelőzheti (kivéve egy másik többesszám-jel, hiszen nem duplikálódhat).

A korpuszvizsgálatból nyert adatokból úgy tűnik, hogy a határozóképző a becéző és kicsinyítő képzővel semmilyen variációban nem fordul elő egymás mellett, de az *-öv* augmentatív képző követheti az adverbiális képzőt a második morfémapozícióban (15). A becéző és kicsinyítő képzők csak a többes szám jelével állhatnak, és ebben az esetben is kizárólag az első morfémapozícióban (16).

(15) *jon-a-öv*
erős-ADV-AUG
'(jóval) erősebben' (Komi Kyv Korpus)

(16) *kös-yńik-öś*
száraz-DIM-PL
'szárazkák' (Komi Kyv Korpus)

Az összesen 49 lehetséges kombinációból a permjákban csak 11-gyel találkozhatunk, melyekből az első morfémapozícióban a hétből hat affixum, míg a másodikban csak négy különböző szuffixum szerepelhet. Ezeket kell sémákba rendezni azon a ponton, amikor definiáljuk, hogy az elemző mely szótő+toldalék(ok) sorozatot ismerje fel jól formált melléknévként.

A melléknevek egy másik problémás aspektusa a főnévként való megjelenésük. Főnévi szerepbe kerülve toldalékolásuk esetraggal vagy személyjellel történik, a melléknevek jelzői szerepükben nem vehetnek fel esetragokat. Az alkalmilag főnevesült melléknéven megjelenhet egyes szám harmadik személyű személyrag, ami gyakran determinatív funkciót lát el (17).

(17) *görd-ys* *pyzan* *vylyn*
piros-3SG asztal -On
'A piros az asztalon van.' (F. Gulyás 2023a)

Ahogy korábban már láthattuk, a többes szám jelét melléknévként attributív és predikatív pozícióban is felveheti. Az ilyen esetekben a POS-tag mindig melléknév marad. Új szerepét a fordítás tükrözi majd.

A szakirodalom nem említi, de a nyelvhasználat alapján úgy tűnik, hogy a főnevekkel használatos többesszám-jel, az *-ez/-jez* szintén megjelenhet mellékneveken anélkül, hogy a melléknév főnevesülne, vagyis szófajváltás nem történik (18).

(18) *Oj* *tijö* *dona-ez* *da* *basök-kez!*
Ó 2PL drága-PL és szép-PL
'Ó, ti [milyen] kedvesek és szépek vagytok!' (Komi Kyv Korpus)

Rédei (1978) szerint a predikátumként megjelenő többes számú melléknév számban egyezik az alannyal és az *-öś* toldalékot veszi fel (1978: 81). A toldalékhasználat tekintetében ez ellentmond a (18)-as példának, de fentebb, a (3)-as példamondat esetében láthattuk, hogy valóban megjelenik az *-öś* morféma a mellékneveken predikátumi helyzetben.

Ezt a jelenséget többféleképpen is kezelhetjük a FLEx nyelvtanában. Az egyik lehetőség, hogy két különböző *-ez/-jez* toldalékot viszünk be a szótárba, és megadjuk, hogy az egyik csak a főneveken, míg a másik csak a mellékneveken jelenhet meg. A másik lehetőség, hogy egy még átfogóbb, névszói kategória bevezetésével megadhatjuk az elemzőnek, hogy a névszókon megjelenő *-ez/-jez* morfémát PL kategóriacímkével lássa el, ebben az esetben elegendő csak egy változatban bevinni a szótárba a toldalékot.[8]

5. Az igékkel kapcsolatos kérdések a FLEx szempontjából

A korpuszépítés során az igék annotálása is számos általános és speciális kérdést vet fel. A komi-permják nyelvben az igék elkülönülnek a nominális szófaji kategóriáktól, például az előző részben említett melléknevektől (F. Gulyás 2023c). Az igék a szótárakban (pl. Batalova – Krivoscsokova-Gantman 1985) a főnévi igenévi alakjukban, a *-ny* infinitívuszi képzővel ellátva jelennek meg: *mun-ny* 'menni'. Az igén többféle kategória jelölhető: az alany száma és személye, az igeidő, az igemód és a polaritás (vö. Ponomarjova 2010: 264–267); a kijelentő mód és a jelen idő jelöletlen. A morfológiai elemzés során figyelembe kell venni mind az igetövek típusait, mind az inflexiós morfémákat.

Az igék két fő tőtípust alkotnak, a ragozási tövükben mássalhangzóra (*pet-ny* 'kimenni') és a magánhangzóra (*košśy-ny* 'keresni') végződők csoportjait (Ponomarjova 2010: 28). Az utóbbi csoportba tartozó tövek általában *-y,* az igék egy zárt csoportjában pedig *-u* (*ju-ny* 'inni') végűek. (A létige jövő idejű formájában a szótő előtt *-o* magánhangzó szerepel.) Két morfonológiai alternáció van a nyelvben, amelyeket az annotálás során figyelembe kell vennünk. Az egyik az ún. *v* ~ *l* váltakozás, melynek következtében az abszolút (tehát szóvégi) helyzetben lévő /v/ fonéma CV

[8] A kérdés tárgyát képező toldalékkal kapcsolatban használt egyes szám szándékos a szerzők részéről, ugyanis a FLEx-ben az *-ez/-jez* többesszámjel valóban egy morféma két allomorfjaként fog szerepelni.

hangkörnyezetben /l/ alakban realizálódik (ld. a 3. táblázatban). A másik váltakozás egyetlen igét, a ’jönni’ jelentésűt érinti, amelynek szótári alakjában nem szerepel /t/, viszont minden inflektált alakjában (az Sg2 felszólító módút kivéve) igen.

3. táblázat: A komi-permják igék ragozása kijelentő mód, jelen időben Ponomarjova (2010) alapján

	Mássalhangzós tő	*v* ~ *l* váltakozós tő	⦰ ~ *t* váltakozós tő	*y* magánhangzós tő	*u* magánhangzós tő
Inf	*mun-ny* ’menni’	*ov-ny* ’élni’	*lok-ny* ’jönni’	*jökty-ny* ’táncolni’	*ju-ny* ’inni’
Sg1	*mun-a*	*ol-a*	*lokt-a*	*jökt-a*	*ju-a*
Sg2	*mun-a-n*	*ol-a-n*	*lokt-a-n*	*jökt-a-n*	*ju-a-n*
Sg3	*mun-ö*	*ol-ö*	*lokt-ö*	*jökt-ö*	*ju-ö*
Pl1	*mun-a-m(ö)*	*ol-a-m(ö)*	*lokt-a-m(ö)*	*jökt-a-m(ö)*	*ju-a-m(ö)*
Pl2	*mun-a-t(ö)*	*ol-a-t(ö)*	*lokt-a-t(ö)*	*jökt-a-t(ö)*	*ju-a-t(ö)*
Pl3	*mun-öny*	*ol-öny*	*lokt-öny*	*jökt-öny*	*ju-öny*

A táblázatból látható, hogy a morfológiai elemző számára formalizált nyelvtan létrehozásakor számolnunk kell a különféle tőtípusok eltérő morfonológiai viselkedésével. Tehát például a *loktö* ’jön’ alak esetében felismeri, hogy ez a *lok-* szótő egyes szám harmadik személyű alakja. Szerencsére a szótárban lehetőség van arra is, hogy megadjuk a szóalak ragozási tövét. Ezt nem külön szótári elemként tünteti fel az elemző, hanem egy adott szóalak variánsaként. Tehát például a ragozási tő ismeretében a FLEx a *lok-* ’jön-’ mellett egy *lokt-* ’jön’ változatot is képes jól formált szó-

alakként azonosítani és elemzést társítani hozzá. Az elemző ugyanígy jár el a *v ~ l* váltakozást mutató tövek esetében is, itt azonban gondot jelent, hogy az alternáció csak speciális hangkörnyezetben aktivizálódik, vagyis nem beszélhetünk egyszerűen csak ragozási tőről. Az efféle problémák kezelésére a FLEx-ben beállíthatók hangkörnyezeti szabályok (*environment*-ek) is, melyek megadásával a külön csoportba (kategóriába) rendezett *-v* végű igék nagy arányban helyes elemzést kapnak az elemzőtől. A továbbiakban áttekintjük az igei kategóriákat korábbi források alapján, a lehetséges elemzésüket, valamint az általunk használni tervezett annotációt.

A korpusz szövegeiben szereplő igealakoknál meg kell határoznunk az igei szuffixumok kategoriális státuszát. Az elemző szempontjából két alapvető kérdésre kell választ adni: 1) mi számít igei kategóriajelölőnek, és 2) ezen elemek hogyan, milyen sorrendben kapcsolódhatnak egymáshoz. A kérdés tehát elsősorban az, hogy a szótő után hány morfémapozíciót adjunk meg a nyelvtanban, a kategóriákat ugyanis felvehetjük külön-külön, például szótő-igemód-igeidő-szám-személy, azaz szótő-slot1-slot2-slot3-slot4 formában, de megadhatunk ennél kevesebb morfémapozíciót is.

A 4. táblázatban szereplő paradigmák alapján látható, hogy a szótő után általában két elem szegmentálható, kérdéses azonban ezek kategoriális státusza. A FLEx nyelvtanában megadhatunk egy olyan szabályt, ami szerint az *-a-* elem az igeidőt jelöli, az utána következő elem pedig az alany számát és személyét. Ebben az esetben egyes szám első személyben csak az igeidő jelölt, az alany száma és személye nem, míg 3. személyben az igeidő jelöletlen, az alany száma és személye viszont jelölt. Emellett pedig feltételezhetünk egy minden alakban zéró morfémás kijelentő módjelet. Ha összevetjük a fenti paradigmát az ún. egyszerű jövő és első múlt idejű paradigmákkal, további kérdések merülnek fel.

4. táblázat: A *munny* 'menni' ige ragozása jelen, egyszerű jövő és első múlt időben Ponomarjova (2010) alapján

	Jelen idő	Egyszerű jövő idő	1. múlt idő
Sg1	*mun-a*	*mun-a*	*mun-i*
Sg2	*mun-a-n*	*mun-a-n*	*mun-i-n*
Sg3	*mun-ö*	*mun-a-s*	*mun-i-s*
Pl1	*mun-a-m(ö)*	*mun-a-m(ö)*	*mun-i-m(ö)*
Pl2	*mun-a-t(ö)*	*mun-a-t(ö)*	*mun-i-t(ö)*
Pl3	*mun-öny*	*mun-a-sö*	*mun-i-sö*

A különféle igeidejű alakok összevetésével látható, hogy 1) a jelen és az egyszerű jövő időt kifejező igék pusztán a 3. személyű alakokban különböznek, továbbá az is, hogy 2) a jövő és az első múlt idejű alakok csak a szótő után álló magánhangzójukban térnek el. Ez úgy értelmezhető, hogy a szótő után következő elem majdnem minden esetben az időjel. Amennyiben azt szeretnénk, hogy az elemző a szótő-igeidő-szám.személy elemzést társítsa egy igealakhoz, akkor külön szabályt kell alkotnunk a jelen idejű, egyes szám harmadik személyű alakok elemzéséhez. Ebben az esetben szintén két lehetőség kínálkozik: a *munö* 'megy' igealakot például vagy a 1) szótő-idő-szám.személy, vagy a 2) szótő-idő.szám.személy címkével szükséges ellátni. Ha az elemzés eredményeként azt szeretnénk látni, hogy az időjelek minden esetben külön szegmentálódjanak a többi toldaléktól, és külön glossza járuljon hozzájuk, akkor több különböző időjellel kell számolnunk: az *-a* mellett egy ⦰ jelen idővel és egy ⦰ jövő idővel. Amennyiben az igealakon nem a testes időjel szerepel, egy legördülő menüből manuálisan kell kiválasztani, hogy az adott zéró milyen igeidőt jelöl. Ennél egyszerűbb az általunk választott második megoldás: a FLEx-ben a nyelvtant úgy adjuk meg, hogy a morfológiai elemző a kategóriákat egy formánsként szegmentálja. Az *-an* szuffixum esetében például a grammatikai kategóriacímke igeidő.szám.személy formátumú.

Az igei kategóriákkal kapcsolatos további kérdés az igemódok státusza, melyeknél nemcsak a használati szabályok algoritmizálása jelent nehézséget, hanem az is, hogy az általános glosszázási elvek megválasztásakor át kell gondolni, hogy mely elvek alkalmazhatóak technikailag úgy, hogy a lehető legmagasabb arányú helyes elemzést kapjuk a FLEx elemzőjétől. A korábbi források egyetértenek abban, hogy a komi-permjákban két morfológiailag jelölt igemód van, a kijelentő és a felszólító (vö. Batalova 2002: 97–104; Ponomarjova 2010: 203–204; Rédei 1978: 78). Emellett természetesen egyéb módokat is ki lehet fejezni, de azok nem (vagy nem csak) inflexiós morfémákkal vannak jelölve (vö. Klumpp 2022: 481). A felszólító módot egyes szám második személyben a puszta igető (*mun* 'menj'), míg többes szám második személyben az *-ö* szuffixum (*munö* 'menjetek') jelöli. Többes szám első személyben a felszólítást a kijelentő mód és a *te* (például *munam te* 'menjünk'), míg harmadik személyben az *aś* és a *med* partikula, valamint a jelen idejű, ragozott igealak (*med munö* 'menjen') fejezi ki. A glosszázáskor indokolt az igemód kategóriájának felvétele, azonban – mivel a maximálisan ragozott igealakon egyszerre vagy csak a mód, vagy csak az idő kategóriája lehet morfológiailag jelölt (vö. F. Gulyás 2023d), a kijelentő és felszólító módú igealakok közül pedig mindösszse a többes szám második személyű, felszólító módú igealaknak van ön-

álló, kitett inflexiós jelölője – az igemódot nincs okunk minden igealakon leválasztani. A ragozott igealakok egyik ígéretes elemzési módja ezért az, hogy az igeragok a szótárba eleve mód.idő.szám.személy jelentéssel kerülnek be, így a glosszasorban az adott szegmentumra jellemző összes kategóriacímke megjelenik, függetlenül attól, hogy az adott kategória zéró vagy testes morfémával van-e jelölve, például:

(19) *mun-am*
megy-IND.PRS.1PL
'megyünk' (elicitált)

Ezen elv alkalmazása egyrészt nagyban megkönnyíti az annotáló munkáját, másrészt pedig segíti a minél pontosabb kereshetőséget a felhasználó oldaláról. A korpuszban lehet keresést végezni 1) végződésre (például *-am*), és 2) grammatikai kategóriacímkére (PRS), vagy azok kombinációjára (IND.PRS) is. Összegezve tehát az állító, ragozott igealakon a kategóriák a korpuszban az igemód, az igeidő, az alany száma és személye. A kategóriák sorrendjét pedig nem kell meghatározni, mivel a szótő után álló elemeket nem szegmentáljuk, azok egyetlen morfémapozícióban jelennek meg.

A továbbiakban bemutatjuk, hogy milyen kérdések merülnek fel egyes ragozott igealakokkal kapcsolatban. A komi-permják igék esetében ugyanis nagyon sok a homonim alak, amit az elemzőnek valamilyen módon kezelnie kell. Sajnos előfordul, hogy a FLEx szabályrendszerében nem adhatók meg olyan szabályszerűségek, amelyek alapján az elemző el tudná dönteni, hogy aktuálisan melyikről van szó a két vagy több azonos alakú toldalék közül. Ezekben az esetekben az a megoldásunk, hogy minden lehetséges igei végződést (tehát a szótő után szereplő elemet vagy elemeket) külön-külön felveszünk a nyelvtanba és megadjuk a glosszáját, végül pedig az előannotálás utáni manuális ellenőrzéskor kiválasztjuk az adott példánál aktuálisan szereplőt. Például az elemző az *-ö* végződést default esetben az IND.PRS.3SG címkével látja el, de emellett szerepel az IMP.PRS.2SG és – ahogy azt később látni fogjuk – a CNG.PL (tehát a konnegatív igető többes száma) címke is, amelyek közül a manuális ellenőrzéskor választjuk ki a megfelelőt. Vannak természetesen olyan végződések is, amelyekhez csak egy címke társítható, pl. az *-in* címkéje: IND.PST.2SG.

Az igeidőkkel kapcsolatban felmerülő további kérdés a jelen és az egyszerű jövő idő viszonya. Az ezekben szereplő *-a-* elem egyszerre kifejezi a jelen és a jövő időt, kivéve 3. személyben, mivel ott csak a jövő időre vonatkozhat:

(20) *Ašyn Polina mun-as kino-ö.*
holnap Polina megy-IND.FUT.3SG mozi-ILL
'Polina holnap moziba megy.' (elicitált)

Azért, hogy a kézi annotációt megkönnyítsük, egy felettes grammatikai kategóriát használunk a nem múlt idejű igealakok címkéjeként, így minden jelen vagy jövő idejű igealakot egyféle jelöléssel (NPST: nonpast, tehát nem múlt idő) látunk el. Ezáltal az elemző az adott végződéshez mindig csak egy kategóriacímkét társít, így nem kell különféle funkciók közül manuálisan választani. A 3. személyű, jövő idejű igealakoknál megtartjuk a hagyományos FUT címkét.

5. táblázat: A nem múlt idejű igealakok kategóriacímkéje a parserben Ponomarjova (2010) alapján

	Szuffixum	Kategóriacímke
Sg1	*-a*	IND.NPST.1SG
Sg2	*-an*	IND.NPST.2SG
Sg3	*-ö*	IND.NPST.3SG
Sg3	*-as*	IND.FUT.1SG
Pl1	*-am(ö)*	IND.NPST.1PL
Pl2	*-at(ö)*	IND.NPST.2PL
Pl3	*-öny*	IND.NPST.3PL
Pl3	*-asö*	IND.FUT.3PL

A komi-permják múlt idők rendszere a korábbi források szerint komplex, általában két szintetikus és három analitikus igeidőt szoktak megkülönböztetni (Ponomarjova 2002: 132–140). Az ún. első múlt idő jele *-i,* a második múlt idő participiumi eredetű jelölője *-öm* (Bartens 2000: 179–215). Utóbbi paradigmája nem teljes, első személyű alakjai nincsenek, az egyes szám második (*-ömyt* és *-ömat*) és harmadik személyben (*-öm* és *-öma*) kétféle alakváltozata van, míg a többes szám második (*-ömnyt* és *-ömaś*) és harmadik személyű (*-ömaś*) alakjai részben megegyeznek. A variánsokat azonos kategóriacímkével látjuk el, a homonimák esetében pedig megállapítunk egy default és egy további funkciót, amelyek közül a manuális

ellenőrzéskor a FLEx felületén legördülő menüből kiválasztjuk az aktuálisat. Emellett az *-öm* és az *-öma* esetében szükség van további címkére is, mivel az előbbi a múlt idejű participium, az utóbbi pedig a múlt idejű predikatív participium jelölője is. Mindezt a 6. táblázatban foglaltuk össze:

6. táblázat: Az *-öm* elemet tartalmazó szóalakok lehetséges kategóriacímkéi Ponomarjova (2010) alapján

	Szuffixum	Kategóriacímke
Sg2	*-ömyt*	IND.PST2.2SG
Sg2	*-ömat*	IND.PST2.2SG
Sg3	*-öm*	IND.PST2.3SG
–	*-öm*	PTCP.PST
Sg3	*-öma*	IND.PST2.3SG
–	*-öma*	PTCP.PST.PRED
Pl2	*-ömnyt*	IND.PST2.2PL
Pl2	*mun-öm-aś*	IND.PST2.2PL
Pl3	*mun-öm-aś*	IND.PST2.3PL

Az analitikus múlt időkben a *völ-* 'van' ige egyes szám harmadik személyű, első (*völi*) vagy második (*völöm*) múlt idejű alakja szerepel egy ragozott főigével. Az összetett múlt idejű alakok a következők: 1) jelen idejű főige + *völi* (például *munö völi*); 2) második múlt idejű főige + *völi* (például *munöm(a) völi*); 3) második múlt idejű főige + *völöm* (például *munöm(a) völöm*). Mivel a FLEx csakis szóhatáron belül tudja elvégezni az előannotációt, így az analitikus múlt időket az elemző nem tudja egy egységként értelmezni. Ez azonban nem jelenti azt, hogy a korpusz felhasználója ne tudna összetett múlt időre keresést végezni. Terveink szerint lehetőség lesz ugyanis több lexéma egyidejű keresésére, akár a szóalakok, akár a nyelvtani kategóriacímkék alapján, valamint ezeket kombinálva.

A múlt időt jelölő morfémák polifunkcionálisak, ahogy azt az *-öm(a)* szuffixum esetében már bemutattuk. Az első és a második szintetikus alakok az idő kategóriáján kívül kifejezhetik az információ forrását is. Az *-i* jelet tartalmazó ige a közvetlen, az *-öm* elemet tartalmazó pedig a közve-

tett forrásból származó információt fejezi ki (Szabó 2023a, 2023b), ez a szembenállás azonban nem következetes (Szabó 2022). Mindezek alapján nem tartjuk indokoltnak az evidenciális kategóriacímkét felvenni. Ugyanezen okból nem adunk meg aspektuális (például imperfektív-perfektív, ld. Batalova 2002: 110–113) címkét sem.

Más rokon nyelvekhez hasonlóan a komi-permjákban is igével fejezik ki a tagadást, amely számban és személyben, valamint igeidőben (és módban) jelölt, ezt követi a konnegatív igető. A jelen, a múlt és az egyszerű jövő idejű tagadó alakok paradigmáját a 7. táblázatban ismertetjük.

7. táblázat: A *munny* 'menni' ige tagadó alakjai jelen, egyszerű jövő és múlt időben Ponomarjova (2010) alapján

	Jelen és egyszerű jövő idő	1. múlt idő	2. múlt idő
Sg1	*og mun*	*eg mun*	–
Sg2	*on mun*	*en mun*	*abu mun-ömyt / munöm-at*
Sg3	*oz mun*	*ez mun*	*abu mun-öm / mun-öma*
Pl1	*og(ö) mun-ö*	*eg(ö) mun-ö*	–
Pl2	*od(ö) mun-ö*	*ed(ö) mun-ö*	*abu mun-ömnyt / mun-ömaś*
Pl3	*oz(ö) mun-ö*	*ez(ö) mun-ö*	*abu mun-ömaś*

A konnegatív igetövek címkéje CNG egyes számban; többes számban, amikor szerepel a szóalakon az *-ö* elem, akkor CNG.PL. Egyes számban ez a tő megegyezik a puszta igetővel, valamint az IMP.2SG alakkal, a többes számú pedig az IND.PRS.3SG és az IMP.PRS.2PL alakokkal. Ebben az esetben az elemző kimenetelében a kijelentő mód lesz a default címke, ha ettől eltérő tagadó vagy felszólító módú alakként akarunk valamit gloszszázni, azt a legördülő menüből kell kiválasztani. Ugyanígy formailag megegyeznek az első múlt idő második személyű tagadó igéi és az azonos szám-személyű tiltó igék. Itt a tagadó ige címke lesz az automatikus elemzés default értéke, a prohibitív pedig a manuálisan kiválasztható variáns.

Az *abu* tagadószó az igeragozási paradigmán belül változatlan alakú, ezért a FLEx szótárában hozzárendelt kategóriacímke a NEG.

Az eddigiekben bemutattuk az igeragozási kategóriákat, azok lehetséges annotálási módjait, valamint az általunk a FLEx számára betanított annotálást. Mindezek nyomán az elemző tehát először azonosítja az ige tövét, amelyhez alapesetben a szótári jelentés van hozzárendelve. Ha az igető grammatikai jelentést is kifejez, akkor ki kell választani egy legördülő menüből, hogy mi az adott morféma lehetséges grammatikai kategóriacímkéje. A szótő után szereplő elemet az elemző egy szuffixumként leválasztja a szótőről, és különféle kategóriacímkéket ajánl fel, amelyek közül az annotáló kiválasztja az adott elemre érvényes variánst.

6. Összefoglalás

Tanulmányunkban egy új komi-permják korpusz létrehozásának aktuális lépéseit ismertettük, azon belül is a FieldWorks Language Explorer (FLEx) szoftver használata közben felmerült általános és specifikus problémákat. A program általános morfológiai elemzőjének segítségével előannotáluk a komi-permják mondatokat. A FLEx általi előcímkézés akkor lesz sikeres, ha a nyelv grammatikáját minél pontosabban adjuk meg a szoftver számára. Ez esetben viszont a komi-permják nyelvtanra nem(csak) nyelvészeti szempontból kell tekintenünk, hanem a program szemszögéből lényeges kérdéseket kellett szem előtt tartanunk. Ennek szemléltetésére több olyan, a melléknevekkel és az igékkel kapcsolatban felmerült kérdést mutattunk be, melyek a FLEx-es nyelvtanépítés során felmerülő különböző problémacsoportokat érintettek. Látható, hogy a technikai korlátok időnként hatással vannak a főbb glosszázási elvek megválasztására (pl. az igei módjelölés vagy az alkalmilag főnevesült melléknevek esetében), máskor többlépcsős szabályrendszerek összeállítására van szükség (pl. a főneveket és mellékneveket egységben kezelő kategória bevezetésére vagy a *v* ~ *l* alternáció kezelésére). Bizonyos kérdések – kimerítő nyelvleírások hiányában – külön korpuszvizsgálatot igényeltek (pl. a mellékneveken megjelenő toldalékok morfotaktikai szabályai), mások esetében nem találtunk módot arra, hogy a FLEx automatikus elemzőjével egyértelmű, emberi döntést nélkülöző kimenetet hozzunk létre (pl. igeragok homonímiája). A tárgyalt témakörök betekintést nyújtanak egy korpusz építésének mozzanataiba és technikai komplexitásába, ám a teljes korpuszt tekintve ennek többszöröse vár még az alkotókra.

Rövidítésjegyzék

1	első személy	GEN	genitívusz
2	második személy	ILL	illatívusz
3	harmadik személy	IND	kijelentő mód
ADV	adverbiális képző	PL	többes szám
AUG	augmentatív képző	PTC	partikula
COMP	komparatívusz	PST	múlt idő
DIM	diminutív képző	SG	egyes szám
ELA	elatívusz	SUPL	szuperlatívusz
FUT	jövő idő		

Irodalom

Antal M. Gergely (2023), Magyar-csángó, komi-permják. Nyelvjárás vagy önálló nyelv? Finnugor Világ 28/4: 43–47.

Bartens, Raija (2000), Permilaisten kielten rakenne ja kehitys. Mémoires de la Société Finno-Ougrienne 238. Suomalais-Ugrilainen Seura, Helsinki. https://doi.org/10.3176/lu.2003.2.10

Batalova, R. M. [Баталова, Р. М.] (1975), Коми-пермяцкая диалектология. Издательство Наука, Москва.

Batalova, R. M. [Баталова, Р. М.] (2002), Кудымкарско-иньвенский диалект коми-пермяцкого языка. Mitteilungen der Societas Uralo-Altaica 23. Moskva – Groningen. https://doi.org/10.3176/lu.1997.2.12

Batalova, R. M. – Krivoscsokova-Gantman, A. S. [Баталова Р. М. – Кривощёкова-Гантман А. С.] (ред.) (1985), Коми-пермяцко-русский словарь. Русский язык, Москва. https://doi.org/10.3176/lu.1986.1.11

Borin, Lars – Forsberg, Markus – Roxendal, Johan (2012), Korp – the corpus infrastructure of Språkbanken. https://gtweb.uit.no/u_korp/?mode=koi#?lang=en

Dryer, Matthew S. – Haspelmath, Martin (eds) (2013), The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info

Eberhard, David M. – Simons, Gary F. – Fennig, Charles D. (eds) (2024), Ethnologue: languages of the world. 27th edition. SIL International, Dallas – Texas. http://www.ethnologue.com

F. Gulyás Nikolett (2023a), Melléknév főnév nélkül (komi-permják). In: Havas et al. 2023. volgatyp.elte.hu/?lang=1&feature=37&code=koi

F. Gulyás Nikolett (2023b), Melléknév mint főnév (komi-permják). In: Havas et al. 2023. volgatyp.elte.hu/?lang=1&feature=38&code=koi

F. Gulyás Nikolett (2023c), Melléknév mint ige (komi-permják). In: Havas et al. 2023. volgatyp.elte.hu/?lang=1&feature=39&code=koi

F. Gulyás Nikolett (2023d), Igei affixumok sorrendje (komi-permják). In: Havas et al. 2023. volgatyp.elte.hu/?lang=1&feature=85&code=koi

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). Budapest: ELTE Finnugor Tanszék. (volgatyp.elte.hu) https://doi.org/10.21862/volgatyp
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. https://doi.org/10.1093/oso/9780198767664.001.0001
Komi Kyv Korpus – Perem Komi Jukön http://perem.komicorpora.ru/
Lewis, Paul M. – Simons, Gary F. – Fennig, Charles D. (eds) (2015), Ethnologue: languages of the world. 18th edition. SIL International, Dallas. http://www.ethnologue.com
Németh Szilvia – Szabó Ditta – F. Gulyás Nikolett (2023), PermCorp: egy komi-permják korpusz létrehozása. Folia Uralica Debreceniensia 30: 181–202. https://real-j.mtak.hu/26994/1/fud30.pdf https://doi.org/10.52401/fud/2021/10
Norvik, Miina – Jing, Yingqi – Dunn, Michael – Forkel, Robert – Honkola, Terhi – Klumpp, Gerson – Kowalik, Richard – Metslang, Helle – Pajusalu, Karl – Piha, Minerva – Saar, Eva – Saarinen, Sirkka – Vesakoski, Outi (2022), Uralic typology in the light of new comprehensive data sets. Journal of Uralic Linguistics 1/1: 4–42. https://uralic.clld.org/ https://doi.org/10.1075/jul.00002.nor
Perepis 2020 = Всероссийская перепись населения (2020–2021) https://www.strana2020.ru/
Ponomarjova, Larisza [Пономарева, Лариса] (2002), Фонетика и морфология Мысовско-лупьинского диалекта Коми-пермяцкого языка. Удмуртский Государственный Университет, Ижевск. Doktori (PhD) értekezés.
Ponomarjova, Larisza (2010), Komi-permják nyelvkönyv. Budapest. Kézirat.
Pusztay János (2022), Az oroszországi 2020. évi népszámlálás uráli (finnugor) szempontból. Folia Uralica Debreceniensia 29: 129–138. https://doi.org/10.52401/fud/2021/22
Rédei Károly (1978), Chrestomathia Syrjaenica. Tankönyvkiadó, Budapest. https://doi.org/10.3176/lu.1980.2.11
Skirgård, Hedvig et al. (2023), Grambank reveals global patterns in the structural diversity of the world's languages. Science Advances 9.
Szabó Ditta (2022), A permi és a török nyelvek evidencialitásának eredetéről. In: Balogné Bérces Katalin – Nemesi Attila László – Surányi Balázs (szerk.), Nyelvelmélet és kontaktológia 5. Pázmány Péter Katolikus Egyetem Bölcsészet- és Társadalomtudományi Kar, Budapest. 87–112. btk.ppke.hu/uploads/articles/2849446/file/7-2022-Szabo_Ditta.pdf

Szabó Ditta (2023a) Evidencialitás (komi-permják). In: Havas Ferenc – Asztalos Erika – F. Gulyás Nikolett – Horváth Laura – Timár Bogáta (2023), A Volga-vidéki finnugor nyelvek tipológiai adatbázisa (VolgaTyp). ELTE Finnugor Tanszék, Budapest. volgatyp.elte.hu/?lang=1&feature=99&code=koi https://doi.org/10.21862/volgatyp

Szabó Ditta (2023b), Evidencialitás kódolása (komi-permják). In: Havas Ferenc – Asztalos Erika – F. Gulyás Nikolett – Horváth Laura – Timár Bogáta (2023), A Volga-vidéki finnugor nyelvek tipológiai adatbázisa (VolgaTyp). ELTE Finnugor Tanszék, Budapest. https://doi.org/10.21862/volgatyp volgatyp.elte.hu/?lang=1&feature=100&code=koi

Zamyatin, Konstantin (2022), Language policy in Russia: The Uralic languages. In: Bakró-Nagy, Marianne – Laakso, Johanna – Skribnik, Elena (eds), The Oxford guide to the Uralic languages. Oxford University Press, Oxford. 79–90. https://doi.org/10.1093/oso/9780198767664.001.0001