DOI: 10.21862/Szociolingvisztika/2024/8122.105

MoMa – Moldvai magyar beszélt nyelvi, nyelvjárási korpusz

Eris Elvira Mária – Huszár Anna Laura – Kalivoda Ágnes – Sass Bálint – Vadász Noémi – Vargha Fruzsina Sára

HUN-REN Nyelvtudományi Kutatóközpont

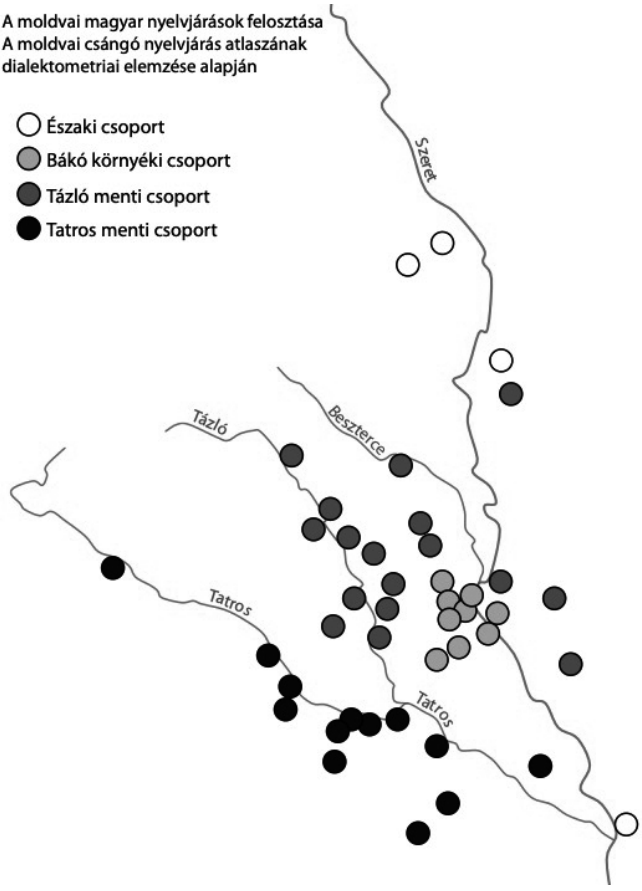
1. Bevezetés

A moldvai magyar nyelvjárások tanulmányozására, a különböző nyelvjárások viszonyrendszerének feltárására alkalmas legjelentősebb adattárunk A moldvai csángó nyelvjárás atlasza (MCsNyA.), amely 43 kutatópontról közöl magyar nyelvjárási adatokat (Bodó – Vargha 2008). Az MCsNyA. kvantitatív nyelvföldrajzi vizsgálata révén van lehetőségünk arra is, hogy dialektometriai térképeken tanulmányozhassuk a moldvai nyelvjárások közti hasonlósági viszonyokat. Az atlaszadatok alapján a terület nyelvjárásait így négy csoportba oszthatjuk: északi, Bákó környéki, Tázló menti és Tatros menti csoportra (Bodó et al. 2012; Vargha 2017: 99–102, l. az *1. ábrá*t). Ahhoz azonban, hogy a moldvai nyelvhasználatot rendszerszerűen tanulmányozhassuk, szintaktikai, pragmatikai vagy diskurzuselemzési vizsgálatokat végezhessünk, nyelvatlaszokon, szótárakon túl leginkább élő, beszélt nyelvi szövegekre van szükségünk. A magyar nyelvészet régi adóssága a moldvai magyar nyelvjárások kutatására alkalmas, megfelelő pontossággal lejegyezett szövegkorpusz létrehozása.

Szervezett dialektológiai gyűjtés, amely reprezentatív hangzó mintát gyűjtött volna a régióban, nem volt. Tánczos Vilmos kolozsvári néprajzkutató viszont évtizedeken át kutatta a moldvai magyarság archaikus népi imáit, és magát a világképet, gondolkodásmódot, vallásgyakorlatot, amely ezekhez a szövegekhez elválaszthatatlanul kapcsolódik. Minden olyan moldvai településen készített hangfelvételeket, ahol azt gondolta, fennmaradhattak még imádságok, esetleg olyan szövegek is, amelyeket még nem sikerült dokumentálni, s így a nyelvcserével és az életforma megváltozásával örökre feledésbe merülhetnek (Tánczos 1995, 1996, 2001).

Tánczos Vilmos gyűjtéséből kiválasztott interjúrészletek feldolgozásával a Veszélyeztetett uráli nyelvek szintaxis-elméleti és magyar nyelvtörténeti tanulságai projektben (NKFIH 129921, a kutatás vezetője É. Kiss Katalin) 2020-ban a Nyelvtudományi Kutatóközpontban megkezdtük egy moldvai nyelvjárási szövegkorpusz kialakítását. A feldolgozást 2021 szeptemberétől a MoMa-projektben folytatjuk. E tanulmány szerzőin kívül az adattárépítés első szakaszában munkatársunk volt Balázs Renáta.

Tanulmányunkban beszámolunk a korpuszépítés folyamatáról, legfontosabb alapelveiről és nehézségeiről, egyszersmind praktikus példákkal szemléltetjük a bárki számára elérhető és kereshető online szövegadatbázisban rejlő kutatási lehetőségeket.



1. ábra. A moldvai magyar nyelvjárások csoportosítása a MCsNyA. dialektometriai elemzése alapján (Vargha 2017: 99)

2. Az alapanyag

A hangfelvételek középpontjában – Tánczos Vilmos kutatási témájából adódóan – a hitélet, az archaikus népi imák szövegei, továbbadása a fiatalabb nemzedéknek, a kántorok tevékenysége és a vallásgyakorlással kapcsolatos témák állnak. A magyar nyelvű imák témája azonban elválaszthatatlan magától a magyar nyelv használatának kérdésétől, így a beszélgetésekben szinte mindig megjelennek nyelvhasználattal, nyelvtudással kapcsolatos kérdések is. A felvételek olykor korrajzok is, számot adnak a 20. század nagyobb eseményeiről, a közösségeket érintő történelmi eseményekről, tragédiákról. Gyakran dokumentálnak család- és helyneveket. Lejegyzéseink 1990 és 2010 között készült interjúkat dolgoznak fel. A hangfelvételek különböző felvételi technikákkal készültek, különböző típusú adathordozókra (magnókazetta, videófelvétel, digitális hangfelvevő). Mohi Sándor 1998-ban *Gyöngyökkel gyökereztél* címmel forgatott filmet a moldvai archaikus népi imákról, a filmben szereplő beszélgetéseket Tánczos Vilmos készítette. Mohi Sándortól megkaptuk a film vágatlan nyersanyagának digitalizált változatát, lejegyzéseink részben ebből a gyűjtésből származnak.

Az interjúk hossza és hangminősége is változó, vannak 1–2 perces, de akadnak több részletben lejegyzett, összesen 1–1,5 órás részletek is, előfordul sok résztvevős utcai beszélgetés, de vannak több órás, zárt térben készült interjúk is. A hangfelvételek általában nem egy interjút rögzítenek, ezért a feldolgozás során vágunk, de minden esetben megőrizzük a kapcsolatot a kivágott részlet és az eredeti hangfelvétel között. A feldolgozás során minden hanganyagot wav-formátumba konvertálunk, és minden esetben feltüntetjük a forrásfájlt és annak formátumát a lejegyzés fejlécében. A lejegyzendő interjúk kiválasztásakor arra törekedtünk, hogy minél teljesebb képet tudjunk adni a moldvai magyar nyelvhasználatról, így olyan interjúk, interjúrészletek kerültek be az adattárba, amelyek zömében magyar nyelvű beszélgetéseket tartalmaznak. Fontos szempont volt az adattár bővítésekor, hogy minél több kutatópontról, minél több beszélőtől legyenek szövegeink.

3. A lejegyzések elkészítése és előfeldolgozása

A hangzó anyag lejegyzését a magyar számítógépes dialektológiában meghonosodott standard szerint végezzük, a Bihalbocs összefoglaló néven ismert nyelvészeti technológiák használatával (a fejlesztések történetéről, alapelveiről l. Vékás 2007). A felvételeket az eredeti hangzó változat és a lejegyzés összekapcsolásával hozzuk létre, így a szövegek olvasása közben egyetlen kattintással elérhetővé válik

a lejegyzett részletek eredeti, hangzó formája. A hangjelölés megfelel a magyar egyezményes hangjelölés szabályainak, attól csak néhány esetben térünk el, annak érdekében, hogy minél egyszerűbb eljárásokkal tudjuk az anyagokat – a különböző kutatói igényekhez alakítva – többféle formába konvertálni. A korpuszépítés folyamatában, a normalizálás előkészítéséhez például automatikusan hozzuk létre a lejegyzés hangtanilag egyszerűsített, a magyar helyesíráshoz közelítő formáját.

A hangminőségek megfelelő pontosságú jelölése lehetővé teszi, hogy elemezzük a nyelvjárások hangtani sajátosságait, azokról kimutatásokat, nyelvföldrajzi térképeket készítsünk. Jelenleg 37 kutatópontról mintegy 20 órányi hangfelvétel lejegyzése készült el, megközelítőleg 120, zömében idősebb vagy középkorú adatközlőtől. Kutatópontjaink között több olyan település is szerepel – főleg északon –, ahol ma már nem beszélnek magyarul. A nyelvvesztés folyamatának utolsó fázisát néhány lejegyzett interjú is példázza, ahol az idősebb beszélők még tudnak magyarul beszélni, de a középkorú résztvevők ugyan még tudják követni a beszélgetést, már nem szólalnak meg magyarul.

A már elkészült lejegyzéseket online is megjelenítjük, így mindenki számára, aki a korpuszt nyelvészeti, néprajzi vagy akár népzenei vonatkozásai miatt böngészni szeretné, már az adattár készítésének folyamatában elérhetők a lejegyzések (https://nlp.nytud.hu/csango/; *letöltés ideje: 2024. szeptember 11.)*.

A lejegyzés során pontosan jelöljük a beszélőt, akihez az adott megnyilatkozás tartozik. Az adattár jellegéből adódóan sok a kötött szöveg, ezeket külön jelöljük, illetve külön jelzéssel látjuk el az énekelt részeket is. A nevetve, köhögve és sírva mondott részleteket kerek zárójelek közé tesszük. Jelöljük az átfedéseket is, vagyis azt, ha két (vagy több) megnyilatkozás egyszerre hangzik el. A lejegyzési standard leírása itt érhető el: https://nlp.nytud.hu/csango/lejegyzes.html *(utoljára megtekintve: 2024. szeptember 11.)*.

A moldvai magyar közösségek beszélői szinte kivétel nélkül kétnyelvűek voltak már az ezredforduló előtt is, az interjúk során sok a román nyelven elhangzó részlet. A lejegyzéseinkben így külön kellett kezelnünk és jelölnünk a román nyelvű megnyilatkozásokat is. Ezeket a román helyesírás szerint jegyeztük le, a későbbiekben pedig nem normalizáljuk, és morfológiailag sem elemezzük. A számos román kölcsönszó miatt olykor nem könnyű feladat eldönteni, hogy egy-egy rövidebb szakasz esetében milyen nyelvűnek tekintsük, ami elhangzik.

Mivel a moldvai nyelvjárások szóhasználatukban, olykor mondatfűzésükben is jelentősen eltérnek a többi magyar nyelvváltozattól, a szövegekben szómagyarázatokat, értelmezéseket is megadunk, így az adattár azok számára is könnyen használható, akik kevésbé járatosak a moldvai nyelvváltozatokban.

A lejegyzési gyakorlatunkkal való játékos ismerkedéshez készítettünk egy kvízt, amely a következő linken elérhető: https://nlp.nytud.hu/csango/lejegyzos_kviz/ *(letöltés ideje: 2024. szeptember 11.).*

4. Normalizálás

A lejegyzett anyagból a hároméves MoMa-projekt keretében kereshető korpusz készül. A folyamatosan bővülő adatbázis az interjúszövegeket az eredeti szövegváltozatuk mellett egy ún. *normalizált* verzióban is tartalmazza. Ez a változat, amelyben mesterségesen ötvöződnek a nyelvjárási és a „sztenderd" nyelvváltozatok sajátosságai, közvetlenül nem alkalmas a moldvai magyar nyelv vizsgálatára, az interjúszövegek kutathatóvá tételében ugyanakkor fontos szerepe van. A normalizálás ugyanis amellett, hogy növeli a feldolgozandó beszélt nyelvi, nyelvjárási, archaikus és román elemekben gazdag anyag gépi elemezhetőségét, a korpusz felhasználói számára is jelentősen megkönnyíti az interjúszövegekben való keresést.

A normalizálást szótáralapú előnormalizálás után kézzel végezzük. Ehhez az anyagot a következőképpen készítjük elő: 1) A lejegyzett szöveget különálló tokenekre (azaz szövegszókra) bontjuk, figyelve arra, hogy a szövegbe a lejegyzéskor bekerülő annotációk (beszélő kódja, magyarázatok, fordítások) megőrződjenek. A románul elhangzó szövegrészeket nem tokenizáljuk és nem elemezzük, ezek megszólalásonként 1–1 tokennek számítanak, és fordításukkal együtt, változatlan formában kerülnek a korpuszba. 2) Az első 98000 kézzel normalizált szövegszóból létrehozott szótár segítségével automatikus módszerekkel előnormalizáljuk a szöveget. Az így előállított alakokat, melyek a normalizálandó tokenek megközelítőleg 60%-át teszik ki, kézzel ellenőrizzük és szükség szerint javítjuk.

A következő lépésben a szöveget szavanként, kézzel normalizáljuk. Az annotátor munka közben a hanganyagot is hallgatja, így a lejegyzés ebben a lépésben is pontosítható, javítható. Normalizáláskor a lejegyzés során kijelölt mondathatárokat és központozást már nem módosítjuk, a helyesírást viszont ellenőrizzük és szükség esetén korrigáljuk. Ez a lépés a szöveg részleges újratokenizálásával is jár. Mivel az eredeti lejegyzésekben a tagmondatok és mondatok nem logikai, hanem beszédfolyam-egységeknek felelnek meg, a korpuszban mondattani vizsgálatok csak ennek tudatában végezhetők.

A normalizálás során az annotátor a lejegyzett szóalakokat egy meghatározott szabályrendszernek megfelelően, kisebb-nagyobb mértékben egységesíti, és a „sztenderd" nyelvváltozathoz közelíti. Az alkalmazott szabályokat, melyek lefektetésekor már létező történeti korpuszok (Simon – Sass 2012; Dömötör és mtsai

2017) normalizálási gyakorlatából indultunk ki, a korpuszhoz tartozó részletes dokumentáció tartalmazza majd. Egy moldvai interjúhoz hasonló nyelvjárási, beszélt nyelvi, kötött szövegektől, kódváltásoktól és archaikus elemektől gazdag anyag normalizálása számos kihívással, de ugyanennyi előnnyel is jár: ellensúlyozni tudja például a nyelvhasználatbeli hatalmas változatosságot, és a lejegyzett alakok sokféleségét (a jelenleg kb. 147 ezer szövegszónyi korpuszban a 'van' jelentésű *vagyon* szó pl. 169-szer, 41-féle különböző alakban jelenik meg: *vadzsǫnd, vajany, vályán, vadon, vagyǫn* stb.).

Az annotátor a tokenekhez a „magyarázat" rétegben megjegyzéseket is fűzhet. Ezeknek csak egy része szigorú értelemben vett magyarázat, itt kap helyet ugyanis minden egyéb, a szót érintő többletinformáció is, például ha 1) bizonytalanság vagy szövegromlás gyanúja merül fel egy szóalakkal kapcsolatban, 2) nyelvi szempontból érdekes alakról van szó (pl. a köznyelvitől eltérő raghasználatot mutat), vagy ha a szó egy másik tőváltozattal is előfordul a korpuszban.

A táj- és jövevényszavakhoz fűzött magyarázatok funkciója kettős: egyrészt segíthetik a felhasználót a szöveg értelmezésében és a korpuszban való navigációban – a *kukorica* szóra keresve például megtaláljuk a *puj* és a *törökbúza* tájnyelvi megfelelőket is[1] –, másrészt szerepet játszanak a morfológiai elemzés pontosságának növelésében is. A magyarázat ugyanis, amelyben legtöbb esetben a tájszó köznyelvi megfelelője az eredetivel azonos módon toldalékolva szerepel, az elemző segítségére lehet például akkor, ha a token normalizált alakjából önmagában nem derül egyértelműen ki, hogy a szóban milyen tőhöz milyen toldalékok kapcsolódnak. Kontextus és megfelelő háttértudás nélkül pl. a *kijött a pensziére* 'nyugdíjba ment' frázis utolsó szava akár birtokos alaknak is tűnhet, holott valójában szublatívuszi ragos főnévről van szó. A magyarázatként megadott *nyugdíjra* alak a pontatlan választást kizárja.

A magyarázat rovat mellett egy ún. „sztenderd alak" réteget is bevezettünk, amely megjelenít és könnyen lekérdezhetővé tesz bizonyos jellegzetes nyelvjárási jelenségeket. Itt jelenik meg többek között az, ha 1) a beszélő egy illatívuszi ragos alakot inesszívuszi funkcióban használ *(házba/házban),* 2) „suksüköl" *(szeressük/szeretjük),* 3) a szóalakon jelöletlen tárgyragot sejtünk *(Mind[et] megtanította?),* vagy ha egy birtokos szerkezetben a sztenderdtől eltérő egyeztetést találunk *(a zsidók kezükbe/kezébe).*

1 Közvetlenül a magyarázatokban való keresésre a NoSketchEngine felületén a Concordance > Advanced > CQL fülön megfogalmazott lekérdezésekkel van lehetőség. Pl.: [magyarazat=».*kukoric[aá].*»] . Fontos megjegyezni azonban, hogy a magyarázat réteg rendeltetése elsősorban nem ez, ezért az ilyen módon kapott találatlistát mindig erősítsük meg további, az eredeti vagy a normalizált szövegváltozatokban végzett keresésekkel.

5. Morfológiai elemzés

A korpusz anyagának egy részén morfológiai elemzést is végeztünk. Az elemzés célja az, hogy lehetővé váljon olyan grammatikai – főként morfoszintaktikai – jelenségek vizsgálata, amely elemzetlen szövegen nagyon időigényes és körülményes volna. A morfológiai elemzés során a szavakat morfémákra bontjuk, ezeket a morfémákat kategorizáljuk és felcímkézzük, majd megállapítjuk a szó lemmáját (szótári alakját). Ezt a folyamatot az *énekeket* szóalakkal szemléltetjük:

1. a szóalakban három morféma azonosítható, ezek határát pontokkal jelöljük: *ének.ek.et*;
2. az egyes morfémák a következő címkéket kapják: *ének* → főnév (jelölése: N), *ek* → többes szám (Pl), *et* → tárgyeset (Acc); az előálló morfológiai elemzés tehát N.Pl.Acc[2];
3. a szó lemmája az *ének* lesz: ez az a szó, amelyre keresve később a különféle toldalékot viselő alakok is megtalálhatók (pl. *énekünk, énekemmel*).

A fenti példában az adott szóalakhoz mindössze egy lehetséges elemzés tartozik, de a korpuszban jóval általánosabb az az eset, amikor egy szóalak többféle elemzést kaphat, és a szövegkörnyezettől függ, hogy melyik a helyes (ld. 1. táblázat). Ekkor szófaji és morfológiai egyértelműsítés is szükséges.

1. táblázat: A *mentek* szóalak háromféle lehetséges elemzése

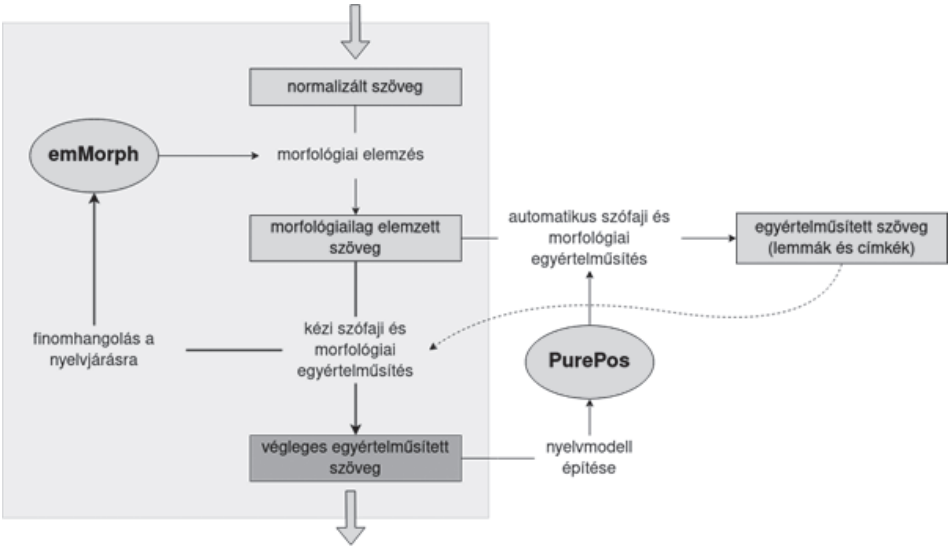
lemma	**címke**	**jelentés**
megy	V.P2	*megy* ige T/2. jelen idő (*ti mentek*)
megy	V.Past.P3	*megy* ige T/3. múlt idő *(ők mentek)*
ment	V.S1	*ment* ige E/1. jelen idő *(én mentek)*

Az automatikus morfológiai elemzéshez az emMorph elemző (Novák 2014; Novák et al. 2016) ó- és középmagyar nyelvállapotra finomhangolt változatát használjuk. Azért döntöttünk emellett, mert a moldvai magyar nyelvjárások sok szempontból ezekhez a korábbi nyelvállapotokhoz állnak közel, így több olyan jelenséget is tudunk kezelni, amelyeket egy mai sztenderd magyarra hangolt elemzővel nem ismernénk fel. Ilyen például a moldvai magyarban gyakori elbeszélő múlt (pl. *mondá, jöve*) vagy az összetett múlt idők (pl. *mondtam volt*). Természetesen szükség volt az elemző további, folyamatos finomhangolására is, hogy le tudja fedni

2 A korpuszban használt morfológiai kódok listája a következő linken érhető el: https://nlp.nytud.hu/csango/morf_kod_lista/ (*Utoljára megtekintve:* 2024. szeptember 10.)

a kizárólag moldvai magyarra jellemző jelenségeket. Ez elsősorban a tőtár bővítését jelentette nyelvjárási szavak hozzáadásával (pl. *csitil* 'olvas', *rozáriu* 'rózsafüzér', *csán* 'csinál').

Az automatikus morfológiai elemzést követően kezdetben kézzel történt a szófaji és morfológiai egyértelműsítés. Az átnézett szövegekből tanítókorpuszt készítettünk a PurePOS (Orosz – Novák 2012, 2013) számára, amely egy statisztikai alapú – rejtett Markov-modellre épülő – címkéző program. A PurePOS a tanítókorpuszból egy nyelvmodellt állít elő, amelynek segítségével az újabb moldvai magyar szövegeket automatikusan egyértelműsíteni tudja az adott kontextusban legvalószínűbb 'lemma – morfoszintaktikai címke' párok kiválasztásával. A PurePOS-sal egyértelműsített szövegeket is szükséges még kézzel átnézni és helyenként javítani, de ezzel az előfeldolgozási lépéssel jelentősen felgyorsítható a munka. A morfológiai elemzés teljes folyamatát a *2. ábra* szemlélteti.



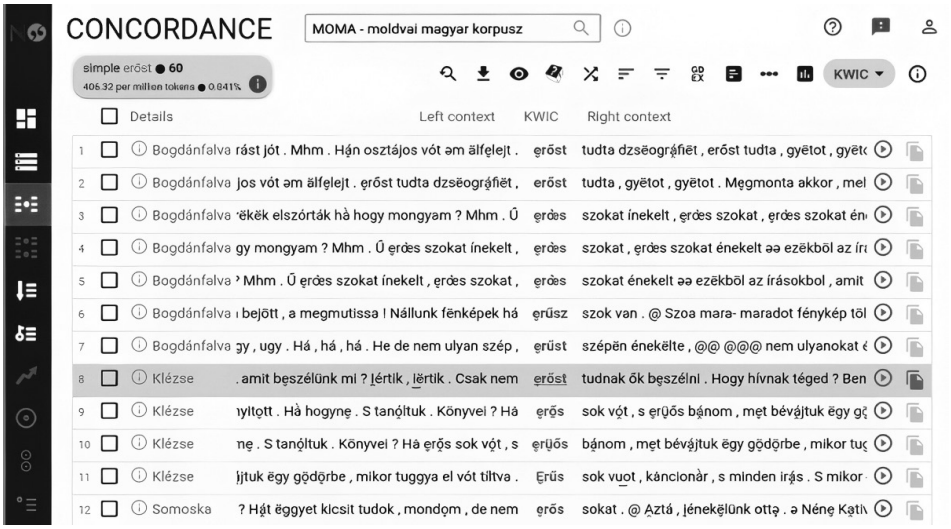
2. ábra. A morfológiai elemzés folyamata

6. Keresőfelület

A lejegyzett, normalizált, morfológiailag elemzett korpuszt a NoSketchEngine (Rychlý 2007; Kilgarriff et al. 2014) korpuszkezelő rendszerben tesszük elérhetővé. Napjainkban a NoSketchEngine a legelterjedtebb, de facto standard korpuszkezelő rendszer, melyet lassan két évtizede folyamatosan fejlesztenek. Számos funkcióval

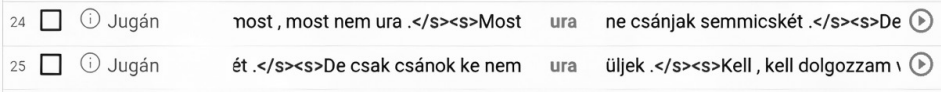
rendelkezik: a klasszikus konkordanciák megjelenítésén túl alkalmas gyakorisági listák készítésére, véletlen mintavételre, kollokációs vizsgálatra. A lekérdezéseinket újabb lekérdezésekkel szűrhetjük. A korpuszban rejlő információ teljességéhez a reguláris kifejezésekre épülő CQL (Corpus Query Language) lekérdezőnyelv használatával férhetünk hozzá (Sass 2022). A rendszer a szöveges annotációs szintek kezelése mellett lehetőséget ad arra is, hogy az egyes szövegrészekhez hozzáillesszük a megfelelő hanganyag-részletet (a 2. részben említett wav-fájlokat), így ezeket az egyes konkordanciasoroknál egy kattintással meghallgathatjuk. A MoMa moldvai magyar beszélt nyelvi, nyelvjárási korpusz a https://moma.nytud.hu *(Utoljára megtekintve:: 2024. szeptember 11.)* internetes címen érhető el szabadon a felhasználók számára. A korpusz lehetőségeit az alábbi négy példával illusztráljuk.

A korpusz megnyitása után a Konkordanciát (Concordance) választva az egyszerű keresésben (BASIC) az *erőst* alakot megadva a 3. ábrán látható konkordanciát kapjuk. Középen pirossal láthatók a találati szavak (kwic-ek), a normalizálásnak köszönhetően az *erőst* összes lejegyzett variációját megkapjuk. A normalizált alak (a 8-as példa esetében például: *Értik, értik. Csak nem erőst tudnak ők beszélni*) a felső menü harmadik (szemecske) gombjának használatával jeleníthető meg. A kapcsolódó hanganyag az adott sor végén lévő piros lejátszás gombra kattintva hallgatható meg, a találat metaadatai pedig a sor elején látható információ gombra kattintva érhetőek el. A metaadatok tartalmazzák a találathoz tartozó interjú azonosítóját, a település nevét és koordinátáit, ahonnan az adat származik, illetve egy, a lejegyzések gyűjteményére mutató linket, amelyen keresztül az adott interjú teljes egészében elolvasható és folyamatában is meghallgatható.



3. ábra. Az *erőst* alak konkordanciája az eredeti lejegyzést megjelenítve

Az anyag gazdagságát mutatja a második példa. Az *ura* lekérdezés 24-25. találatában *(4. ábra)* két rövid mondatban négy jellegzetes moldvai jelenségre is példát találunk: (1) a *(nem) ura vmit csináljak* speciális szerkezet, jelentése: '(nem) tudok valamit csinálni'; (2) a *semmicskét* tipikus példa a gyakran használt kicsinyítésre; (3) a *ke* 'mert' egy szókincset érintő specialitás; a (4) *csán* pedig a *csinál* szokásos moldvai változata, ami természetesen a kibővített tőtárba is belekerült (ld. az 5. részt).



4. ábra. Jellegzetes moldvai nyelvi jelenségek az *ura* konkordanciájában. A szöveg kisebb egységeit <s> tagek jelölik, az ábrán a normalizált alak szerepel.

A harmadik példában, annak vizsgálata kapcsán, hogy milyen gyakori az igék között ez a bizonyos *csán* tő, egy bonyolultabb, normalizálásra és morfológiai elemzésre is építő, CQL-t alkalmazó lekérdezést mutatunk be. Az *5. ábrán* látható gyakorisági lista a következő módon áll elő:

1. Válasszuk az egyszerű helyett most az összetett keresést (advanced);
2. kattintsunk a CQL-re, majd a CQL BUILDER-re, és válasszuk a normal tokent;

3. az ige mint szófaj kiválasztásához a bal oldalon az Attribute-nál válasszuk az ELEMZÉS-t a jobb oldalon, a menüben pedig görgessünk le az IGE bejegyzésig, majd pipa és USE THIS CQL;

4. kattintsunk a GO-ra, így megkapjuk az igék konkordanciáját, a morfológiai elemzésnek köszönhetően az összes ragozott alakot;

5. gyakorisági listát a felső menü gombjára, majd a KWIC alatti LEMMAS gombra kattintva készíthetünk. Az eredményt az *5. ábrá*n látjuk.

	Szótő ('lemma')	Frequency	Relative ?	
1	van	218	2,039.82	•••
2	tud	140	1,309.98	•••
3	mond	129	1,207.05	•••
4	megy	47	439.78	•••
5	imádkozik	37	346.21	•••
6	lesz	33	308.78	•••
7	csán	28	262.00	•••
8	jön	28	262.00	•••
9	énekel	22	205.85	•••
10	tanul	22	205.85	•••

5. ábra. Igetövek gyakorisági listája a MoMa-korpuszban. Az anyag tematikáját jól mutatja az *imádkozik* és az *énekel* gyakorisága. A jellegzetes moldvai *csán* a hetedik helyen található.

A kutatások során a legtöbb esetben fontos, hogy kizárjuk a terepmunkás által mondottakat, és csak az adatközlőkre korlátozzuk a keresést. Ezt a [beszelo="ak"] CQL kifejezéssel való KWIC-re vonatkoztatott szűrés révén érhetjük el, amit szintén a felső menüből érünk el.

Végül, a negyedik példában egy pragmatikai jelenséget, a „magázást" vizsgáljuk meg röviden. A *maga* szótőre keresve (ADVANCED/LEMMA), és a fenti [beszelo="ak"], majd külön lekérdezésben a terepmunkás szövegrészeit elkülönítő

[beszelo=”tm”] CQL kifejezéssel szűrve azt látjuk, hogy míg a terepmunkához tartozó számos találat mind „magázó” forma, az adatközlők megszólalásaiban a „magázás” ezen módja csak elszórtan, feltehetően a köznyelvvel való érintkezés hatására fordul elő. Olyan példákat, amelyek a nyelvváltozatra valóban jellemző „magázódó” formákat tartalmaznak, a *kend*, illetve a *kelmed* lemmákra keresve találhatunk.

7. Összefoglalás

A Tánczos Vilmos moldvai gyűjtéséből épített nyelvjárási korpusz több szempontból is egyedülálló a magyar dialektológia történetében. Ez a legnagyobb hanggal összekapcsolt, spontán beszélgetéseket és kötött szövegeket is tartalmazó, fonetikai igényességgel lejegyzett magyar nyelvjárási adattár. A szövegek normalizálása és morfológiai annotálása révén a korpusz kereshetőség és kutathatóság szempontjából is példaértékű. Tanulmányunkban konkrét példákkal igyekszünk segíteni a felhasználókat abban, hogy minél könnyebben tudjanak saját kutatási témájuk szempontjából releváns kérdéseket megfogalmazni a https://moma.nytud.hu *(letöltés ideje: 2024. szeptember 11.)* keresőfelületen.

Mivel a teljes anyag szabadon hozzáférhető, jól használható más tudományterületek kutatói számára, sőt oktatási és ismeretterjesztő célokra is. Bízunk benne, hogy projektünk lendületet adhat további nyelvjárási, néprajzi gyűjtések hasonló feldolgozásához, és így minél több adattár válhat elérhetővé és kutathatóvá. Amennyiben kutatásához felhasználja a korpuszt, kérjük hivatkozzon a jelen cikkre, illetve magára a korpuszra is (Eris et al. 2023).

Köszönetnyilvánítás

A tanulmány elkészítését *A moldvai magyar nyelv: adatbázisépítés és grammatikai kutatások* projekt támogatta (vezető kutató: Surányi Balázs, ELKH SA-53/2021).

Felhasznált szakirodalom

Bodó Csanád – Vargha Fruzsina Sára (2008): Régi nyelvatlaszok – új módszerek. *Magyar Nyelv* 104. 335–351.

Bodó, Csanád – Vargha, Fruzsina Sára – Vékás Domokos (2012): Classifications of Hungarian dialects in Moldavia. In: Lehel, Peti – Tánczos, Vilmos (eds.), *Language Use, Attitudes, Strategies: linguistic identity and ethnicity in the Moldavian Csángó villages*. The Romanian Institute for Research on National Minorities, Cluj-Napoca. 51–69.

Dömötör Adrienne – Gugán Katalin – Novák Attila – Varga Mónika (2017): Kiútkeresés a morfológiai labirintusból – korpuszépítés ó- és középmagyar kori magánéleti szövegekből. *Nyelvtudományi Közlemények* 113. 85–110 https://doi.org/10.15776/NYK.2017.113.3

Eris Elvira – Huszár Anna Laura – Kalivoda Ágnes – Sass Bálint – Vadász Noémi – Vargha Fruzsina Sára (2023): *Moldvai magyar korpusz – részletek Tánczos Vilmos gyűjtéséből*. https://m2.mtmt.hu/api/publication/34205989 *(Letöltés ideje: 2024. szeptember 11.)*

Kilgarriff, Adam – Baisa, Vít – Bušta, Jan – Jakubíček, Miloš – Kovář, Vojtěch – Michelfeit, Jan – Rychlý, Pavel – Suchomel, Vít (2014): The Sketch Engine: Ten Years on. *Lexicography* 1(1). 7–36. https://doi.org/10.1007/s40607-014-0009-9

Novák, Attila (2014): A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In: Calzolari, Nicoletta –Choukri, Khalid – Declerck, Thierry – Loftsson, Hrafn – Maegaard, Bente – Mariani, Joseph – Moreno, Asuncion – Odijk , Jan – Piperidis, Stelios (eds.): *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14).* European Language Resources Association, Reykjavík. 1068–1073.

Novák, Attila – Siklósi, Borbála – Oravecz, Csaba (2016): A New Integrated Open-source Morphological Analyzer for Hungarian. In: Calzolari, Nicoletta –Choukri, Khalid – Declerck, Thierry – Goggi, Sara – Grobelnik, Marko – Maegaard, Bente – Mariani, Joseph – Mazo, Helene – Moreno, Asuncion – Odijk, Jan – Piperidis, Stelios (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association, Portorož. 1315–1322.

Orosz, György – Novák, Attila (2012): PurePos – an open source morphological disambiguator. In: Sharp, Bernadette – Zock, Michael (eds.): *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science.* Wroclaw. 53–63. https://doi.org/10.5220/0004090300530063

Orosz, György – Novák, Attila (2013): PurePos 2.0: a hybrid tool for morphological disambiguation. In: Mitkov, Ruslan – Angelova, Galia – Bontcheva, Kalina (eds.): *Proceedings of the International Conference on Recent Advances in Natural Language Processing RANLP 2013.* Incoma, Hissar. 539–545.

Rychlý, Pavel (2007): Manatee/Bonito-A Modular Corpus Manager. In: Sojka, Petr – Horák, Aleš (eds.): *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN 2007.* Masaryk University, Brno. 65–70.

Sass, Bálint (2022): Principles of corpus querying: A discussion note. *Acta Linguistica Academica* 69(4). 599–614. http://doi.org/10.1556/2062.2022.00581

Simon Eszter – Sass Bálint (2012): Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből. *Általános Nyelvészeti Tanulmányok 24.* 243–264.

Tánczos Vilmos (1995): *Gyöngyökkel gyökereztél. Gyimesi és moldvai archaikus imádságok*. Pro Print, Csíkszereda.

Tánczos Vilmos (1999): *Csapdosó angyal. Moldvai archaikus imádságok és életterük*. Pro Print, Csíkszereda.

Tánczos Vilmos (2001): *Nyiss kaput, angyal! Moldvai csángó népi imádságok. Archetipikus szimbolizáció és élettér.* Püski, Budapest.

Vargha Fruzsina Sára (2017): *A nyelvi hasonlóság földrajzi mintázatai. Magyar nyelvjárások dialektometriai elemzése*. Magyar Nyelvtudományi Társaság, Budapest.

Vékás Domokos (2007): Számítógépes dialektológia. In: Guttmann Miklós – Molnár Zoltán (szerk.), *V. Dialektológiai Szimpozion*. Berzsenyi Dániel Főiskola, Szombathely. 289–293.