PG70

SOSEMVOLT CIGÁNY SZÓTÁR: a latin–cigány szójegyzéktől a lehetséges romani spacy-ig[1]

ROSENBERG MÁTYÁS | PhD-hallgató
matyas.rosenberg@gmail.com | DOI: 10.18135/PG70.2024.12

Bevezetés

A romani nyelvvel szemben minden korban komoly elvárásokat támasztanak. Az elvárásokat az emberek alapvetően olyan nyelvek alapján fogalmazzák meg, amelyek több beszélővel bírnak, és amelyeknek használói nyelvi, társadalmi és gazdasági szempontból eltérő helyzetben vannak a romanit beszélőkhöz képest, ezért a saját nyelvükre is másképp tekintenek, annak más funkciókat tulajdonítanak. Ezek az elvárások gyakran olyan nyelvek mintájára épülnek, amelyek széles körű beszélői bázissal rendelkeznek, továbbá jelentős kulturális és gazdasági erőforrást képviselnek, mint amilyen a magyar, a román vagy a szlovák nyelv. Az ilyen nyelvek elitjei gyakran olyan ambiciózus célokat tűznek ki nyelvük fejlesztésére, mint például a minőségi szótárak készítése, inspirálódva más nagyobb nyelvek hasonló törekvéseiből, ezzel is emelve nyelvük presztízsét. Ez az elit aztán általában a kisebb nyelvek felé fordul és megpróbálja azt is „felzárkóztatni" – jelentsen a kisebb nyelv ez esetben bármit is, ami miatt kevésbé vetül rá figyelem.

A romani nyelvű elit nem így tesz: más, kisebbségben élő népek iskolázottabb tagjaihoz hasonlóan elsősorban nem a saját igényeiket fogalmazzák meg, hanem a többségi társadalom a romani nyelvvel és általában a cigányokkal szemben támasztott igényeinek kívánnak megfelelni, még ha adott esetben interiorizálják is azokat. Ezek az elvárások pedig rendre nem teljesíthetők, jellemzően olyan tényezők miatt, amelyeket – hegemón szemszögből – régóta cipelt hiányosságnak is tekinthetnénk. Írásom amellett kíván érvelni, hogy a romani nyelv helyzete, bár számos kihívást rejt, nem tekinthető reménytelennek. Az előttünk álló akadályok ellenére léteznek lehetőségek és stratégiák, amelyek révén a romani nyelv és annak beszélői megerősíthetik pozícióikat a társadalomban, elősegítve ezzel nyelvük megőrzését és fejlődését.

[1] Intertextuális utalás Bartos (1958) által szerkesztett *Sosemvolt cigányország – Szegkovács cigány történetek* c. gyűjteményére, melyben megjelenik az önrendelkezésért küzdő, állam nélküli nemzet fogalma, valamint Szuhay (2012) *Sosemlesz Cigányország* c. művére, amely bemutatja, hogy a magát cigánynak/románok mondó vagy a mások által cigánynak mondott csoportok hogyan szerkesztik meg saját cigány/roma vagy a cigánytól elhatárolódó képüket, vagyis azt, amelyet magukra érvényesnek tartanak, illetve hogyan próbálják érvénytelenné tenni azt a képet, amelyet akár mások fogalmaztak meg róluk, akár korábban maguk fogalmaztak meg önmagukról.

A lemaradás

A magyarországi írott romani források rövid történeti áttekintése során egyfajta állandó lemaradást észlelhetünk. A legkorábbi ismert dokumentum, a Vistai Farkas (1797–1798) által összeállított latin–cigány–magyar szójegyzék keletkezési ideje sem biztosan tudható, de a kutatók egyetértenek abban, hogy a szöveget ténylegesen lejegyző személy – akinek kiléte máig bizonytalan – nem rendelkezett megfelelő romani nyelvtudással. A következő két évszázad folyamán – Sztojka (1886) 90%-ban tükörfordításokat tartalmazó munkáját leszámítva – nem jelent meg anyanyelvi szerző által írt nyelvi témájú munka, és az egyetlen bizonyosan elkészült szótár sem maradt fenn. A többségi társadalom politikai és vallási irányzataihoz és mintáihoz való igazodás, valamint a restancia ledolgozása tükröződik abban, amikor magas presztízsű(nek vélt) szövegek fordításai látnak napvilágot, így a cigány nyelvű Kommunista kiáltvány (Choli, 1975), amely 16 évvel előzte meg Máté evangéliumának romani fordítását (Choli, 1991), és amelyet hamar a „Szentírás cigány nyelvre történő lefordításaként" értelmeztek, annak ellenére, hogy a teljes bibliafordítás csak 2008-ban készült el (Vesho-Farkas, 2008), és amellyel szintén komoly szakmai problémák merülnek fel.

A 2000-es évek óta jelentős növekedés mutatkozik a romani szótárak kiadásában, így az elmúlt kb. 230 év során összesen 49 önálló művet tartunk számon. Az első időszak szótárai inkább amatőr gyűjtemények, olyan kuriózumok, melyeket szigorú tudományos kritériumok alapján nehéz értékelni: kevés, sokszor egyetlen adatközlőre támaszkodnak, ortográfiájuk inkonzisztens, és szerkesztőik gyakran nem rendelkeznek megfelelő romani nyelvtudással. Előnyük, hogy több nyelvjárást is lefednek és deskriptív jellegűek, a magyarizmusok módszeres kipurgálásán túl nem alakították a szóanyagot. Az 1967–1984 közötti rövid időszak jelenti Magyarországon a romani nyelv tudományos csúcspontját: képzett szakemberek kezdenek a romani nyelvvel foglalkozni, akik szisztematikusabb és tudományos alapokon nyugvó terepmunkát végeznek, ezáltal megbízhatóbb adatok láttak napvilágot, még ha a megvalósítás minősége olykor kívánnivalót is hagyott maga után.

1984-től a leíró szemléletet a preskriptív váltja fel, ahol a szótárak neologizmusok bevezetésével próbálják a romani nyelvet a többi európai nyelvhez közelíteni, ezzel bővítve használatának lehetőségeit. Ebben az időszakban már az oláh cigány dialektusok dominálnak, és bár a korábbi kutatásokra támaszkodnak, azok feldolgozása közel sem megfelelő. Az utóbbi két évtizedben a szanszkrit bőrbe bújt hindi kölcsönzések beáramlása a szótárak minőségének további csökkenését eredményezte. A korábbi időszak terepmunkájával ellentétben a szótárírók nem végeznek gyűjtő tevékenységet, nem rendelkeznek elegendő szakmai tudással és az alapvető romani ismereteik is megkérdőjelezhetők.

A nehézségek és az igények

Az elmúlt évszázadok során nem jött létre olyan romani szótár, amely legalább az alapvető igényeket kielégítené, ami számos jól ismert tényezőre vezethető vissza: a

romani nyelvváltozatok heterogenitása, a cigányság transznacionális jellege, az ortográfiai dilemmák, a nyelvi asszimiláció, az identitásbeli különbségek közösségenként és beszélőnként egyaránt, a romani beszélőkkel szembeni diszkrimináció és alacsony iskolázottságuk stb. mind hozzájárulnak ehhez. Az adatgyűjtést megnehezíti továbbá a nyelv titkos funkciója miatti bizalmatlanság (ld. *te na dasla avri le gaženge* 'ne adjuk ki [a nyelvet] a gázsóknak'), mely az anyanyelvűt egyébként épp úgy sújtja, mint a közösségen kívüli kutatót. A minőségi szótárakhoz elengedhetetlen az alapos terepmunka és a megfelelő korpusz, a korábbi hiányosságok – többek között az adatközlők kiválasztása, a lejegyzői kompetencia, valamint lejegyzési és nyomdai hibák – számos problémát vetnek fel. A jó minőségű korpusz és a nyelvújítás végsősorban elengedhetetlen, de ezek a feladatok, bár kihívást jelentenek, nem reménytelenek, mindössze sosem kezdődtek el, az elenyésző számú próbálkozással pedig komoly szakmai kifogások merültek fel.

A szükségletek a romani nyelvű oktatási, kulturális és más anyagok tekintetében igen sokrétűek, mégis hiányoznak az alapvető eszközök, mint például általános iskolai és gimnáziumi tankönyvek, szótárak, nyelvtanok és nyelvtankönyvek, sőt alapvetően nincsenek romani nyelvű könyvek. Az 1987 és 2010 közötti időszakot kivéve romani nyelvű újságot nem adtak ki, hírportálok pedig ezt követően sem állnak rendelkezésre, és a romani nyelvű filmek vagy filmfeliratok is hiányoznak. Ez különösen visszás, hiszen az oktatásban egyre több bilingvis diák tanul egyre tovább, akik minél tovább maradnak az oktatási rendszerben, annál valószínűbb, hogy monolingvis magyar beszélővé válnak. Emellett számos olyan újságíró ismert, aki képes lenne romani nyelven publikálni, ám nincsenek meg a szükséges platformok vagy források, hogy munkájukat piacképessé tegyék.

A lehetőségek

A romani nyelv egy pozitívuma, hogy számos beszélője van, ha a kutatások számára sokszor láthatatlanok is. Tapasztalataink szerint a romani nyelv titkos használata is addig tart, míg fel nem ismerik, hogy több száz éve léteznek szabadon hozzáférhető, írott romani források, illetve, hogy jó promptolás esetén akár még a ChatGPT 4 is rávehető a cigány nyelvű kommunikációra és fordításra – még ha igen kezdetleges és változatos módon is teszi azt. Az elmúlt 11 év terepmunkája során számos helyszínen megfigyeltem, hogy az ott-tartózkodásom végére mind a több száz településen több hasznot láttak a nyelvről való nyílt diskurzusban, mint a nyelv elrejtésében, így ez biztosan nem jelenthet komoly és tartós problémát, hiszen minél több pozitív visszacsatolás érkezik, annál kevesebb a negatív prekoncepció a többségi társadalommal szemben és fordítva.

A gyűjtési és adatfeldolgozási technikák jelentős fejlődésen mentek keresztül. Már nem szükséges súlyos fonográfot cipelni, hogy aztán a nagy melegben megolvadjon a viaszhenger, és minden tönkremenjen, mint a Csenki testvérek esetében. A táska méretű audiokazettás diktafonok és a vállra vehető VHS kamerák ideje is letűnt, a mai eszközök kiváló minőségű adatgyűjtést tesznek lehetővé. Ráadásul az utóbbi időben,

részben a COVID–19-járvány következtében egyes cigány közösségek maguk is megosztanak bárki számára elérhető videókat online platformokon, amelyek sokszor többórányi értékes nyelvi anyagot tartalmaznak.

A digitális korban, bár sok eszköz áll rendelkezésre, nem minden megoldás elérhető automatikusan. A nyelvtechnológia, mint például a szövegfeldolgozás és a gépi fordítás, kulcsfontosságú a kommunikációhoz és az információhoz való hozzáférésben. A tokenizálás az egyik alapvető lépés az NLP-ben. Bőséges forrásokkal rendelkező nyelveknél ez a folyamat viszonylag egyszerű, míg a kevesebb írott forrással bíró nyelveknél ez komoly kihívást jelent. Azonban az első lépések megvalósítása elérhető:

```
greeting = "Baxtalo rakhadjimasko djes le eftavardeše beršenge Prószéky Gáboreske!"
word_tokenize(greeting)
['Baxtalo', 'rakhadjimasko', 'djes', 'le', ' eftavardeše', ' beršenge', ' Prószéky', ' Gáboreske', '!']
```

Az ezt követő lépések, így például a PoS tagging a romani nyelv esetében még nem megoldott, mivel a gépitanulás-alapú modellek és algoritmusok hatékony működése nagy mennyiségű adatra épül, ami a romanihoz hasonló kevésbé dokumentált nyelveknél akadályt jelent (Bird, 2006), ráadásul a romani nyelvű közösségek főként szóbeli hagyományokon alapulnak, ami korlátozza az írott anyagok mennyiségét és a nyelvi regiszterek változatosságát (Grenoble & Whaley, 2006). Mindez megnehezíti a nyelvmodellezéshez, beszédfelismeréshez, gépi fordításhoz vagy hangulatanalízishez szükséges, következetes írásbeli formák meghatározását, a nyelvi és dialektális diverzitás pedig további kihívásokat jelent (Crystal, 2000). Megjegyzendő, hogy a romani nem rendelkezik olyan összetett morfológiai szerkezetekkel, mint például a magyar vagy a török, mégis összetettebb, mint az angol, ami többletmunkát jelent modellalkotás során (Comrie, 1989). Mindezek ellenére a jövőben nagy valószínűséggel egy lehetséges AI fordítás eredményét is olvashatjuk majd:

```
['Boldog', 'hetvenedik', 'születésnapot', 'Prószéky', 'Gábornak!']
```

Következtetések

Az NLP a kevésbé dokumentált, kisebb nyelvek számára jelentős kihívásokat állít a nyelvtechnológia terén, azonban ezek leküzdése hozzájárulhat a nyelvi sokszínűség digitális korszakbeli megőrzéséhez és fejlesztéséhez. Ez azt követeli meg, hogy a romani nyelv esetében is kifejlesszünk adaptált NLP-eszközöket, hasonlóan a magyar nyelvre szabott HunPoS, HuSpaCy nyelvelemző eszközökhöz. Fontos lenne továbbá a már létező nyelvmodellek (pl. PULI) továbbtanítása/adaptálása a romani nyelvre. Ezen új eszközök hozzájárulhatnának a modern és részletesebb romani nyelvű adatbázisok és korpuszok előállításához. Emellett az AI-fordítás során felmerülő lexikai hiányok azono-

sítása elősegítheti a nyelvújítást, és a közösségek bevonása a kutatásba aktivizálhatja őket, hangsúlyozva a tanulás fontosságát. Ezáltal a romani nyelv vélt vagy valós „lemaradása" ha nem is válna egy csapásra „ledolgozhatóvá", de nagy lépést tehetnénk abba az irányba.

Irodalom

Bartos T. (szerk.). (1958). *Sosemvolt cigányország – Szegkovács cigány történetek.* Budapest: Európa Könyvkiadó.

Bird, S. (2006). NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69–72).

Choli Daróczi, J. (1975). *Le kommunishtenge partosko cipipe.* Rom Som.

Choli Daróczi, J. (1991). *Nyevo teshtamento. Amare Rajesko le Jesusesko Nyevo Jekkethanipe. Le shtar evandyeliumura.* Budapest: Romano Kulturalno haj sittyimasko Jekhipe.

Comrie, B. (1989). *Language Universals and Linguistic Typology.* Chicago: University of Chicago Press.

Crystal, D. (2000). *Language Death.* Cambridge: Cambridge University Press.

Grenoble, L. A., & Whaley, L. J. (2006). *Saving Languages: An Introduction to Language Revitalization.* Cambridge: Cambridge University Press.

Sztojka (Nagyidai) F. (1886). *Ő császári és magyar királyi fensége József főherceg magyar és czigány nyelv gyök-szótára. Románé álává. Iskolai és utazási használatra.* Kalocsa: Malatin Nyomda.

Szuhay P. (2012). *Sosemlesz cigányország.* Budapest: Osiris.

Vesho-Farkas, Z. (2008). *Biblia – Dulmutano thaj nyevoteshtamenticko suntoiskiripe.* Budapest: Szent Jeromos Katolikus Bibliatársulat.

Vistai Farkas M. (1797–1798). *Cigány–magyar szótár.* (Kézirat).