Helena Grochola-Szczepanek 0000-0002-1511-0486

Instytutu Języka Polskiego PAN

helena.grochola@ijp.pan.pl

Ruprecht Von Waldenfels 0000-0001-5822-5040

Friedrich-Schiller-Universität, Jena

ruprecht.waldenfels@uni-jena.de

Rafał L. Górski 0000-0003-4727-2639

Instytutu Języka Polskiego PAN

rafal.gorski@ijp.pan.pl

Michał Woźniak 0000-0001-9018-2204

Instytutu Języka Polskiego PAN

michal.wozniak@ijp.pan.pl

A szepességi lengyel nyelvjárás korpusznyelvészeti elemzése*

A tanulmány a lengyel Szepesség nyelvjárásának korpuszát létrehozó projektet ismerteti. A lengyelországi dialektológiai kutatások többségétől eltérően korpuszunk a fiatal és a középgeneráció beszédét is tartalmazza, mivel célja a régió nyelvjárásának, szociolingvisztikai helyzetének dokumentálása is. A felvételeket nem fonetikusan, hanem a sztenderd lengyel ortográfiával írtuk át, ami nemcsak a korpuszban való egyszerű keresést teszi lehetővé, hanem azt is, hogy a meglévő eszközökkel lemmatizáljuk és morfoszintaktikai annotációval egészítsük ki a szövegeket. A fonetika iránt érdeklődő felhasználók a felvételeket mondatonként érhetik el. A cikk ismerteti a korpusz összeállításának lépéseit, és tárgyalja a lehetséges alkalmazásokat. A szerzők amellett kívánnak érvelni, hogy egy nagy korpusz, amely egy kis, homogén területet fed le, sokkal értékesebb forrás a dialektológusok számára, mint egy sor kisebb korpusz, amely egy nagyobb régiót dokumentál.

Kulcsszavak:
korpusz, beszélt nyelv, dialektológia, szepességi nyelvjárás

* Eredeti megjelenés: Helena Grochola-Szczepanek, Ruprecht Von Waldenfels, Rafał L. Górski i Michał Woźniak, „Korpus języka mówionego mieszkańców Spisza," *LingVaria* 14, 1. sz. (2019): 165–180.

1. Bevezetés

A nyelvjárási szövegek lejegyzése során mindig nehéz meghozni a döntést, hogy vajon arra kell-e törekednünk, hogy a lehető legnagyobb területen történjen a dokumentáció, beérve ezáltal egy meglehetősen felületes eredménnyel, vagy arra, hogy csak kiválasztott helyszínek elmélyült felmérését végezzük el. Képletesen szólva: tárjuk-e fel a felszín egészét, vagy végezzünk inkább helyszíni mélyfúrásokat? E dilemma nem csupán a dialektológiát vagy az areális nyelvészetet érinti, de tágabb értelemben a művészettörténetet, a botanikát, a zoológiát, a geológiát stb. is. Erre a kérdésre persze nem lehet kizárólagos választ adni, az elkövetkezőkben mégis annak bizonyítására teszünk kísérletet, hogy egy kisebb terület nyelvjárásának részletes feltérképezése tudományos szempontból kiemelten értékes lehet; annak ellenére, hogy ezáltal számos más területet hagyunk felfedezetlenül.

Az alábbi tanulmány a szepességi lengyel nyelvjárás nagy méretű elektronikus korpuszának létrehozását és annak szempontjait hivatott bemutatni.[1] E tudományos munka forradalmian újnak számít a lengyel dialektológiában; habár nyelvjárási szövegek több mint száz éve kerülnek kiadásra nyomtatott formában, ahogyan bizonyos hangfelvételek is már fél évszázada elérhetők a kutatók számára.[2] A szóban forgó munka meglehetősen nagy léptékű, digitális szövegkorpuszát hangfelvételek és azok lejegyzései alkotják. A létrehozott korpusz keresője továbbá lehetővé teszi, hogy egy adott szó előfordulásait, valamint ragozott alakjait másodpercek alatt kikereshessük. Elérhetők ugyan lengyel köznyelvi (például *A lengyel nyelv nemzeti korpusza,* a továbbiakban: NKJP[3]) és lengyel nyelvtörténeti korpuszok is,[4] a *Szepességi korpusz* az első

[1] A „Język mieszkańców Spisza. Korpus tekstów i nagrań gwarowych" ['A lengyel Szepesség lakóinak nyelve. Nyelvjárási szövegek és hangfelvételek korpusza'] című tudományos kutatás 2015–2019 között a Narodowy Program Rozwoju Humanistyki [Nemzeti Bölcsészettudományi Fejlesztési Program] finanszírozásával valósult meg (1bH 15 0166 83).

[2] Ma már léteznek teljesen digitalizált nyelvjárási gyűjtemények is, mint például *A mazóviai nyelvjárások akusztikus adatbázisa* vagy a *Lengyel dialektusok és nyelvjárások. Internetes kézikönyv.* (*Akustyczna baza danych gwar mazowieckich. Wokalizm,* 2013–2017, hozzáférés: 2021.09.22, `http://www.bazamazak.uw.edu.pl`; *Dialekty i gwary polskie. Kompendium internetowe,* pod. red. Haliny Karaś, hozzáférés: 2021.09.22, `http://www.dialektologia.uw.edu.pl/index.php`.) Míg az előbbi kizárólagosan a hanganyagon alapszik, addig az utóbbi, bár szöveget is rendel a hangfelvételekhez, csekély mérete miatt nem képezheti elmélyült kutatás alapját, csupán a dialektológiai ismeretterjesztést célozhatja. A szakirodalomban szó esik Maćkowce falu nyelvjárási korpuszának összeállításáról, e gyűjtést azonban a mai napig nem publikálták: Aleksandra Krawczyk-Wieczorek, „Automatyczna lematyzacja tekstu w zapisie fonetycznym: Korpus polskiej gwary południowokresowej," *Język Polski* 92, 1. sz. (2012): 11–19. Ugyanígy A *Lengyel Nyelvjárások Korpuszának* összeállítása szintén a tervezés fázisában van. Halina Karaś, Monika Kresa i Aleksandra Krawczyk-Wieczorek, „Towards a Corpus of Polish Dialect Texts," *Prace Filologiczne* 63 (2012): 129–145.

[3] *NKJP: Narodowy Korpus Języka Polskiego* (red. A. Przepiórkowski, M. Bańko, R. L. Górski, B. Lewandowska-Tomaszczyk, Varsó, 2012), hozzáférés: 2021.09.22, `http://nkjp.pl./`.

[4] *KorBa: Elektroniczny korpus tekstów polskich z XVII I XVIII w. (do 1772 r.),* hozzáférés: 2021.09.22, `http://korba.edu.pl/`. Lásd még Włodzimierz Gruszczyński, Dorota Adamiec i Maciej Ogrodniczuk, „Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) – prezentacja projektu

 DOI 10.31400/dh-hun.2021.5.3148

elektronikus lengyel nyelvjárási gyűjtemény, amely a mai kutatási eszközök minden kihívásának megfelel. A projekt az alábbi kutatásokkal áll szoros összefüggésben:

- az Usztja-folyóvidék lakóinak nyelvi korpusza (Arhangelszki terület, Oroszország) `http://parasolcorpus.org/Pushkino/login.php`;
- ruszin nyelvi korpusz `https://www.russinisch.uni-freiburg.de/`;
- Litvánia, Fehéroroszország és Oroszország határvidéki nyelvjárásainak korpusza `http://www.trimcocorpus.de/spoco/`.

Ezek a munkák szakmódszertani megközelítésükben, technikai megoldásaikban és (bizonyos mértékig) információs infrastruktúrájukban is megegyeznek egymással.

2. Földrajzi kiterjedés

A korpusznyelvészeti kutatás a Szepesség lengyelországi részét érinti (15 falut), és nem vonatkozik a Szepesség szlovákiai részére, amelynek kiterjedése jelentősen nagyobb a lengyelországinál. A teljes Szepesség feltérképezése szükségszerű volna ugyan, e vállalkozás a munka jelen fázisában azonban erős pénzügyi és munkaerőbeli korlátokba ütközne.

3. A korpusz összetétele

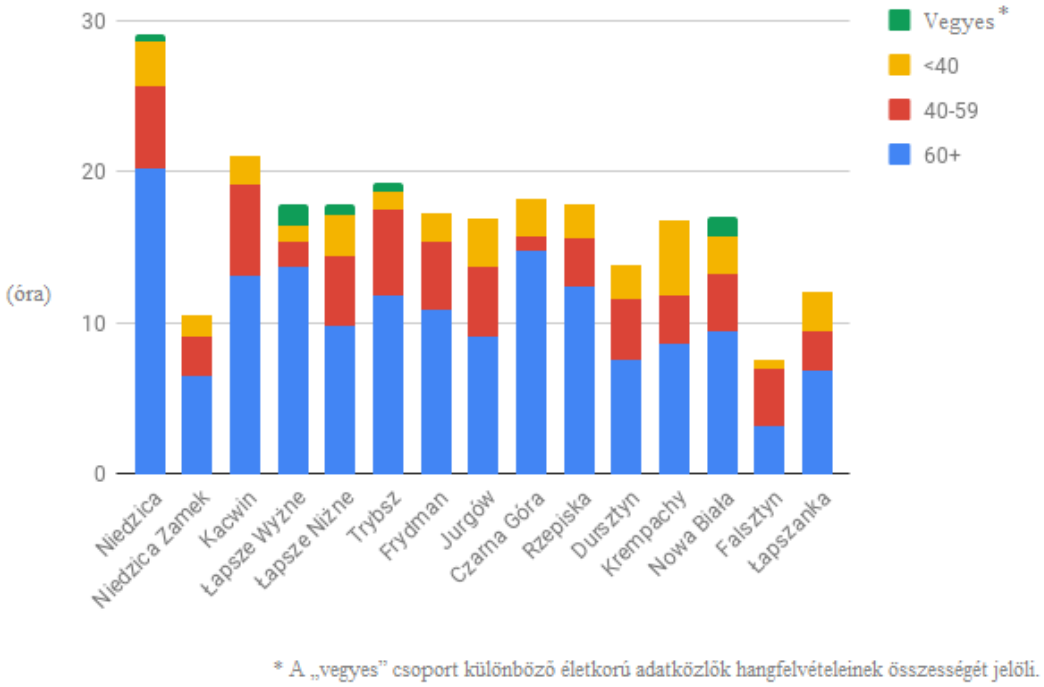
A *Szepességi korpusz* hangfájlok és lejegyzéseik egymáshoz rendelt rendszereként írható le: minden hangfájlhoz egy XML-formátumú szöveges fájl tartozik. A gyűjtemény 340 adatközlőtől származó, több mint 320 felvételből áll, ezek összesen nagyjából 250 órányi anyagot tesznek ki.[5]

A korpusz szövegszinten nem kevesebb mint kétmillió szóból, pontosabban tokenből áll.[6] Token lehet egy szó vagy egy írásjel, bár bizonyos szavak három különálló részre osztva kerültek rögzítésre, például a *chciałbyś* [~'szeretnél'] ige *chciał* [~'szeret, akar'], *by* [a feltételes mód morfémája] és *ś* [E/2 személyrag] egységekre tagolva szerepel a korpuszban, amely elemek a *chcieć* ['akarni'], *by* és *być* [létige] szótári alakokra (lemmákra) vezethetők vissza. Ez utóbbi – az NKJP terminológiája szerint – az úgynevezett agglutináns: olyan mozgó morféma, amely nemcsak az igéhez, de más szófajú szavakhoz is képes kapcsolódni. A korpusz keresési eredményei a szövegek hosszabb részleteibe ágyazva jelennek meg a keresőfelületen, ezeket a lejegyzés *szegmenseinek* nevezzük, és megközelítőleg mondatértékűnek tekinthetők. A korpusz közel kilencvenezer ilyen szegmensből áll.

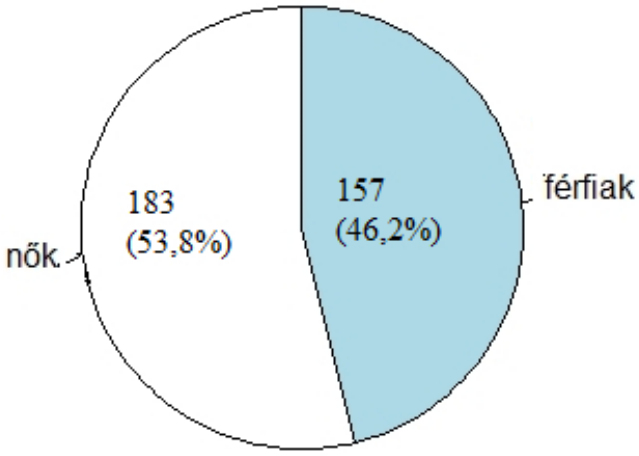
badawczego," *Polonica* 33 (2013): 311–318; Magdalena Derwojedowa, Witold Kieraś, Dorota Skowrońska i Robert Wołosz, „Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych," *Polonica* 34 (2014): 21–27.

5 A felvételek és adatközlők száma közti különbség azzal magyarázható, hogy bizonyos hangfelvételek két vagy több adatközlő közreműködésével készültek – lásd „vegyes" csoport *(1. ábra)*.

6 Adam Przepiórkowski, *Korpus IPI PAN. Wersja wstępna* (Varsawa: IPI PAN, 2004).



1. ábra. A hangfelvételek hosszúsága lakhely és korcsoport függvényében.



2. ábra. A férfiak és nők aránya a hangfelvételeken.

3.1. A szemléletesség és a hitelesség értékei közötti kompromisszum megtalálása

Egy nyelvi korpusznak reprezentatívan kell tükröznie egy adott nyelvi közösség beszédét. Ennek során figyelmen kívül hagyjuk az írott nyelvi szövegek kérdését (e feladat jóval bonyolultabb volna), és kizárólag beszélt nyelvi adatokra szorítkozunk.[7] Ebben

[7] Vö. Rafał L. Górski i Marek Łaziński, „Reprezentatywność i zrównoważenie korpusu," in Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski i Barbara Lewandowska-Tomaszczyk, red., *Narodowy Korpus Języka Polskiego,* 25–36 (Warsawa: Wydawnictwo Naukowe PWN, 2012).

 DOI 10.31400/dh-hun.2021.5.3148

az esetben a reprezentativitást hasonlóképpen értelmezhetjük, mint egy közvéleménykutatásnál, amennyiben egy korpusztól is azt várjuk, hogy abban a férfiak és a nők (vö. *2. ábra*), a fiatalok és az idősek, a felsőfokú képesítéssel rendelkezők, illetve nem rendelkezők aránya megegyezzen a magában a közösségben megfigyelhető arányokkal. Ilyen megközelítés adhat „átlagolt" képet az adott nyelvről. A korpusz létrehozásakor kismértékben el kellett térnünk ettől a módszertől, méghozzá úgy, hogy az alanyok kiválasztásánál arra törekedtünk, hogy az adatközlők között nagyobb számban legyenek az idősebbek, akik körében a nyelvjárás korábbi állapotában őrződött meg; beszédükön kevésbé figyelhető meg a lengyel köznyelv hatása. E csoport mérsékelt túlsúlya olyan adathalmazt eredményezhet, amely a nyelvjárási sajátosságokat a lehető legtisztább formájában teszi elérhetővé. Ez azonban nem lehet az egyedüli szempont. A középkorú és fiatal generációk felvételeinek segítségével ugyanis más, hasonlóan fontos jellegzetességeket is megfigyelhetünk, például a nyelvjárási elemek eltűnésének és a nyelv sztenderdizációjának folyamatát. Továbbá fontos az is, hogy a nyelvjárást hitelesen, aktuális állapotában őrizzük meg egy ilyen kutatásban, beleértve ebbe a fiatalabb korosztályok beszédét is.

Mindezeket figyelembe véve e korpusznyelvészeti kutatás legfontosabb célkitűzése, hogy a szepességi falvak lakóinak beszédét kortól, végzettségtől és egyéb tényezőktől függetlenül archiváljuk – és csak ezt követi az autentikus nyelvhasználókat előnyben részesítő szempont. Ez a megközelítés eltér a hagyományos dialektológiai kutatásoktól, amelyek a nyelvjárás legidősebb rétegének rögzítésére törekszenek, figyelmen kívül hagyva a fiatalabb, illetve a magasabb iskolai végzettséggel rendelkező adatközlőket.[8] A szepességi nyelvjárás rögzítésének vonatkozásában ezen kívül lényeges szempont, hogy az adatközlők kötetlenül beszéljenek a felvétel ideje alatt azon a nyelven, amelyet a mindennapjaikban használnak. Ezen elsősorban a nyelvjárást, illetve a regionális köznyelvet értjük.

Az alanyok kiválasztását ily módon két szempont határozta meg: egyszerre kellett biztosítani a felmérés reprezentatív voltát, valamint dokumentálni a szepességi nyelvjárás jellegzetességét. Tisztában vagyunk vele, hogy e két kritérium bizonyos mértékben kölcsönösen kizárja egymást, mindazonáltal igyekeztünk olyan kompromisszumot találni, amely segítségével a korpusz mindkét szempontból kutathatóvá válik. Így tehát, bár a felvételeken minden generáció képviseltette magát, mégsem tekinthető arányosnak a korcsoportok eloszlása. A falusi lakosság nyelvének archiválása esetében csupán törekedhetünk e kompromisszum megtalálására, amelyet valóban elérni csaknem lehetetlen (vö. *1. táblázat*).

1. táblázat. A korcsoportok megoszlása a tokenek számában.

Korcsoport	40 év alatt	40–59 év	59 év felett
Tokenek száma	281 632 (14,4%)	510 337 (26%)	1 168 900 (59,6%)

Érdemes kiemelni, hogy jelen korpusz a 2015 és 2018 közötti nyelvállapotot dokumentálja. Nem tartalmaz korábbi szövegeket, még ha léteznek is ilyen felvételek, és az is

[8] AJPP: Mieczysław Małecki i Kazimierz Nitsch, *Atlas językowy polskiego Podkarpacia. Cz. I: Mapy. Cz. II: Wstęp, objaśnienia, wykazy wyrazów* (Kraków: Polskiej Akademji Umiejętności, 1934), 18.

valószínű, hogy ez a kiterjedt dokumentációs munka később sem fog megismétlődni. E korpusz tehát szigorúan véve szinkrón természetű, a nyelvjárásnak nem múltbeli állapotát, hanem fejlődésének aktuális fázisát tárgyalja.

4. A kutatómunka állomásai

A nem sztenderd beszélt nyelvi korpusz feldolgozásának folyamata a következő állomásokra osztható: adatgyűjtés (az adatközlőkkel folytatott beszélgetések rögzítése és archiválása), adatfeldolgozás (a lejegyzés elkészítése, a morfoszintaktikai annotáció hozzáadása) és végül az eredmények rögzítése egy korpuszkeresővel ellátott adatbázis formájában.

4.1. Terepmunka

A nyelvjárási anyag rögzítését kutatók egy csoportja végezte el a terepmunka során, amely a lengyelországi Szepesség mind a 15 települését érintette. Ez a feladat semmiben nem különbözik a nyelvjárásgyűjtő kutatók hagyományos munkájától. A felvételeket az adatközlők beleegyezésével rögzítették; magától értetődik tehát, hogy a felvett párbeszédek mellőzik a spontaneitást; épp ellenkezőleg, a felvett szövegek jobbára narratív jellegű megszólalások, az alanyok közötti legcsekélyebb interakcióval. Az adatközlőktől írásbeli hozzájárulást kértünk a kutatásban való részvételhez, a felvételek felhasználásához és a korpuszban való elhelyezésükhöz. A hangfelvételeket Olympus LS-12 és Olympus LS-14 típusú diktafonok segítségével WAV-formátumban rögzítették. Tudatosan mellőzzük az MP3-formátum használatát, amely lényegesen kisebb fájlméretben menthető, éppen ezért nem használható fonetikai kutatás céljaira.

A 2015 és 2018 közötti terepmunka folyamán mintegy 600 szepességi lakossal rögzítettünk felvételt, ez összesen nagyjából 400 órányi hanganyagot tesz ki. Ebből 250 órányit szántunk lejegyzésre. A legtöbb beszélgetést az 1940-es években született személyekkel vettük fel (137 adatközlő). A legidősebb felvételi alanyunk egy 1915-ben, Frydman faluban született nő, a legfiatalabb egy 2008-ban született tanuló volt, Nowa Biała községből. Az adatközlők átlagéletkora 58 év (medián: 61, szórás: 22,3).

Az adatfelhalmozás során a társadalom nyelvre gyakorolt hatását vizsgáló szocilingvisztikai módszer került a kutatás homlokterébe.[9] Ezt a megközelítést a falusiak nyelvhasználatában megfigyelhető nagy fokú eltérések alapozzák meg,[10] amelyek jelen esetben olyan tényezők mentén alakulnak, mint például az életkor, a nem, a végzettség, a származás vagy a hosszabb életvitelszerű tartózkodás a falu határain kívül. Ezért

[9] Wladyslaw Lubaś, *Społeczne uwarunkowania współczesnej polszczyzny*: *Szkice socjolingwistyczne* (Kraków: Wydawn. Literackie, 1979). Illetve: Bogusław Dunaj, „Dialektologia a socjolingwistyka," *Acta Universitatis Lodziensis. Folia Linguistica* 12 (1986): 15–23.

[10] Helena Grochola-Szczepanek, „Badanie języka mieszkańców wsi w kontekście przemian społecznych," *Socjolingwistyka* 27 (2013): 43–53; Bogusław Wyderka, „Problemy teoretyczne współczesnej dialektologii," in Maciej Rak i Kazimierz Sikora red., *Badania dialektologiczne: Stan, perspektywy, metodologia, Biblioteka LingVariów 17,* 13–21 (Kraków: Księgarnia Akademicka, 2014).

 DOI 10.31400/dh-hun.2021.5.3148

minden adatközlő a felvétellel egy időben kitöltött egy szociológiai kérdőívet is, amely a nyelvhasználatot befolyásoló lehetséges tényezőket hivatott feltárni.[11]

4.2. A lejegyzés

Egy nem sztenderd nyelvváltozat annotációja során a köznyelvtől eltérő nyelvi rendszerrel kell szembenézni, azaz a sztenderd nyelvváltozatban ismeretlen lexémák előfordulásával, valamint fonetikai vagy morfológiai változatokkal. A lejegyzés módját az a tényező is meghatározta, hogy e szövegek egy nyelvi korpusz létrehozása céljából készülnek. Tehát a hangfelvételek annotációját a korpusz összeállíthatóságának igényeihez kell szabni, különböző nyelvi és műszaki szabályoknak megfelelően.[12]

A lejegyzést meghatározó elvek kidolgozása kulcsfontosságú már a korpuszkészítés tervezési folyamatában. A készítőkben felmerülő problémákra a következő három lejegyzési mód adható válaszként:

1. fonetikus lejegyzés (a szlavisztikának megfelelő IPA-variáns);
2. félig fonetikus lejegyzés;
3. a sztenderd lengyel nyelvváltozat helyesírásának követése.

Esetleges negyedik megoldásként egy önálló szepességi helyesírás kialakítása is felmerülhet, ez azonban a korpusz esetében nem jöhetett szóba.

Az első megoldás, érthető módon, több akadályba ütközik. Először is, ez a teljesen fonetikus lejegyzés meglehetősen munkaigényesnek bizonyulna, mivel a lejegyzőknek minden esetben el kellene dönteniük, hogy az adott szóelőfordulás pontosan milyen kiejtésben realizálódott. Ez a döntés ráadásul gyakran vitatható. Ugyanakkor mivel a lejegyzés mellett a hangfelvétel is rendelkezésre áll, a fonetika iránt érdeklődők vagy akár a fonetikus ábécét nem ismerő laikusok is tanulmányozhatják az adatközlők kiejtésének sajátosságait.

Miközben a harmadik pontban említett normalizált lejegyzés hátránya a fonetikus írásmóddal szemben az, hogy közel sem azonosítható a tényleges kiejtéssel vagy annak idealizált változatával. Így például a selypes fogréshangokat <s> vagy <sz> alakban is lejegyezhetjük, annak függvényében, hogy az adott hang hogyan szerepel a köznyelvi szóban – bár az adatközlő kiejtésében ugyanarról a hangról beszélhetünk.

Ezen kívül viszont két tényező is a normalizált lejegyzési mód mellett szól. Először is ily módon megkönnyíthetjük a korpuszban történő keresést, leginkább abban az esetben, ha a hangérték a meghatározó (fonetika, fonológia, morfológia, esetleg ezeknek a szociolingvisztikához kapcsolódó határterületei), hozzásegítve a kutatót ahhoz, hogy a

[11] A terepmunka módszertani leírása egy külön szaktanulmányban kapott helyet. Helena Grochola-Szczepanek, „Nowe badania języka mieszkańców wsi regionu polskiego Spisza," in Błażej Osowski, Paulina Michalska-Górecka, Justyna Kobus i Agnieszka Piotrowska-Wojaczyk, red., *Język w regionie, region w języku 2,* 103–119 (Poznań: Wydawnictwo „Poznańskie Studia Polonistyczne", 2017).

[12] A lejegyzés irányelveinek kialakításáról külön szaktanulmányban esik szó: Helena Grochola-Szczepanek i Michal Woźniak, „Transkrypcja języka mieszkańców wsi w aplikacji ELAN w Korpusie Spiskim," Renata Przybylska, Maciej Rak i Agata Kwaśnicka-Janowicz, red., *Historia języka, dialektologia i onomastyka w nowych kontekstach interpretacyjnych,* 267–278 (Kraków: Wydawnictwo Uniwersytetu Jagiellonskiego, 2018).

normalizált fonetikus lejegyzés segítségével találja meg a megfelelő hangfájlt. Másodszor, a normalizált lejegyzés eszközéül szolgálhatnak a lengyel köznyelv morfológiai jelölései is (vö. 4.3. alfejezet).

A fenti okokból döntöttünk a harmadik, és bizonyos esetekben az első megoldás mellett (ekkor mindkettőt alkalmaztuk). Azért nem használtunk félig fonetikus lejegyzést, mert az a módszer egyesíti magában a két bemutatott megoldás hátrányait, ellenben nem jár együtt azok előnyeivel. Egyrészt e megközelítés nem teszi lehetővé a már meglévő nyelvelemzési eszközök használatát, és nem könnyíti meg a keresést, hiszen tükrözi a kiejtésbeli következetlenségeket, másrészt – ahogy már a neve is sugallja – csupán megközelíti, sőt meglehetősen gyengén adja vissza a tényleges kiejtést.

A morfológia esetében hasonló helyzettel állunk szemben. Azok a morfémák, amelyek csak alakjukban különböznek a sztenderd nyelvváltozatban használatostól, a nyelvi sztenderd készletével jelölhetők. Más szóval, ha a nyelvjárási szó morfémája visszavezethető a sztenderd nyelvváltozatban szereplő morfémára, akkor a sztenderd változat jelenik meg a lejegyzésben. Ha azonban a szepességi morféma a sztenderdben használthoz képest más morfémára vezethető vissza, akkor a szepességi változatot visszaadó írásmódhoz folyamodunk. Például a *chodzić* 'jár' ige egyes szám első személyű múlt idejű alakját *chodziyłef* alakban, nem pedig *chidziłem* alakban tüntetjük fel, hiszen az első az *-ech* régies múlt idejű igevégződésre vezethető vissza, míg a második, az *-em,* a ma használatos múlt idejű személyrag továbbélése. A *biała* 'fehér' melléknév ugyanakkor a sztenderd alakban szerepel, mivel a köznyelvi és a szepességi változat is ugyanarra a morfémára vezethető vissza; az eltérésre rendszerszerű hangváltozás ad magyarázatot.

Tehát négy különböző esettel állunk szemben, amelyek mindegyike a lengyel köznyelvhez való közeledés különböző fokozatának tekintendő, ennélfogva különbözőképp írhatók le:

1. A köznyelvnek megfelelő vagy szabályszerűen módosult alakváltozatok (például a zártabb magánhangzók) lejegyzése a sztenderd szerint történt.
2. Morfológiailag eltérő szavak: a lengyel köznyelvhez közvetlenül kapcsolódó, de egy-egy morfémában vagy ragozott alakban eltérő nyelvi egységeket a lejegyzésben mindkét – sztenderd és nyelvjárási – változatban feltüntetjük // jellel elválasztva.
3. Azok az alakváltozatok, amelyek megfelelői a lengyel köznyelvben megtalálhatók, de szemantikájukban eltérnek attól, a sztenderd szerint szerepelnek a lejegyzésben ^ szimbólummal jelölve.
4. Azok a lexémák, amelyek a lengyel köznyelvben nem, kizárólag a nyelvjárásban jelennek meg, hangzás utáni fonetikus alakban szerepelnek # szimbólummal ellátva. A kutatás későbbi munkafázisaiban ezek egységesítését is elvégezzük.

 DOI 10.31400/dh-hun.2021.5.3148

2. táblázat. A lejegyzés és a szóosztályok lejegyzésben szereplő annotációjának példái, valamint ezek előfordulásának száma a korpuszban.

Szóosztályok száma	Példák (hangzás utáni alakban)	A lejegyzésben szereplő adat	Annotáció	Tokenek száma
1	*mlyko* 'tej'	mleko	nincs	1 844 353
2	*dałak* 'adtam'	dałam//dałak	//	116 516
3	*ślafrok* 'köpeny, köntös'	szlafrok	^	35 086
4	*odziywacka* 'szabó'	*odziywacka* (végleges formában: *odziewaczka*// *odziywacka*)	#	70 812

A 2-es és 4-es számú osztályokba tartozó szavakat tehát egyszerre két alakban, a köznyelvi és a nyelvjárási változatban is feltüntettük, az érvényes helyesírási szabályoknak megfelelően. A köznyelvi változat annotációja mesterséges, míg a nyelvjárási változaté megközelítőleg tükrözi a szepességi nyelvjárási kiejtés sajátosságait. Mindkét változat feltüntetése lehetővé teszi, hogy a sztenderd és a nyelvjárási alakváltozatot egyszerre lássuk. A lejegyzésben jelöltünk egyéb nyelvjárási változatokat is: elkülöníthetők több szóból álló kifejezések (*młodzi panowie* a *państwo młodzi* vagy a *nowożeńcy* alakok helyett ['friss házasok, ifjú pár']), a sztenderdtől eltérő szószerkezetek (*ku moście* ['a hídhoz'] a *do mostu* vagy a *w stronę mostu* alakok helyett) és jövevényszavak<*pridi*> (a *przyjdź* ['jön'] alak helyett – szlovák hatás), <*dawaj sało*> (a *słonina* szó ['szalonna'] helyett – orosz hatás).

Az összes hangfelvétel visszajátszása és a lejegyzések elkészítése a korpuszkészítés folyamatának legkimerítőbb része. Egyórányi hanganyag átdolgozása, dokumentációja és a lejegyzés nagyjából 40 munkaórányi feladat. A hangfelvétel és a lejegyzés összevetését néhány munkatárs több ízben elvégezte. Kiemelendő továbbá, hogy a tisztázat elkészítése nem esik egybe a lejegyzés befejezésével: a hallás utáni leírás során számos hiba merülhet fel, ezért lehetőséget kell biztosítani azok későbbi módosítására.

4.3. A morfoszintaktikai annotáció

Az elkészült lejegyzéseket XML-formátumba kódolva mentjük el a felvételekhez külön fájlként hozzárendelve. Ekkor zárul le a fájlok kézi szerkesztése, innen automatikus rendszerek veszik át a munkát. Az első ilyen munkafolyamat a morfoszintaktikai annotáció elvégzése, tehát a tokenek szótári alakjának (lemma) és nyelvtani sajátosságainak azonosítása. Irányadóként az NKJP egységes jelölőelem-készletét (a nyelvtani kategóriák és a hozzájuk tartozó jelölések címkézőrendszerét) határoztuk meg. Például az *akordeonie* ('harmonika') token a következő annotációt kapja subst:sg:loc:m3, ahol a kettősponttal tagolt egységek megfeleltethetők a szófajnak, a számnak, a nyelvtani esetnek és a nemnek.

Az automatikus nyelvi elemzést két lépésben végeztük. A köznyelvi szavak elemzésére a *Pantera* programot használtuk. Ezek közé tartoznak a sztenderddel megegyező

szófajú és homonim jelentésű nyelvjárási szavak is (lásd a 4.2. alfejezetben tárgyalt osztályozás hármas számú osztályát). Azoknak a szavaknak az elemzései pedig, amelyeket a köznyelvre adaptált szoftver nem képes kezelni, a *Kuźnia* nevű program kiegészítő adatbázisán alapulnak.

4.4. A korpusz elkészítése

A következő lépés a szövegek és a hangfájlok feldolgozása és átalakítása egy konkordanciaprogram segítségével. Ez a fázis ugyancsak teljesen automatikus, és a következő szakaszra tagolható:

1. a hangfájlok feldarabolása a lejegyzésének megfelelően;
2. a szöveges fájlok átalakítása a CWB[13] programnak megfelelő formátumra;
3. a metaadatok hozzárendelése a fájlokhoz;
4. a személyes adatok anonimizálása;
5. magyarázatok hozzárendelése a sztenderdből hiányzó szavakhoz;
6. az adatok mentése a CWB-adatbázis formátumában;
7. az adatbázis csatlakoztatása a webes felülethez.

5. Szabványos eszközök

Egy beszélt nyelvi korpusz többlépcsős munkával állítható össze, minden egyes lépés során jelentős erőfeszítésre és gondosságra van szükség, hogy megőrződjön az adatok konzisztenciája. A rendelkezésre álló eszközök nagymértékben felgyorsították és megkönnyítették a munkafolyamatot. A *Szepességi korpusz* az alábbi eszközök felhasználásával valósulhatott meg:

> 1. ELAN – a Max Planck Pszicholingvisztikai Intézet multimédiás források annotációjára használható programja.[14] A programot széles körben alkalmazzák beszéd rögzítésére, adatok feldolgozására és archiválására. Az ELAN a hangfelvételeket feldolgozók és lejegyzők számára létrehozott speciális munkakörnyezet. Lehetővé teszi az adatközlések szegmentálását és többszintű címkézését. Az ELAN a lejegyzést XML-formátumban rögzíti, amely a legnépszerűbb és legszélesebb körben használt formátum az efféle kutatásokban, ezzel egy időben a hangfelvételeket WAV-formátumba írja át, amely a hangfelvételek rögzítésének egyik legnépszerűbb módja.
>
> 2. *Pantera* nyelvi elemző. A morfoszintaktikai annotációra alkalmazott szoftver igényei döntő szerepet játszottak abban, hogy az adatok sztenderd nyelvváltozatban történő lejegyzése mellett döntsünk. Ez az alkalmazás

[13] The IMS Open Corpus Workbench, hozzáférés: 2021.09.22, `http://cwb.sourceforge.net`.

[14] Hennie Brugman and Albert Russel, „Annotating Multimedia/Multi-modal Resources with ELAN," in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva, eds., *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC'4),* 2065–2068 (Lisbon: European Language Resources Association [ELRA], 2004), `http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf`.

 DOI 10.31400/dh-hun.2021.5.3148

lehetővé teszi, hogy egy adott szó összes előfordulására rákereshessünk, alakváltozattól függetlenül. Ugyanígy elérhető bármely szó a ragozott alakja alapján is. A lengyel fejlesztésű elemzők kizárólag a lengyel köznyelvre adaptálva működnek, ezért nem használhatók olyan szavak esetében, amelyek nem felelnek meg egyetlen lengyel köznyelvi szó szótári alakjának sem. A szepességi kutatási projektben azért használtuk a *Panterá*t, mert kimagaslóan pontos annotációt tesz lehetővé.

3. A Lengyel Tudományos Akadémia Számítástechnikai Kutatóintézet (IPI PAN) által lengyel nyelvű ragozási szótárak összeállítására kifejlesztett *Kuźnia* program. Mivel a meglévő elemzők nem boldogulnak a nyelvjárás lexikai elemeivel, azok morfoszintaktikai annotációját kézzel, a *Kuźnia* szoftver segítségével végeztük el. Ez lehetővé teszi ragozási paradigmák, szótári alakok, illetve (a sztenderdből hiányzó szavak szótárához használt) magyarázatok hozzárendelését a lengyel köznyelvben ismeretlen lexémákhoz.

A rendelkezésre álló digitális eszközök használata számos jól látható előnnyel bír: felgyorsítja a munkát, és lehetővé teszi a feldolgozott adatok következetes kezelését, valamint olyan normatív szabályozások eszközlését (például XML-formátum, az NKJP annotáló rendszerének alkalmazása), amelyek növelik az adatok egységességét, ezáltal könnyítve azok felhasználhatóságát más hasonló jellegű kutatások során, illetve biztosítva azok nagyobb stabilitását a hosszú távú adatmegőrzés számára.[15] Mindazonáltal meg kell jegyeznünk, hogy a szabványos eszközök használata nem sztenderd nyelvváltozatok feldolgozása során szükségszerű módosításokat követel meg. Ezen módosítások a következő két csoportra bonthatók:

a) Ami az adatok sokféleségét illeti: ahogy fent említettük, ahhoz, hogy a nyelvi elemző szoftver használatát lehetővé tegyük, a lejegyzés során normalizálnunk kellett a szavak helyesírását.

b) Az eszközöket illetően: a *Kuźnia* a nyelvi sztenderd ragozási rendszerének kezelésére készült. Mivel a nyelvjárás több helyen eltér ettől, elengedhetetlenné vált bizonyos módosítások bevezetése, amelyek lehetővé tették a kizárólag nyelvjárási lexémák paradigmáinak kezelését. Minthogy ezek a szavak két (normalizált és nyelvjárási) alakváltozatban is szerepelnek a korpuszban, indokolttá vált a *Kuźnia* szoftver oly jellegű bővítése, amely megengedi, hogy egy adott lexéma két változatát is tárolja az adatbázisban. Ezeket a módosításokat az tette lehetővé, hogy a *Kuźnia* nyílt forráskódú alkalmazás, amelynek forráskódja BSD-licenc alatt érhető el.

6. Hatás

A kutatási projekt elsődleges eredményének a beszélt nyelvi szövegekből álló korpusz létrejötte tekinthető, amely a `https://spisz.ijp.pan.pl` címen érhető el az

[15] Ruprecht von Waldenfels and Michał Woźniak, „SpoCo – A Simple and Adaptable Web Interface for Dialect Corpora," *Journal for Language Technology and Computational Linguistics* 31 (2016): 155–170.

interneten. Ehhez kapcsolódik a korpusz igényeit kiszolgáló online felület is, amely a korábbi Spoco-projekt módosított verziójaként jött létre.

A korpusz lehetővé teszi a szövegegységek szerinti keresést, amelyek a hangfelvétel egy részletéből és a lejegyzésükből állnak. Egy CWB-alapú keresőről van szó, amelyet széles körben használnak korpuszkészítésre, mivel ez a keretrendszer engedi az összetett keresést, valamint lehetőséget teremt a kvantitatív és a kvalitatív elemzés számára is. A korpuszban a CQL formális lekérdezőnyelv segítségével lehet keresni, ez nagy fokú szabadságot biztosít a felhasználónak, ugyanakkor a lekérdezőnyelv ismeretét is feltételezi. A felhasználóbarát kialakítás érdekében ugyanakkor a lekérdezőnyelv használata helyett egyszerű szerkezeti egységekből álló kereséseket indíthatunk a felületen – elegendő bevinni a kívánt szót valamelyik keresőmezőbe. A felhasználói felület négyféle keresőmezőt kínál: szóalak *(token)* – az adott szó normalizált alakjának szövegszintű keresését teszi lehetővé; lemma – az adott szótári alakú szó minden alakváltozata kikereshető; nyelvjárási alak – kikeresi a kívánt szó nyelvjárási alakváltozatát; és grammatikai tulajdonság – nyelvtani sajátosságok szerinti szókeresést tesz lehetővé. A lekérdezési eredményeket szűrők segítségével tudjuk korlátozni (szűkíteni lehet: nem, nemzetiség, végzettség, lakhely, születési év és adatközlő szerint).

Az imént tárgyalt modul az egyszerű lekérdezési feltételek szerinti keresést teszi lehetővé. Az összetett keresés (például az összes olyan szó megkeresése, amelynek eltér a köznyelvi és a nyelvjárási alakja) bár jóval nagyobb szabadságot biztosít a felhasználónak, megkívánja a CQL ismeretét. A keresési eredmények a lekérdezési feltételek által meghatározott szövegegységekbe ágyazva jelennek meg. Az összes szegmenshez tartozó hangfájl meghallgatható, ezzel egy időben a vonatkozó normalizált és nyelvjárási szövegrészletet is meg lehet tekinteni. A keresési eredmények kétféleképpen jeleníthetők meg: a szegmensek egyszerűsített megjelenítése, valamint KWIC-listaként (Key Word in Context – kontextusos kulcsszó), amely a szövegegységeket három oszlopra tagolja: a keresett elemet megelőző kontextus, a keresett elem és az azt követő kontextus. Mindkét megjelenítési mód lehetővé teszi a keresési eredmények csoportosítását, a szegmensekhez rendelt metaadatok, valamint a szélesebb (hét szövegegységből álló) kontextus megtekintését. Ezenfelül a keresési eredmények (a lejegyzések és a hangértékek egyaránt) letölthetők.

A szövegkorpuszhoz és a hangfelvételek gyűjteményéhez továbbá egy szótár is kapcsolódik. E szótár olyan szavakból és kifejezésekből áll, amelyek a hangfelvételen szerepelnek, viszont a lengyel köznyelv számára ismeretlenek, illetve amelyek a nyelvjárásban más jelentéshez köthetők, mint a sztenderd nyelvváltozatban. A valódi tájszók a szócikkekben köznyelviesített vagy nyelvjárási alakváltozatban szerepelnek. A jelentésbeli tájszók szócikkei a sztenderd és a nyelvjárási alakváltozatot is mutatják. A szótár továbbá olyan kifejezéseket is tartalmaz, amelyek szerepelnek lengyel nyelvű szótárakban, ám régies vagy nyelvjárási szónak számítanak.

7. Alkalmazás

A korpusznyelvészet mindenekelőtt olyan módszertan, amely számos lehetőséget kínál, ugyanakkor korlátokat is állít a kutatók elé. Az egyik ilyen korlátot a hiányzó adatok jelentik: az a tény, hogy valamely kifejezés nem szerepel a korpuszban, nem

 DOI 10.31400/dh-hun.2021.5.3148

jelenti egyértelműen azt, hogy ez a szó vagy alak ne képezné a nyelvhasználat részét. A korpusz összeállítói gyakorta találkoztak ezzel a problémával a ragozott alakok kapcsán, hiszen néhány szó egészen sajátos paradigmával rendelkezik, amelynek nem minden eleme jelenik meg az adatközlők szövegeiben. Az ilyen alakok korpuszba való felvételéhez a kutató nyelvi kompetenciájára kell hagyatkoznunk.

7.1. Szociolingvisztika és nyelvföldrajz

A lekérdezések eredményeinek szociológiai adatokkal történő szűrhetősége lehetővé teszi az adatközlők bármely csoportjának részletes elemzését. Lekérdezhetjük például a *dom* 'ház' szó összes előfordulását azokban a közlésekben, amelyeket 1950 előtt született, Niedzica (Nedec) községben élő nőkkel vettek fel. Az ilyen jellegű szűrések lehetővé teszik az olyan demográfiai és társadalmi tényezők nyelvre gyakorolt hatásának korpuszalapú vizsgálatát, mint az életkor vagy a nem kategóriája.[16]

7.2. Nyelvtan : ragozás és szintaxis

Ahogy azt már fentebb említettük, a korpusz elemeit morfológiai annotációval láttuk el. Ez a jelölőrendszer egy sor nyelvtani kutatás elvégzését biztosítja ragozás, szóalkotás és szintaxis viszonylatában egyaránt. Meg kell jegyeznünk azonban, hogy az annotáció az adott szóhoz tartozó nyelvtani kategóriákra vonatkozik, és nem a külön morfémákra ; morfémák keresésére karakterszekvenciák[17] megadásával van lehetőség, amely csupán megközelítőleges keresési eredményt biztosít. Hasonló helyzet áll fenn a megadott ragozási jellemzők szerinti szerkezetekre irányuló keresés esetében is.[18] Ekkora méretű korpusz esetén ugyanakkor az adott nyelvtani jelenségek előfordulásának száma már elegendő ahhoz, hogy mennyiségi mérést végezhessünk, és elkülöníthessük egymástól a marginális és a tipikus eseteket.

7.3. Pragmatika

Meg kell jegyeznünk, hogy a *Szepességi korpusz* beszélt nyelvi szövegek olyan gyűjteménye, amely méretében megközelíti az NKJP társalgási alkorpuszát. Az adatbázisban történő keresés lehetővé teszi a szélesebb kontextus vizsgálatát is, ami a pragmatika területén végzett kutatómunka alapjául is szolgálhat. A szélesebb szövegkörnyezet vizsgálatának lehetősége elengedhetetlen a téma-réma szerkezetek tanulmányozása esetén, illetve akkor, amikor e szerkezetek szintaxisra gyakorolt hatását kívánjuk elemezni. (Miközben ez a jelenség még a sztenderd nyelvváltozat esetében sem gyakran kerül az elemzések középpontjába.) További feladat, amelynek megoldására a *Szepességi korpusz* alkalmas lehet, a diskurzusjelölők korpuszalapú kutatása.

[16] Külön megjelenő cikkben tárgyaljuk azt, miként befolyásolják a metaadatokba felvett jellemzők az adatközlők kódját.

[17] A kicsinyítő képzős szavakat például *-eczek, -eczka, -eczko* végződéssel lehet lekérdezni [lemma=".+*ecz*(*ek*|*ka*|*ko*)"].

[18] A folyamatos jövő idő a *być* (lét)ige és a főnévi igenév vagy az ige múlt idejű alakjának egymást követő sorrendjével fejezhető ki. Ez a séma azonban nem alkalmazható az olyan szekvenciák esetén, mint például *będzie szybko szedł* [a *być* ige ragozott alakja és az *iść* – 'megy' – ige múlt idejű alakja között határozószó – *szybko* – áll, jelentése : 'ő (hímnem) gyorsan fog menni' – *a ford.*].

7.4. Fonetika

A hanganyag kiváló alapul szolgálhat fonetikai és prozódiai kutatásokhoz. Éppen ezért már a kutatási projekt kezdeti szakaszában meghoztuk a döntést: a hangfelvételeket veszteségmentes formátumban fogjuk rögzíteni, nem pedig a legnépszerűbb MP3-formátumban, amelynek vitathatatlan előnye, hogy kisebb fájlméretben menthető – ugyanakkor a veszteségmentes formátum jóval alkalmasabb a kutatásokra.

Az adatközlések részleteit (általánosan 15 másodperc) le lehet tölteni és fel lehet dolgozni fonetikai elemzőprogramok (például *Praat*) segítségével. Mindazonáltal nem hallgathatjuk el, hogy nem minden felvétel minősége felel meg fonetikai kutatás igényeinek. Elsődleges feladatunknak azt tekintettük, hogy minél nagyobb mennyiségű adatot gyűjtsünk; a munkatársak nem töröltek egy felvételt sem a gyenge hangminőség miatt.

Megemlítendő még, hogy a keresőrendszer lehetővé teszi például a két zárhang között álló magánhangzók kigyűjtését is. A metaadatok segítségével pedig tovább szűkíthető a lekérdezett eredmények listája nem, életkor vagy lakóhely szerint. A hangfelvételek ösztönözhetik a prozódia területén végzett kutatásokat is, amelyek a szintaxis vizsgálatának fontos elemét képezik.

7.5. Szókészlet

Bár egy ilyen méretű korpusz nem meríti ki az alapos lexikai kutatás igényeit, a mintegy 10 000 különböző szót[19] számláló gyűjtemény mégis tekintélyesnek nevezhető. Igaz, a szókészlet gazdagsága függ a felvételeken hallható beszédtémáktól is, amelyek gyakran érintik a hagyományok, szokások, gyermekjátékok, mondókák és a mezőgazdasági munka stb. területeit. Ezen felül a korpusz lehetőséget biztosít a szókapcsolatok, többszavas kifejezések megfigyelésére, továbbá konkordanciaszótár elkészítésére.

Mindehhez hozzátehetjük, hogy a korpusz alkalmas lehet a szepességiek kultúráját és szokásait illető ismereteink bővítésére is.

8. Összegzés

A fent leírt problémák egyike sem oldható meg nagy méretű, digitalizált korpusz létrehozása nélkül. Nem helyettesítheti sem szótár, sem szöveglejegyzések, sem hangfelvételek puszta gyűjteménye.

Hangsúlyoznunk kell, hogy a leírt kutatási projekt érdeme a feldolgozott anyag mennyiségében keresendő. Egy kis méretű korpusz még viszonylag gyakori jelenségekre is csak kevés előfordulást tartalmaz, így csupán meglehetősen korlátozott kutatásokat tenne lehetővé. A lengyel dialektológiának a *Szepességi korpusz* révén olyan eszköze jött létre, amely bár pontszerű, mégis lehetővé teszi a Kis-Lengyelország déli részén található nyelvjárások egyikének alapos és sokrétű tanulmányozását.

Fordította: Bali Farkas Péter

[19] Összehasonlításként az *Árva mente nyelvjárásának szótára* (KąśSGO) megközelítőleg 28 000 szócikkből áll, beleértve a sztenderd nyelvvel közös szavakat is.

 DOI 10.31400/dh-hun.2021.5.3148

A Spoken Corpus of Inhabitants of Polish Spisz

The article describes a dialect corpus project that documents the dialect of Polish Spisz. In contrast to the majority of dialectological research in Poland, our corpus also includes the speech of the youngest and middle generations, as its aim is also to document the sociolinguistic situation of the dialect of the region. Recordings have been transcribed into standard Polish orthography, not phonetically, which makes it possible not only to easily search the corpus but also to use existing tools to lemmatize and add morphosyntactic annotation to the texts. Users interested in the phonetic layer can access the recordings on a per-utterance basis. The article describes the stages of compiling the corpus and discusses its potential applications. The authors argue that a large corpus which covers a small, homogeneous area is a more valuable resource for dialectologists than a series of small corpora documenting a larger region.

Keywords:
corpus, spoken language, dialectology, Spisz dialect