

Uráli nyelvek nyílt hozzáférésű magyarországi korpuszai

Horváth Csilla¹

¹MTA Nyelvtudományi Intézet
Budapest, Benczúr u. 33., e-mail: naj.agi@gmail.com

Körülmények

- Az érintett uráli nyelvek Oroszország területén beszélt veszélyeztetett kisebbségi őshonos nyelvek, alacsony presztízzsel, folyamatosan csökkenő beszélői létszámmal.
- Az Oroszországban beszélt uráli nyelvek nyelvtechnológiai vizsgálata mindössze pár évtizedes múltra tekint vissza. A legtöbb uráli nyelvre nem készült szisztematikusan épített, konzekvensen lejegyzett, morfológiailag annotált szövegtörzs.
- Az elkészült szövegtörzsek a nehezen hozzáférhető korábbi, illetve a teljesen új szempontok alapján gyűjtött új szövegek segítségével nem csak az adott nyelvekről rendelkezésre álló legkorszerűbb információkat tartalmazzák, de további nyelvészeti kutatásokat is lehetővé tesznek.

Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben (ERC_15_HU, OTKA 118079)

Szinjai hanti adatbázis

- készítő: Arkagyij Longortov, Ruttkay-Miklán Eszter
- tartalom: 4 szöveg W. Steinitz gyűjtéséből, 3 szöveg Ruttkay-Miklán E. gyűjtéséből
- feldolgozás: lejegyzés, transzkripció, morfológiai elemzés, fordítás (angol, német, magyar)

Szurguti hanti adatbázis

- készítő: Gugán Katalin (MTA NyTI), Schön Zsófia (LMU München)
- tartalom: 5 szöveg H. Paasonen gyűjtéséből, 2 szöveg A. Sz. Pesikova gyűjtéséből, 8 szöveg Csepregi M. és Gugán K. gyűjtéséből
- feldolgozás: lejegyzés, transzkripció, annotálás, fordítás (angol, német, orosz, magyar)

Tundrai nyenyec adatbázis

- készítő: Mus Nikolett (MTA NyTI)
- 13 szöveg T. Lehtisalo gyűjtéséből, 31 szöveg K. Labanauskas gyűjtéséből, 16 szöveg a Наръяна нерм újságból, 11 szöveg E. T. Puskarjova és L. V. Homics gyűjtéséből, 5 szöveg Mus N. gyűjtéséből
- feldolgozás: lejegyzés, transzkripció, morfológiai elemzés, fordítás (angol, német, magyar, orosz)

Udmurt adatbázis

- készítő: Asztalos Erika (MTA NyTI), Tánczos Orsolya (MTA NyTI)
- tartalom: 6 szöveg Munkácsi B. gyűjtéséből, 12 szöveg Y. Wichmann gyűjtéséből, 6 szöveg a Марајко blogról, 5 szöveg a Мынам малпанъёсы blogról
- feldolgozás: lejegyzés, transzkripció, morfológiai elemzés, fordítás (angol, német)

Udmurt korpusz

- készítő: Ekaterina Georieva (MTA NyTI), Tyimofej Arhangelszkij (Universität Hamburg)
- tartalom: 19 szöveg E. Georieva gyűjtéséből
- feldolgozás: lejegyzés, morfológiai elemzés, fordítás (orosz)

A többnyelvűség megnyilvánulásai finnugor nyelvű közösségekben (OTKA FNN 107793)

Északi manysi adatbázis:

- készítő: Németh Szilvia (ELTE), Anasztaszija Szajpaseva
- tartalom: 11 szöveg Németh Sz. és A. Szajpaseva gyűjtéséből
- feldolgozás: lejegyzés, annotáció, fordítás (orosz, magyar)

Baskíriai mari adatbázis

- készítő: Anasztaszija Szajpaseva
- 7 szöveg A. Szajpaseva gyűjtéséből
- feldolgozás: lejegyzés, transzkripció, annotálás, fordítás (orosz)

Udmurt adatbázis

- készítő: Salánki Zsuzsa (ELTE), Antal Gergely (ELTE)
- tartalom: 4 szöveg Salánki Zs. gyűjtéséből
- feldolgozás: lejegyzés, annotáció, fordítás (orosz, angol)

Déli manysi korpusz

- készítő: Szilágyi Norbert (SZTE)
- tartalom: 10 szöveg Munkácsi B. gyűjtéséből, 19 szöveg A. Kannisto gyűjtéséből
- feldolgozás: digitalizáció, annotálás, fordítás (magyar, angol)

Hivatkozások

1. Simon, E., Mus, N.: Languages under the influence: Building a database of Uralic languages. In: M. Tyers, F., Riessler, M., Pirinen, T., Trosterud, T. (szerk.) Proceedings of the Third International Workshop on Computational Linguistics for Uralic Languages, Saint Petersburg. (2017) 10–24.
2. Arkhangelskiy, A., Georieva, E.: Sound-aligned corpus of Udmurt dialectal texts. In: Pirinen, T., Riessler, M., Rueter, J. & Trosterud, T. (szerk.), Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (2018) 26–38.
3. Szilágyi N., Horváth, Cs., Vincze, V., Nagy, Á.: A szó elszáll, az írás megmarad? Nyelvtechnológiai eszközök a déli manysi nyelvre. XIV. In: Vincze, V., szerk.: Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2018) 413–419.