Online morfológiai elemzők és szóalak-generátorok kisebb uráli nyelvekhez

Bakró-Nagy Marianne[2], Endrédy István[1], Fejes László[2], Novák Attila[1], Oszkó Beatrix[2], Prószéky Gábor[1], Szeverényi Sándor[2], Várnai Zsuzsa[2], Wagner-Nagy Beáta[3]

[1]MorphoLogic
1116 Budapest, Kardhegy utca 5.
{endredy, novak, proszeky}@morphologic.hu
[2]MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr utca 33.
{bakro, fejes, oszko, szeverenyi, varnai}@nytud.hu
[3]Universität Hamburg
Institut für Finnougristik/Uralistik
Johnsallee 35, 20148 Hamburg
beata.wagner-nagy@uni-hamburg.de

Kivonat: Cikkünkben egy olyan webhelyet mutatunk be, amelyen több korábbi projekt keretein belül számos kisebb uráli nyelvre készített morfológiai elemzőket, szóalak-generátorokat és korpuszokat tettünk elérhetővé. Az elemzések a webes felületen egy morfológiai és szemantikai egyértelműsítő eszköz formájában jelennek meg és minden nyelvhez virtuális billentyűzet segíti a szövegbevitelt.

1 Bevezetés

Az MTA Nyelvtudományi intézetének Finnugor és Nyelvtörténeti Osztálya és a MorphoLogic közötti együttműködés számos veszélyeztetett kisebb uráli nyelv morfológiájának számítógépes feldolgozására 2001-ben kezdődött a NKFP-5/135/01 számú *Komplex uráli nyelvészeti adatbázis* című projektum keretében. Ezt három OTKA projekt követte, amelyeknek keretében a finnugor nyelvcsalád permi ágához tartozó komi és udmurt nyelvre[1], az északi szamojéd nganaszan nyelvre[2] és a két obi-ugor nyelv, a manysi és a hanti három nyelvjárására[3] készített számítógépes morfológiákat sikerült olyan szintre fejleszteni, hogy azokat a tudományos közösség számára publikálhatónak éreztük.

A morfológiákat, valamint a kipróbálásukhoz használható szövegeket webes felületen keresztül tettük elérhetővé, amelyet a MorphoLogic üzemeltet, és a `http://www.morphologic.hu/urali/index.php` címen érhető el.

[1] OTKA T 048309 *Permi nyelvészeti adatbázisok*
[2] OTKA K 60807 *A nganaszan nyelv számítógépes morfológiai elemzése*
[3] OTKA NF 71707 *Obi-ugor morfológiai elemzők és korpuszok*

2 A morfológiák

A finnugor nyelvek (komi, udmurt, manysi, hanti) elemzésére a MorphoLogic *Humor* elemzőjét használjuk. A nganaszan morfológiát a Xerox *xfst* eszközének felhasználásával készítettük el.

A komi és az udmurt beszélőinek száma a többi itt bemutatott nagyon erősen veszélyeztetett nyelvvel ellentétben viszonylag jelentős, a permi nyelvek számottevő irodalommal és könyvkiadással rendelkeznek, ezért ezekre a nyelvekre olyan elemzőket készítettünk, amelyek a sztenderd cirill helyesírással írott szövegek elemzésére képesek. A kisebb, elsősorban csak beszélt nyelvként élő nyelvekre olyan elemzőket készítettünk, amelyek latin betűs fonologikus átírást használnak. Az utóbbiak esetében komoly problémát jelentett a lejegyzések következetlensége, illetve az, hogy a különböző szövegkiadásokban jelentősen különböző átírásokat használtak. A manysi esetében ugyanazon nyelvjárás (az északi) három korpuszának (ChVog[4], WT[5], VNGY[6]) feldolgozásához három különböző elemzőt kellett készítenünk.

A weboldal létrehozását elsősorban az a cél motiválta, hogy ezekkel a nyelvekkel kapcsolatos nyelvi adatokat hozzáférhetővé tegyük a tudományos kutatóközösség minél szélesebb köre számára. Ezért lehetőség szerint glosszákkal együtt jelenítjük meg az elemzéseket, hogy azok a nyelvet nem beszélő kutatók számára is értelmezhetők legyenek. A közeljövőben befejeződő obi-ugor OTKA projektben már kifejezetten cél volt az elemzők tőtárának angol glosszákkal való ellátása is. Az egyértelműsítő felület ily módon egyben szemantikai egyértelműsítésre is használható.

A weboldalon jelenleg az alábbi morfológiák érhetők el:

nyelv	glosszázás	tőtár	tőtár	toldaléktár mérete[7]
		lemma	jelentés	
nganaszan	magyar	4200	4775	310
komi (zürjén)	nincs (orosz glosszák hozzáadását tervezzük)	36000		193
udmurt	magyar	13500	18500	286
északi manysi (WT)	magyar, német, angol	3820	4200	376

Az év végéig az alábbi morfológiák kerülnek még fel az oldalra:

északi manysi (ChVog)	magyar, német, angol	1250	1530	271
északi manysi (VNGY)	magyar, német, angol	15600	16500	297

[4] Kálmán Béla: Chrestomathia Vogulica. Tankönyvkiadó, Budapest. (1989)

[5] Kálmán Béla: Wogulische Texte mit einem Glossar. Akadémiai Kiadó, Budapest. (1976)

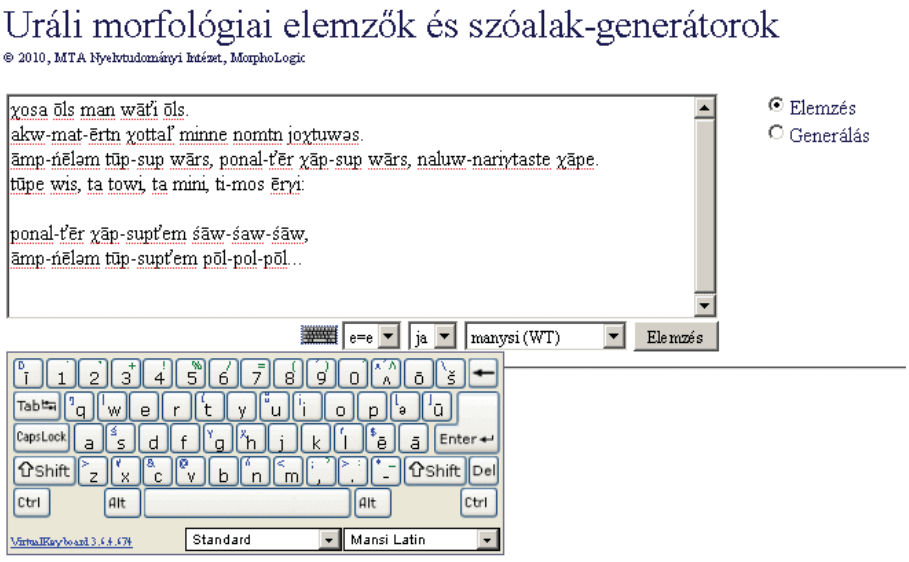
[6] Munkácsi Bernát: Vogul népköltési gyűjtemény. 1–4. Budapest. (1892–1921)

[7] morféma, ill. lexikalizált morfémakombináció

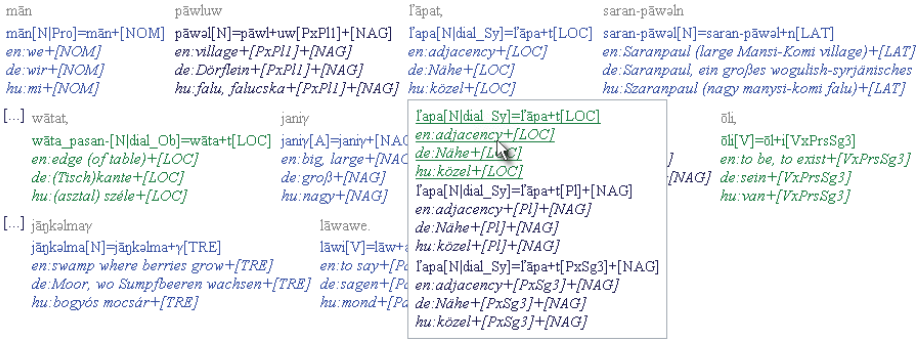
szinjai hanti	magyar és angol	2300	2500	138
kazimi hanti	magyar és angol	1750	1950	151

3 A webes felület

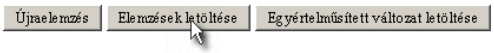
Az elemzők esetében a kiválasztott szöveget a megfelelő ablakba másolva a felhasználó megkapja a szövegben szereplő szavak lehetséges morfológiai elemzéseit és az elemzésekben szereplő tőmorfémák jelentését. A webes felületen valamennyi nyelven hozzáférhetőek olyan példaszövegek, amelyekkel az elemző kipróbálható. Virtuális billentyűzet segítségével a felhasználó maga is gépelhet be szövegeket.



Az elemzéseket megjelenítő webes felület egyben kézi egyértelműsítő eszközként is szolgál: a többértelmű szavak elemzései pop-up ablakban jelennek meg, ha az egeret egy többértelmű szó fölé mozgatjuk, ezek közül egérrel választhatunk.

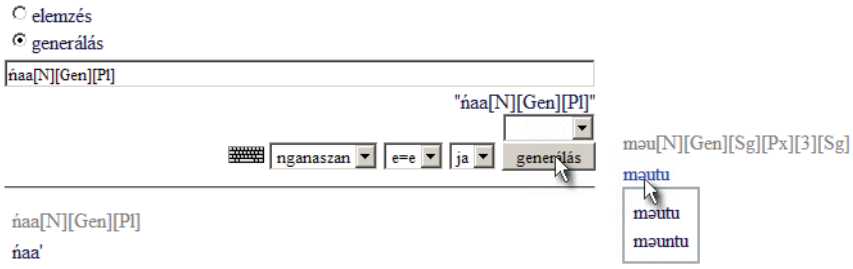


Az elkészült elemzések, illetve azok egyértelműsített változata elmenthető, az elmentett változatot a böngészőbe betöltve, az esetlegesen félbehagyott egyértelműsítő munka később folytatható.



Egyértelműsítés közben javíthatóak az elemzendő szövegben előforduló elgépelések is, ezután a szöveg újraelemeztethető. Ilyenkor nem vesznek el a korábban hozott egyértelműsítési döntések.

A webes felületen keresztül nemcsak morfológiai elemzők, hanem szóalakgenerátorok is elérhetők az egyes nyelvekhez. Ha egy adott morfémasorozat több formában is megjelenhet, akkor a generátor kimenete az elemző többértelmű kimenetének megjelenítéséhez hasonlóan jelenik meg a webes felületen, a lehetséges szóalakváltozatok itt is az egérmutatót a generált szóalak fölé mozgatva megjelenő pop-up ablakban láthatóak.



Terveink között szerepel, hogy a weblapot egyértelműsített korpuszokkal és korpuszkereső szolgáltatással egészítjük ki.

Bibliográfia

1. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications. Stanford University, Stanford (2003)
2. Novák A.: Milyen a jó Humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, Szeged (2003) 138–145
3. Novák, A.: Language resources for Uralic minority languages. In: Proceedings of the SALTMIL Work-shop at LREC-2008: Collaboration: interoperability between people in the creation of language resources for less-resourced languages. Marrakech (2008) 27–32
4. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H., Yli-Jyrä, A. (szerk.): *Inquiries into Words, Constraints and Contexts Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*. Gummerus Printing, Saarijärvi/CSLI Publications, Stanford (2005) 116–125