# A Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály jelentése a 2016. évről

# Kiemelkedő kutatási és más jellegű eredmények

**Nyelvtechnológiai Kutatócsoport**

Lezárult a *Nyílt, integrált magyar nyelvtechnológiai kutatási infrastruktúra fejlesztése* című projekt. Az MTA 2015. évi Infrastruktúra-fejlesztési Pályázat 2. kategóriájában elnyert támogatás segítségével elkészült az *e-magyar* digitális kutatási infrastruktúra, amely a magyar nyelvtechnológia eddig elkészített szövegelemző eszközeit egy egységes technológiai láncba integrálja. Az AITIA Zrt., SZTE, PPKE, BME, SZTAKI közreműködésével készült nyelvelemző rendszer nyílt forráskódú, kutatásfejlesztési célokra ingyenesen használható, és egyaránt szolgálja a a szakmai fejlesztők, a digitális bölcsészet valamint az érdeklődő nagyközönség igényeit. Az e-magyar rendszer tartalmaz még egy szabadon letölthető beszédarchívumot továbbá a beszédtechnológiai kutatásokat támogató nyílt beszédtechnológiai szoftvereszközöket. Az új infrastruktúra fő célja a munkában részt vevő műhelyekben eddig előállított különböző eszközök továbbfejlesztése, egységesítése és egyetlen koherens technológiai láncba szervezése volt.

Az e-magyar rendszer most elkészült szövegfeldolgozó eszközlánca a szövegben rejlő információkat automatikus módon fedi fel, teszi explicitté. Egymásra épülő automatikus szövegfeldolgozó eszközökből áll: az eszközlánc tetszőleges magyar nyelvű szövegrészt feldolgozva elvégzi a szöveg elemeinek a szegmentálását (az emToken eszköz révén), megállapítja az egyes szavak tövét és teljes morfológiai elemzését (emMorph, emLem és emTag), majd ezek után megadja a mondatok összetevős (emCons), valamint függőségi elemzését (emDep); de ha csak egy gyors elemzésre van szükségünk, felismeri a mondatban szereplő frázisokat (emChunk), továbbá a szövegben előforduló tulajdonneveket (emNer).

A projekt keretében újonnan készült el az emToken szövegszegmentáló eszköz, és az emMorph morfológiai elemző eszköz, az eszközlánc további tagjai korábban meglévő eszközök legújabb verziói. Az emToken egy magyar nyelvű szövegek mondatrabontását és tokenizálását végző szoftver. A program támogatja a jelenleg standardnak tekinthető UTF-8 karakterkódolást, moduláris felépítésénél fogva alkalmas új, egyedi igényekhez szabott változatok létrehozására, valamint kiterjedt és szintén testreszabható tesztelési képességekkel rendelkezik. Az emMorph morfológiai elemző egyesíti az eddigi magyar morfológiai elemzők jó tulajdonságait, egy nemzetközi szabványoknak megfelelő, részleteiben kidolgozott teljes új magyar morfológiai kódkészletet vezet be.

Az e-magyar rendszer elemeinek integrálása a nemzetközileg széles körben használt GATE nyelvfeldolgozó keretrendszerben valósult meg. A különféle felhasználói csoportok igényeinek megfelelően a rendszer elérhető online, a GATE saját grafikus felületén, valamint forráskód szinten a github-on keresztül. A rendszer online a http://e-magyar.hu címen található meg. A honlap bemutatja az egyes nyelvelemző eszközöket, valamint lehetővé teszi, hogy a szövegfeldolgozó eszközlánc funkciói egy online felületen keresztül is használhatóak legyenek: itt egy tetszőleges magyar szövegrészleten lefuttathatjuk a rendszert, és visszakapjuk a szöveget a hozzárendelt elemzésekkel együtt.

Remélhető, hogy a jövőben további magyar szövegfeldolgozó eszközök épülnek majd be ebbe a rugalmasan bővíthető keretrendszerbe, így egyre gazdagabb elemzési lehetőségek válnak elérhetővé, és az e-magyar a magyar szövegfeldolgozó eszközök közös platformjává válik. Az e-magyar rendszerről készült publikációk 2017-ben jelennek meg.

Igen jelentős folyóiratpublikáció jelent meg a *számítógépes lexikográfia* területén a fordítási reláció feltételes valószínűség alapú megközelítéséről. A cikk azt állítja, hogy a fordítási relációt célszerű *P(b|a)* feltételes valószínűségként felfogni, amely párhuzamos korpuszok alapján egy becslést nyújt arra vonatkozóan, hogy *a* forrásnyelvi kifejezést hányszor fordították *b* célnyelvi kifejezésre. Ez egy lexikográfiailag motivált definíciót jelent, és fordítási párokat kinyerő automatikus algoritmusoknak is megfelelő alapját képezheti. Az ötlet számos előnnyel jár. Az így definiált fordítási reláció egyrészt kvantifikálható is lesz, tükrözve azt az intuíciót, hogy a fordítási reláció fokozatos; másrészt képes megragadni azt is, hogy a fordítási reláció általában aszimmetrikus; továbbá, általa számot tudunk ad arról a speciális esetről is, amikor a fordítási reláció -- a tökéletes fordítási ekvivalencia esetében -- szimmetrikus.

Tovább folytak a kutatások a *Multimodális kommunikáció időszerkezete* című OTKA projekben. Külföldi konferenciapublikációk születtek, melyek bemutatták a kutatásban használt adatokat, eszközöket és módszereket.

A *Magyar Generatív Szintaxis 2* című OTKA-projektben az év folyamán megtörtént a Jordánszky-kódex normalizálása és majdnem teljesen elkészült a morfológiai egyértelműsítése is. Elkészült a Heltai-féle Újszövetség-fordítás OCR-ezése, a nyers OCR-kimenet összeolvasása pedig folyamatban van. A Sylvester- és a Károli-féle Újszövetség-fordítások betűhű szövege teljes egészében készen van. Káldi Újszövetség-fordításának OCR-ezése folyamatban van. Az elkészült szövegek elérhetőek a Régi Magyar Konkordancia keresőfelületén keresztül. Ezen felül készült egy újfajta korpuszlekérdező eszköz, amellyel különböző korokból származó, illetve különböző nyelvű bibliafordításokat lehet párhuzamosan megjeleníteni. Megtörtént a morfológiailag elemzett ómagyar szövegek konverziója a Universal Dependencies and Morphology formátumára, amely egy nemzetközi sztenderd, melyet már 50 nyelvre alkalmaztak. Azzal, hogy ezeket az ó- és középmagyar korból származó bibliafordításokat kereshető szöveges formában, illetve mai magyar átírásban és nyelvészeti elemzéssel ellátva elérhető tesszük, jelentősen hozzájárulunk a nemzeti kulturális örökség megőrzéséhez.

A *Finn-OTKA* projekt keretein belül többféle szótárépítési módszer lett kipróbálva, melyek mindegyike alkalmas volt több száz fordítási jelöltet tartalmazó protoszótár létrehozására a következő nyelvpárokra: {komi-zürjén, komi-permják, hegyi mari, mezei mari, udmurt, északi számi} - {angol, finn, magyar, orosz}. A különböző módszerekkel előállított protoszótárak uniója szolgált bemenetül a kézi validálást végző anyanyelvi beszélőknek. Jelenleg még zajlik a kiértékelési folyamat.

2016-ban indult el *Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben* című projekt, melynek egyik célja egy annotált korpusz létrehozása udmurt, tundrai nyenyec, szinjai és szurguti hanti nyelvű, írott és beszélt nyelvi szövegekből, amely lehetővé teszi az uráli–orosz kontaktushatás kutatását. Annak érdekében, hogy nyomon követhessük a kisebbségi nyelvek orosz hatásra végbemenő szintaktikai változásait, különböző korokból gyűjtött szövegek feldolgozására kerül sor. Az adatbázis minden szöveget legalább az eredeti átírásában, amelyet a lejegyző használ, valamint IPA-átírásban tartalmaz. Továbbá – mivel az érintett nyelvek írásrendszere a cirill ábécén alapszik – megőrizzük az eredeti cirill írást, amennyiben van ilyen. Az adatbázisban elérhetőek az eredeti szöveganyag mondat szinten párhuzamosított angol, magyar, német és orosz fordításai is. A korpusz egy része morfológiai szintű annotációt is tartalmaz.

2016 őszén elindult az *Új, innovatív turisztikai szolgáltatás alapjainak megteremtése NLP módszer segítségével* című GINOP projekt.

A felhasználói visszajelzések alapján folyamatosan történik a *helyesiras.mta.hu* helyesírási tanácsadó portál fejlesztése, a hibák javítása. 2015 szeptemberében megjelent az akadémiai helyesírási szabályzat 12. kiadása, a változások alapján módosításra kerültek az érintett eszközök javaslatai. Mivel a 11. kiadás 2016 szeptemberéig még érvényben volt, a portál mind a 11., mind a 12. kiadás szerinti írásmódot megadja a kiadás feltüntetésével.

Kísérleti szótárak készültek a *NooJ* program számára kezelhető formátumban. Megtörtént a NooJ segítségével előállítható különféle ragozási típusú szavak valamennyi alakjának generálása.

2016 őszén jelentős méretű közösségi média anyaggal kiegészülve a Magyar Nemzeti Szövegtár mérete elérte az egymilliárd szót. Regisztráció után bárki számára szabadon hozzáférhető ez a magyar nyelvet reprezentáló nagy méretű nyelvi adatbázis. Új, modern korpuszkezelő rendszerre került és új, gazdag funkcionalitású lekérdezőfelületet kapott a Magyar Nyelv Nagyszótára munkálatainak alapját képező Magyar Történeti Szövegtár. Az elmúlt 240 év magyar nyelvi anyagából merítő gazdag annotációjú adatbázis szabadon elérhető a http://clara.nytud.hu/mtsz címen. 2016 nyarán az intézetben sikeres oktató előadásra került sor a korpuszokban való keresésről, a korpuszok használatáról.

**Matematikai Nyelvészeti Kutatócsoport**

A 4lang fogalmi gráfokkal folytatott újabb kísérletek eredményeképp létrejött a *wordsim*: egy szavak szemantikai hasonlóságát mérő, gépi tanulás alapú rendszer (https://github.com/recski/wordsim), mely az általunk épített definíciós gráfok segítségével minden korábbi rendszernél magasabb pontosságot ér el a hasonló rendszerek összehasonlítására leggyakrabban használt SimLex adathalmazon (ld. https://www.cl.cam.ac.uk/~fh295/simlex.html).

Az e-magyar rendszer keretén belül létrejött a beszédarchívum (http://e-magyar.hu/hu/speechmodules/speechoverview). Létrehozásával három fő célunk volt. Az első és legfontosabb a magyar beszédtechnológiára annak kezdetei óta jellemző zárt kutatási és publikációs modell felváltása egy szabad, nyílt forrású (Free and Open Source Software, FOSS) modellel. Második célunk a hagyományos, gondosan felcímkézett és mind artikulációsan mind akusztikailag tiszta adatokon alapuló felügyelt tanulási módszerek felváltása gyengén felügyelt illetve felügyeletlen (weakly supervised, unsupervised) módszerekkel. Harmadik, az első kettőtől nem mindig könnyen elválasztható célunk pedig egy a digitális bölcsészeti munkát, elsősorban a szociológiát, történelemtudományt, folklorisztikát, és néprajzot beszédtechnológiai oldalról támogató platform alapjainak megteremtése.

A mesterséges neuronhálók által tanult szóreprezentációknak, az úgynevezett szóembeddingeknek a nyelvészeti tartalmára, szemantikus hálókkal való kapcsolatára vonatkozó kutatás középpontjában idén a szavak többértelműsége állt. Nyílt forráskódú eszközök annotálatlan szövegből (felügyeletlen tanulással), a szavak előfordulásait csoportosítva képesek megállapítani hogy egy szónak hány jelentése van, és a a jelentéseket reprezentálni. A projekt az így tanult többjelentéses szóembeddingek szemcsésségét vizsgálta. A kutatás fő tanulsága, hogy míg a hagyományos szótárakban a legtöbb szó egyjelentésű, a többértelműek nagy részének csak két jelentése van, és így tovább, a felügyeletlen eszközök ezt a gyors lecsengést nem minden esetben tükrözik. A jelentéskészlet felbontásának vizsgálatára egy olyan módszer is kidolgozásra került, mely azon alapszik, hogy a különböző jelentések fordítása egy másik nyelvre gyakran különböző.

**Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport**

A csoport munkatársai előadásokat tartottak és tanulmányokat készítettek az alábbi témakörökben: (1) helyesírás, nyelvi tanácsadás, stilisztika; (2) névtan, névkultúra, névjog. Kiemelendő a "Köznyelvi és szaknyelvi helyesírási kérdések a gyakorlatban" címmel a gödöllői szakmai napon tartott előadás.

A kutatócsoport speciális feladata az utónév-szakvéleményezés: a névviselésről szóló 2010. évi I. törvény az anyakönyvi eljárásról 44.§-a alapján. A csoport havi 30-50 utónévkérelem szakvéleményének elkészítése kapcsán működik együtt a Bevándorlási és Állampolgársági Hivatal Anyakönyvi Felügyeleti Osztályával, valamint a honosítást végző munkatársakkal, osztályvezetővel. A csoport 2016-ban is az Intézetben erre a célra létrehozott Utónévbizottsággal és osztályon belül a Nyelvtechnológiai Kutatócsoporttal együttműködve látta el a hatósági

anyakönyvezésre beterjesztett új utónevek (női és férfi keresztnevek) nyelvi szakvéleményezését és a bejegyzésre alkalmasnak tartott nevek listájának gondozását. A bejegyzésre alkalmasnak minősített nevek listája az intézet honlapján hozzáférhető (http://www.nytud.hu/oszt/nyelvmuvelo/utonevek), havonta frissül. 2016-ban a lista 96 új névvel bővült (59 női, 37 férfi) így 2017. január 1-jén 3860 bejegyzésre alkalmasnak minősített utónév (2203 női és 1657 férfi) alkotja a jegyzéket. A nevekkel kapcsolatos tanácsadó szolgálat 2016-ban kb. 2500 emailben (nevtanacs@nytud.mta.hu) érkezett, és kb. 500 telefonos névadással kapcsolatos, névhasználati kérdésre adott választ, 10 névhasználati szakvéleményt készített, 15 családnevekkel kapcsolatos szakvéleményt készített, valamint BÁH egyszerűsített honosítási eljárás során kért 100 keresztnevet véleményezett szakértőként.

# Párbeszéd a tudomány és a társadalom között

**Nyelvtechnológiai Kutatócsoport**

**Médiamegjelenések.** Az osztály munkatársa két alkalommal szerepelt nyelvi ismeretterjesztéssel foglalkozó televíziós műsorban: a Duna Televízió Család-barát, illetve az M5 Felsős c. műsorában ismertette a főbb helyesírási változásokat, bemutatta a helyesiras.mta.hu online tanácsadó portált. Az osztály egy munkatársa heti rendszerességgel állandó vendége a Lánchíd Rádió nyelvi, nyelvészeti ismeretterjesztéssel foglalkozó műsorának.

**Rendszeres tud. ism. tevékenység (weben, sajtóban).** A helyesiras.mta.hu portál munkatársai több csatornán végeznek tudományos ismeretterjesztő tevékenységet. A hivatalos Facebook-oldalon (és Twitteren) napi egy-két poszt jelenik meg rajta különböző helyesírási, nyelvhelyességi, illetve egyéb érdekes, a magyar nyelvvel kapcsolatos témában. Az osztály egyik munkatársa rendszeresen publikál tudományos ismeretterjesztő írásokat helyesírás témában egy orvosoknak szóló szaklapban.

**Eseti tud. ism. tevékenység (ea-k, sajtótájékoztatók stb).** Az osztály több munkatársa ismeretterjesztő előadást tartott 2016. február 9-én a Csodák Palotája Tudományos Csopa Cafe rendezvényén, amely minden hónap második keddjén különféle területek kutatóit látja vendégül egy beszélgetés és egy ismeretterjesztő előadás erejéig. A „Sokszínű nyelvi ismeretterjesztés” című beszélgetésben szó volt a helyesírás.mta.hu keretein belül működő Helyes blogról, a helyesírás szabályainak 2015-ös változásáról és általában a nyelvészeti ismeretterjesztésről.

Az osztály munkatársával interjút készített a Szövegkovács blog a helyesírási szabályzatról, változásokról. Az interjú a blog egyik legnagyobb nézettségű bejegyzése lett. Ugyenezen interjú megjelent a Reggeli Újság című erdélyi napilapban is.

**Egyéb.** A *helyesiras.mta.hu helyesírási tanácsadó portál* új lehetőséget teremt a széles közvélemény számára a magyar helyesírás területein való tájékozódáshoz. Továbbra is nagy népszerűségnek örvend a portál: 2016-ban 850000 látogató (60%-uk rendszeres látogató) több mint 1,8 millió tanácsot kért itt. A lekérdezések száma munkanapokon rendszeresen eléri a 9000-et. Az MTA Nyelvtudományi Intézete egy olyan tudományos alapokon álló szolgáltatást üzemeltet immár három és fél éve, mely valóban sokakhoz eljut, sikeresen működik.

A csoport által fejlesztett nyelvi adatbázisok társadalmi szempontból is jelentősek. Ezen adatbázisok az anyanyelvi kulturális örökség digitális formában őrzött részei, melyek referenciapontként szolgálnak nemcsak a tudományos kutatásban, hanem a közgondolkodásban, az érdeklődő laikusok körében is. A több mint 11000 regisztrált felhasználóval bíró Magyar Nemzeti Szövegtár, a Ómagyar Korpusz, a BUSZI és a Magyar Történeti Szövegtár új felülete is elérhető a Nemzeti Korpuszportálon egyben érhetők el a szakma és a nagyközönség számára. A Magyar Nemzeti Szövegtárban 2016-ban 85000 lekérdezést futtattak.

A frissen megnyílt e-magyar.hu weboldal által nyújtott szövegelemző szolgáltatás nem csak a

szakma, ill. tudományterület szereplőit célozza meg, hanem olyan laikus felhasználókat is, akik érdeklődnek a magyar nyelv gépi feldolgozásának lehetőségei iránt, esetleg a saját kutatásukhoz szükségük van valamilyen gépi szövegelemző megoldásra. Mindemellett az oktatás területén is hasznos segítséget nyújthat.

A csoport esetenként számítógépes nyelvészeti támogatást ad különböző nyelvészeti kutatásoknak, nyelvi adatokat szolgáltat, egyéni kéréseket teljesít, a Magyar Nemzeti Szövegtár, más korpuszok, és az e-magyar rendszer vonatkozásában.

#### Matematikai Nyelvészeti Kutatócsoport

Megalakult az akadémia dolgozók Stádium 28 köre, melynek az osztály néhány munkatársa is tagja.

#### Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport

**Médiamegjelenések.** A névadással, névkultúrával kapcsolatos tájékoztatásra folyamatosan nagy az igény a média részéről. 2016-ban ez számos rádiós és tv-s, illetve online és papír alapú sajtóban megjelent interjú formájában valósult meg. Helyesírási kérdésekről két rádió-, egy sajtó- és egy tv-riportban adott tájékoztatást a csoport munkatársa 2016-ban.

**Rendszeres tud. ism. tevékenység (weben, sajtóban).** A csoport munkatársa részt vesz az Édes Anyanyelvünk ismeretterjesztő folyóirat szerkesztésében, együttműködik az Anyanyelvápolók Szövetségével: a szövetség választmányi tagjaként az MTA Nyelvtudományi Intézete és az Anyanyelvápolók Szövetsége szakmai kapcsolatának elősegítése.

**Eseti tud. ism. tevékenység (ea-k, sajtótájékoztatók stb).** A csoport munkatársa 4 alkalommal tartott tudományos ismeretterjesztő előadást nagyvállalalatoknak aktuális helyesírási és egyéb nyelvi kérdésekről.

**Egyéb.** A folyamatosan működő nyelvi tanácsadó szolgálat 2016-ban kb. 2500 emailben (tanacs@nytud.mta.hu) érkezett, és kb. 1200 telefonos helyesírási, nyelvhasználati kérdésre adott választ, 11 nyelvészeti szakvéleményt készített. A tanácsadás fő témaköre az AkH.12. bevezetése, megismerése, a változások tudatosítása, az új szabályrészek értelmezése volt. A 2015-ben megjelent „új helyesírási szabályzat”, az AkH.12. megismerése és érvényesítése – különösen a nyilvánosság elé kerülő szövegek esetében – jelentős társadalmi és szakmai igény volt 2016-ban. Mintegy 120–130 írásbeli kérdés és telefonhívás irányult a jelenlegi szabályzatot érintő változásokra. A partnerek között megtalálhatók állami és európai uniós hivatalok, minisztériumok, cégek, más szervezetek, fordítók, tanárok, újságírók, szerkesztők, magánszemélyek. A szolgálat emléktáblák (32 db) szakvéleményezésével is foglalkozott önkormányzatok és magánszemélyek kérésére. A csoport részt vesz az MTA Nyelvhelyességi Tanácsadó Testületének munkájában, az intézeti Utónévbizottságban és a helyesiras.mta.hu Helyesírási tanácsadó portál működtetésében, folyamatosan egyedi válaszokat ad az érkező kérdésekre. Az utóneveket bemutató Utónévportál 2016-ban több mint 35000 látogató 100000 lekérdezését szolgálta ki.

# A kutatóhely hazai és nemzetközi kapcsolatai 2016-ban

## Hazai kapcsolatok

#### Nyelvtechnológiai Kutatócsoport

A kutatócsoport több tagja részt vett a Kuna Ágnes által vezetett Magyar Orvosi Nyelvi Kutatócsoport munkájában (MTA Társadalomtudományi Kutatóközpont -- Pázmány Péter Katolikus Egyetem).

A kutatócsoport egy tagja részt vesz a Tolcsvai Nagy Gábor vezette Stíluskutató csoport

munkájában (Eötvös Loránd Tudományegyetem).

A kutatócsoport egy tagja szakmai kapcsolatot ápol a Szegedi Tudományegyetem Bölcsészettudományi Karával.

A kutatócsoport több tagja szakmai kapcsolatot ápol a Károli Gáspár Református Egyetem Magyar Nyelvtudományi Tanszékével. A tanszék részéről együttműködő felek: Fóris Ágota terminológiai témákban illetve Dér Csilla Ilona és Csontos Nóra pragmatikai témákban.

Szakmai kapcsolat épült ki a Pragmatika Centrum Országos Kutatóközponttal (Szegedi Tudományegyetem).

A kutatócsoport tagjai különböző előadásokat tartottak az Eötvös Loránd Tudományegyetemen, a Károli Gáspár Református Egyetemen, a Pázmány Péter Katolikus Egyetemen és a Semmelweis Egyetemen.

A FinnOTKA projekt keretén belül a kutatócsoport együttműködött a Szegedi Tudományegyetem Angol-Amerikai Intézetének és Mesterséges Intelligencia kutatócsoportjának egyes tagjaival. A projektpartner vezető kutatója Fenyvesi Anna. Az együttműködés célja a kisebbségi finnugor nyelvek revitalizációjának nyelvtechnológiai támogatása.

Az e-magyar projekt a magyar nyelvtechnológiai közösség összefogásaként jött létre. A koordinátor a Nyelvtudományi Intézet volt, a munkában részt vett még a Szegedi Tudományegyetem, az MTA SZTAKI, a Pázmány Péter Katolikus Egyetem, az AITIA International Zrt., valamint a MorphoLogic Kft.

A GINOP projekt keretein belül a kutatócsoport a következő cégekkel áll kapcsolatban: CARTOUR Idegenforgalmi Szolgáltató Kft, TRAVELWEB Informatikai, Kereskedelmi és Szolgáltató Kft

Az MGTSz2 projekt keretein belül együttműködés jött létre Vincze Veronikával a Szegedi Tudományegyetem Informatikai Intézetéből. Az együttműködés célja a morfológiailag elemzett ómagyar szövegek konverziója a Universal Dependencies and Morphology formátumára.

Ludányi Zsófia: MTA Magyar Nyelvi Osztályközi Állandó Bizottság -- tag, MTA Orvosi Nyelvi Munkabizottság -- tag, MTA Nyelvhelyességi Tanácsadó Tesület -- titkár, Magyar Orvosi Nyelv -- a szerkesztőbizottság tagja

Pajzs Júlia: MTA Szótári Munkabizottság -- tag

Simon Eszter: XIII. Magyar Számítógépes Nyelvészeti Konferencia -- a programbizottság tagja

Váradi Tamás: MTA Alkalmazott Nyelvészeti Munkabizottság -- tag

**Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport**

Heltainé Nagy Erzsébet: MTA Magyar Nyelvi Osztályközi Állandó Bizottság -- tag, MTA Magyar Nyelvészeti Munkabizottsága -- tag, Magyar Nyelvtudományi Társaság -- tag, MANYE -- tag, Termini Egyesület -- alapító tag, Anyanyelvápolók Szövetsége -- választmányi tag, Magyar Nyelvőr -- a szerkesztőbizottság tagja, Édes Anyanyelvünk -- a szerkesztőbizottság tagja

Raátz Judit: MTA Magyar Nyelvi Osztályközi Állandó Bizottság -- tag, MTA köztestület -- tag, Anyanyelvápolók Szövetsége -- választmányi tag, International Council of Onomastic Sciences -- tag, Magyar Nyelvtudományi Társaság -- tag, Magyar Nyelvtudományi Társaság Névtani Tagozat -- tag, Magyar Nyelvtudományi Társaság Magyartanári Tagozat -- tag, Szemere Gyula anyanyelv-pedagógiai kutatócsoport -- tag, HUNRA Magyar Olvasástársaság -- tag

## Nemzetközi kapcsolatok

### Nyelvtechnológiai Kutatócsoport

Az Intézet a CLARIN európai kutatási infrastruktúra-hálózat magyar koordinátoraként tagja lett a CLARIN ERIC szervezetnek.

A FinnOTKA projekt keretén belül együttműködés jött létre a University of Helsinki Institute of Behavorial Sciences tanszékének munkatársaival. A projektpartner vezető kutatója Kristiina Jokinen. Az együttműködés célja a kisebbségi finnugor nyelvek revitalizációjának nyelvtechnológiai támogatása.

Az uráli projekt keretein belül együttműködés jött létre az Ob-Ugric Database: analysed text corpora and dictionaries for less described Ob-Ugric dialects című projekt résztvevőivel a müncheni Ludwig Maximilian Egyetemről. Schön Zsófiával közös munka folyik a szurguti hanti szövegek IPA-átírásán és morfológiai elemzésén.

Számos nemzetközi kutatóval közösen beadásra került egy COST pályázat, amelynek címe: Historical Cryptology -- Unlocking Europe's Encrypted Heritage. A projekt egyik célja, hogy feltérképezze, hogy milyen nyelvtechnológiai módszerek alkalmazhatók a történeti titkosírások megfejtésében, illetve hogy ez utóbbiaknak egy strukturált adatbázisát felépítse. A közvetlen együttműködő partnerek: Megyesi Beáta (University of Uppsala), Láng Benedek (BME).

Meghívott előadás Challenges of processing cultural heritage data -- Introducing the Old Hungarian Corpus címmel az Uppsalai Egyetemen, a Seminar in Computational Linguistics sorozatban, 2016. ápr. 29-én Uppsalában. Meghívó: Megyesi Bea.

A kutatócsoport szoros kapcsolatot ápol az ACL Special Interest Group on Uralic Languages alapító tagjaival, név szerint Tommi A. Pirinennel (Universität Hamburg) és Francis Tyersszel (UiT Norgga árktalaš universitehta). Közös szervezésben lett megtartva a Second International Workshop on Computational Linguistics for Uralic Languages című nemzetközi konferencia. Az együttműködés további eredménye egy társszerkesztett Acta Linguistica Hungarica különszám a fenti témában, továbbá rendszeres a konzultáció az uráli nyelvek nyelvtechnológiai támogatását érintő témákban.

Gregory Grefenstette 2016. október 11-e és október 14-e között az Intézet meghívására előadást és szakmai megbeszéléseket tartott. Gregory Grefenstette a Paris-Sud Egyetem professzora, illetve az Inria (Inventors for the digital world) TAO kutatócsoport tagja. Előadásának címe Organizing Personal Data with Personal Semantics and Natural Language Processing volt, a megbeszélések pedig a terminológiakivonatolás kapcsolatos kutatásairól szóltak, amely kutatási terület a Nyelvtechnológiai Osztály jövőbeni tervei között szerepel.

Váradi Tamás: Language Resources and Evaluation Conference (LREC) 2016 -- a programbizottság tagja, Computational Linguistics (COLING) -- a programbizottság tagja, Text Speech and Dialogue (TSD) 2016 -- a programbizottság tagja.

Simon Eszter: Third International Workshop on Computational Linguistics for Uralic Languages -- a programbizottság tagja, Second International Workshop on Computational Linguistics for Uralic Languages -- a programbizottság tagja, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities 2016 -- a programbizottság tagja, Association for Computational Linguistics -- tag, ACL Special Interest Group on Uralic Languages -- tag

### Matematikai Nyelvészeti Kutatócsoport

Jelentések készültek a Linguistic Data Consortium (LDC) számára különböző nyelvekről világnyelvektől erőforrásszegény nyugat-afrikai nyelvekig, melyekben a demográfiai és nyelvészeti jellemzés mellett nagy hangsúlyt kapott a számítógépes támogatottság is. A LDC-vel való jó kapcsolat nagyban hozzájárul az intézetben zajló világszínvonalú adatvezérelt kutatáshoz. A téma

aktualitása kettős: egyrészt a közösségi média korában olyan nyelveken is lehetővé válik a gépi szövegfeldolgozás, amelyekhez nincs dedikált intézet korpuszszerkesztésre, ugyanakkor a ritkán tanított nyelvek felkarolására az is sürget, hogy a ma beszélt mintegy hétezer nyelvnek az optimistább becslések szerint is közel a harmada veszélyeztetett, és ha a digitális használatot nézzük, akkor csak 5% túlélését jósolhatjuk.

**Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport**

Raátz Judit: International Council of Onomastic Sciences -- tag

## Felsőoktatási tevékenység

**Nyelvtechnológiai Kutatócsoport**

Ludányi Zsófia: a KRE BTK, Magyar Nyelvtudományi Tanszék meghívott előadója, Nyelvészeti filológia címen tartott gyakorlati kurzust.

összes opponensi feladat: 5; PhD témavezetés: 2; BA/BSc témavezetés: 1

**Matematikai Nyelvészeti Kutatócsoport**

összes opponensi feladat: 1

**Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport**

Heltainé Nagy Erzsébet: a KRE BTK, Magyar Nyelvtudományi Tanszék oktatója, 8 gyakorlati és 2 elméleti kurzust tartott stilisztika, szöveg-és stíluselemzés, jelentéstan, retorika, terminusalkotás, és helyesírás témákban.

Raátz Judit: az ELTE Mai Magyar Nyelvi Tanszékének főállású docense, 22 kurzus oktatója.

összes opponensi feladat: 4; PhD témavezetés: 5; MA/MSc témavezetés: 7; BA/BSc témavezetés: 1

## Az intézettel szerződéses kapcsolatban álló cégek

**Nyelvtechnológiai Kutatócsoport**

e-magyar: AITIA International Zrt

GINOP: CARTOUR Idegenforgalmi Szolgáltató Kft, TRAVELWEB Informatikai, Kereskedelmi és Szolgáltató Kft

## Részvétel nemzetközi konferenciákon

**Nyelvtechnológiai Kutatócsoport**

| előadó | előadás címe | konferencia elnevezése | helye (város) | ideje (hónap) |
|---|---|---|---|---|
| Ludányi Zsófia | „Az antibiotikum (talán) nem összetett szó”. A magyar orvosi szaknyelv idegen előtagos összetételeiről. | A magyar nyelvészeti kutatások újabb eredményei V. | Kolozsvár | április |
| Váradi Tamás | Language technology tools | LREC 2016 | Portorož | május |

| | | | | |
|---|---|---|---|---|
| (Hunyadi László, Szekrényes István) | and resources for the analysis of multimodal communication. | | | |
| Benyeda Ivett, Koczka Péter | Creating seed lexicons for under-resourced languages. | GLOBALEX workshop, LREC 2016 | Portorož | május |
| Simon Eszter (Vincze Veronika) | Universal Morphology for Old Hungarian. (poszter) | 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities | Berlin | augusztus |
| Váradi Tamás, Benyeda Ivett, Koczka Péter | Automatic lexicon creation to support the digital vitality of endangered Uralic languages. | The Tenth International Conference on Natural Language Processing (HrTAL2016) | Dubrovnik | szeptember |
| Simon Eszter | Uralic Languages Under the Influence Database (UraLUID). | 6th International Conference on Samoyedology | Moszkva | szeptember |
| Simon Eszter (Mus Nikolett) | Languages under the Influence: Building a database of Uralic languages. | 6th International Conference on Samoyedology | Moszkva | szeptember |
| Váradi Tamás (Hunyadi László, Szekrényes István) | Language technology tools and resources for the analysis of multimodal communication in digital humanities. | COLING 2016 | Oszaka | december |

**Matematikai Nyelvészeti Kutatócsoport**

| előadó | előadás címe | konferencia elnevezése | helye (város) | ideje (hónap) |
|---|---|---|---|---|
| Kornai András | Computational linguistics of borderline vital languages in the Uralic family. | Second International Workshop on Computational Linguistics for Uralic Languages | Szeged | január |
| Kornai András, Recski Gábor (Nemeskey Dávid Márk) | Detecting optional arguments of verbs. (poszter) | LREC 2016 | Portorož | május |
| Makrai Márton | Filtering Wiktionary triangles by linear mapping between distributed word models. | LREC 2016 | Portorož | május |
| Recski Gábor, Kornai András (Iklódi Eszter, Pajkossy Katalin) | Measuring semantic similarity of words using concept networks. (poszter?) | The First Workshop on Evaluating Vector Space Representations for NLP, Annual Meeting of the Association for Computational Linguistics. | Berlin | augusztus |

| | | | | |
|---|---|---|---|---|
| Kornai András, Makrai Márton (Borbély Gábor, Nemeskey Dávid) | Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation. (poszter) | The First Workshop on Evaluating Vector Space Representations for NLP, Annual Meeting of the Association for Computational Linguistics. | Berlin | augusztus |

## Részvétel hazai konferenciákon

### Nyelvtechnológiai Kutatócsoport

| előadó | előadás címe | konferencia elnevezése | helye (város) | ideje (hónap) |
|---|---|---|---|---|
| Ludányi Zsófia | Szaknyelvi helyesírási változások az AkH. 12. tükrében, különös tekintettel az orvosi nyelvre. | Szakfordító Szakmai Nap | Gödöllő | január |
| Falyuna Nóra, Kovács Réka | A terminológia szakos hallgatók elhelyezkedési lehetőségei - tapasztalatok a munka erőpiacon. Kerekasztal-beszélgetés. | Terminológiai csütörtük. III. Szakmai nap és öregdiák-találkozó | Budapest | február |
| Ludányi Zsófia (Kuna Ágnes, Kocsis Zsuzsanna) | A Magyar orvosi nyelv 16–17. századi alkorpusza. Tervezet, átírás, annotálás. | A nyelvtörténeti kutatások újabb eredményei IX. | Szeged | április |
| Sass Bálint | A kibővített Magyar történeti szövegtár új keresőfelülete. | A nyelvtörténeti kutatások újabb eredményei IX. | Szeged | április |
| Falyuna Nóra, Kovács Réka, Ludányi Zsófia (Kuna Ágnes) | Poétika és korpusz. Hogyan nyújthat segítséget a korpusznyelvészet a poétika vizsgálatában? | Nyelv, poétika, kogníció konferencia | Eger | május |
| Simon Eszter | Metaforák és metonímiák kezelése a számítógépes nyelvészetben. | Nyelv, poétika, kogníció konferencia | Eger | május |
| Ludányi Zsófia | Online segédeszközök a helyesírás tanításának szolgálatában. | Találkozások az anyanyelvi nevelésben 3. | Pécs | június |
| Dodé Réka | Az automatikus terminuskivonatolás új perspektívái. | Félúton konferencia | Budapest | október |

### Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport

| előadó | előadás címe | konferencia elnevezése | helye (város) | ideje (hónap) |
|---|---|---|---|---|
| Heltainé Nagy Erzsébet | Köznyelvi és szaknyelvi helyesírási kérdések a gyakorlatban. | Szakfordító Szakmai Nap | Gödöllő | január |

## Rendezett konferenciák

### Nyelvtechnológiai Kutatócsoport

| szervező | konferencia elnevezése | helye (város) | ideje (napra!) | (társszervező, ha volt) |
|---|---|---|---|---|
| Simon Eszter | Second International Workshop on Computational Linguistics for Uralic Languages | Szeged | 2016. január 20. | Tommi A. Pirinen, Francis M. Tyers, Vincze Veronika, Nagy Ágoston, Horváth Csilla |
| Váradi Tamás, Ludányi Zsófia, Dodé Réka, Falyuna Nóra | X. Alkalmazott Nyelvészeti Doktoranduszkonferencia | Budapest | 2016. február 5. | |

# A 2016-ban elnyert fontosabb hazai és nemzetközi pályázatok rövid bemutatása

### Nyelvtechnológiai Kutatócsoport

Az Intézet alvállakozóként közreműködik az "Új innovatív turisztikai szolgáltatás alapjainak megteremtése NLP módszer segítségével" című, GINOP-2.1.1-15-2015-00517 azonosítószámú pályázatban, amelynek célja a CARTOUR Idegenforgalmi Szolgáltató Kft. és a TRAVELWEB Informatikai, Kereskedelmi és Szolgáltató Kft. mint konzorcium együttműködésével az utazási irodák számára nyelvtechnológia felhasználásával készített intelligens ajánlórendszer kifejlesztése. Az 2016. október 1. és 2018 április 30. között futó projektben az Intézet 15m Ft+ÁFA értékben végez kísérleti K+F tevékenységet.

| pályázó vezető kutató neve | pályázat címe | pályázat kiírója | pály. azonosító száma | a projekt futamideje (tól-ig) | teljes támogatási összeg | együttműködő partnerek (ha vannak) | a projekt fő célja egy mondatban megfogalmazva |
|---|---|---|---|---|---|---|---|
| Váradi Tamás | Új innovatív turisztikai szolgáltatás alapjainak megteremtése NLP módszer segítségével. | Nemzetgazdasági Minisztérium | GINOP-2.1.1-15-2015-00517 | 2016. 10. 01. -- 2018. 04. 30. | 15 000 000 + ÁFA | TrawelWeb Kft. | Nyelvtechnológiai módszereket felhasználó turisztikai ajánlórendszer fejlesztése. |

# A 2016-ban megjelent jelentősebb tudományos publikációk

### Nyelvtechnológiai kutatócsoport

Benyeda, Ivett, Koczka, Péter, Váradi, Tamás. Creating seed lexicons for under-resourced languages. In: Proceedings of GLOBALEX2016. Portorož, Slovenia, pp. 52–56. (2016) URL: http://ailab.ijs.si/globalex/files/2016/06/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf

Falyuna Nóra. Az arcvédő hazugság. A szívességkérés elutasításában megjelenő hazugság vizsgálata. Jelentés és Nyelvhasználat 3: 21–48. (2016) URL:

http://www.jeny.szte.hu/jeny/images/issues/2016/JENY-2016-FalyunaN.pdf

Héja, Enikő. Revisiting Translational Equivalence: Contributions from Data-Driven Bilingual Lexicography. International Journal of Lexicography (2016), doi: 10.1093/ijl/ecw032 URL: http://ijl.oxfordjournals.org/content/early/2016/07/11/ijl.ecw032.full.pdf?keytype=ref&ijkey=FRNMmy2CrsF4Q7z

Hunyadi, László, Váradi, Tamás, Szekrényes, István. Language technology tools and resources for the analysis of multimodal communication in digital humanities 2016. In: Proceeding of the Workshop on Language Technologies for Digital Humanities. COLING 2016. 117–124. (2016) URL: http://www.clarin-d.de/images/lt4dh/pdf/LT4DH16.pdf

Ludányi Zsófia. Szaknyelvi helyesírási változások az AkH.[12] tükrében, különös tekintettel az orvosi nyelvre. Szaknyelv és szakfordítás 16: (16) 34–42. (2016) URL: http://tti.gtk.szie.hu/sites/default/files/upload/page/szie_2016-v3.pdf

Sass Bálint. Nyelvészeti szövegkeresők, Nemzeti Korpuszportál. Magyar Tudomány (7): 798–808. (2016)

Eszter Simon, Veronika Vincze. Universal Morphology for Old Hungarian. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, Berlin, 2016. 118–127. URL: http://www.aclweb.org/anthology/W/W16/W16-2115.pdf

**Matematikai nyelvészeti kutatócsoport**

Borbély, Gábor, Makrai, Márton, Nemeskey, Dávid, Kornai, András. Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. Association for Computational Linguistics. 83–89 (2016) Berlin, Germany. URL: http://anthology.aclweb.org/W16-2515

§Recski Gábor, Iklódi Eszter, Pajkossy Katalin, Kornai András 2016. Measuring Semantic Similarity of Words Using Concept Networks. In: Ph. Blunsom, et al. (szerk.) Proceedings of the 1st Workshop on Representation Learning for NLP: The 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics. (2016) 193–200. URL: http://aclweb.org/anthology/W/W16/W16-1622.pdf

**Alkalmazott nyelvészeti kutatócsoport**

Heltainé Nagy Erzsébet. A nyelvi tanácsadás mint az anyanyelvi nevelés és közművelődés lehetősége. In: Kas Bence (szerk.). „Szavad ne feledd!” Tanulmányok Bánréti Zoltán tiszteletére. MTA Nyelvtudományi Intézet, Budapest. (2016) 247–256.