A Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály jelentése a 2017. évről

Kiemelkedő kutatási és más jellegű eredmények

Nyelvtechnológiai Kutatócsoport

Folytatódott az *Új, innovatív turisztikai szolgáltatás alapjainak megteremtése NLP módszer segítségével* (GINOP-2.1.1.-15) című projekt, amelynek keretein belül a nyelvtechnológiai kutatócsoport szakmai tanácsokkal, háttértanulmányok készítésével segítette az új turisztikai weboldal és a mögötte levő keresőmotor fejlesztőit.

A projekt célja a magyar turisztikai szektorban egyedinek számító szolgáltatás kifejlesztése, amely rugalmasan, felhasználóbarát módon képes felmérni az ügyfelek igényeit és azokra testreszabott megoldást nyújt, beleértve a szálláshoz kapcsolódó szolgáltatások csomagban történő ajánlását. Egy ilyen szolgáltatás létrehozása két ponton is megkívánja a nyelvtechnológiai megoldásokon alapuló intelligens nyelvfeldolgozást: egyrészt az ügyfelek kívánságainak a megértésében, továbbá a szálláshelyek és turisztikai szolgáltatást kínáló ügynökségek kínálatának elemzésében.

Mindkét problémában közös az a feladat, hogy a gépi rendszernek valamilyen szinten meg kell értenie a természetes nyelvi inputot. Az ügyfél által megadott kötetlen keresőkifejezésekből pontosan meg kell tudnia állapítania azokat a jellemzőket, amelyekkel az ügyfél szálláshelyre, utazásra vonatkozó igényét szakszerűen meg lehet határozni. Másrészt az ügynökségek, turisztikai szolgáltatók által adott szöveges leírásokból egy szakszerű értelmezést kell kinyerni a turisztikai kínálat pontos tartalmáról, ahhoz, hogy az igények és a kínálat illesztéséből egy csomag ajánlatot lehessen megfogalmazni. A tanácsadás keretein belül különféle szólisták és nyelvi háttéranyagok fejlesztésére és tanulmányok írására került sor.

A *Magyar Generatív Szintaxis 2* című OTKA-projektben az év folyamán megtörtént a Káldi György- és a Pesti Gábor-féle Újszövetség-fordítás OCR-ezése és a nyers OCR-kimenet összeolvasása, így elkészült a betűhű szövegváltozat. A Károli-féle bibliafordítás újszövetségi részének a kézi normalizálása majdnem teljes egészében készen van, amelynek támogatásához egy automatikus előnormalizáló eszköz is kifejlesztésre került. A Sylvester-féle Újszövetség-fordítás normalizálása elkezdődött. Elkészült a Heltai-féle fordítás nyers OCR-kimenetének az összeolvasása, így előállt a teljes betűhű szöveg.

Az elkészült szövegek elérhetőek a Régi Magyar Konkordancia (http://omagyarkorpusz.nytud.hu) keresőfelületén keresztül. Ezen felül készült egy újfajta korpuszlekérdező eszköz, amellyel különböző korokból származó, illetve különböző nyelvű bibliafordításokat lehet párhuzamosan megjeleníteni: http://parallelbible.nytud.hu/. Ennek az adatbázisában jelenleg 4 ómagyar, 3 középmagyar, 3 modern magyar és egy angol bibliafordítás található, és a további források feltöltését lehetővé tevő infrastruktúra is kidolgozásra került. Emellett a lekérdező felület újabb lehetőségekkel gazdagodott, amelyek a kutatók számára könnyebben elérhetővé teszik a különböző bibliafordítások tanulmányozását.

Az, hogy ezek az ó- és középmagyar korból származó bibliafordítások kereshető szöveges formában, illetve mai magyar átírásban és nyelvészeti elemzéssel ellátva elérhetővé váltak, jelentős hozzájárulás a nemzeti kulturális örökség megőrzéséhez.

Elkészült az ó- és középmagyar szövegek morfológiai elemzésére használható új morfológiai elemző, amely az újonnan kifejlesztett mai magyar morfológiai elemzőnek, az emMorph-nak a régi magyarra adaptált változata. A forráskóddal együtt az eddig ismeretlen új szavak és morfológiai konstrukciók felvétele is elérhetővé vált.

A *FinnOTKA* projekt korábbi szakaszaiban előállított protoszótárak kézi validálása, javítása és kiértékelése történt meg a projektnek ebben a szakaszában. Az összevont szótárak kézi kiértékelését az adott nyelv anyanyelvi beszélői és nyelvész szakértők végezték. A teljes validálás és kiértékelés munkafolyamata ki lett dolgozva, és a validátorok el lettek látva a megfelelő instrukciókkal. Ezekből az instrukciókból születtek meg a kiértékelésnél használt kategóriák. Az automatikusan létrehozott protoszótárak kézi kiértékelése és javítása több célt is szolgált. Egyrészt lehetőséget adott a szótárépítési módszerek összehasonlítására, amelynek eredményei több előadásban és cikkben is publikálva lettek. Másrészt megadta azoknak a szópároknak a számát, amelyek a megfelelő nyelvi információkkal kibővítve feltölthetők a Wiktionarybe. A Wiktionary-szócikkek generálása teljesen automatikusan történt; a feltöltés még folyamatban van: a 2017-es év alatt a magyar Wikiszótár 3909, a finn Wikisanakirja 763 új szócikkel bővült a projekt jóvoltából.

A projekt során előállított kétnyelvű szótárak a kisebbségi finnugor nyelvek revitalizációját próbálják támogatni úgy, hogy az addig csak szegényes digitális tartalommal rendelkező nyelvek számára újabb szópárokkal gazdagítják az interneten fellelhető fordítási párok számát. A szótári elemek a Wiktionary különböző nyelvű változataiban összekapcsolhatók, az interwiki linkek pedig a Wikipédia felé biztosítják az átjárást. Ez lehetővé teszi, hogy a nyelvközösségek gazdag lexikai anyaghoz férjenek hozzá. Emellett olyan új adatok is elérhetők lesznek a lexikai elemekhez, mint a szófaji információ vagy a fordítási megfelelők. Mindez egyfajta hozzájárulás kíván lenni a nyelvi sokszínűség fenntartásához.

Idén zárult *Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben* című projekt, amelynek célja egy annotált korpusz létrehozása volt udmurt, tundrai nyenyec, szinjai és szurguti hanti nyelvű, írott és beszélt nyelvi szövegekből, amely lehetővé teszi az uráli–orosz kontaktushatás kutatását. Az adatbázis különböző korokból gyűjtött szövegeket tartalmaz — minden szöveg legalább a lejegyző által használt eredeti átírásában, valamint IPA-átírásban szerepel. 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. Ezen felül minden írott, illetve lejegyzett hangzó anyaghoz készült egy .eaf fájl, amely minden token- és mondatszintű információt tartalmaz, továbbá illesztve van a hangzó anyaghoz. Szintén idén készült el a projekt honlapja, amely részletes információt tartalmaz az adatbázisról, valamint innen érhetők el az elkészült fájlok is: http://www.nytud.hu/oszt/elmnyelv/urali/adatbazisok.html

A nyelvtechnológia bevett szóreprezentációi a szóbeágyazások (word embeddings), különösen a skip-gram modell (Mikolov et al 2013), amelyek gépi tanulása felügyeletlen módon, szövegkorpuszokból történik, és a szokásos esetben szóalakokat reprezentálnak, tehát érzéketlenek a szavak többértelműségére. Vannak többjelentésű szóbeágyazások is (multi-sense embeddings, MSE), ezeknek azonban mind a gyarkolati hasznossága (Li és Jurafsky, 2015), mind az elméleti magyarázóereje korlátozott. Utóbbi téren az lenne az ideális, ha a különböző vektorok különböző jelentéseknek (a gyakorlatban homonímáknak) felelnének meg, de a mostani modellekben gyakran duplikálódnak a jelentések. Az intézetben szófordításban vannak vizsgálva az MSEk. Az idei kutatások kísérleti úton mutatták ki azt a nem meglepő jelenséget, hogy a szemantikai felbontás mentén csereviszony van: specifikusabb jelentésreprezentációkat könnyebb fordítani, azon az áron, hogy a különbözőnek jósolt jelentések fordítása gyakran egybeesik.

2017-ben született meg két, az *e-magyar* Digitális Nyelvfeldolgozó Rendszert bemutató publikáció, ezek a Magyar Számítógépes Nyelvészeti Konferencián hangzottak el. Az egyik magát a projektet és az integrált nyelvtechnológiai eszközöket mutatja be, a másik pedig a rendszer használatát a célközönség számára. Nemzetközi színtéren 2018-ban mutatkozik be a rendszer.

Az osztály egy tagja 3 évre szóló Bolyai János Kutatási Ösztöndíjat nyert. A pályázat címe: "Igei szerkezetek algebrai struktúrája". Megkezdődött a kutatás, megtörtént az első átfogó tanulmány összeállítása, mely a modell általános tulajdonságait írja le.

A igék bővítményszerkezetének vizsgálatára szolgáló *Mazsola* eszköz szabadon elérhetővé vált. Ezzel megnyílt a lehetőség, hogy a saját adatai használatával bárki egyszerűen készítsen hasonló

funkcionalitású kutatóeszközt. A Mazsola alkalmas nem csak magyar nyelvű, hanem más nyelvű szövegek; nem csak egynyelvű, hanem párhuzamos korpusz; sőt nem csak igék és bővítmények, hanem más struktúrák kezelésére is.

Kutatás folyt arról, hogy hogyan lehetne alkalmazni a Barabási Albert László-féle skálafüggetlen hálózatok módszertanát a szövegek vizsgálatára. Publikáció 2018-ban esedékes.

Megjelent egy tanulmány, mely egyrészt bemutatja a *Magyar történeti szövegtár* új keresőfelületét, másrészt ennek segítségével ismerteti az alapvető, általános, minden korpuszra alkalmazható korpuszkeresési módszereket.

Az osztály egy tagja jelentős nemzetközi folyóiratpublikációjával kiérdemelte az intézet évente egy kutatónak kiosztott publikációs díját.

Elkészült a "Cognitive Infocommuniations and Computing" könyvbe szánt "Using deep rectifier neural nets and probabilistic sampling for topical unit classification" című fejezet, melynek megjelenése 2018-ra várható.

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.

A kutatócsoport speciális feladata az utónév-szakvéleményezés: a névviselésről szóló 2010. évi I. törvény az anyakönyvi eljárásról 44.§-a alapján. A csoport a Miniszterelnökségtől érkező 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 2017-ben 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. 2017-ben a lista 95 új névvel bővült (40 női, 55 férfi) így 2018. január 1-jén 3955 bejegyzésre alkalmasnak minősített utónév (2243 női és 1712 férfi) alkotja a jegyzéket. A nevekkel kapcsolatos tanácsadó szolgálat 2017-ben kb. 2500 emailben (nevtanacs@nytud.mta.hu) érkezett, és kb. 500 telefonos névadással kapcsolatos, névhasználati kérdésre adott választ, 10 névhasználati szakvéleményt készített, 25 családnevekkel kapcsolatos szakvéleményt készített, valamint BÁH egyszerűsített honosítási eljárás során kért kb. 110 keresztnevet véleményezett szakértőként.

Párbeszéd a tudomány és a társadalom között

Nyelvtechnológiai Kutatócsoport

Eseti tud. ism. tevékenység (ea-k, sajtótájékoztatók stb). Az MTA SZTAKI-val való szoros együttműködésben zajló kutatásokat a műkedvelő nagyközönség számára a *Számítógépes társadalomtudomány* témacsoport Magyar Tudomány ünnepén megrendezett alkuló workshopján elhangzott előadás helyezte kontextusba. Két nyelvtechnológiai paradigma ötvözéséről van szó, nevezetesen a diszkrét szimbólumkezelésnek a statisztikai alapú, gépi tanulásos modellekkel való hibridizációjáról, vagyis hogy miképp találhatjuk meg a diszkrét struktúrát a folytonos, zajos adatokban, illetve hogyan tudjuk a struktúráról való ismereteinket hatékonyabb algoritmusok építésében kamatoztatni. Konkrétabban a szavak többértelműségének szóbeágyazásokban való reprezentálásával a *Fiatal kutatók félidőben* című konferencián ismerkedhetett a műkedvelő nagyközönség. Az osztály munkatársai előadóként részt vettek Debreceni Egyetem által szervezett *Az emberi viselkedés rejtett mintázatai és a gép* című workshopján, szintén a "Magyar Tudomány

ünnepe 2017" rendezvénysorozat keretében.

Egyéb. Az osztály üzemelteti a nagy népszerűségnek örvendő *helyesiras.mta.hu helyesírási tanácsadó portál*t. 2017-ben 1,5 millió látogató (54%-uk rendszeres látogató) több mint 3 millió tanácsot kért itt. Ez jelentős növekedés az elmúlt évhez képest, a lekérdezések száma munkanapokon rendszeresen meghaladja a 10000-et. Az MTA Nyelvtudományi Intézete egy olyan tudományos alapokon álló szolgáltatást üzemeltet 2013 óta, mely valóban hasznos, sokakhoz eljut és 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, az 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 2017-ben 593000 lekérdezést futtattak. A 2016-ban megnyílt *e-magyar.hu* weboldal által nyújtott szövegelemző szolgáltatás nem csak a szakma szereplői, hanem a laikus felhasználók számára is hasznos.

A csoport esetenként számítógépes nyelvészeti támogatást ad különböző hazai és külföldi nyelvészeti kutatásoknak, nyelvi adatokat szolgáltat, egyéni kéréseket teljesít, a Magyar Nemzeti Szövegtár, szógyakorisági adatok, más korpuszok és az e-magyar rendszer vonatkozásában. A korpusznyelvészeti támogatás idén hozzájárult egy jelentős folyóiratpublikáció megjelenéséhez.

Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport

Médiamegjelenések. A névadással, névkultúrával kapcsolatos tájékoztatásra folyamatosan nagy az igény a média részéről. 2017-ben ez számos rádiós és tv-s, illetve online és papír alapú sajtóban megjelent interjú formájában valósult meg. Az osztály egy munkatársa heti rendszerességgel állandó vendége a Lánchíd Rádió nyelvi, nyelvészeti ismeretterjesztéssel foglalkozó műsorának. Helyesírási és nyelvhasználati kérdésekről két rádióriportban adott tájékoztatást a csoport munkatársa 2017-ben.

Rendszeres tud. ism. tevékenység (weben, sajtóban). A csoport egy 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 érdekében.

Eseti tud. ism. tevékenység (ea-k, sajtótájékoztatók stb). A csoport munkatársa 3 alkalommal tartott tudományos ismeretterjesztő előadást nagyvállalalatoknál aktuális helyesírási és egyéb nyelvi kérdésekről.

Egyéb. A folyamatosan működő nyelvi tanácsadó szolgálat – melyet a csoport három munkatársa lát el – 2017-ben kb. 2000 emailben (tanacs@nytud.mta.hu) érkezett, és kb. 800 telefonos helyesírási, nyelvhasználati kérdésre adott választ. A kérdések 90 százaléka helyesírási és szövegértelmezési jellegű volt, a többi pedig nyelvhelyességi, szókészlettani, szemantikai, stilisztikai. Összességében elmondható, hogy a hozzánk fordulók elfogadták, megismerték és alkalmazták a 2015-ben megjelent helyesírási szabályzat, az AkH.12 módosulásait, változásait; ezekben alkalmanként megerősítést kértek. A helyesírási szabályzat ezen kiadása a mai írásgyakorlat szerves részévé vált. A kérdezők visszajelzései alapján tovább nőtt a helyesírás.mta.hu portál ismertsége, különösen a *Helyes-e így?* és a *Külön vagy egybe?* eszközök használata. Jelentős segítségnek bizonyult a felhasználók számára az oldalon elérhető helyesírási szabályzatok (a 11. és a 12. kiadás) teljes szövege, különös tekintettel a változások megjelenítésére. A korábbiakhoz képest érzékelhető egyfajta változás a kérdések minőségét illetően: az utóbbi időben főleg olyan "profik" (szerkesztők, fordítók, lektorok, korrektorok) kérdeznek, akik a helyesiras.mta.hu automatikus válaszaival elégedetlenek. A nyelvi tanácsadó szolgálatban nőtt a külhoni kérdezők

száma. 2017-ben számos helyesírási, nyelvhasználati kérdés érkezett új ügyfelektől, különösen helynevek, utcanevek, egyéb földrajzi nevek írásgyakorlatával kapcsolatban. A partnerek között megtalálhatók állami és európai uniós hivatalok, minisztériumok, cégek, kiadók, más szervezetek, fordítók, tanárok, újságírók, szerkesztők, magánszemélyek. A szolgálat 6 nyelvészeti szakvéleményt készített és emléktáblák (23 db) nyelvi lektorálásával 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. Az utóneveket bemutató Utónévportál 2017-ben több mint 40000 látogató 110000 lekérdezését szolgálta ki.

A kutatóhely hazai és nemzetközi K+F kapcsolatai 2017-ben

Hazai kapcsolatok

Nyelvtechnológiai Kutatócsoport

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.

A FinnOTKA és az uráli projektek keretein belül a kutatócsoport jó kapcsolatot ápol az ELTE Finnugor Tanszékével. Az együttműködés célja egy közös konferencia szervezése a társprojektek finnugor nyelveket érintő eredményeinek ismertetése céljából.

Egy másik OTKA projekt a Debreceni Egyetemmel közös munkában folyik.

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

Megalakult a Computational Social Science - Számítógépes Társadalomtudomány témacsoport, amelynek a kutatócsoport is aktív tagja. Célunk, hogy nyelvtechnológiai támogatást nyújtsunk a társadalom- és bölcsészettudományi kutatásokhoz.

Együttműködés alakult az ELTE Digitális Bölcsészet Kutatócsoportjával. A cél elsősorban az irodalmi, stilometriai, filológiai és egyéb digitális bölcsészeti kutatások nyelvtechnológiai támogatása. Az együttműködő kutató Palkó Gábor.

Az intézet igazgatója az általa megnyert és 2017.07.01-én induló kutatócsoport-pályázatban (30203 sz. MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport) tanácsadóként vesz részt.

Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport

Együttműködés a KRE BTK Nyelvtudományi Tanszékével a Terminológia MA szakmai gyakorlatok támogatásában.

Együttműködés az Anyanyelvápolók Szövetségével.

Együttműködés a PeLi Oktatásnyelvészeti kutatócsoporttal (Eszterházy Károly Egyetem Magyar Nyelvészeti Tanszéke, vezető kutató: Domonkosi Ágnes). Az oktatásnyelvészet az alkalmazott nyelvészet meghatározó részterülete, amelynek vizsgálati körébe tartozik minden olyan oktatási, pedagógiai kérdés, amelyek megoldásában, feltárásában a nyelvtudomány eredményei és módszerei hasznosulhatnak. A PeLi Oktatásnyelvészeti kutatócsoport az EKE Nyelv-és Irodalomtudományi Intézete Magyar Nyelvészeti Tanszékének tudományos munkaközössége, amelynek munkájában a tanszék oktatóin kívül más intézmények munkatársai, illetve az OFI kutatói is részt vesznek.

Nemzetközi kapcsolatok

Nyelvtechnológiai Kutatócsoport

A FinnOTKA projekt keretén belüli együttműködés továbbra is fennáll 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 belüli együttműködés az Ob-Ugric Database: analysed text corpora and dictionaries for less described Ob-Ugric dialects című projekt résztvevőivel a müncheni Ludwig Maximilian Egyetemről ebben az évben is folytatódott. Az együttműködés célja szurguti hanti szövegek morfológiai elemzése és beszélt nyelvi anyagok lejegyzésének IPA-konverziója volt.

A kutatócsoport szoros kapcsolatot ápol az ACL Special Interest Group on Uralic Languages vezetőségi tagjaival, név szerint Tommi A. Pirinennel (Universität Hamburg), Francis Tyersszel (UiT Norgga árktalaš universitehta) és Vincze Veronikával (Szegedi Tudományegyetem). Az együttműködés eredménye egy társszerkesztett Acta Linguistica Hungarica különszám az uráli nyelvek számítógépes nyelvészeti támogatásáról, valamint több közös publikáció.

COST Independent External Expert (OC-2017-1: 2 pályázat).

Sikeresen befejeződött az Európai e-lexikográfiai hálózat (European Network of e-lexicography, ENEL) COST projekt, amelyben az Szótári Osztály és a Nyelvtechnológiai kutatócsport aktívan részt vett. A projekt 2017. februári plenáris konferenciájának az Intézet volt a helyi szervezője. A projekt eredményeként létrejött egy európai lexikográfiai portál (www.e-lexcography.eu) és elkészült egy H-2020 pályázat European Lexicographic Infrastructure (ELEXIS) címmel, amelynek elkészítésében az Intézet is részt vett. A pályázat sikeres lett és az ELEXIS projekt 2018. február 1-én megkezdi munkáját.

A CLARIN európai kutatási infrastruktúra (clarin.eu) 2017 idén Budapesten tartotta éves konferenciáját, amelynek az Intézet volt a helyi szervezője. Erre annak nyomán került sor, hogy az Intézet által koordinált HUN-CLARIN hálózat 2016-ban tagja lett az európai CLARIN ERIC szervezetnek. A 160 fő részvételével tartott konferencia alkalmat adott a magyar CLARIN műhelyek bemutatására illetve a magyar és külföldi kollégák személyes megbeszéléseire is. Ennek nyomán együttműködés is indult a Wroclawi Egyetemmel, az általuk fejlesztett WEBSty irodalmi stíluselemző rendszer magyar szövegekre való alkalmazása témájában.

Felsőoktatási tevékenység, tagságok

Nyelvtechnológiai Kutatócsoport

Makrai Márton: Számítógépes szemantika gyakorlat, ELTE BTK, 2017. őszi félév.

Sass Bálint: 2 MA szakdolgozat bírálás (PPKE BTK), 2 MA záróvizsgáztatás (PPKE BTK). Digitális Bölcsészet folyóirat -- szerkesztőbizottsági tag.

Simon Eszter: 2 MA szakdolgozat bírálás (PPKE BTK), 1 MA záróvizsgáztatás (PPKE BTK). Szakmai tagságok: XIII. Magyar Számítógépes Nyelvészeti Konferencia — programbizottsági tag, Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature — programbizottsági tag, Syntax Of Uralic Languages — programbizottsági tag, Third International Workshop on Computational Linguistics for Uralic Languages — programbizottsági tag, ACL Special Interest Group on Uralic Languages -- tag

Prószéky Gábor: A PPKE ITK-n egyetemi tanári minőségben oktatói tevékenységet folytat (két állandó kurzusa van). 6 PhD-hallgatója és egy posztdoktora van. Szakmai tagságok: A PPKE ITK Doktori és Habilitációs Tanácsának elnöke. Az MTA MNYOÁB (Magyar Nyelvi Osztályközi

Állandó Bizottság) elnöke, a Magyar Alkalmazott Nyelvészek és Nyelvtanárok Egyesületének elnöke, a Magyar Nyelvészeti Diákolimpia Bizottság elnöke, tagja az MTA Szótári és Alkalmazott Nyelvészeti Munkabizottságának és a Nyelvtudományi Bizottságnak.

Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport

Heltainé Nagy Erzsébet: a KRE BTK, Magyar Nyelvtudományi Tanszék oktatója, 18 gyakorlati és 2 elméleti kurzust tartott stilisztika, szöveg-és stíluselemzés, jelentéstan, retorika, terminusalkotás, és helyesírás témákban. MA szakdolgozat vezetése: 2. MA szakdolgozati opponens: 2. PhD-védésen bizottsági tag: 1. Szakmai tagságok: MTA Magyar Nyelvi Osztályközi Állandó Bizottság -- tag, MTA Magyar Nyelvészeti Munkabizottsága -- tag, MTA Nyelvhelyességi Tanácsadó Testület -- tag, Magyar Nyelvtudományi Társaság -- tag, MANYE -- tag, Termini Egyesület -- alapító tag, Anyanyelvápolók Szövetsége -- választmányi tag Szerkesztőbizottsági tagságok: Magyar Nyelvőr -- a szerkesztőbizottság tagja, Édes Anyanyelvünk -- a szerkesztőbizottság tagja

Kardos Tamás: 2 x 2 óra helyesírási szeminárium tartása a Károli Gáspár Református Egyetemen.

Ludányi Zsófia: Helyesírási gyakorlatok (szeminárium), Károli Gáspár Református Egyetem, 2017. tavaszi félév. Beszédművelés, nyelvi norma (szeminárium), Nyelvművelés (szeminárium), Eszterházy Károly Egyetem, 2017. őszi félév. Szakmai tagságok: MTA Magyar Nyelvi Osztályközi Állandó Bizottság -- tag, MTA Orvosi Nyelvi Munkabizottság -- tag, MTA Nyelvhelyességi Tanácsadó Tesület -- titkár Szerkesztőbizottsági tagságok: Magyar Orvosi Nyelv

Raátz Judit 20 gyakorlat az ELTE BTK-n magyar nyelv tanításának módszertana, tanári kommunikáció és retorika, kritériumvizsga témákban. 3 BA, 2 MA és 5 PhD témavezetés. Szakmai tagságok: MTA Magyar Nyelvi Osztályközi Állandó Bizottság tag, MTA köztestületi tag, Anyanyelvápolók Szövetsége választmányi tag, International Council of Onomastic Sciences (ICOS) tagja, Magyar Nyelvtudományi Társaság, Magyar Nyelvtudományi Társaság Névtani Tagozat, Magyar Nyelvtudományi Társaság Magyartanári Tagozat, Szemere Gyula anyanyelv-pedagógiai kutatócsoport, HUNRA Magyar Olvasástársaság

KKV

Nyelvtechnológiai Kutatócsoport

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

Részvétel nemzetközi konferenciákon

Nyelvtechnológiai Kutatócsoport

előadó	előadás címe	konferencia elnevezése	helye (város)	ideje (hónap)
Kovács György	Classification of Formal and Informal Dialogues Based on Turn-Taking and Intonation Using Deep Neural Networks	19th International Conference on Speech and Computer (SPECOM 2017	Hatfield (UK	szeptember
Sass Bálint	Tools for corpus-driven research of Hungarian: the Hungarian Gigaword Corpus and the Verb	CLARIN 2017 conference -- CLARIN bazaar	Budapest	szeptember

	Argument Browser			
Simon Eszter, Mus Nikolett	Languages under the influence -- Building a database of Uralic languages	Third International Workshop on Computational Linguistics for Uralic Languages	Szentpétervár, Oroszország	január
Simon Eszter, Mittelholcz Iván	Automatic creation of bilingual dictionaries for Finno-Ugric minority languages	Trans-Linguistica 4.	Marosvásárhely, Románia	május
Kristiina Jokinen, Graham Wilcock, Tamás Váradi, Eszter Simon	Finno-Ugric Digital Natives -- Linguistic Support for Finno-Ugric Digital Communities	Digital Revolution for Under-Resourced Languages	Stockholm, Svédország	augusztus
Simon Eszter, Mittelholcz Iván	Evaluation of Dictionary Creating Methods for Under-Resourced Languages	Text, Speech and Dialogue 2017	Prága, Csehország	augusztus
Simon Eszter, Mus Nikolett, Kalivoda Ágnes, Ruttkay-Miklián Eszter	UraLUID: Supporting data-driven (prosodic) research	Second workshop on Uralic prosody	Budapest	szeptember
Simon Eszter, Kalivoda Ágnes	Introducing the UraLUID database	Some results of the project Languages under the Influence. Uralic syntax changing in an asymmetrical contact situation	Budapest	november
Prószéky Gábor	A Functional Computational Approach to Human Text Comprehension	Computational linguistics from science to ubiquitous utility	Hamburg	június

Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport

előadó	előadás címe	konferencia elnevezése	helye (város)	ideje (hónap)
Ludányi Zsófia	Nyelvi ideológiák napjaink orvosi szaknyelvi sztenderdizációs tevékenységében.	A nyelv közösségi perspektívája III.	Nagyvárad	július

Rendezett konferenciák

Nyelvtechnológiai Kutatócsoport

szervező	konferencia elnevezése	helye (város)	ideje (napra!)	(társszervező, ha volt)
Ludányi Zsófia	XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia	Budapest	2017. február 3.	Váradi Tamás, Tóth Bianka, Kuti Judit
Simon Eszter	XIII. Magyar Számítógépes Nyelvészeti Konferencia	Szeged	2017. január 26-27.	Vincze Veronika, Farkas Richárd, Csirik János
Váradi Tamás	Final Conference of the ENEL COST Action	Budapest	2017. február 24-25.	Nyelvtudományi Intézet
Váradi Tamás	CLARIN Annual Conference 2017.	Budapest	2017. szeptember 18-20.	HUN-CLARIN
Prószéky Gábor	CiCLing-2017	Budapest	2017. április 17-23.	Novák Attila, Siklósi Borbála

A 2017-ben elnyert fontosabb hazai és nemzetközi pályázatok rövid bemutatása

Nyelvtechnológiai Kutatócsoport

Az Intézet osztály tagja az ELEXIS nevű H-2020-as pályázatot elnyerő nemzetközi konzorciumnak. A 2018-ban induló pályázat célja az európai lexikográfia számára egy egységes infrastruktúra megteremtése.

Számítógépes pszicholingvisztikai modell támogatása nagyméretű, gazdagon annotált korpusz és neurális hálók alkalmazásán alapuló megoldásokkal. A webről gyűjtött magyar nyelvű szövegeket tartalmazó Pázmány Korpuszt tovább bővítve az így kapott, legnagyobb magyar nyelvű szöveges anyagot a neurális hálózatok által implementált disztribúciós modell alapjaként használva a nyelvben releváns szemantikai jegyek olyan gazdag halmazát tudjuk meghatározni, aminek felhasználásával az emberi agyban történő nyelvi feldolgozás minden eddiginél jobban közelíthető lesz számítógépes módszerekkel. Prószéky Gábor 2017. április 1-től az MTA NYTI igazgatója lett. Az általa megnyert és 2017.07.01-én induló kutatócsoport-pályázatban a kinevezés miatt tanácsadóként működik tovább.

Az osztály egy munkatársa 2017-ben Bolyai János Kutatási Ösztöndíjat nyert. A pályázat címe: "Az igei szerkezetek algebrai struktúrája." A pályázat célkitűzése egy olyan általános modell kidolgozása, mely alkalmas a különféle igei szerkezetek reprezentálására, illetve lehetőséget ad a modellre épülő lexikai kinyerő eljárások kidolgozására.

pályázó vezető kutató neve	pályázat címe	pály. azonosító száma	a projekt futamideje (tól-ig)	teljes támogatási összeg	együttműködő partnerek (ha vannak)	a projekt fő célja egy mondatban megfogalmazva
Váradi Tamás	European Lexicographic Infrastructure	731015	2018. febr. 1 - 2021 január 31.	130895 euró	A szlovén Jozef Stefan Institute által vezett 17 fős konzorcium	Az európai lexikográfia számára egy egységes infrastruktúra

						megteremtése.
Prószéky Gábor	Számítógépes pszicholingvisztikai modell támogatása nagyméretű, gazdago annotált korpusz és neurális hálók alkalmazásán alapuló megoldásokkal	30203 sz. MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport	2017.07.01 -- 2022.06.30.	125 millió Ft	PPKE ITK	Az emberi agyban történő nyelvi feldolgozás számítógépes közelítése neurális hálózatok által implementált disztribúciós modellek segítségével.
Sass Bálint	Az igei szerkezetek algebrai struktúrája	ügyszám: BO/00064/17/1	2017.09.01 -- 2020.08.31.	4,5 millió Ft	--	Az igei szerkezetek általános modelljének kidolgozása

A 2017-ben megjelent jelentősebb tudományos publikációk

Nyelvtechnológiai kutatócsoport

Kuna Ágnes, Kocsis Zsuzsanna, Ludányi Zsófia. A Magyar orvosi nyelvi korpusz 16–17. századi alkorpusza: Tervezet, átírás, annotálás. In: Forgács Tamás, Németh Miklós, Sinkovics Balázs (szerk.) A nyelvtörténeti kutatások újabb eredményei IX. Szeged: Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék. 239–253. (2017) URL: https://drive.google.com/open?id=128eKlUFVpJH2zYogImwZ2QnaMUk7yGgW

Sass Bálint. Keresés korpuszban: a kibővített Magyar történeti szövegtár új keresőfelülete. In: Forgács Tamás, Németh Miklós, Sinkovics Balázs (szerk.) A nyelvtörténeti kutatások újabb eredményei IX. Szeged: Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék. 267–277. (2017) URL: http://www.nytud.hu/oszt/korpusz/resources/sb_kereses_korpuszban.pdf

Eszter Simon, Iván Mittelholcz. Evaluation of Dictionary Creating Methods for Under-Resourced Languages. In: Kamil Ekštein, Václav Matoušek (szerk.): Text, Speech and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings. Cham: Springer. 246–254. (Lecture Notes in Computer Science; 10415.) (2017) URL: https://www.academia.edu/34690587/Evaluation_of_Dictionary_Creating_Methods_for_Under-Resourced_Languages

Szekrényes István, Kovács György. Classification of Formal and Informal Dialogues Based on Turn-Taking and Intonation Using Deep Neural Networks. Proc. Specom 2017, pp. 233–243, Hatfield (2017) URL: https://drive.google.com/open?id=1_IcUVOziVID1-21jcu2UCOu3ltmr8Qr_

Alkalmazott nyelvészeti kutatócsoport

Fercsik Erzsébet, Raátz Judit. Örök névnaptár. Budapest–Piliscsév: Műszaki Kiadó. (2017)