Szegedi Tudományegyetem
Juhász Gyula Pedagógusképző Kar
Magyar és Alkalmazott Nyelvészeti Tanszék

Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért

HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban

Összefoglalók és programfüzet

Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért
HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban
Összefoglalók és programfüzet

Közreműködő szervezetek:

Szegedi Tudományegyetem
Juhász Gyula Pedagógusképző Kar
Magyar és Alkalmazott Nyelvészeti Tanszék

Magyar Alkalmazott Nyelvészek és Nyelvtanárok Egyesülete

Emberi Erőforrások Minisztériuma
Emberi Erőforrás Támogatáskezelő
Nemzeti Együttműködési Alap

CLARIN – HunCLARIN

A rendezvény és a kötet létrejöttét a Nemzeti Együttműködési Alap (pályázati azonosító: NEA-KK-18-SZ-0653) és a CLARIN ERIC támogatta.

Szerkesztette: Sulyok Hedvig, Juhász Valéria, Erdei Tamás

ISBN: 978-615-5455-92-6

Kiadja:
SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszék
Szeged, 2018

Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért

HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban

2018. október 19., Szeged

Program:

10:00	Megnyitó: dékáni köszöntő (SZTE JGYPK); és Váradi Tamás, a HunCLARIN koordinátora
10:15	Vincze Veronika: Bevezetés a korpuszok és nyelvészeti adatbázisok világába
11:00	Sass Bálint: Keresés korpuszban
11:45 – 12:45	Ebédszünet
12:45	Simon Eszter: Magyar nyelvű történeti korpuszok
13:15	Mittelholcz Iván: Bevezetés az e-magyar programcsomag használatába
13:45	Juhász Valéria: A MAXQDA tartalomelemző szoftver lehetőségei
14:15 – 14:45	Kávészünet
14:45	Babarczy Anna: Gyermeknyelvi korpuszok és erőforrások
15:15	Péter Róbert: A big data kihívás a bölcsészettudományokban: néhány digitális bölcsészeti kutatási eszköz bemutatása
15:45	Zárszó

Vincze Veronika

tudományos munkatárs
Szegedi Tudományegyetem
Természettudományi és Informatikai Kar
MTA-SZTE Mesterséges Intelligencia Kutatócsoport

A Mesterséges Intelligencia Kutatócsoport projektjeinek nyelvészeti vonatkozásait felügyeli és koordinálja. Érdeklődési körébe tartozik a korpusz- és ontológiaépítés, a jelentésegyértelműsítés és a többszavas kifejezések számítógépes kezelése. Ezenkívül foglalkozik számítógépes morfológiával és szintaxissal, valamint információkinyeréssel, kiemelten a tagadott és/vagy bizonytalan szövegrészek gépi felismerésével. Az évente megrendezett Magyar Számítógépes Nyelvészeti Konferencia szervezőbizottságának tagja. A konferencia előadói mind a felsőoktatási és tudományos intézményekben, mind a vállalatoknál zajló legfrissebb beszéd- és nyelvtechnológiai kutatások eredményeiről beszámolnak. A Szegedi Tudományegyetemen a nyelvtechnológiával és a számítógépes nyelvfeldolgozással kapcsolatos számos kurzust tartott.

Bevezetés a korpuszok és nyelvészeti adatbázisok világába

Az előadás célja, hogy a hallgatóságot megismertesse néhány korpusszal és egyéb nyelvészeti adatbázissal, továbbá a korpusznyelvészet alapjaival. A legfontosabb alapfogalmak után ismertetjük a különféle korpusztípusokat, létrehozási módjukat, továbbá néhány példán keresztül megmutatjuk, milyen nyelvészeti jellegű információkat (annotációkat) tudunk a szövegekben kódolni. Arra is hozunk példát, hogy a nyers szövegállományból miként tudunk automatikusan annotált adatbázist előállítani. A korpuszok gyakorlati felhasználására is külön figyelmet fordítunk: bemutatjuk, hogy a korpuszokból származó adatokat hogyan lehetséges kigyűjteni, majd azokat nyelvészeti vagy más bölcsészettudományi kutatásra felhasználni.

kulcsszavak: korpusz, nyelvészeti adatbázis, annotálás, korpusznyelvészet

Sass Bálint

tudományos munkatárs
Magyar Tudományos Akadémia
Nyelvtudományi Intézet
Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
Nyelvtechnológiai Kutatócsoport

Fő kutatási területei a korpuszlekérdező felületek és a korpuszépítés. Egyik megalkotója a Magyar Nemzeti Szövegtárnak, a Magyar Történeti Szövegtárnak, az Ómagyar Korpusznak és a Budapesti Szociolingvisztikai Interjúból készített számítógépes korpusznak, illetve ezek lekérdezőfelületének. A magyar igei bővítményszerkezetek vizsgálatára kifejlesztett egy lekérdezőeszközt Mazsola néven, mellyel számos sikeres kutatást folytattak. Foglalkozik még számítógéppel segített szótáralkotással és a magyar Braille-rövidírással is. Egyetemi hallgatóknak több nyelvtechnológiai kurzust is tartott.

Keresés korpuszban

Az előadás során ismertetjük a NoSkE korpuszkezelő rendszer funkcióit, a szűrési lehetőségeket, és azt, hogy miként kaphatunk a céljainknak megfelelő gyakorisági listákat. Betekintést adunk a reguláris kifejezések és a CQL lekérdezőnyelv, valamint a Magyar Történeti Szövegtár és a Magyar Nemzeti Szövegtár használatába. Bemutatjuk, hogyan hajtható végre keresés elemzetlen és elemzett korpuszban, és hogy a korpuszkeresés során milyen elveket érdemes szem előtt tartani. Végül mindezekre példákat mutatunk a cigány eredetű szavaktól az „automatikus” versírásig.

kulcsszavak: korpusz, szövegtár, keresés, lekérdezés, CQL

Simon Eszter

tudományos főmunkatárs
Magyar Tudományos Akadémia
Nyelvtudományi Intézet
Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
Nyelvtechnológiai Kutatócsoport

A nyelvtechnológia számos ágával foglalkozik, kutatási területei közé tartozik a tulajdonnév-felismerés, a morfológiai elemzés, a korpuszépítés és -annotáció, a történeti korpuszok fejlesztése, a digitális bölcsészet és az uráli nyelvek számítógépes nyelvészeti támogatása. Több nagyszabású számítógépes nyelvészeti projekt és kutatócsoport résztvevője és koordinátora, mint a Magyar Generatív Történeti Szintaxis, vagy az Uráli Nyelvek Szintaxisa. Egyik létrehozója az Ómagyar Korpusznak és az e-magyar.hu digitális nyelvfeldolgozó rendszernek. Több nyelvtechnológiai kurzus oktatója.

Magyar nyelvű történeti korpuszok

A nyelvi kulturális örökség elérhetővé tételében kulcsfontosságú szerep jut a nyelvtechnológiának, melynek módszereivel a kutatók egységes, következetes, nyelvi információval ellátott adatbázisokhoz juthatnak. A nyelvtörténészek és nyelvtechnológusok egyik legfontosabb együttműködési terepe a történeti korpuszok építése, melyek kiváló alapanyagot szolgáltatnak az elméleti és történeti nyelvészeti kutatásoknak. Az elmúlt évtizedekben számos történeti korpuszt fejlesztettek – elsősorban indoeurópai nyelvekre, de a magyarra is készült néhány. Időrendi sorrendben haladva ezek a következők. Az Ómagyar Korpusz tartalmazza az összes ómagyar korból fennmaradt szövegemléket és néhány középmagyar kori bibliafordítást is. A Történeti Magánéleti Korpusz az ó- és középmagyar kor magánéleti nyelvi regiszteréhez közelebb álló műfajokat tartalmazza: 1772 előtti magánlevelekből és peres eljárások jegyzőkönyveiből épül fel. A Magyar Történeti Szövegtár pedig 1772-től, vagyis az újmagyar kor kezdetétől egészen a 20. század végéig tartalmaz szövegeket. Előadásomban ezeket a korpuszokat és a hozzájuk tartozó lekérdezőfelületeket fogom ismertetni, és néhány példán keresztül azt is illusztrálom, hogy milyen kutatási kérdésekre hogyan tudunk választ kapni ezeknek az adatbázisoknak a segítségével.

kulcsszavak: történeti korpuszok, nyelvtechnológia, történeti nyelvészet, korpuszok

Mittelholcz Iván

tudományos segédmunkatárs
Magyar Tudományos Akadémia
Nyelvtudományi Intézet
Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
Nyelvtechnológiai Kutatócsoport

2006 óta foglalkozik nyelvtechnológiával, számos kutatási projektben vett részt. Többek között tokenizálással, helyesírás-ellenőrzéssel, ontológiaépítéssel és felügyelt gépi tanulással foglalkozik. Egyik létrehozója az e-magyar.hu digitális nyelvfeldolgozó rendszernek és a helyesiras.mta.hu oldalnak. A közelmúltban több egyetemi kurzust is tartott a logika, a nyelvtechnológia és a programozás témaköreiben.

Bevezetés az e-magyar programcsomag használatába

Előadásunkban átfogó ismertetést adunk az e-magyar rendszerről, amely a magyar nyelvtechnológiai műhelyek összefogásával jött létre. Az új infrastruktúra fő célja: írott szövegek sokrétű automatikus nyelvi elemzése egyetlen koherens technológiai lánccal. A programcsomag egyes moduljainak ismertetésén túl az e-magyar.hu weboldalon keresztül a gyakorlati használat is bemutatásra kerül.

kulcsszavak: tokenizálás, morfológia, szófaji egyértelműsítés, szintaktikai elemzés, tulajdonnév-felismerés

Juhász Valéria

főiskolai docens
Szegedi Tudományegyetem
Juhász Gyula Pedagógusképző Kar
Alkalmazott Humántudományi Intézet
Magyar és Alkalmazott Nyelvészeti Tanszék

Az SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszékének docense és tanszékvezetője. Kutatási területei az anyanyelv-pedagógia, az olvasás tanítása és tanulása, az olvasási készség fejlesztése, valamint a médiában és számítógépeken zajló kommunikáció tartalomelemzése. A közelmúltban a MAXQDA tartalomelemző szoftver használati lehetőségeivel, és használatának népszerűsítésével is foglalkozott.

A MAXQDA tartalomelemző szoftver lehetőségei

A MAXQDA, az Atlas.ti, a NVivo, Dedoose, Nudist, QDA Miner és társai kvalitatív és kvantitatív kutatásokhoz nyújtanak olyan technikai segítséget és hátteret, amely nélkül ma már szinte kőkorszakinak tűnhet tartalomelemzéseket végezni. Az utóbbi évtizedig erőteljesen lehetett érzékelni a társadalomtudományi kutatási módszerekben egy olyan éles elkülönülést, amely mögött vagy a kvalitatív („puha”), vagy a kvantitatív („kemény”) elemzés hívei sorakoztak fel. Mára azonban egyértelműen kirajzolódik egy olyan kutatási módszer, egy olyan kutatói kör, amely e kettőt elválaszthatatlannak, sőt, egymás validitását erősítőnek tartja. A kvalitatív elemzés során, az értelmezések által kinyert adatok, kategóriák tömbösülhetnek, ezek alkalmasak lesznek kvantitatív vizsgálatokra, összevetésekre, melyeket aztán ismét kvalitatív értelmezések alá vethetünk. Ezt a munkafolyamatot segíti például a MAXQDA szoftver is.

A MAXQDA szoftvert számos területen sikeresen alkalmazzák szociológusok, politológusok, pszichológusok, egészségügyi kutatók, antropológusok, piackutatók, közgazdászok stb. Kiemelkedő tulajdonságai közül említik elsősorban a magas hatékonysági fokát, megbízhatóságát, stabilitását, jól kidolgozott funkcionalitását és – nem utolsósorban – felhasználóbarát felületét. A program egyszerűen kezelhető és világos struktúrával rendelkezik. Használhatjuk PDF-fájlok, képek elemzésére is. A kinyert adatok statisztikai programba exportálhatók.

kulcsszavak: MAXQDA, tartalomelemző szoftver, kvalitatív kutatás, kvantitatív kutatás

Babarczy Anna

tudományos munkatárs
Magyar Tudományos Akadémia
Nyelvtudományi Intézet
Pszicho-, Neuro- és Szociolingvisztikai osztály
Pszicho- és Neurolingvisztikai Kutatócsoport

A Budapesti Műszaki és Gazdaságtudományi Egyetem Kognitív Tudományi Tanszékének egyetemi docense. Elismert kutatója és oktatója a gyermeknyelvnek és a pragmatikának, melyekhez korpusznyelvészeti eszközöket is használ. Érdeklődési körébe tartozik a kísérleti pragmatika, az absztrakció pszicholingvisztikája, és a szó szerinti, valamint a nem szó szerinti jelentések automatikus gépi azonosítása. Több nagyszabású projekt vezetője és résztvevője volt ezekben a témákban, illetve egyetemi oktatóként számos kurzust tartott ezekkel kapcsolatban.

Gyermeknyelvi korpuszok és erőforrások

Az előadás két gyereknyelvi korpuszt és a hozzájuk tartozó nyelvtechnológiai eszközöket mutatja be. Az egyik, a CHILDES (Child Language Data Exchange System, https://childes.talkbank.org) egy nemzetközi vállalkozás Brian MacWhinney, a Carnegie Mellon University pszicholingvista professzora szervezésében. A CHILDES kutatási célokra szabadon hozzáférhető; egy folyamatosan bővülő, néhol hanganyaggal vagy videóval is összekötött gyermeknyelvi spontánbeszéd és történetmesélés korpuszból, továbbá a korpusz elemzését, valamint újabb anyagok átírását segítő számítógépes eszközökből áll. Az angol és néhány más nyelvű anyag egyértelműsített morfológiai elemzéssel együtt érhető el, a magyar anyag jelenleg elemzetlen. A rendszer kínál előre kidolgozott elemzési paradigmákat, és lehetőséget ad saját annotációs rendszerek kidolgozására és alkalmazására.

A másik bemutatott korpusz a MONYEK (Magyar Óvodai Nyelvi Korpusz), Mátyus Kinga doktori disszertációjának anyaga, melyet az én témavezetésemmel hozott létre, és amely az MTA Nyelvtudományi Intézetének honlapjáról érhető el (http://www.nytud.hu/oszt/korpusz/monyek.html). A MONYEK óvodás gyerekekkel szigorú forgatókönyv szerint folytatott interjú és képleírási feladat átírt anyagából áll. Az átírás a CHILDES szabályai szerint történt; a szöveg automatikus morfológiaelemzését a HuMor végezte, melynek kimenetét a PurePos egyértelműsítő rendszer gyermeknyelvre adaptált változata egyértelműsítette utólagos kézi ellenőrzéssel.

kulcsszavak: gyereknyelv, CHILDES, kötetlen annotáció, computer-aided annotáció, kép-hang-szöveg szinkronizáció

Péter Róbert
egyetemi adjunktus
Szegedi Tudományegyetem
Bölcsészet- és Társadalomtudományi Kar
Angol-Amerikai Intézet
Angol Tanszék

Főként brit történelmet és digitális bölcsészetet oktat. Egyik fő kutatási területe a szabadkőművesség, több – a témával foglalkozó – kiadvány, folyóirat szerzője és szerkesztője. Digitális bölcsészettel is foglalkozik: kiemelten olyan kvantitatív módszerek kifejlesztésével és alkalmazási lehetőségeivel, amelyekkel nagy mennyiségű bibliográfiai és metaadat elemzésével lehet hosszú távú tendenciákat és mintázatokat megfigyelni történelmi és kulturális folyamatokban. Részt vett a TANIT és az AVOBMAT digitális bölcsészeti eszközök kifejlesztésében.

A big data kihívás a bölcsészettudományokban: néhány digitális bölcsészeti kutatási eszköz bemutatása

Az előadás célja, hogy bemutasson néhány, digitális bölcsészeti kutatások során használt, jól ismert eszközt (pl. Google N-Gram Viewer, Bookworm, Voyant, Juxta Commons), valamint a Szegedi Tudományegyetemen az elmúlt évben kifejlesztett – és fejlesztés alatt lévő – két webszolgáltatást (TANIT és AVOBMAT). A TANIT (Text ANalysIs Tools) rendszer célja, hogy magyar nyelvű szövegek számítógépes nyelvészeti feldolgozásával dokumentumok összehasonlító elemzéséhez szükséges statisztikákat kigyűjtsön. Ez a webszolgáltatás létező nyelvtechnológiai elemzőlánc kimenetére épülő aggregált statisztikákat számít ki, témamodelleket épít, és ezeket olyan formátumban adja át a digitális bölcsész felhasználónak, aki utána programozói ismeretek nélkül is fel tudja ezt használni kutatásaiban. Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) segítségével nagy mennyiségű metaadatot és szöveget tudunk elemezni és vizualizálni. Az AVOBMAT-ba saját fájlokat és könyvtári repozitóriumokat tölthetünk fel többek között Zoteroból exportált csv és rdf valamint EP3 xml formátumokban. A feltöltés után tudjuk az adathalmazt szűkíteni fazettás, összetett és CCL kereséssel. A metaadatokat számtalan módon tudjuk interaktív módon vizualizálni, amelynek segítségével új, eleddig ismeretlen összefüggéseket és trendeket fedezhetünk fel digitális bölcsészeti elemzések során. Az AVOBMAT az egyszerű vizualizációk segítségével tudja például modellezni szerzők, kiadók, kulcsszavak kapcsolatát és időbeni eloszlását. Az AVOBMAT működését a szegedi Acta repozitórium és brit sajtócikkek elemzésével fogom demonstrálni.

kulcsszavak: digitális bölcsészet, big data, TANIT, AVOBMAT