Magyar Tudomány 179(2018)7, 1017–1026
DOI: 10.1556/2065.179.2018.7.9

MAGYAR NYELVTECHNOLÓGIAI INFRASTRUKTÚRA A TÁRSADALOMTUDOMÁNYOK SZOLGÁLATÁBAN

HUNGARIAN LANGUAGE TECHNOLOGY INFRASTRUCTURE FOR THE SOCIAL SCIENCES

Simon Eszter[1], Váradi Tamás[2]

[1] tudományos főmunkatárs, MTA Nyelvtudományi Intézet
simon.eszter@nytud.mta.hu
[2] tudományos főmunkatárs, MTA Nyelvtudományi Intézet
varadi.tamas@nytud.mta.hu

ÖSSZEFOGLALÁS

Tanulmányunkban bemutatjuk, hogy mivel foglalkozik a nyelvtechnológia, és hogy melyek azok a területek, ahol – jellemzően nagyobb rendszerekbe beépítve – nyelvtechnológiai eszközöket használunk. Ismertetjük az *e-magyar* szövegfeldolgozó eszközláncot, amelyet hatékonyan tudnak használni a társadalomtudományok területén tevékenykedő kutatók is, illetve bárki, aki magyar nyelvű szöveget szeretne feldolgozni további kutatási céljainak megfelelően. A nyelvtechnológiai fejlesztések használhatóságát három alkalmazás bemutatásával igyekszünk megvilágítani. Hangsúlyozzuk, hogy az MTA Nyelvtudományi Intézet koordinálásával, a HunCLARIN égisze alatt a magyar nyelvtechnológia kész és képes hatékonyan segíteni a társadalomtudományi kutatásokat.

ABSTRACT

The present paper aims to introduce the field of human language technology and the areas where language technology can be usefully applied, typically as embedded in a complex system. We present the *e-magyar* language processing pipeline, which can be used with ease by researchers in social sciences and indeed by anyone who wants to process large amounts of Hungarian text material. As examples of how to use language technology tools, we describe three applications. As a take-away message, we emphasize that mobilizing the resources and services of the HunCLARIN research infrastructure the Hungarian language technology community with the coordination of the Research Institute for Linguistics of the Hungarian Academy of Sciences is ready and able to efficiently support research in the social sciences.

Kulcsszavak: nyelvtechnológia, nyelvi erőforrások, társadalomtudományok, szövegfeldolgozás, kutatási infrastruktúra, HunCLARIN

Keywords: language technology, language resources, social sciences, text processing, research infrastructure, HunCLARIN

BEVEZETÉS: MI A NYELVTECHNOLÓGIA?

A nyelvtechnológia célja, hogy a gépi rendszereket felruházza azzal a páratlan képességgel, nyelvi intelligenciával, amellyel mi emberek rendelkezünk, és amit anyanyelvünk esetében olyan rendkívüli könnyedséggel használunk. Ezt a mentális kapacitást természetesen a gépekről nem feltételezhetjük, de távlati célként erre vagy legalábbis az emberi nyelvtudás olyan szintű gépi modellálására lenne szükség, ahol a gépi rendszerek (itt nemcsak a számítógépekre, hanem egyre inkább az okostelefonokra gondolunk) az emberhez hasonló gyorsasággal, könnyedséggel és intelligenciával kezelik a nyelvet. Azaz megértik a beszédet és az írott szöveget, és azt a tárgyi tudást, amellyel rendelkeznek, természetes emberi nyelven tudják kommunikálni.

Ebben az értelemben a nyelvtechnológia a mesterséges intelligencia azon területe, ami a gépi rendszerek nyelvi készségét, intelligenciáját hivatott kiépíteni. E távoli jövőbe nyúló kutatásfejlesztési program nem jelenti azt, hogy a nyelvtechnológia a sci-fi világába tartozó terület lenne. Ellenkezőleg, bár nincs nagyon a köztudatban, a nyelvtechnológia eredményeit szinte naponta használjuk az alkalmazások egész sorában. Elég csak a helyesírás-ellenőrzést, a karakterfelismerést, a gépi szövegfelolvasást, a diktálást vagy a gépi fordítást említeni.

Bár a nyelvtechnológia nagy fejlődésen ment keresztül az utóbbi években, az emberi nyelv komplexitásának köszönhetően távolról sem tekinthető megoldottnak a nyelvfeldolgozás minden lépése. A feladat egyik fő nehézségét az adja, hogy az ember az értelmezés során számos nehezen formalizálható információt is figyelembe vesz, amelyeket egy gép számára csak korlátozott módon lehet elérhetővé tenni. Ilyenek többek között a megnyilatkozás körülményei (hol, mikor, kikkel), valamint azok többletjelentése (például: ígéret, fenyegetés), amely szintén hatással van arra, hogy hogyan értelmezünk egy üzenetet. A nyelvtechnológia feladata azonban egyelőre nem az ilyen jellegű többletinformáció figyelembevétele, hanem csakis a szövegfolyamban detektálható releváns információ adott célnak megfelelő feldolgozása. Mint minden, a nyelvben tárolt információ – bizonyos fokú – megértését magában foglaló, tehát szemantikai célkitűzéssel bíró feladat, ez is rengeteg, önmagában is kihívást jelentő részfeladatot tartalmaz.

A nyelvtechnológiai fejlesztések tipikusan nagyobb alkalmazásokba beépítve jelennek meg, segítve bennünket, amikor például információt keresünk internetes keresővel, helyesírást és nyelvtant ellenőrzünk szövegszerkesztőben, termékajánlásokat olvasunk *online* vásárláskor, egy navigációs szoftver szóbeli utasításait hallgatjuk, vagy online szolgáltatással fordítunk weboldalakat. A nyelvtechnológia segítségével elérhetővé válik az automatikus fordítás és tartalom-előállítás, az információfeldolgozás és a tudásmenedzsment – akár több nyelven is. Emellett elősegíti az intuitív, természetes nyelv alapú interfészek fejlesztését a háztartási

elektronika, a gépészet, a járműgyártás, a számítástechnika és a robotika területén is. Nagy lehetőségek rejlenek a nyelvtechnológiának az oktatásba és a szórakoztatóiparba, például játékokba, oktatóprogramokba, szimulációs környezetekbe való integrálásában is. A számítógéppel támogatott nyelvtanulás, az *e-learning,* az önellenőrző eszközök és a plágiumszűrő szoftverek csak kiragadott példák arra, hogy hány helyen játszik fontos szerepet a nyelvtechnológia. A közösségi alkalmazások terjedésével felmerül az igény a kifinomultabb nyelvtechnológiai alkalmazásokra is, amelyek figyelemmel követik a bejegyzéseket, összegzik a vitákat, ajánlásokat tesznek, kiszűrik az érzelmi tartalmú válaszokat, szerzői jogi szabálytalanságokat vagy visszaéléseket.

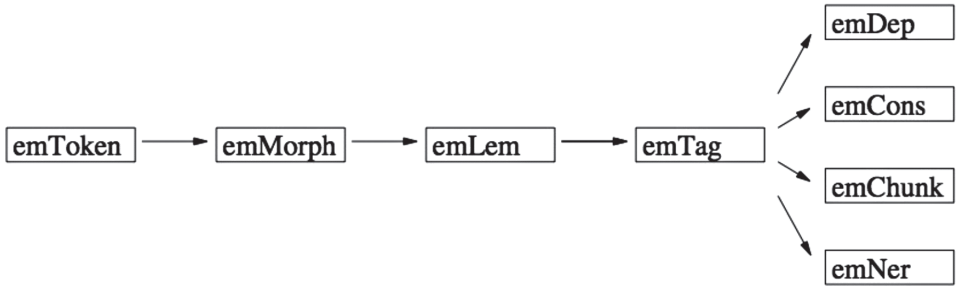
A nyelvtechnológia egyik legfontosabb módszere, hogy a nyelvi információt hordozó digitális adatfolyamokat (nyomtatott oldalak beszkennelt képét, írott szöveget, beszédet tartalmazó hangfelvételt, videót) automatikus módszerekkel feldolgozva olyan – az eredeti anyagban explicit formában nem szereplő – további információval lássa el, amely lehetővé teszi a szövegben kódolt információ, illetve tudás minél többféle szempont szerinti megtalálását (az intelligens keresést), strukturált adatbázisokba szervezését és a felhasználó számára optimális prezentációját. Az optimális prezentálás magában foglalja többek között a legrelevánsabb információ kiemelését és annak a felhasználónak leginkább megfelelő modalitásban (írás, beszéd, esetleg animált szájmozgással kísérve, jelelés stb.) és nyelven (például automatikus fordítással) való visszaadását. A nyelvi tartalmak hatékony szűrése, a lényeges információ megtalálása és kiemelése az adattengerből, a nehezen érthető információnak a felhasználó számára jobban értelmezhető formában való visszaadása alapvető fontosságú feladatok, melyek megoldásával a nyelvtechnológia nélkülözhetetlen háttér-infrastruktúrát ad a többi tudományterületnek. Az intézmények, a vállalkozások számára a (szöveges-, hangzó- és videóanyagokból automatikusan kinyert) tudásbázisok komoly versenyelőnyt jelentenek információs társadalmunkban, és az állampolgárok számára is az élet megjobbításának alapvető eszközeivé válhatnak. Kiemelkedő szerepet játszhatnak különböző hátrányos helyzetű csoportok (siketek, gyengénlátók, baleset következtében beszédkészségüket elvesztők, idegen nyelveket nem tudók) életminőségének javításában.

Jelenleg a sokrétű előfeldolgozási feladatoknak a megoldása áll a magyar nyelvtechnológia homlokterében: a szöveg alkotóelemeinek azonosítása (mondatokra és szavakra bontás), különböző szó- és mondatszintű nyelvtani információk hozzárendelése (morfológiai elemzés és egyértelműsítés, szintaktikai elemzés) már megoldottnak tekinthető, de folyamatban van a mondatok közötti összefüggések felismerése, a világról szóló tudásunkat reprezentáló ontológiák építése, valamint az érzelmek detekciója is.

AZ E-MAGYAR NYELVFELDOLGOZÓ ESZKÖZLÁNC

Tanulmányunkban ismertetjük az *e-magyar* rendszert (Váradi et al., 2017), amely a fent említett alapvető szövegfeldolgozási lépéseket valósítja meg a munkában részt vevő műhelyekben eddig előállított különböző eszközök továbbfejlesztésével, egységesítésével és egyetlen koherens technológiai láncba való szervezésével. Az *e-magyar* rendszer a Magyar Tudományos Akadémia támogatásával készült a 2015-ben kiírt infrastruktúra-fejlesztési pályázat keretében. A munkálatok a pályázat kedvezményezettje, a Nyelvtudományi Intézet koordinálásával széles körű együttműködés keretében folytak, melyben részt vett a hazai nyelvtechnológia számos vezető kutatásfejlesztő műhelye.

A központi gondolat az *e-magyar* kialakításában az integráció volt. A magyar nyelvtechnológiai közösség külön-külön, de az utóbbi évtizedben egyre inkább együttműködve számos erőforrást és eszközt hozott létre. Ezek a Nyelv- és Beszédtechnológiai Platform, illetve a META-NET (URL1) hálózaton keresztül is publikálásra kerültek. Az eszközök egy része nyílt forráskódú (ilyen például a *hun** eszközcsalád [URL2]), mások csak bináris formában érhetők el kutatásfejlesztési célokra (ilyen például a Humor morfológiai elemző [Novák, 2003]). A mostani infrastruktúra interoperábilissá tette ezeket az eszközöket abban az értelemben, hogy az infrastruktúra egyes eszközei modulárisan egymásra épülnek, vagyis önállóan is működnek, de olyan elemzési láncba is szervezhetők, amelyben zökkenőmentesen halad az adat a különböző eszközökön át. Ez azt jelenti, hogy a nyers szövegből kiindulva az *e-magyar* szövegfeldolgozó eszközlánca elvégzi a szöveg elemeinek a szegmentálását *(emToken),* 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 sekély elemzésre van szükségünk, felismeri a mondatban szereplő frázisokat *(emChunk),* továbbá a szövegben előforduló tulajdonneveket (*emNer*). Az eszközök egymásra épülése az *1. ábrá*n látható.



1. ábra. Az *e-magyar* szövegfeldolgozó lánc elemeinek egymásra épülése

Fontos megemlíteni, hogy magyarra már létezik egy szövegfeldolgozó eszközlánc, a *magyarlanc* (Zsibrita et al., 2013), amely szintén megvalósítja ezt a moduláris architektúrát, de egy zárt rendszeren belül. Az *e-magyar* fejlesztésénél fontos szempont volt, hogy nyílt rendszer legyen, vagyis, hogy az infrastruktúra egésze és annak minden eszköze külön-külön is elérhető, letölthető, világos licenccel publikált és kutatásfejlesztési célra, de adott esetben üzleti felhasználásra is ingyenesen használható legyen.

A láncban részt vevő szoftverek licence GNU GPLv3 vagy GNU LGPLv3, a nem-szoftver elemekre vonatkozó licenc pedig Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA), kivéve az *emMorph* morfológiai elemző alatt működő adatbázist, amelyre az üzleti felhasználást kizáró Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA) licenc vonatkozik.

Az *e-magyar* nem csak a nyelvtechnológiai szakma vagy a nyelvtechnológiát használó ipari fejlesztők igényeit kívánja szolgálni. A szakmabeli felhasználók mellett támogatjuk a számítógépes eljárások iránt fogékony, de a nyelvtechnológiában nem jártas diákok és kutatók körét is: a bölcsészet- és társadalomtudományok művelőit, valamint akár az érdeklődő nagyközönséget is. Nekik két szinten kínálunk támogatást. Egyrészt egyedi igények alapján különböző elemzőláncokat állíthatnak össze az *e-magyar* eszközeinek a felhasználásával a GATE szövegelemző rendszer (Cunningham et al., 2013) keretein belül. Másrészt igénybe vehetik az e-magyar.hu oldal webszolgáltatását, amely rövidebb szövegek azonnali elemzését végzi.

Az eszközlánc négyféleképpen használható. A honlapon keresztül egy rövidebb szöveget egyszerűen bemásolva kipróbálhatjuk az eszközláncot. Szövegelemzéshez, társadalomtudományi vagy digitális bölcsészeti kutatáshoz a GATE-rendszer GATE Developer nevű grafikus felhasználói felületét ajánljuk, amelybe az e-magyar lánc egyszerűen telepíthető. A telepítés leírása a teljes rendszerrel együtt elérhető az URL3 oldalon. Itt lehetőség van a rendszer továbbfejlesztésére, vagyis az elemzőlánchoz saját készítésű modulok is hozzáadhatók. Nagyobb korpuszok feldolgozásához a GATE parancssori hozzáférését ajánljuk, ennek használata szintén az említett honlapon található, szükséges hozzá a *github* repozitórium használata. Negyedik módszerként használatba vehető az ún. gate-server is, ez szintén parancssori technológia, és ez az egyébként, amely a honlap mögött is üzemel. Az integrációról és a rendszer használatáról részletesebben lásd még Sass Bálint és szerzőtársai ismertetését (2017).

A projekt célkitűzései között szerepelt, hogy az elemzőlánc hozzáférhető és érdemben használható legyen olyan felhasználók körében is, akik nem feltétlenül járatosak az informatika területén. Ennek az igénynek igyekszik megfelelni az *e-magyar.hu* webes szövegelemző szolgáltatása (URL4), amely lehetővé teszi, hogy egy webes interfészen keresztül bárki egyszerűen kipróbálhassa az egyes

elemző modulokat vagy akár a teljes elemzőláncot anélkül, hogy ehhez a böngészőn kívül bármilyen egyéb szoftvert használnia kellene.

A szövegelemző egy olyan webszolgáltatásra épül, amely a GATE-es könyvtárakat használja, bemenetként az elemzett szöveget és a futtatni kívánt elemző modulok listáját várja, kimenetként pedig a GATE által generált, az annotációkat tartalmazó XML-t adja vissza. A weboldal a visszakapott XML-t feldolgozza, és a kinyert adatokat megjeleníti egy könnyen értelmezhető, vizualizált formában. Az elemzés eredményét a felhasználó le is töltheti magának további felhasználásra.

TOVÁBBI ERŐFORRÁSOK ÉS FELHASZNÁLÁSI LEHETŐSÉGEK

A nyelv- és beszédtechnológia területén sikerrel alkalmazható módszerek és eljárások jellegéből következik, hogy magas színvonalú kutatási eredmények és alkalmazások nem jöhetnek létre a megfelelő erőforrások, írott és beszélt nyelvi adatbázisok, alapvető sztenderdizált feldolgozó eszközök nélkül; ezek a nyelvtechnológia elengedhetetlen szükségletei a fejlesztésben és az elért eredmények kiértékelésében is. A magyarországi nyelv- és beszédtechnológia fennállása óta a különböző műhelyekben szép számú adatbázist és szövegfeldolgozó eszközt fejlesztettek ki.

Ezek a műhelyek, melyek sokáig elszigetelten működtek, az utóbbi években felismerték az együttműködés fontosságát, és annak szükségét, hogy egységes, mindenki számára elérhető és könnyen kiterjeszthető kutatási infrastruktúrákat hozzanak létre. Ennek a törekvésnek az eredménye a magyarországi nyelv- és beszédtechnológiai műhelyeket tömörítő Nyelv- és Beszédtechnológiai Platform, amely az akadémiai/egyetemi kutatásfejlesztés vezető műhelyei és ipari partnerek stratégiai szövetségéből jött létre. A platform kidolgozott egy középtávú stratégiai kutatási tervet (URL5), valamint egy megvalósítási tervet (URL6). A platform kutatásfejlesztő központjai részt vettek a META-NET kiválósági hálózatban, ahol a CESAR- (URL7) projekt keretében értékes nyelvi erőforrásokat és eszközöket tettek egységes alakban elérhetővé a META-SHARE- (URL8) hálózaton keresztül. A Nyelvtudományi Intézet egyik alapítója volt a CLARIN (URL9) európai kutatásiinfrastruktúra-hálózatnak. A HunCLARIN, a CLARIN magyar hálózata, összhangban a jelenleg húsz országot összefogó európai szervezet céljaival, küldetésének tekinti a bölcsészet- és társadalomtudományi kutatások támogatását a nyelvtechnológia, a nyelvi erőforrások könnyen elérhetővé tételével.

Az MTA Nyelvtudományi Intézete számos olyan nyelvi erőforrást és nyelvfeldolgozó eszközt fejlesztett az elmúlt években, amelyek hatékonyan tudják támogatni a társadalom- és bölcsészettudományban dolgozó kutatók munkáját. Az alábbiakban bemutatunk néhány példaalkalmazási területet, társítva hozzájuk azokat az erőforrásokat, amelyek jól használhatók ezeken a területeken.

Információkinyerés és trendelemzés a munkaerőpiacon

Az elektronikus formában rendelkezésre álló feldolgozatlan állományok sok hasznos információt rejtenek, amelyekből szövegbányászati technikák segítségével statisztikák gyűjthetők, vagy következtetések vonhatók le. Emellett éppoly fontos az ontológiák, tudástárak építése is, amelyekkel a világismereti és nyelvi tudásunkat szimulálhatjuk. Ezekre támaszkodva megvalósíthatóvá válnak a jelenleginél jobb eredményt nyújtó trendelemző szoftverek, amelyek segítségével, különböző aggregált statisztikákra támaszkodva az érdeklődők reális képet kaphatnak olyan speciális területekről, mint például a munkaerőpiac.

Az álláshirdetések és önéletrajzok számos értékes információt rejtenek, de ezek nem mindig érhetők el strukturált formában, így gépi feldolgozásukhoz jelentős támogatást nyújthat bizonyos nyelvtechnológiai eredmények felhasználása. Szakontológiák és szinonimaszótárak beépítésével, valamint információkinyerő eszközök alkalmazásával könnyebben elérhetők lehetnek olyan, egyes iparágakra, szakterületekre jellemző adatok, amelyek nagyban megkönnyíthetik az álláskereső dolgát. Többek között választ kaphatnak az elvárható fizetések nagyságáról, vagy arról, hogy melyik régióban keresett egy adott szakterület. Lehetőség nyílhat a különböző végzettségek összehasonlítására, így nem utolsósorban az egymással versengő felsőoktatási intézmények rangsorolására is kiváló eszköz lehet. Másrészről a humánerőforrás-szakemberek munkáját jelentősen segítheti egy, az önéletrajzokból automatikusan felépített adatbázis, amely igen nagy mértékben megkönnyíti a megfelelő jelöltek kiválasztását.

Ehhez a feladathoz szükség van szakontológiákra és szinonimaszótárakra, melyek közül a legismertebb a WordNet (Miller, 1995). A Magyar WordNet (Miháltz et al., 2008) egy szemantikailag strukturált, általános célú fogalomtár a magyar nyelvre, amely bizonyos szavak közötti lexikai relációkat ír le, így reprezentálva a világ bizonyos elemei között fennálló kapcsolatokat. A WordNethez hasonló szemantikai fogalomhálók kinyerésére az elmúlt években elsősorban neurális hálókra épülő, ún. mély tanulást megvalósító technikákat használnak.

A magyar nyelvváltozatok feltérképezése és adatbázisba szervezése

A magyar nyelv különböző, még élő változatainak digitális rögzítése fontos feladat, hiszen a ritkább nyelvváltozatok beszélői kiöregednek, és a vidéki lakosság városokba áramlásából fakadóan egyre inkább megfigyelhető a dialektusok eltűnése, illetve a határon túli nyelvváltozatok esetében a nyelvvesztés folyamata. A magyar nyelvváltozatok kutatása túlnyomórészt nyelvtechnológiai támogatás nélkül zajlott, az elmúlt években azonban a szociolingvisztikai és szociológiai kutatások számos területén kezdődött meg az együttműködés különböző kutatócsoportok között. A nyelvtechnológiai eszközökkel segített értékőrző tevékenység nagy része

az írott szövegek digitalizálására terjed ki, de a beszélt nyelv hangzó formában való rögzítése és kereshetővé tétele is nagyon fontos. A magyar nyelv különböző nyelvváltozatainak rögzítésével digitális lenyomatot készíthetünk a már eltűnőben lévő dialektusokról, megnyitva az utat ezzel a szociolingvisztikai vizsgálatok széles spektruma előtt. A felvett anyagok hangzó és szövegesen lejegyzett változata egyaránt fontos, hiszen más és más vizsgálatok alapjául szolgálhatnak.

A Nyelvtudományi Intézet számos olyan kutatásban vett részt, amelynek céljai egybeesnek a fent megfogalmazottakkal. Az egyik ezek közül a Budapesti Szociolingvisztikai Interjú (BUSZI) (Váradi, 2003), amely egy nagyszabású felmérés a magyar nyelv Budapesten beszélt változatairól. Az 1987 óta több fordulóban felvett adatokat tartalmazó adatbázis segítségével megbízható adatok és elemzések kaphatók az élő nyelvhasználat és a szociológia számos fontos kérdésére. A hanganyagok digitalizálása után került sor az anonimizálásra, a hangzó anyag szöveges lejegyzésére, a morfológiai elemzésre és egyértelműsítésre, a szóalak fonetikai reprezentációjának hozzáadására, mindennek strukturált adatbázisba szervezésére és a lekérdező rendszer kifejlesztésére, mely utóbbin keresztül az adatbázis hozzáférhetővé vált az érdeklődő kutatók számára. Ehhez és minden megelőző lépéshez szükség volt nyelvtechnológiai támogatásra, amely nélkül az információ nem lenne explicit, egyértelmű, számítógéppel egyszerűen és hatékonyan kiolvasható és feldolgozható formában tárolva, vagyis nem lenne alkalmas további kutatásokra.

Egy másik projekt, amely a magyarországi és a határon túli nyelvváltozatok feltérképezését segíti, a Magyar Nemzeti Szövegtár (MNSz) (Oravecz et al., 2014) építése volt. Az MNSz – szándékai szerint – reprezentatívan tartalmazza a mai magyar nyelv jellegzetes megnyilvánulásait. A korpusz nagyobb része magyarországi forrásokból származik, de jelentős mennyiségű szlovákiai, kárpátaljai, erdélyi és vajdasági szöveget is tartalmaz. Az MNSz lényegi tulajdonsága, hogy minden szó mellett feltünteti a szótövet, a szófajt és a szó morfológiai elemzését is. A szótő, szófaj és elemzés megállapítása és az elemzések egyértelműsítése automatikus gépi eszközökkel történt. A rendszer megbízhatósága kb. 97,5%-os, így az összes szóalak kb. 2,5%-a van hibásan elemezve. Ennél jobb eredményt csak a kézi elemzés biztosíthatna, ami ekkora méretű anyag esetén megvalósíthatatlan. Az MNSz aktuális verziója 1,04 milliárd szövegszót tartalmaz.

Pszichodiagnosztikai vizsgálatok nyelvtechnológiai támogatása

A pszichodiagnosztika és a nyelvtechnológia integráns összefüggésének logikai alapja az, hogy az egyének és a csoportok pszichológiai állapotai és folyamatai (érzelmek, gondolkodásmód, szándékok stb.) nem csupán a fizikai, hanem a verbális viselkedésben is kódolódnak. E kódok nyelvi markerek formáját öltik. Ekképp az elektronikusan rögzített kommunikáció tartalomelemzése révén az

egyének és a csoportok pszichológiai folyamatai diagnosztizálhatók, ezek időbeli változásai statisztikailag mérhetők és feltérképezhetők, továbbá az egyes egyének és csoportok egymással kvantitatíven összehasonlíthatók. A nyelvi markerek és a pszichológiai állapotok és folyamatok összefüggéseit mintegy hatvan éve kutatják. Klasszikus példa, hogy az egyes szám első személyű igék, névmások és a tagadószavak együttes túlzott használata a depresszió nyelvi tünete lehet, illetve a veszélyes küldetést teljesítő legénységek csoporton belüli konfliktusa, illetve a távoli irányító személyzettel való szembefordulása a kommunikációból előrejelezhető.

Az MTA Nyelvtudományi Intézete többéves együttműködés keretén belül segíti az MTA Pszichológiai Intézet munkatársainak kutatásait, melyeknek célja automatikus tartalomelemző módszerek fejlesztése, melyek részben kiválthatják a személyiség- és klinikai pszichológiában jelenleg használatos teszteket, továbbá lehetővé teszik célvezérelt kiscsoportok távoli, automatikus pszichodinamikai monitorozását.

ÖSSZEFOGLALÁS

Tanulmányunkban bemutattuk, hogy mivel foglalkozik a nyelvtechnológia, és hogy melyek azok a területek, ahol – jellemzően nagyobb rendszerekbe beépítve – nyelvtechnológiai eszközöket használunk. Prezentáltuk az *e-magyar* szövegfeldolgozó eszközláncot, amelyet hatékonyan tudnak használni a társadalomtudományok területén tevékenykedő kutatók is, illetve bárki, aki magyar nyelvű szöveget szeretne feldolgozni további kutatási céljainak megfelelően. A nyelvtechnológiai fejlesztések használhatóságát három példaalkalmazással igyekeztünk megvilágítani. Hangsúlyozzuk, hogy az MTA Nyelvtudományi Intézet koordinálásával, a HunCLARIN égisze alatt a magyar nyelvtechnológia kész és képes hatékonyan segíteni a társadalomtudományi kutatásokat.

IRODALOM

Cunningham, H. – Tablan, V. – Roberts, A. – Bontcheva, K. (2013): Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, 9, 2: e1002854. DOI:10.1371/journal.pcbi.1002854, http://tinyurl.com/gate-life-sci/

Miháltz M. – Hatvani Cs. – Kuti J. et al. (2008): Methods and Results of the Hungarian WordNet Project. In: Tanács A. – Csendes D. – Vincze V. et al. (eds.): *Proceedings of the Fourth Global WordNet Conference GWC 2008*. Szeged: University of Szeged, 310–320. http://www.inf.u-szeged.hu/projectdirs/gwc2008/GWC2008_Proceedings_Final.pdf

Miller, G. A. (1995): *WordNet: A Lexical Database for English. Communications of the ACM*. 38, 11, 39–41. DOI: 10.1145/219717.219748, http://nlp.cs.swarthmore.edu/~richardw/papers/miller1995-wordnet.pdf

Novák Attila (2003): Milyen a jó Humor? In: *I. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: SZTE, 138–144. http://www.morphologic.hu/downloads/publications/na/2003_mszny_Humor_na.pdf

Oravecz Cs. – Váradi T. – Sass B. (2014): The Hungarian Gigaword Corpus. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC-2014). Reykjavik: European Languages Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf

Sass B. – Miháltz M. – Kundráth P. (2017): Az e-magyar rendszer GATE környezetbe integrált magyar szövegfeldolgozó eszközlánca. In: Vincze V. (szerk.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem. 79–90. http://real.mtak.hu/72474/1/kotet.pdf

Váradi T. (2003): A Budapesti Szociolingvisztikai Interjú. In: Kiefer Ferenc (szerk.): *A magyar nyelv kézikönyve*. Budapest: Akadémiai Kiadó, 339–360. http://www.nytud.hu/oszt/elonyelv/adat/buszi.pdf

Váradi T. – Simon E. – Sass B. et al. (2017): Az e-magyar digitális nyelvfeldolgozó rendszer. In: Vincze V. (szerk.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, 49–60. http://real.mtak.hu/72361/

Zsibrita J. – Vincze V. – Farkas R. (2013): magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing* RANLP 2013. Shoumen, Bulgaria: INCOMA Ltd., 763–771. http://www.aclweb.org/anthology/R13-1099

URL1: META-NET http://www.meta-net.eu

URL2: hun* eszközcsalád http://hlt.bme.hu/en/resources/hun-toolchain

URL3: https://github.com/dlt-rilmta/hunlp-GATE

URL4: az *e-magyar.hu* webes szövegelemző szolgáltatása http://e-magyar.hu/parser

URL5: a Nyelv- és Beszédtechnológiai Platform Stratégiai Kutatási Terve nih.gov.hu/download.php?docID=19926

URL6: a Nyelv- és Beszédtechnológiai Platform Megvalósítási Terve http://nkfih.gov.hu/download.php?docID=23049

URL7: Central and South-East European Language Resources http://cesar.nytud.hu

URL8: META-SHARE http://www.meta-share.org

URL9: CLARIN https://www.clarin.eu